

Article

Auditing Per-User Reliability in Webcam-Based Gaze Estimation: Subject- and Pose-Stratified Analysis with an Exploratory Multi-Criteria Decision Scaffold

Kaveti Pavan ^{1,2} , Sota Asahara ², Nagarajan Ganapathy ¹ and Hiroyuki Sugimori ^{3,*} ¹ Department of Biomedical Engineering, Indian Institute of Technology Hyderabad, Kandi, Hyderabad 502284, Telangana, India² Graduate School of Health Sciences, Hokkaido University, Sapporo 060-0812, Japan³ Department of Biomedical Science and Engineering, Faculty of Health Sciences, Hokkaido University, Sapporo 060-0812, Japan

* Correspondence: sugimori@hs.hokudai.ac.jp

Abstract

Low-cost webcam gaze estimation is increasingly proposed as a front-end sensor for self-monitoring of attention and visual well-being, yet average accuracy hides where and for whom such a sensor is reliable. Using only three public gaze datasets (MPIIFaceGaze, Columbia Gaze, and RT-GENE; no new data collection), we make three contributions. First, taking a naive sample-wise split (shuffling samples, not subjects) that places every one of 84 subjects across train/validation/test as a deliberate reference baseline, we show apparent error rising from 4.47 deg on seen subjects to 11.63 deg on subject-disjoint unseen subjects (each a five-seed mean; 95% CI: ± 0.11 deg and ± 4.64 deg, respectively), with an optimism of +7.15 deg (~62%); every naive seed lies far below every subject-disjoint seed, and because the unseen estimate varies widely across seeds, its magnitude should be read as indicative rather than precise. Second, we audit reliability per subject and head pose stratum without demographic labels, revealing a $4.9\times$ spread across users and a Monte Carlo dropout uncertainty that provides a partial, borderline per-subject reliability signal (Pearson $r = 0.48$; Spearman $\rho = 0.50$, with a bootstrap interval including zero); bias-only recalibration lowers error only modestly (subject-macro: 11.79 to 11.50 deg) and does not help the least-reliable users. Third, we convert this measured reliability profile into a per-user decision (trust/recalibrate/abstain) using objective-weight multi-criteria decision making (averaged CRITIC and entropy weights with TOPSIS; the largest weight falls on MC-dropout uncertainty at 0.42, and then pose sensitivity at 0.23); the decision layer disagrees with a single-metric threshold for 35% of users (six of 17) while remaining stable under $\pm 20\%$ weight perturbation (no action changes). This study provides an exploratory, retrospective, public-data-only implementation of a per-user reliability decision scaffold for low-cost gaze-based self-monitoring.



Academic Editor: Stanislav Dadelo

Received: 1 July 2026

Revised: 28 July 2026

Accepted: 29 July 2026

Published: 1 August 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

Keywords: appearance-based gaze estimation; per-user reliability; subject-disjoint evaluation; data leakage; uncertainty estimation; multi-criteria decision making; TOPSIS; CRITIC; personalized decision support

1. Introduction

Low-cost webcam gaze estimation is an attractive sensing modality for AI-driven health and well-being because it requires no specialized hardware: a commodity laptop

camera can, in principle, turn everyday screen time into a continuous, passive signal for self-monitoring. Appearance-based deep models map a face or eye image directly to a gaze direction without dedicated eye-tracking rigs [1], and the same webcam stream that supports video calls could plausibly support longitudinal behavioral self-tracking. This vision motivates the present work: we treat a webcam gaze model not as an end in itself but as a candidate self-monitoring instrument whose readings might one day feed well-being applications. Our scope, however, is deliberately conservative. No health labels exist in the public gaze datasets we use, and we make no clinical claim whatsoever; our contribution is instrument validation plus a decision scaffold, and we reframe the special issue's 'early detection' pillar as a preliminary step toward deployment evaluation rather than as any diagnostic capability.

The well-being relevance of such a sensor hinges on a structural rather than purely statistical problem: the prediction-to-decision gap. Recent work on AI-driven digital health argues that systems are routinely over-engineered on the predictive/inference side and under-engineered on the decision side, and proposes the Decision Integration Deficit Index (DIDI) to quantify this architectural imbalance [2]. A gaze model that emits a (pitch, yaw) estimate is a prediction; it does not, by itself, tell a deployer whether to trust that estimate for a given person, recalibrate, or abstain. Trustworthy-AI guidance for healthcare makes the same demand from the governance side: the FUTURE-AI international consensus organizes deployable medical AI around Fairness, Universality, Traceability, Usability, Robustness, and Explainability, and asks for auditable, per-deployment evidence rather than a single headline accuracy [3]. Our paper instantiates this guidance concretely for one webcam gaze sensor, turning reliability evidence into an explicit per-user action.

The central methodological problem is that average accuracy hides per-user reliability. A single mean angular error reported over a pooled test set can be dominated by easy subjects and easy poses, masking large individual disparities that are exactly what a per-person self-monitoring tool must respect. In our subject-disjoint evaluation, the spread between the best and worst subjects is wide (Section 4.2). A deployer who reads only the aggregate would systematically overstate reliability for the users who need protection most. This concern echoes fairness audits in gaze that reveal demographic performance disparities [4], but our unit of analysis is the individual subject and head pose in a retrospective per-user audit rather than a coarse demographic group.

Worse, headline numbers are easily inflated by subject leakage. The reproducibility literature catalogs data leakage as a pervasive driver of over-optimistic machine learning results across many scientific fields and identifies train-test contamination as a recurring culprit [5], while cross-subject work in adjacent biometric-style domains argues that strictly subject-disjoint evaluation is the only valid metric of utility on unseen users, since subject-dependent protocols leak identity and inflate accuracy dramatically [6,7]. As a deliberate reference baseline, we take a naive sample-wise split that shuffles samples rather than subjects, placing every subject in train, validation, and test. Under an identical protocol on a subject-disjoint re-split with zero subject overlap, estimated symmetrically over the same five seeds, the unseen-subject error more than doubles, and every subject-disjoint seed far exceeds every naive seed (Section 4.1).

Two further observations sharpen what 'reliability' must mean for this sensor and show that recalibration is not a blanket fix. First, training-free Monte Carlo dropout uncertainty ($T = 20$), obtained with no retraining from dropout already present in the gaze head, is a partial, borderline per-subject reliability signal, whose correlation with per-subject error has a bootstrap interval that includes zero (Section 4.3); its main value is separating a stable majority of users from a cluster whose predictions are highly unstable, making it a useful but imperfect reliability proxy that requires no extra model. Second, bias-

only recalibration that subtracts a per-subject bias estimated from nine calibration samples lowers the subject-macro error only modestly and does not help the least-reliable users, so the residual error is largely variance- rather than bias-driven. Head pose sensitivity, measured as the absolute slope of error against pose magnitude, adds a second axis of per-user risk. Together, these mean that a deployer cannot simply ‘calibrate away’ the gap; reliability must be characterized per user and per pose and then acted upon, not assumed.

This exposes the explicit gap our work fills. Prior gaze research either chases the cross-dataset domain gap with stronger estimators and pre-training [8–10] or characterizes fairness and uncertainty without issuing an action [4,11,12], or, at most, performs per-frame selective prediction from a single uncertainty signal [13,14]. Multi-criteria decision analysis in health, meanwhile, operates at model- or cohort-level granularity, selecting which model or program to adopt for everyone [15,16]. To our knowledge, no prior gaze work converts a label-free, subject- and pose-stratified reliability audit into a per-user multi-criteria decision (MCDM) trust action. We address this gap by feeding four reliability-derived cost criteria (mean error, p90 error, absolute pose-sensitivity slope, and MC-dropout uncertainty), weighted objectively by averaged CRITIC and Shannon entropy weights, into a TOPSIS closeness score that maps each user to trust, recalibrate, or abstain (Section 4.4).

We make the following contributions. (1) Using a naive sample-wise split that shuffles samples rather than subjects as a deliberate reference baseline, we quantify subject leakage optimism of +7.15 deg ($\sim +62\%$) between the seen-subject (4.47 $\pm$ 0.11 deg) and the subject-disjoint (11.63 $\pm$ 4.64 deg) (mean: $\pm$ 95% CI), each averaged over the same five seeds, a concrete instance of the reproducibility-eroding leakage the literature warns about [5–7], here in appearance-based gaze regression. (2) We deliver a subject- and pose-stratified reliability audit that surfaces per-user disparity (subject-macro: 11.79 deg; best: 5.16, worst: 25.16 deg; $4.9\times$ spread; Gini: 0.245) and per-domain difficulty (sample-weighted unseen MPII: 7.71, Columbia: 12.50, and RT-GENE: 15.69 deg), going beyond a single average. (3) We characterize a training-free MC-dropout uncertainty signal as a partial, borderline per-subject reliability proxy (Pearson $r = 0.48$; Spearman $\rho = 0.50$ to error, with a bootstrap interval that includes zero; it separates a stable majority of users from an unstable cluster) and show, via bias-only recalibration that lowers error only modestly (subject-macro: 11.79 $\rightarrow$ 11.50 deg) and leaves the least-reliable users unimproved, that the residual error is largely variance- not bias-driven. (4) We introduce a per-user MCDM decision layer with objective CRITIC + entropy weights and TOPSIS aggregation that emits trust/recalibrate/abstain (5/6/6), disagrees with a naive single-metric (mean-error tercile) threshold for 6/17 (35%) of users, and is stable under $\pm 20\%$ weight perturbation (no action change across 136 decisions). (5) We frame the whole pipeline as a transparent, public-data-only operationalization of the prediction-to-decision gap [2] and trustworthy-AI expectations [3], with no clinical claim.

The three public datasets we re-use span complementary capture regimes and together stress the reliability of a webcam gaze sensor. MPIIFaceGaze [17] provides 37,667 full-face images from 15 participants recorded during everyday laptop use over several months, giving wide, natural variation in illumination, distance and head pose. Columbia Gaze [18] contains 5880 high-resolution studio photographs of 56 subjects captured on a discrete grid of five head poses and 21 gaze directions. RT-GENE [19] offers free-viewing imagery in natural environments at large camera-to-subject distances, with ground-truth gaze obtained from mobile eye-tracking glasses that are subsequently removed by semantic inpainting; we use its 13 official evaluation sessions (s001–s013; 92,595 samples), which we re-split into our own subject-disjoint partitions rather than reproducing the dataset’s official three folds. Re-using all three lets us probe reliability across in-the-wild, controlled-grid and free-viewing conditions with a single model before any downstream decision is made.

2. Related Work

2.1. Appearance-Based Gaze Estimation and Cross-Dataset Generalization

Appearance-based gaze estimation maps face or eye images to gaze direction with deep networks, and the field is well surveyed and benchmarked. The review and benchmark of Cheng et al. systematizes feature extraction, network design, calibration, and platforms, and provides reference implementations, while explicitly warning that inconsistent pre/post-processing and 2D-versus-3D evaluation make cross-method and cross-dataset comparisons unfair [1]. That review codifies standard within-dataset benchmarking and names the cross-dataset comparison pitfall; we operationalize one such pitfall by contrasting a naive sample-wise split, in which subjects overlap across train, validation, and test, with a subject-disjoint split under which the unseen-subject error more than doubles (Section 4.1), and then go beyond accuracy entirely to a per-user trust decision this review does not address. The large controlled benchmarks the community relies on, ETH-XGaze (>1 M images, 110 participants, extreme head pose, standardized protocol) [20], deliberately push estimator robustness; we instead use only three smaller webcam-scale datasets and, rather than chase extreme-pose accuracy, audit pose sensitivity, quantified as the absolute error-vs-pose slope $|\beta|$, as one of four trust criteria.

The dominant response to the cross-dataset domain gap is to build better, more generalizable estimators. UniGaze applies masked-autoencoder self-supervised pre-training on large-scale in-the-wild faces with a ViT backbone and fine-tunes for gaze, substantially improving leave-one-dataset-out and joint-dataset generalization while reducing reliance on labeled gaze data [8]. PureGaze purifies the gaze feature, removing gaze-irrelevant factors such as illumination, identity, and expression via a self-adversarial framework requiring no target-domain samples, reaching state-of-the-art cross-domain accuracy [9]. Cross-Dataset Evidential Inter-Intra Fusion trains a bespoke evidential architecture with NIG regressors and Mixture-of-NIG fusion to improve accuracy on both source and unseen domains while emitting uncertainty [10]. Our work is orthogonal to all three: we make no attempt to close the gap with a better model or new features. We accept a fixed, off-the-shelf GazeNet, quantify the gap as subject-level leakage optimism and per-user error spread (Section 4.1), and convert per-user reliability into a trust/recalibrate/abstain decision. The closest competitor here is the evidential method [10]: it and we both attach uncertainty to cross-dataset gaze, but it builds a custom architecture to boost accuracy, whereas we attach training-free MC-dropout to a frozen model, characterize it, and consume it as the highest objective-weight MCDM criterion rather than to improve the estimator.

Our three source datasets anchor the audit. MPIIGaze/MPIIFaceGaze provides real-world laptop use imagery and the person-independent paradigm we extend [17]; we re-use it not to introduce data but to quantify leakage optimism and measure subject-stratified reliability. Columbia Gaze contributes studio images on a discrete five-pose grid [18]; its degenerate head pose metadata (yaw = roll = 0) required recovering horizontal pose from the filename tokens, and it is among our harder domains, directly motivating abstain/recalibrate rather than a single global accuracy claim. RT-GENE supplies natural-environment gaze with wide pose ranges [19]; we incorporate its 13 official evaluation sessions into our own subject-disjoint splits rather than reproducing the official folds, making it our hardest domain and directly illustrating the per-subject reliability gaps our scaffold is built to surface.

2.2. Fairness, Uncertainty, and Selective Prediction in Gaze

A growing body of gaze work audits disparity. Investigations of bias and fairness in appearance-based gaze reveal significant performance gaps across ethnicity and gender and find existing mitigation strategies of limited effectiveness, with companion work document-

ing age bias and the absence of age demographics in public datasets [4]. These works audit across coarse demographic attributes but stop at measurement: they produce audit metrics only and recommend no per-user action. Our work also audits disparity (Section 4.2), but it stratifies by subject and head pose rather than demographic labels and crucially feeds the reliability evidence into a multi-criteria per-user trust/recalibrate/abstain decision, addressing one aspect of the decision integration gap identified by these audits. Subject debiasing via stochastic subject-wise adversarial learning (SAZE) instead attacks variation in training time, adversarially removing identity to raise average cross-subject accuracy [21]; this yields a single point estimate and no per-user reliability signal, whereas we treat subject identity as the unit of audit in a retrospective per-user audit and convert per-subject reliability into an action rather than erasing subject variation.

Uncertainty in gaze has been studied chiefly as something to calibrate or suppress. One line improves the calibration of uncertainty estimates, arguing probabilistic gaze models overfit and build biased uncertainties, and proposing strict coverage-based evaluation and a calibration correction [11]; this produces a better-calibrated scalar but issues no decision. SUGE estimates per-sample uncertainty from triplet-label consistency and then suppresses it via reweighting and label correction to reach state-of-the-art accuracy [12], treating uncertainty as a nuisance and a means rather than an output. We do the opposite: we preserve and surface per-subject uncertainty as a reliability signal, assess it against actual error, and turn it into an explicit trust/recalibrate/abstain action. Our uncertainty estimator follows the foundational Monte Carlo dropout approximation to Bayesian inference [22], which lets us run $T = 20$ stochastic passes through the existing dropout (0.3/0.2) with no retraining; we prefer this to deep ensembles, which require training multiple models [23], because instrument auditing should not demand re-training, and the cheaper signal still offers a partial reliability cue and becomes our highest objective-weight criterion, driven by its cross-user dispersion.

The closest competitor in spirit is selective prediction in gaze. Factor-Informed Uncertainty Distillation (FIUD) distills interpretable image quality factors into a lightweight uncertainty head and uses the resulting scalar to reject high-uncertainty frames, an explicit action [13]. This is genuinely action-producing, but FIUD acts per-frame from a single learned uncertainty signal; our work acts per-user, aggregating subject- and pose-stratified reliability into a multi-criteria decision (four criteria, CRITIC/entropy weights, and TOPSIS), yielding a richer trust/recalibrate/abstain governance decision that is stable under $\pm 20\%$ weight perturbation (0% action change). The reject option machinery itself rests on selective classification foundations: SelectiveNet learns abstention end-to-end inside the predictor [14], and the risk–coverage trade-off and reject option were formalized by El-Yaniv and Wiener [24]. We adapt the coverage-versus-reliability idea but keep the gaze regressor fixed and compute abstention externally as one action of a per-user MCDM layer driven by stratified reliability, rather than learning a coverage head or drawing a global classifier coverage curve.

2.3. MCDM/MCDA and Decision Support in Health

Our trust layer is built from established multi-criteria decision methods. TOPSIS ranks alternatives by relative closeness to a positive-ideal and distance from a negative-ideal solution in weighted normalized space [25]; we adopt it as the aggregation step, mapping each user's closeness score to a trust/recalibrate/abstain action rather than ranking design alternatives as originally intended. VIKOR is the canonical compromise-ranking alternative in the same closeness-to-ideal family [26]; we cite it as the established counterpart but choose TOPSIS for its straightforward distance-to-ideal score that maps cleanly to a three-way action. Objective weights come from CRITIC, which derives weights from

each criterion's contrast intensity and inter-criteria conflict [27], averaged with Shannon entropy weights [28]; this makes uncertainty the dominant criterion, with weights that are data-driven rather than analyst-chosen, and the resulting actions are stable under $\pm 20\%$ weight perturbation (0% change). The way we couple a trained model to MCDM follows the sequential ML-to-MCDM integration mode of Kocaman and Asan's taxonomy [29]; we explicitly avoid a bolted-on, decorative ranking by deriving objective weights from the reliability data itself and stress-testing action stability.

MCDM/MCDA is well established in health decision support, but at a coarser granularity than ours. Drissi et al. propose an AHP-based MCDA framework to select deep-learning models in healthcare over criteria such as accuracy, sensitivity, specificity, and complexity, validated on COVID-19 chest X-ray model comparison [15]; both they and we reject single-metric model judgment, but they rank candidate models globally, producing one decision for all patients, whereas we make a per-user, retrospective reliability decision over operational reliability signals, and we use objective CRITIC + entropy weights rather than subjective AHP pairwise judgments. The fuzzy MCDM evaluation of AI-driven mental health solutions by Ojo et al. applies fuzzy TOPSIS and fuzzy ARAS over criteria like feasibility, scalability, and clinical impact, ranking which solution to adopt at the program level under expert-elicited fuzzy judgments [16]; we share the TOPSIS family machinery and the personalization emphasis, but our decision object is how reliable a specific webcam gaze sensor is for a specific user, driven by empirical reliability measurements rather than expert fuzzy ratings, and we make no clinical claim. The unifying point against both closest competitors is granularity: MCDM-in-health papers operate at the model or cohort level, never at per-user-sensor granularity, which is exactly what we add.

The broader healthcare MCDA literature also supplies a methodological warning that we answer directly. The updated systematic review by Gongora-Salazar et al. finds the value measurement approach (rating scales and AHP scoring/weighting) most common, but reports substantial under-reporting of methods and high within-context variation in criteria selection [30]. We respond with objective data-driven weights (CRITIC + Shannon entropy), fully reported criteria and a TOPSIS procedure, and explicit sensitivity analysis via action-stability under weight perturbation. This rigor connects to the deficit our work targets: DIDI names and measures the prediction-to-decision gap at the architecture level [2], and we operationalize a concrete remedy for one instrument; FUTURE-AI sets the trustworthy-deployable principles [3], several of which we instantiate empirically (robustness via subject-disjoint re-splitting that quantifies leakage optimism, universality/fairness via per-subject spread, and usability via an auditable per-user decision).

2.4. Data Leakage and Subject-Independent Evaluation

The justification for our leakage audit comes from the reproducibility literature. Kapoor and Narayanan surveyed 17 scientific fields, found data leakage affecting 294 papers and often producing wildly over-optimistic results, defined a taxonomy of eight leakage types, and proposed model info sheets to detect and prevent them; in a civil war prediction case study, they claimed that ML superiority vanished once leakage was corrected [5]. As a deliberate reference baseline, we construct a naive sample-wise split that shuffles samples rather than subjects, so all 84 subjects appear in train, validation, and test (zero test-only subjects); this is a concrete instance of train-test contamination in appearance-based gaze regression rather than classification, and, relative to our subject-disjoint re-split, it produces the substantial leakage optimism quantified in Section 4.1. We additionally honor public-data reuse reporting norms by reporting both the naive and the subject-disjoint numbers rather than only the optimistic figure.

The principle that strictly subject-disjoint evaluation is the only valid generalization metric for biometric-style sensing is well argued in adjacent fields. A survey of cross-subject EEG decoding holds that keeping train/test disjoint at the subject level is the only valid metric of utility on unseen subjects, noting subject-dependent protocols inflate accuracy dramatically (e.g., >95% vs. ~60%) by leaking identity [6]. We operationalize exactly this for webcam gaze with a 55/12/17 subject-disjoint split (zero overlap), exposing the large seen-versus-unseen gap and substantial per-subject spread (Sections 4.1 and 4.2) that subject-dependent splits hide. The leave-one-subject-out precedent from affective computing further motivates testing on never-seen subjects [7], and we extend it with a finding specific to gaze: even a generalizing model needs per-user reliability auditing, because bias-only recalibration yields only a small overall gain and does not help the least-reliable users, our error being largely variance- rather than bias-dominated, unlike the subject bias correction common in EEG.

Finally, our paradigm and one of our datasets descend directly from the person-independent appearance-based gaze line of Zhang et al., which introduced MPIIGaze and CNN-based gaze, evaluating cross-dataset and leave-one-person-out to measure generalization to unseen users without person-specific calibration [1]. That work reports a single cross-person mean error; we extend it from one average number to a subject- and pose-stratified reliability assessment (per-subject MC-dropout uncertainty partially associated with error; head pose sensitivity measured as the absolute error-vs-pose slope) feeding a per-user multi-criteria trust decision. Across all four subsections, the consistent thing we add over every closest competitor is the same: prior work audits, debiases, calibrates, or selects, and MCDM-in-health work ranks models or programs, but none converts a label-free, subject- and pose-stratified reliability audit into a per-user MCDM trust action that disagrees with a naive single-metric threshold for roughly a third of the users while remaining stable under weight perturbation. Table 1 summarizes this positioning against the closest prior art.

Table 1. Positioning against the closest prior art and the remaining contribution.

Prior Work	Overlap with the Present Study	Remaining Contribution of This Study
Gaze bias and fairness audits	Characterization of performance disparity in gaze estimation	Stratification by subject and head pose without demographic labels, mapped to an explicit per-user decision
Subject-wise debiasing	Label-free treatment of subject-to-subject variation	Conversion of measured per-user reliability into a multi-criteria decision rather than its removal at training time
Gaze uncertainty calibration	Uncertainty estimated on comparable data	An explicit decision layer (trust, recalibrate, and abstain) rather than a calibrated scalar alone
Selective prediction for gaze	Accept and reject actions	Per-user rather than per-frame granularity, aggregating four reliability criteria
MCDM model selection in health	Multi-criteria decision machinery	Per-user granularity over measured sensor reliability rather than model- or cohort-level selection

3. Materials and Methods

3.1. Datasets, Unification, and Head Pose Recovery

We use three public appearance-based gaze datasets and introduce no new data, in line with reporting expectations for secondary analyses of public repositories [31]. MPIIFaceGaze contributes 37,667 samples from 15 subjects captured during everyday laptop use [17]; Columbia Gaze contributes 5880 high-resolution samples from 56 subjects

spanning a discrete 5-pose by a 21-gaze-direction grid [18]; and RT-GENE contributes 92,595 samples from the 13 sessions of its official three-fold evaluation set (s001–s013), captured in natural environments with mobile eye-tracking ground truth [19]. The merged corpus contains 84 subjects and 136,142 samples. Figure 1 gives an overview of the full audit pipeline. Each unified sample exposes a full-face image, left/right eye crops, a gaze label as (pitch, yaw) in radians, and a head pose label as (pitch, yaw, roll) in radians.

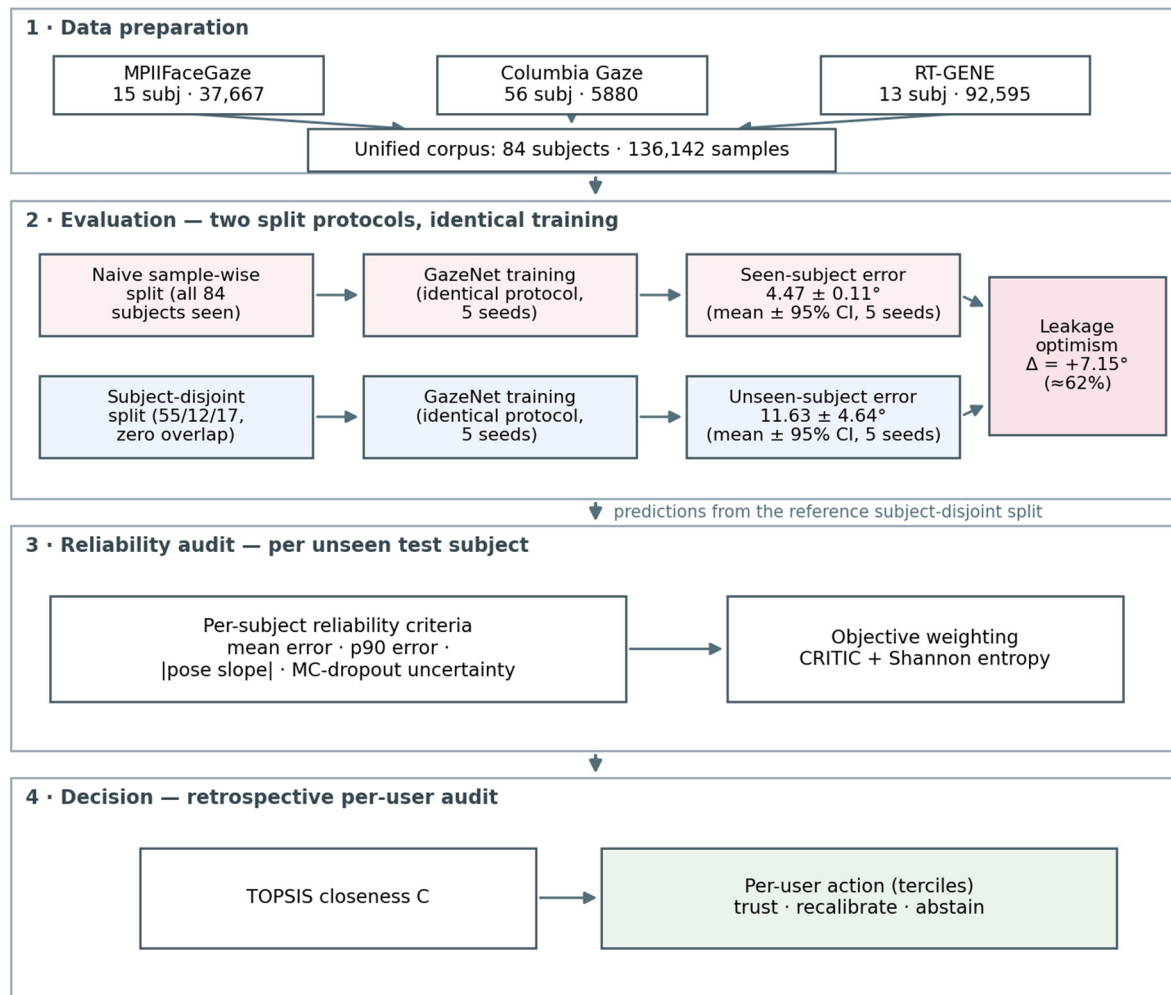


Figure 1. Functional overview of the pipeline in four stages. (1) Data preparation: three public gaze datasets are unified into an 84-subject, 136,142-sample corpus. (2) Evaluation: two parallel branches train GazeNet under an identical five-seed protocol—a naive sample-wise split in which all subjects are seen (red-tinted row, 4.47 deg) and a subject-disjoint split with zero subject overlap (blue-tinted row, 11.63 deg)—and their comparison yields the leakage optimism indicator (+7.15 deg, approximately 62%). (3) Reliability audit: per-subject criteria (mean error, p90 error, absolute pose slope, and MC-dropout uncertainty) are combined with objective CRITIC + Shannon entropy weights, computed from the reference subject-disjoint split. (4) Decision: TOPSIS closeness is discretized by terciles into the per-user actions trust, recalibrate (a candidate for further calibration), or abstain (green box).

We evaluate the corpus under two complementary partitions, following the established principle that only strictly subject-disjoint evaluation measures generalization to new users [5–7]. As a deliberate leaky reference baseline, we construct a naive sample-wise split that shuffles samples rather than subjects; because images of the same person are distributed across folds, all 84 subjects appear simultaneously in train, validation, and test, so this split measures memorization of seen identities rather than generalization to

unseen users. We keep it only to quantify the resulting optimism relative to the subject-disjoint split of Section 3.3, in which no subject appears in more than one fold. Both partitions are redrawn and retrained over the same five seeds, so the seen- and unseen-subject errors are estimated symmetrically. Two dataset-specific notes apply. For RT-GENE, we re-split the 13 evaluation sessions s001–s013 into our own subject-disjoint partitions rather than reproducing the official three-fold protocol; the official validation sessions s014–s016 (of which s014 and s015 are repeat recordings of participants 002 and 005) are not used. For Columbia, the supplied head pose field carries the horizontal head rotation in its first component, while the yaw and roll fields are degenerate (both zero), so we take the head pose from that first component and verify it against the filename tokens, as described next.

To confirm the Columbia head pose geometry, we parsed the canonical filename tokens, in which the P token encodes the head pose (horizontal rotation), and the remaining tokens encode gaze direction. All 5880 filenames parse consistently and reproduce the published capture design of 5 head poses by 21 gaze directions, that is, 105 images per subject across the 56 subjects. The angle recovered from the signed P token, $\theta_P = (\pi/180) \times P_{\text{token}}$, agrees with the supplied head pose first component to within 3.6×10^{-15} rad; therefore, we use that first component as the horizontal head pose and discard the degenerate yaw = roll = 0 fields. This matters because head pose is later used both as a model input and as the independent variable of our pose sensitivity reliability criterion.

$$N = N_{MPII} + N_{Col} + N_{RT} = 37,667 + 5880 + 92,595 = 136,142$$

$$\theta_P = \frac{\pi}{180} P_{\text{token}}$$

3.2. Gaze Estimation Model and Training Protocol

The estimator is GazeNet, an off-the-shelf multi-branch convolutional regressor with approximately 11.6 M parameters; we deliberately do not modify the architecture or its features, since our objective is instrument validation, rather than building a better estimator [8,9]. The network comprises a ResNet-18 face encoder, two shared-weight eye encoders applied to the left and right eye crops, and a head pose fusion module that concatenates the encoded face, eye, and head pose features before a multilayer gaze head that regresses the 2D gaze direction $\hat{g} = (\hat{\text{pitch}}, \hat{\text{yaw}})$. The gaze head contains two dropout layers with rates of 0.3 and 0.2; retaining these active at inference enables Monte-Carlo dropout uncertainty with no retraining and no architectural change [22].

The model is trained with an L1 (mean absolute error) loss on the angular components, which is robust to the heavy-tailed label noise present in webcam gaze data. Optimization uses AdamW with a learning rate of 1×10^{-4} and weight decay of 1×10^{-4} , a cosine-annealing learning rate schedule, gradient clipping at a global norm of 5, and 10 training epochs. Each training batch holds 128 samples. The face input is a 224×224 pixel crop, and each eye input is a 64×64 pixel crop, normalized with ImageNet channel statistics; light photometric augmentation is applied to the training stream only (color jitter with a brightness of 0.2, a contrast of 0.2, and a saturation of 0.1 on the face crop and a brightness of 0.15 and a contrast of 0.15 on the eye crops, plus random grayscale with a probability of 0.05). The ResNet-18 face encoder is initialized from ImageNet-pretrained weights (IMAGENET1K_V1), while the eye encoders, fusion module, and gaze head start from PyTorch's default Kaiming-uniform initialization. The naive sample-wise split allocates 80/10/10% of samples (108,913 training, 13,614 validation, and 13,615 test samples); under the subject-disjoint protocol, the per-fold sample counts vary across seeds because subjects contribute unequal numbers of samples. The identical optimizer, schedule, regularization,

batch size, and epoch budget are reused for both splits, so any difference in measured error is attributable to the data partition rather than the training recipe. All training and evaluation used PyTorch v2.10.0 under Python 3.11.9.

Reliability is reported in degrees of 3D angular error. Each (pitch, yaw) prediction is converted to a unit gaze direction vector in Cartesian coordinates, and the angular error is the angle between the predicted and ground-truth direction vectors via the arccosine of their dot product. This 3D angular metric is the field-standard evaluation and is invariant to the choice of 2D parameterization [1,20].

$$\mathcal{L} = \frac{1}{N} \sum_{i=1}^N (|\hat{p}_i - p_i| + |\hat{y}_i - y_i|)$$

$$d(p, y) = (-\cos p \sin y, -\sin p, -\cos p \cos y)$$

$$\theta_{err} = \arccos \left(\frac{\hat{d} \cdot d_{gt}}{\|\hat{d}\| \|d_{gt}\|} \right) \cdot \frac{180}{\pi}$$

3.3. Subject-Disjoint Evaluation Protocol and Leakage Optimism Definition

Because a webcam gaze sensor is a biometric-style instrument, the only valid measure of utility for a new user is performance on subjects never seen in training [6,7]. Alongside the naive reference baseline of Section 3.1, we construct a subject-disjoint partition of the 84 unified subjects into train/validation/test groups of 55/12/17 subjects with zero subject overlap across the three folds. The identical training protocol of Section 3.2 is applied. To quantify stochastic variability, the re-split and training are repeated over 5 random seeds, and we report the unseen-subject error as mean plus or minus a 95% confidence interval computed with Student's t-distribution.

We formally define leakage optimism as the difference between the unseen-subject error obtained under the subject-disjoint protocol and the seen-subject error obtained under the naive sample-wise split, holding the model and protocol fixed. On the naive split, GazeNet attains 4.47 deg on seen subjects (SD: 0.09 deg; 95% CI: ± 0.11 deg), averaged over five sample-wise partitions in which both the split and the training seeds are redrawn (seeds 1–4 and 42; per-seed: 4.42, 4.61, 4.52, 4.42, and 4.40 deg). Under the subject-disjoint protocol, the identical five-seed procedure yields an unseen-subject error of 11.63 deg (SD: 3.74 deg; 95% CI: 6.99 to 16.26 deg), giving a leakage optimism of +7.15 deg (computed from the unrounded five-seed means; the difference of the rounded means is 7.16 deg); the naive protocol, thus, underestimates the unseen-subject error by approximately 62% [5]. The two estimates are symmetric—five seeds per protocol—yet behave very differently: the naive estimate is tight across seeds because every subject is seen in training, whereas the subject-disjoint estimate varies widely (7.29 to 15.22 deg) because each redraw changes which few subjects are held out. Every subject-disjoint seed exceeds every naive seed. This is a concrete instance of subject-level train–test contamination in appearance-based gaze regression.

For all detailed per-subject and per-pose analyses, we fix a single reference split (seed 42), for which the sample-mean unseen error is 13.93 deg and the subject-macro error (the mean over per-subject means, which equally weights each user) is 11.79 deg. Reported as sample-weighted means, the per-domain unseen-subject error is 7.71 deg for MPIIFaceGaze, 12.50 deg for Columbia, and 15.69 deg for RT-GENE, whereas the overall 11.79 deg is a subject-macro mean; RT-GENE and Columbia are, thus, the harder domains, motivating per-user decisions rather than a single global accuracy claim [18]. This

subject-disjoint design directly instantiates the robustness and universality principles of trustworthy-AI guidance for deployable systems [3].

$$\Delta_{opt} = \theta_{unseen} - \theta_{seen}$$

$$\bar{\theta} \pm CI_{95} = \bar{\theta} \pm t_{0.975, S-1} \frac{s}{\sqrt{S}}$$

3.4. Per-Subject and Per-Pose Reliability Criteria

From the reference split we compute, for every test subject m , four reliability criteria together characterize per-user reliability beyond a single mean. The first is the per-subject mean angular error $e_{\text{mean}}(m)$, the average of θ_{err} over the n_m samples of subject m . The second is the per-subject 90th-percentile error $e_{\text{p90}}(m)$, the empirical quantile capturing tail (worst-case) behavior that a mean conceals; per-subject means range from a best of 5.16 deg to a worst of 25.16 deg, with a $4.9\times$ spread, with a between-subject Gini coefficient of 0.245 [4].

The third criterion is head pose sensitivity, quantified per subject by the ordinary least-squares slope $\beta(m)$ of angular error against absolute head pose magnitude, where $\beta(m)$ is the covariance of pose magnitude and error divided by the variance of pose magnitude. Because a strong dependence of either sign signals pose fragility, the cost criterion is the absolute slope $|\beta(m)|$ (larger is worse). Throughout this paper, ‘pose-stratified’ refers to this continuous stratification of each subject’s error along the head-pose-magnitude axis via the per-subject regression slope, rather than to discrete pose bins. Across subjects, the absolute slope has a median of about 0.13 deg per degree but reaches roughly 1.0 deg per degree for the most pose-sensitive users [20]. The fourth criterion is the per-subject MC-dropout predictive angular spread: with dropout active, we draw $T = 20$ stochastic forward passes per sample, average them to a mean direction, and define the sample uncertainty as the mean angular deviation of the T passes from that mean; the per-subject uncertainty $u(m)$ is the average over the subject’s samples [22]. This training-free uncertainty is only borderline associated with actual error across the 17 subjects: the Pearson correlation between $u(m)$ and $e_{\text{mean}}(m)$ is $r = 0.48$ (asymptotic $p = 0.051$; percentile-bootstrap 95% CI: [0.00, 0.92]) and the Spearman correlation is $\rho = 0.50$ (asymptotic $p = 0.041$; bootstrap 95% CI: [−0.01, 0.87]). Because the Spearman bootstrap interval includes zero, we treat the association as borderline rather than statistically significant. Its main role is not to track the magnitude of the error but to separate users whose predictions are unstable under MC-dropout from users whose error is large yet stable; subject s001, for instance, has the largest mean error (25.16 deg) with low uncertainty (1.79 deg), whereas subject 0017 pairs a comparable mean error with high uncertainty (22.28 deg). We treat it, therefore, as a partial reliability signal [10,11].

To test whether the residual per-user error can be removed by lightweight personalization, we evaluate four calibration schemes, fitting each on k randomly drawn calibration samples per subject and evaluating on that subject’s remaining held-out samples, averaged over 200 random draws. The schemes are bias, which subtracts the mean calibration residual, $p' = p - \text{mean}(p_{\text{cal}} - g_{\text{cal}})$; scale, $p' = \mu_g + s(p - \mu_p)$ with s the calibration regression slope clipped to [0,3]; affine, a per-axis fit $p'_j = a_j p_j + b_j$ by least squares; and linear2d, a full two-dimensional map $p' = [p, 1] W$ by least squares. A bias-only (constant-offset) correction from $k = 9$ calibration samples lowers the subject-macro error only marginally, from 11.79 to 11.50 deg, indicating the residual error is variance-dominated rather than bias-dominated [7,21]. This motivates an abstain action, alongside trust and recalibrate, for users whose error a simple offset cannot fix, in the spirit of se-

lective prediction [13,14,24]; the full per-scheme and per-k calibration results are reported in Tables S3 and S4.

$$e_{mean}(m) = \frac{1}{n_m} \sum_{i \in m} \theta_{err,i}$$

$$e_{p90}(m) = Q_{0.90}\{\theta_{err,i} : i \in m\}$$

$$\beta(m) = \frac{\sum_i (x_i - \bar{x})(\theta_i - \bar{\theta})}{\sum_i (x_i - \bar{x})^2}$$

$$\bar{g}_i = \frac{1}{T} \sum_{t=1}^T \hat{d}_i^{(t)}, u_i = \frac{1}{T} \sum_{t=1}^T \arccos(\hat{d}_i^{(t)} \cdot \bar{g}_i)$$

$$G = \frac{\sum_a \sum_b |e_a - e_b|}{2M^2 \bar{e}}$$

3.5. Multi-Criteria Per-User Trust Decision Layer

We aggregate the four reliability criteria into a per-user decision using a sequential machine learning-to-MCDM design, in which reliability metrics from the trained model feed an objective multi-criteria decision layer rather than a decorative post hoc ranking [15,29]. The decision matrix X has $M = 17$ rows (one per test user) and four columns, with all cost criteria (lower is better): mean error e_{mean} , p90 error e_{p90} , absolute pose-sensitivity slope $|\beta|$, and MC-dropout uncertainty u . Columns are vector-normalized so that each criterion contributes on a comparable scale. Because the mean error and p90 columns are computed from ground-truth gaze already collected for each user, the resulting decision is a retrospective per-user audit rather than a deployment-time decision for an unseen user.

Objective weights are derived by two complementary data-driven schemes and averaged, avoiding analyst-chosen weights that undermine reported MCDA in health applications [30]. The Shannon entropy method first converts each normalized column into a probability vector p_{ij} , computes the entropy E_j , and assigns weight proportional to the divergence $(1 - E_j)$, so that criteria with greater variability receive larger weight [28]. The CRITIC method weights each criterion by the product of its contrast intensity (standard deviation σ_j) and its conflict with other criteria (the sum over k of one minus the Pearson correlation r_{jk}), and then normalizes [27]. The two weight vectors are averaged element-wise. The resulting weights are an uncertainty of 0.422, an absolute pose slope of 0.232, a p90 error of 0.175, and a mean error of 0.170; the dominant weight, thus, falls on MC-dropout uncertainty [11].

The weighted normalized matrix is then ranked by TOPSIS [25,26]. Because all criteria are cost-type, the positive-ideal solution A^+ is the per-column minimum of the weighted values, and the negative-ideal A^- is the per-column maximum. For each user, we compute the Euclidean separation d^+ to A^+ and d^- to A^- , and the closeness coefficient $C_i = d^- / (d^+ + d^-)$ in $[0,1]$, where higher C_i denotes a more reliable user. Users are mapped to actions by terciles of C_i : the top tercile receives trust, the middle recalibrate, and the bottom abstain, yielding a 5/6/6 split over the 17 users. The tercile boundaries discretize the continuous closeness score into three governance actions; they are not calibrated operating thresholds, so at this pilot scale, the counts illustrate the mechanism rather than fixing deployment cut-offs [14,24]. The recalibrate label designates a candidate for further calibration or adaptation—a user whose composite profile makes calibration worth attempting—not a prediction that calibration will improve that user; it is assigned by tercile rank, not by a validated improvement criterion.

We benchmark this against a naive single-metric baseline that terciles users on mean error alone, and against weight perturbation sensitivity. The MCDM action disagrees with the naive mean error threshold for 6 of 17 users (35%), showing that the multi-criteria signal materially changes per-user governance relative to accuracy alone. Perturbing each of the four weights

one at a time by plus or minus 20% (eight scenarios, with the weights renormalized after each perturbation) leaves the assigned action unchanged in all cases (0 of 136 decisions), indicating the entropy/CRITIC weighting yields stable actions [28,29]. We make no clinical claim; the contribution is instrument validation plus a decision scaffold that provides an exploratory implementation of the prediction-to-decision framework [2,3,16,32].

$$r_{ij} = \frac{x_{ij}}{\sqrt{\sum_{a=1}^M x_{aj}^2}}$$

$$E_j = -\frac{1}{\ln M} \sum_{i=1}^M p_{ij} \ln p_{ij}, w_j^{ent} = \frac{1 - E_j}{\sum_k (1 - E_k)}$$

$$w_j^{CRITIC} = \frac{\sigma_j \sum_k (1 - r_{jk})}{\sum_l \sigma_l \sum_k (1 - r_{lk})}$$

$$w_j = \frac{1}{2} (w_j^{ent} + w_j^{CRITIC})$$

$$v_{ij} = w_j r_{ij}, A_j^+ = \min_i v_{ij}, A_j^- = \max_i v_{ij}$$

$$d_i^+ = \sqrt{\sum_j (v_{ij} - A_j^+)^2}, d_i^- = \sqrt{\sum_j (v_{ij} - A_j^-)^2}$$

$$C_i = \frac{d_i^-}{d_i^+ + d_i^-}$$

4. Results

4.1. Subject Leakage Substantially Underestimates Unseen-Subject Error

Our naive sample-wise split—shuffling individual samples rather than subjects, so all 84 subjects appear in train, validation, and test—is a deliberate reference baseline; under it, the model attains 4.47 deg on seen subjects (SD: 0.09 deg; 95% CI: ± 0.11 deg), averaged over five sample-wise partitions redrawn and retrained per seed (per-seed: 4.42, 4.61, 4.52, 4.42, and 4.40 deg). On subject-disjoint unseen subjects, the identical five-seed protocol yields 11.63 deg (SD: 3.74 deg; 95% CI: ± 4.64 deg), so the naive split underestimates unseen-subject error by +7.15 deg (~62% of the unseen error) (Figure 2). The two protocols separate cleanly: the largest naive seed (4.61 deg) lies far below the smallest subject-disjoint seed (7.29 deg). The subject-disjoint estimate itself is unstable across seeds (7.29, 13.81, 7.88, 15.22, and 13.93 deg; 95% CI: 6.99 to 16.26 deg), a width driven by the small number of held-out subjects per fold that we return to in Section 5. This is consistent with documented leakage effects in ML-based science [5] and with the need for subject-independent evaluation [6].

4.2. Reliability Is Highly Non-Uniform Across Users

The per-subject error ranges from 5.16 deg to 25.16 deg, a $4.9\times$ spread, with a between-subject Gini of 0.245 (Figure 3); the subject-macro-averaged unseen error is 11.79 deg, and head pose sensitivity is mild for most users (median absolute slope: 0.130 deg/deg; Table 2) but pronounced for a minority. The per-domain unseen-subject error is 7.71 deg for MPIIFaceGaze, 12.50 deg for Columbia Gaze, and 15.69 deg for RT-GENE; these per-domain values are sample-weighted means, whereas the overall figure is a subject-macro mean (the mean of per-subject means, weighting each user equally), so the two are not directly comparable. The RT-GENE left/right-eye disagreement is small (0.17 deg) and is used only as an auxiliary consistency check.

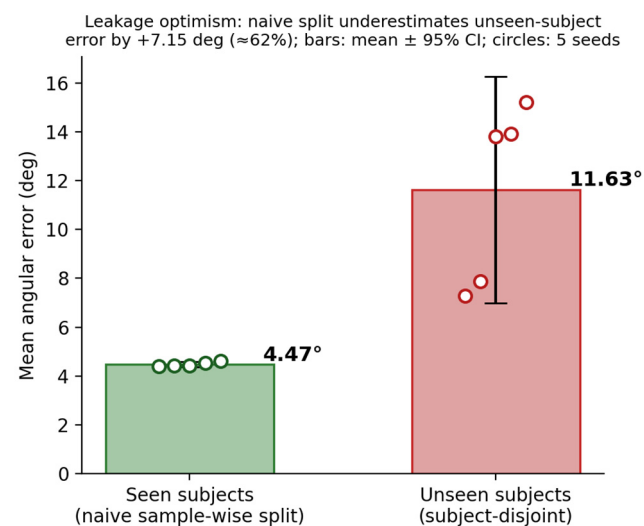


Figure 2. Leakage optimism: seen-subject error on the naive sample-wise reference splits versus unseen-subject error on subject-disjoint splits under an identical protocol (bars: five-seed means $\pm$ 95% CI; open circles: the five individual seed values of each protocol, showing the tight naive spread and the wide subject-disjoint spread).

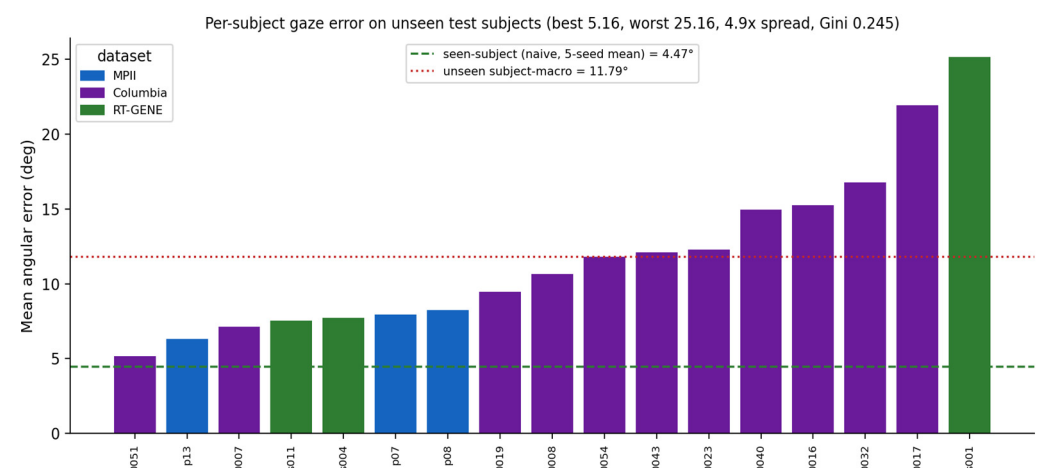


Figure 3. Per-subject mean gaze error on unseen test subjects (blue: MPIIFaceGaze, purple: Columbia Gaze, and green: RT-GENE). Dashed line: seen-subject error; dotted line: unseen subject-macro mean.

Table 2. Per-user MCDM trust decisions on unseen subjects, sorted by TOPSIS closeness (all criteria are cost criteria; lower is better).

Subject	Mean Err. (deg)	p90 Err. (deg)	Pose Slope	Uncertainty (deg)	TOPSIS C	Action
p13	6.30	10.89	0.0547	1.85	0.968	Trust
p07	7.93	14.04	0.0077	1.59	0.955	Trust
s004	7.71	13.39	0.0605	1.48	0.952	Trust
s011	7.53	15.09	0.0100	1.95	0.951	Trust
0007	7.14	12.54	0.0830	2.25	0.943	Trust
p08	8.25	14.65	0.0825	1.72	0.937	Recalibrate
0051	5.16	9.58	0.0948	2.92	0.933	Recalibrate
0008	10.64	16.86	0.1333	1.92	0.896	Recalibrate

Table 2. Cont.

Subject	Mean Err. (deg)	p90 Err. (deg)	Pose Slope	Uncertainty (deg)	TOPSIS C	Action
0043	12.09	20.60	0.0786	2.48	0.879	Recalibrate
s001	25.16	47.60	1.0407	1.79	0.548	Recalibrate
0023	12.30	22.62	0.1299	16.67	0.498	Recalibrate
0054	11.81	27.21	0.5248	14.12	0.465	Abstain
0019	9.48	19.96	0.2392	20.10	0.425	Abstain
0016	15.24	37.06	0.6334	16.83	0.336	Abstain
0040	14.97	50.33	0.8528	17.17	0.252	Abstain
0032	16.76	42.77	0.6563	20.86	0.233	Abstain
0017	21.94	59.39	1.0276	22.28	0.043	Abstain

4.3. Uncertainty Flags Instability; Recalibration Helps Some Users but Not the Least Reliable

Monte Carlo dropout uncertainty is only a partial predictor of per-subject error (Pearson $r = 0.48$, asymptotic $p = 0.051$; Spearman $\rho = 0.50$, asymptotic $p = 0.041$; Figure 4); because the percentile-bootstrap interval for Spearman includes zero, we treat the association as borderline rather than firmly significant. It does not linearly track mean error but flags users with highly unstable predictions, a label-free instability signal. Bias-only recalibration (per-subject constant-offset subtraction from nine calibration samples) gives only a small overall improvement (subject-macro $11.79 \rightarrow 11.50$ deg) and helps the recalibrate group most ($12.27 \rightarrow 11.55$ deg) while leaving the least-reliable abstain group no better ($15.03 \rightarrow 15.23$ deg), indicating variance- rather than bias-dominated error; abstain or defer is, therefore, the meaningful action when reliability is low.

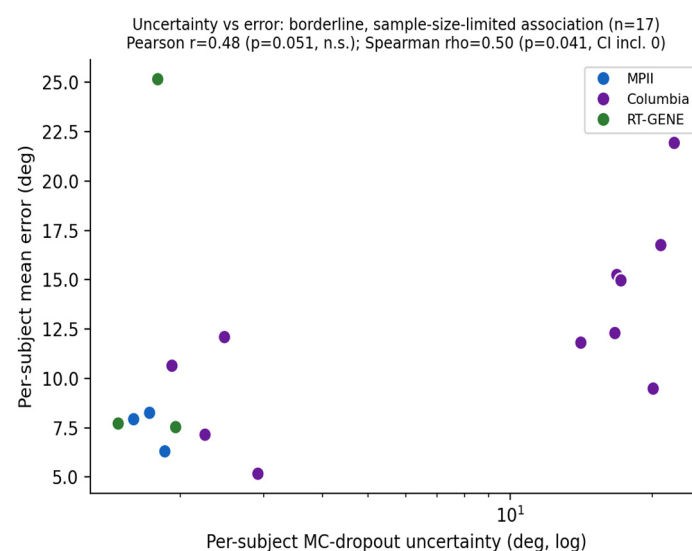


Figure 4. Per-subject Monte Carlo dropout uncertainty versus mean angular error (log x -axis). The association is borderline and sample-size-limited ($n = 17$; Pearson asymptotic $p = 0.051$; Spearman bootstrap 95% CI includes zero) and should not be read as a robust correlation.

4.4. A Multi-Criteria Per-User Trust Decision

Objective weighting assigns the largest weight to Monte Carlo dropout uncertainty (0.422), followed by absolute pose sensitivity (0.232), worst-case p90 error (0.175), and mean error (0.170); the pose criterion is the absolute pose sensitivity slope $|\beta|$, with larger values incurring higher cost. TOPSIS yields per-user actions (trust 5, recalibrate 6, and

abstain 6; Table 2). The MCDM decision disagrees with a naive mean error tercile threshold for six of 17 users (35%), and no decision changes under $\pm 20\%$ weight perturbation (0 of 136 decisions: each of the four weights perturbed one at a time across the 17 users, with eight scenarios per user), showing the layer is both consequential and stable (Table 3).

Table 3. MCDM trust decision summary on unseen subjects.

Metric	Value
Unseen subjects (test)	17
Actions (trust/recal/abstain)	5/6/6
Disagreement vs. naive threshold	6/17 (35%)
Action change under $\pm 20\%$ weights	0.0% (0/136)
Uncertainty–error correlation	0.48 ($p = 0.051$)

4.5. Qualitative Behavior and Operating Envelope

Figure 5 shows predictions qualitatively on unseen subjects. Good cases exhibit near-coincident ground-truth and predicted gaze vectors, whereas hard cases show clear divergence; the prediction scatter and the yaw-pitch error map indicate where the sensor is reliable. To comply with the datasets' NonCommercial licenses and subject privacy constraints, no source face images are reproduced; all panels are the authors' own plots computed on the datasets [17–19].

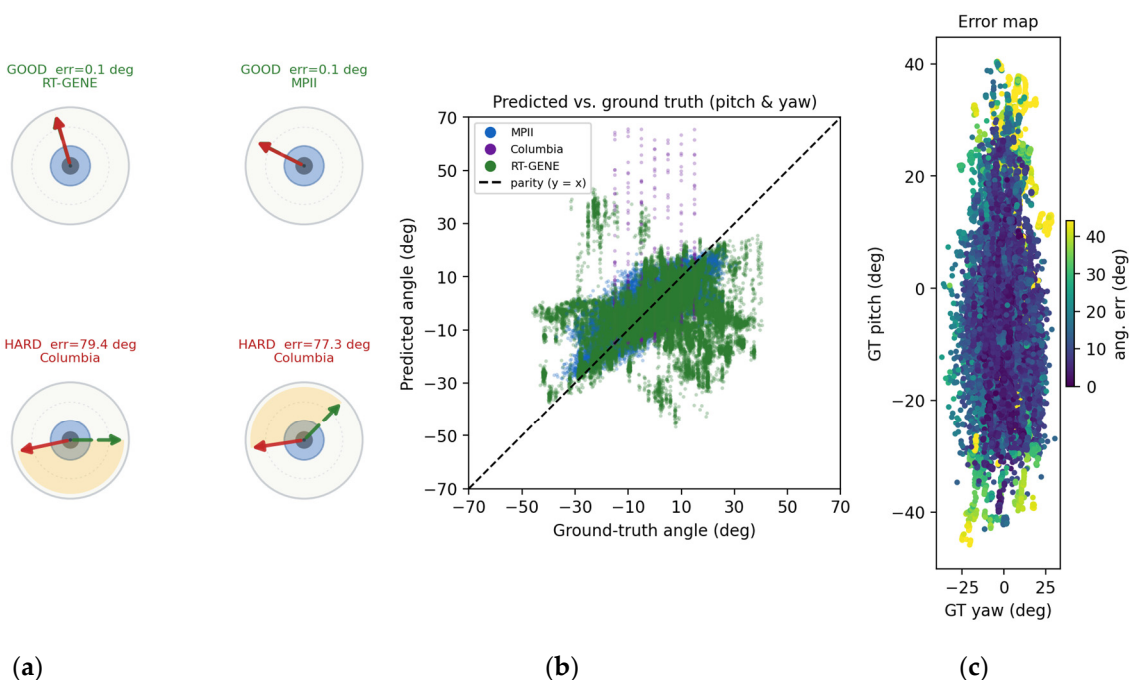


Figure 5. Qualitative results on unseen test subjects, shown as the authors' own plots (no dataset face images are reproduced, in compliance with the datasets' NonCommercial licenses and subject privacy). (a) Ground-truth (green dashed) versus predicted (red) gaze vectors for three good and three hard cases; (b) predicted versus ground-truth gaze angles (pitch and yaw); and (c) angular error map over the gaze direction space. Evaluated on MPIIFaceGaze, Columbia Gaze and RT-GENE.

5. Discussion

Our central empirical finding concerns how the evaluation split is constructed. A naive sample-wise split that shuffles individual samples rather than subjects places all 84 subjects simultaneously in train, validation, and test, leaving no unseen subjects; we use it as a delib-

erate reference baseline. On this naive split, GazeNet appears markedly more accurate than under an identical-protocol subject-disjoint re-split with zero subject overlap; the resulting leakage optimism, reported in Section 4.1, means the naive protocol substantially underestimates unseen-subject error. The comparison is symmetric, five seeds per protocol, and the separation is complete. Every subject-disjoint seed lies well above every naive seed. The optimism this produces is the train–test contamination that [5] catalogs across 294 papers in 17 fields; we reproduce a concrete instance in appearance-based gaze regression rather than the classification science it surveys. The subject-independent evaluation literature makes the same demand from the biometric side: [6] argues that strict subject-level disjointness is the only valid measure of utility to unseen users, reporting >95% vs. ~60% accuracy gaps from leaked identity, and the person-independent gaze paradigm itself was founded on leave-one-person-out evaluation [1]. We do not present this unseen-subject error as a state-of-the-art accuracy claim: GazeNet is small (~11.6 M params), and our purpose is instrument validation, so the number is a decision input rather than a leaderboard entry.

Our unseen-subject error should be read against the cross-dataset gaze literature, where stronger methods exist precisely to close the domain gap we instead audit. The study in [8] scales masked-autoencoder pre-training on a ViT backbone to improve cross-dataset, leave-one-dataset-out, and joint-dataset generalization; [9] purifies gaze features to remove illumination/identity/expression and reaches state-of-the-art cross-domain results with no target samples; and the review-and-benchmark of [1] codifies fair within- and cross-dataset comparison. Our work is deliberately orthogonal to all three. We make no attempt to shrink the gap with a better estimator; we accept a fixed off-the-shelf model and quantify the gap as subject-level leakage optimism and per-user error spread, as reported in Section 4. The per-domain unseen errors are sample-weighted means, whereas the overall figure is a subject-macro mean, so the two are not directly comparable. Where [1] notes cross-dataset comparison pitfalls, we operationalize one such pitfall end to end, contrasting the naive sample-wise and subject-disjoint splits, reporting the subject-disjoint error, and then going beyond accuracy to a per-user decision this review does not address.

The second, larger novelty contrast is that we convert reliability evidence into a per-user action, whereas the fairness, uncertainty, and selective prediction gaze literature largely stops at characterization or calibration. Bias/fairness audits such as [4] document significant disparities across ethnicity, gender, and age but produce only audit metrics and recommend no per-user action; we also audit disparity but stratify by subject and head pose rather than by coarse demographic labels, and we feed the result into a decision. Subject debiasing in [21] is a training-time mitigation that erases identity to raise average accuracy yet yields a single point estimate with no deployment signal; we instead treat subject identity as the unit of audit in a retrospective per-user audit. On the uncertainty side, [11] improves the calibration of an uncertainty scalar but issues no decision, and [12] suppresses uncertainty during training as a nuisance to gain accuracy, the opposite of our stance, which preserves and surfaces per-subject uncertainty as a reliability signal. The closest prior work is [13], which does output an action (selective accept/reject) via a learned image quality uncertainty head, but it acts per-frame on a single distilled signal, whereas we act per-user over four criteria. Likewise, the selective prediction foundations in [14,24] place abstention inside the predictor or in global coverage curves; we keep the regressor fixed and compute abstention externally as one action of a governance layer, adapting the reject option idea to per-user reliability auditing.

Our third contribution is the MCDM trust layer itself, which we argue materially shapes the decisions rather than being bolted on, the recurring critique of MCDM-plus-ML pipelines. The study in [29] taxonomizes five integration modes; we occupy the sequential ML-to-MCDM mode, and we guard against the decorative-ranking failure in three concrete

ways. First, the layer demonstrably changes decisions: against a naive single-metric (mean-error tercile) threshold, our TOPSIS action disagrees for over a third of users, so it is not a relabeling of accuracy. Second, the weights are data-driven, not analyst-chosen: averaging CRITIC [27] and Shannon entropy [28] weights makes uncertainty the dominant criterion (Section 4.4), with the weights emerging from the data rather than from subjective AHP-style pairwise judgments. Third, the decisions are stable: perturbing each of the four weights one at a time by $\pm 20\%$ and renormalizing (eight scenarios) changes none of the 136 resulting decisions (0%). This directly answers the field-level diagnosis of [30], which found healthcare MCDA dominated by under-reported, subjective value measurement weighting. We report the full criterion set, the TOPSIS procedure [25], objective weights, and an explicit sensitivity analysis. Compared with MCDA-in-health, which ranks candidate DL models globally for all patients [15] or selects AI mental-health programs under expert fuzzy ratings [16], our decision object is finer-grained and empirically grounded: how reliable one webcam gaze sensor is for one specific user. The weights, being objective and interpretable, also double as an explanation of why a user is judged more or less reliable; uncertainty and pose sensitivity dominate.

Fourth, the uncertainty–error coupling and the behavior of recalibration together justify abstain/defer as the appropriate action for the least-reliable users. Per-subject MC-dropout uncertainty ($T = 20$, exploiting the head’s 0.3/0.2 dropout with no retraining [22]) is only a partial predictor of per-subject mean error, with a bootstrap interval that includes zero, so the association is borderline (Section 4.3); it earns the highest objective weight not from this correlation but from its large cross-user dispersion under the entropy/CRITIC schemes, and a training-free signal still suffices in place of costlier deep ensembles [23]. Bias-only (constant-offset) recalibration—subtracting a per-subject bias estimated from nine calibration samples—improves the subject-macro error only modestly overall, mainly within the recalibrate group, while the abstain group is left no better (Section 4.3, Table S4). Richer few-shot calibration that adjusts scale or an affine map outperforms a pure offset from as few as five calibration samples (Table S3), consistent with variance- rather than bias-dominated error; yet even the best affine calibration reduces the abstain group only to 9.53 deg (from 15.04 deg), still above the trust group’s uncalibrated 7.33 deg, so calibration does not raise the least-reliable users to the trust level (these group-specific uncalibrated values from the $k = 36$ affine experiment differ slightly from the $k = 9$ bias-only baselines in Table S4 because the held-out evaluation samples vary with the calibration budget). This differs from the EEG/affective-computing tradition, where subject bias correction under LOSO is standard and effective [7]. For our gaze sensor, a high-uncertainty user is not rescued by an offset. The action semantics distinguish wrong-but-stable from wrong-and-unstable: the highest-mean-error user (s001, 25.16 deg) is assigned to the middle (recalibrate) tercile rather than abstain because its low uncertainty (1.79 deg) partly offsets its high mean error and pose sensitivity in the composite score, an assignment that follows from the tercile construction and should not be read as evidence that recalibration will necessarily improve this individual; recalibrate designates a candidate for further calibration or adaptation, not an expected benefit, whereas user 0017 pairs a comparably high mean error (21.94 deg) with high predictive instability (uncertainty: 22.28 deg) that an offset cannot fix and is assigned abstain; the abstain set is, thus, driven by predictive instability and pose sensitivity, not by mean error alone. This instantiates trustworthy-AI guidance in concrete terms: against [3]’s FUTURE principles, we address Robustness (the leakage optimism revealed by the subject-disjoint split), Universality/Fairness (the per-user disparity audit), and Usability (an auditable per-user decision). And it engages the prediction-to-decision gap named by [2]: rather than only diagnosing the decision integration deficit, we address it for one instrument by routing stratified reliability into an explicit per-user decision.

We are explicit about limitations, several of which bound the claims tightly. First and foremost, no health labels exist in any of the three public datasets; therefore, we make no clinical claim. The contribution is instrument validation plus an exploratory decision scaffold, and the special-issue “early detection” pillar is reframed as a preliminary step toward deployment evaluation, a prerequisite for, not a demonstration of, early detection. Second, the subject-disjoint evaluation rests on 17 unseen test subjects, which is a pilot scale; the wide 95% CI on the unseen-subject error reflects this, and the per-user action counts should be read as illustrative of the mechanism rather than as population statistics. Third, although the leakage comparison is symmetric (five naive and five subject-disjoint seeds under one protocol), the subject-disjoint estimate itself is unstable across seeds: its 95% CI spans 6.99 to 16.26 deg (SD: 3.74 deg) because each test fold holds only 17 subjects and a few subjects per domain. The direction of the optimism is unambiguous—every subject-disjoint seed exceeds every naive seed—but its magnitude should be read as indicative rather than as a precisely estimated constant. Fourth, the per-domain unseen errors are sample-weighted means and are, therefore, not directly comparable with the subject-macro overall; RT-GENE, drawing on the 13 sessions of its official evaluation set re-split into our own subject-disjoint partitions (three RT-GENE subjects per test fold), is the hardest domain. Fifth, all data were obtained from previously collected benchmark datasets rather than from a prospective real-webcam deployment; their capture regimes range from everyday laptop use (MPIIFaceGaze) and free-viewing natural environments (RT-GENE) to a controlled studio grid (Columbia’s discrete five-pose grid with degenerate head pose metadata (yaw = roll = 0), with horizontal pose recovered from filename tokens). A prospective deployment would introduce additional variability that our per-subject pose sensitivity criterion only partially anticipates. Sixth, the uncertainty criterion should be read as an instability flag rather than a calibrated error predictor: its association with mean error is borderline (Pearson $r = 0.48$, asymptotic $p = 0.051$; Spearman $\rho = 0.50$ with a percentile bootstrap interval that includes zero), and it earns its 0.422 weight from a cross-user dispersion that a minority of high-uncertainty users dominate; with only 17 users, this dispersion structure is a sensitivity worth flagging even though the action distribution was weight-stable. Seventh, and most important for interpretation, the trust decision is a retrospective per-user audit rather than a deployment-time decision on a brand-new user: two of the four MCDM criteria are each user’s actual mean and p90 error, which require the ground-truth gaze already collected for that user, so the layer characterizes reliability for users whose errors are known rather than predicting it for an unseen user. Finally, as a secondary analysis of public repositories, the study aligns with transparent reuse reporting expectations [31] by reporting the subject-disjoint error rather than the optimistic naive-split of 4.47 deg.

Future work follows directly from these limits. The most pressing need is scale: many more unseen subjects, multiple subjects per domain, and ideally additional in-the-wild webcam captures, so the per-user action distribution and the minority of high-uncertainty users can be characterized statistically rather than illustratively. Our few-shot analysis already indicates that variance-oriented calibration (adjusting scale or an affine map) outperforms a pure bias offset from as few as five calibration samples (Table S3); extending this toward online per-user adaptation and re-auditing it with the same MCDM layer—to test whether it moves users out of the abstain set—is a natural next step, recognizing that the best calibration we tried still did not raise the least-reliable users to the trust group’s uncalibrated level. The decision layer can be extended with additional criteria (e.g., temporal stability and image quality factors in the spirit of [13]) and benchmarked against compromise-ranking alternatives such as VIKOR [26] to verify action stability beyond TOPSIS. Stronger backbones or domain-generalizing estimators [8,9]

could be dropped in unchanged, since the scaffold is estimator-agnostic, letting us test whether a better model narrows the per-user spread or merely shifts the abstain threshold. A further step is to move from the present retrospective per-user audit toward a deployment-time variant that estimates reliability for a brand-new user without first collecting that user's ground-truth error. Most importantly, the path toward the special issue's health framing is to couple this audited, decision-integrated sensor to an actual labeled well-being task, at which point the trust/recalibrate/abstain decision can be evaluated for its downstream value—the formal trust assessment that XAI/CDSS work is shown to rarely perform [32]—rather than as instrument validation alone.

6. Conclusions

Using only public gaze datasets and no new data, we showed that the naive split underestimates unseen-subject error by approximately 62% (a five-seed symmetric comparison whose wide seed-to-seed interval we disclose), that reliability is strongly user-specific, and that an objective-weight MCDM layer turns a measured reliability profile into a per-user trust/recalibrate/abstain decision that was stable under the tested one-at-a-time $\pm 20\%$ weight perturbations, providing an exploratory decision scaffold for reliability-aware, low-cost gaze-based self-monitoring; the scaffold is validated only as a retrospective audit on public datasets, not in a real-world or health-related self-monitoring setting.

Supplementary Materials: The following supporting information can be downloaded at <https://www.mdpi.com/article/10.3390/app16157630/s1>: Table S1: Per-seed unseen-subject error under the subject-disjoint protocol; Table S2: Per-subject error with bootstrap 95% confidence intervals (reference split, seed 42); Table S3: Per-user calibration schemes as a function of the calibration budget k (subject-macro error, deg); Table S4: Effect of bias-only (constant-offset) recalibration within each MCDM action group (subject-macro, seed 42, 200-draw hold-out); Table S5: Association between per-subject MC-dropout uncertainty and per-subject mean error.

Author Contributions: Conceptualization, H.S. and K.P.; methodology, K.P. and H.S.; software, K.P. and S.A.; validation, K.P., S.A. and N.G.; formal analysis, K.P.; investigation, K.P. and S.A.; resources, H.S. and N.G.; data curation, K.P. and S.A.; writing—original draft preparation, K.P.; writing—review and editing, H.S., N.G. and S.A.; visualization, K.P.; supervision, H.S. and N.G.; project administration, H.S.; funding acquisition, H.S. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the Japan Science and Technology Agency (JST) through the FY2025 LOTUS Program (India–Japan Circulation of Talented Youths in Science Program) (grant number: 25152939).

Institutional Review Board Statement: Not applicable. This study analyzed only previously published, publicly available datasets and did not involve any new experiments on humans or animals.

Informed Consent Statement: Not applicable. No new human participants were involved; only previously published public datasets were analyzed, and no dataset images of identifiable persons are reproduced.

Data Availability Statement: This work used three publicly available gaze datasets—MPIIFaceGaze/MPIIGaze [17], Columbia Gaze [18] and RT-GENE [19]—obtained from their official sources and used under their respective non-commercial research terms; no new data were collected. The analysis code and derived result files that support the findings of this study are available from the corresponding author upon reasonable request. Third-Party Material and Licensing: The source datasets are distributed under NonCommercial terms (MPIIFaceGaze and RT-GENE under CC BY-NC-SA 4.0; Columbia Gaze under a non-commercial research grant) that are not compatible with this article's CC BY 4.0 license. Accordingly, no dataset face images are reproduced; all qualitative figures are the

authors' own plots computed on these datasets. Reproducing any source image would require written permission from the dataset rights holders and confirmation of subject consent for publication.

Acknowledgments: The authors thank the creators of the MPIIFaceGaze/MPIIGaze, Columbia Gaze and RT-GENE datasets for making their data publicly available for research.

Conflicts of Interest: The authors declare no conflicts of interest.

References

- Cheng, Y.; Wang, H.; Bao, Y.; Lu, F. Appearance-Based Gaze Estimation With Deep Learning: A Review and Benchmark. *IEEE Trans. Pattern Anal. Mach. Intell.* **2024**, *46*, 7509–7528. [\[CrossRef\]](#) [\[PubMed\]](#)
- Dadelo, S. From Prediction to Decision: The Decision Integration Deficit Index (DIDI) and Structural Imbalance in AI-Driven Digital Health Systems. *Appl. Sci.* **2026**, *16*, 3380. [\[CrossRef\]](#)
- Lekadir, K.; Frangi, A.F.; Porras, A.R.; Glocker, B.; Cintas, C.; Langlotz, C.P.; Weicken, E.; Asselbergs, F.W.; Prior, F.; Collins, G.S.; et al. FUTURE-AI: International Consensus Guideline for Trustworthy and Deployable Artificial Intelligence in Healthcare. *arXiv* **2025**, arXiv:2309.12325.
- Akgül, B.; Şahin, E.; Kalkan, S. Investigating Bias and Fairness in Appearance-Based Gaze Estimation. *arXiv* **2026**, arXiv:2604.10707. [\[CrossRef\]](#)
- Kapoor, S.; Narayanan, A. Leakage and the Reproducibility Crisis in Machine-Learning-Based Science. *Patterns* **2023**, *4*, 100804. [\[CrossRef\]](#) [\[PubMed\]](#)
- Li, T.; Yan, Y.; Dou, F.; Song, W.; Zhang, X. Cross-Subject Generalization for EEG Decoding: A Survey of Deep Learning Methods. *Prog. Biomed. Eng.* **2026**, *8*, 022013. [\[CrossRef\]](#) [\[PubMed\]](#)
- Su, J.; Zhu, J.; Song, T.; Chang, H. Subject-Independent EEG Emotion Recognition Based on Genetically Optimized Projection Dictionary Pair Learning. *Brain Sci.* **2023**, *13*, 977. [\[CrossRef\]](#) [\[PubMed\]](#)
- Qin, J.; Zhang, X.; Sugano, Y. UniGaze: Towards Universal Gaze Estimation via Large-Scale Pre-Training. In *Proceedings of the 2026 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV 2026)*, Tucson, AZ, USA, 6–10 March 2026; IEEE: New York, NY, USA, 2026. [\[CrossRef\]](#)
- Cheng, Y.; Bao, Y.; Lu, F. PureGaze: Purifying Gaze Feature for Generalizable Gaze Estimation. *Proc. AAAI Conf. Artif. Intell.* **2022**, *36*, 436–443. [\[CrossRef\]](#)
- Wang, S.; Huang, Y.; Xie, J.; Tian, Y.; Chen, F.; Wang, Z. Cross-Dataset Gaze Estimation by Evidential Inter-Intra Fusion. *arXiv* **2024**, arXiv:2409.04766. [\[CrossRef\]](#)
- Zheng, Q.; Zhang, J.; Zhang, X. Enhancing Accuracy of Uncertainty Estimation in Appearance-Based Gaze Tracking with Probabilistic Evaluation and Calibration. *arXiv* **2025**, arXiv:2501.14894. [\[CrossRef\]](#)
- Wang, S.; Huang, Y. Suppressing Uncertainty in Gaze Estimation. *Proc. AAAI Conf. Artif. Intell.* **2024**, *38*, 5581–5589. [\[CrossRef\]](#)
- Jamalifard, M.; Lei, Y.; Fumanal Idocin, J.; Azizinezhad, P.; Foulsham, T.; Andreu-Perez, J. Factor-Informed Uncertainty Distillation for Gaze Estimation. In *Proceedings of the 2026 Symposium on Eye Tracking Research and Applications*; Association for Computing Machinery: New York, NY, USA, 2026.
- Geifman, Y.; El-Yaniv, R. SelectiveNet: A Deep Neural Network with an Integrated Reject Option. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, Long Beach, CA, USA, 9–15 June 2019*; PMLR: Cambridge, MA, USA, 2019; pp. 3768–3776.
- Drissi, N.; El-Kassabi, H.; Serhani, M.A. A Multi-Criteria Decision Analysis Framework for Evaluating Deep Learning Models in Healthcare Research. *Decis. Anal. J.* **2024**, *13*, 100523. [\[CrossRef\]](#)
- Ojo, Y.; Makinde, O.A.; Babatunde, O.V.; Babatunde, G.; Okeowo, S. Evaluating AI-Driven Mental Health Solutions: A Hybrid Fuzzy Multi-Criteria Decision-Making Approach. *AI* **2025**, *6*, 34. [\[CrossRef\]](#)
- Zhang, X.; Sugano, Y.; Fritz, M.; Bulling, A. MPIIGaze: Real-World Dataset and Deep Appearance-Based Gaze Estimation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *41*, 162–175. [\[CrossRef\]](#) [\[PubMed\]](#)
- Smith, B.A.; Yin, Q.; Feiner, S.K.; Nayar, S.K. Gaze Locking: Passive Eye Contact Detection for Human-Object Interaction. In *Proceedings of the 26th Annual ACM Symposium on User Interface Software and Technology (UIST 2013)*, St. Andrews, Scotland, UK, 8–11 October 2013; Association for Computing Machinery (ACM): New York, NY, USA, 2013; pp. 271–280.
- Fischer, T.; Chang, H.J.; Demiris, Y. RT-GENE: Real-Time Eye Gaze Estimation in Natural Environments. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops, Munich, Germany, 8–14 September 2018*; Lecture Notes in Computer Science; Springer: Cham, Switzerland, 2018; Volume 11214, pp. 339–357. [\[CrossRef\]](#)
- Zhang, X.; Park, S.; Beeler, T.; Bradley, D.; Tang, S.; Hilliges, O. ETH-XGaze: A Large Scale Dataset for Gaze Estimation Under Extreme Head Pose and Gaze Variation. In *Proceedings of the European Conference on Computer Vision (ECCV 2020)*, Glasgow, UK, 23–28 August 2020; Lecture Notes in Computer Science; Springer: Cham, Switzerland, 2020; Volume 12350, pp. 365–381.

21. Kim, S.; Nam, W.J.; Lee, S.W. Appearance Debaised Gaze Estimation via Stochastic Subject-Wise Adversarial Learning. *Pattern Recognit.* **2024**, *152*, 110441. [\[CrossRef\]](#)
22. Gal, Y.; Ghahramani, Z. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. In *Proceedings of the 33rd International Conference on Machine Learning, ICML 2016, New York, NY, USA, 19–24 June 2016*; PMLR: Cambridge, MA, USA, 2016; Volume 3, pp. 1651–1660.
23. Lakshminarayanan, B.; Pritzel, A.; Blundell, C. Simple and Scalable Predictive Uncertainty Estimation Using Deep Ensembles. In *Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017*; Curran Associates, Inc.: Red Hook, NY, USA, 2017; pp. 6403–6414.
24. El-Yaniv, R.; Wiener, Y. On the Foundations of Noise-Free Selective Classification. *J. Mach. Learn. Res.* **2010**, *11*, 1605–1641.
25. Tzeng, G.H.; Huang, J.J. *Multiple Attribute Decision Making: Methods and Applications*; Springer: Berlin/Heidelberg, Germany, 2011; ISBN 9781439861585.
26. Opricovic, S.; Tzeng, G.H. Compromise Solution by MCDM Methods: A Comparative Analysis of VIKOR and TOPSIS. *Eur. J. Oper. Res.* **2004**, *156*, 445–455. [\[CrossRef\]](#)
27. Diakoulaki, D.; Mavrotas, G.; Papayannakis, L. Determining Objective Weights in Multiple Criteria Problems: The Critic Method. *Comput. Oper. Res.* **1995**, *22*, 763–770. [\[CrossRef\]](#)
28. Shannon, C.E. A Mathematical Theory of Communication. *Bell Syst. Tech. J.* **1948**, *27*, 379–423. [\[CrossRef\]](#)
29. Kocaman, H.; Asan, U. Integration Modes Between MCDM Methods and Machine Learning Algorithms: A Structured Approach for Framework Development. *Mathematics* **2025**, *14*, 33. [\[CrossRef\]](#)
30. Gongora-Salazar, P.; Rocks, S.; Fahr, P.; Rivero-Arias, O.; Tsiachristas, A. The Use of Multicriteria Decision Analysis to Support Decision Making in Healthcare: An Updated Systematic Literature Review. *Value Health* **2023**, *26*, 780–790. [\[CrossRef\]](#) [\[PubMed\]](#)
31. Rudan, I.; Song, P.; Adeloye, D.; Campbell, H. Journal of Global Health’s Guidelines for Reporting Analyses of Big Data Repositories Open to the Public (GRABDROP): Preventing ‘Paper Mills’, Duplicate Publications, Misuse of Statistical Inference, and Inappropriate Use of Artificial Intelligence. *J. Glob. Health* **2025**, *15*, 01004. [\[CrossRef\]](#) [\[PubMed\]](#)
32. Abbas, Q.; Jeong, W.; Lee, S.W. Explainable AI in Clinical Decision Support Systems: A Meta-Analysis of Methods, Applications, and Usability Challenges. *Healthcare* **2025**, *13*, 2154. [\[CrossRef\]](#) [\[PubMed\]](#)

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.