

Coordinate Singularities Break Conformal Coverage for Gaze and Head Pose

Mohammadreza Jamalifard[✉], Yaxiong Lei^{✉*}, Parastoo Azizinezhad[✉], and
Javier Andreu-Perez^{✉*†}

Smart Health Technologies Group, Centre for Computational Intelligence, School of
Computer Science and Electronic Engineering, University of Essex, Colchester, UK
{m.jamalifard,yaxiong.lei,p.azizinezhad,j.andreu-perez}@essex.ac.uk

*Equal contribution. †Corresponding author.

Abstract. Conformal prediction provides distribution-free reliability guarantees for vision systems, but these guarantees depend on how prediction errors are measured in the output space. Many vision tasks produce outputs on curved spaces (*e.g.* gaze directions on the sphere or 3D head rotations), yet intermediate prediction heads, residuals, uncertainty estimates, or conformal scores are often defined in flat coordinate charts such as yaw-pitch or Euler angles. We show that this scoring choice introduces systematic geometric distortion near coordinate singularities (large pitch angles on the sphere and poses approaching gimbal lock in 3D rotations). Across four datasets (ETH-XGaze, Gaze360, BIWI, AFLW2000-3D), slice-conditional coverage at a nominal 90% target drops by 30–50 percentage points in these regions, falling to 38.9% on ETH-XGaze and 42.0% on Gaze360 at gaze pitch above 70° , and to 57.5% on BIWI and 55.2% on AFLW2000-3D at head pose pitch above 60° near gimbal lock, despite marginal coverage remaining near 90%. We prove that this is structural. Scalar thresholding changes the size of chart-coordinate prediction sets but leaves their distorted axis ratios unchanged. To diagnose this hidden failure mode, we show that a simple geometric quantity, the Riemannian volume density, strongly correlates with where coverage collapse occurs. Finally, we show that coordinate-free geodesic scoring removes this distortion. It requires no retraining and adds negligible computational cost.

Keywords: Conformal prediction · Gaze estimation · Head pose estimation · Riemannian geometry

1 Introduction

Consider a driver monitoring system using conformal prediction to guarantee 90% coverage of a driver’s gaze direction. Marginal conformal evaluation appears nominal: overall coverage is 90.8%. Yet at extreme pitch angles associated with distraction and drowsiness [29], coverage drops to roughly 39%, so fewer than half of the prediction sets contain the true gaze direction. Such viewing angles

are not rare: combined eye and cervical rotation regularly exceeds 70° [7,21] during mirror checks, instrument-panel glances, and drowsiness-related head drops, consistent with natural driving variation [10]. The failure persists across architectures, training procedures, and calibration set sizes. Its source is more fundamental: the coordinate system used for the conformal nonconformity score.

Gaze directions lie on the sphere $\mathbb{S}^2$ [40], and head poses are rotations in $\text{SO}(3)$ [13,26]. Vision pipelines often represent intermediate predictions, labels, or downstream calibration residuals using yaw–pitch or Euler angles, even when final benchmark accuracy is reported with angular error. Our analysis concerns the nonconformity score used inside conformal calibration. If this score is computed from chart-coordinate residuals, the resulting prediction sets inherit the chart geometry. These charts distort distances: they compress regions near coordinate singularities (the poles of $\mathbb{S}^2$ or gimbal lock in $\text{SO}(3)$) and stretch others [15,45]. A fixed threshold corresponds to a large region near the equator but only a narrow sliver near the pole (Fig. 1). Because marginal coverage is enforced by construction, coverage is redistributed across the manifold. Well-conditioned regions overcover while singular regions undercover [9,37]. Global evaluation metrics therefore do not reveal this failure.

A natural response is to apply adaptive conformal methods [22,33], which rescale thresholds based on local difficulty. However, these methods address heteroscedastic model error, not geometric distortion. The mismatch is one of *shape*, not *scale*. Near the pole, yaw–pitch residuals form an ellipse whose eccentricity is fixed by the coordinate geometry. Scalar rescaling can change the size of this ellipse but not its shape. On ETH-XGaze [40], normalised yaw–pitch scoring improves near-pole coverage only from 39% to 50%. Switching to geodesic scoring with the same normalisation raises coverage to 89%. These observations suggest that conformal prediction on manifold-valued outputs requires respecting the intrinsic geometry of the output space.

Contributions. We formalise this geometric failure mode and show how to eliminate it.

- C1. Coverage audit for gaze and head pose.** Across four datasets and two manifolds ($\mathbb{S}^2$, $\text{SO}(3)$), we observe drops of 30–50 pp in slice-conditional coverage for yaw–pitch and Euler-angle conformal scores despite correct marginal calibration. A controlled experiment on $\text{SO}(3)$ with synthetic isotropic noise isolates coordinate geometry as the sole cause.
- C2. Impossibility of scalar chart correction.** We prove (Proposition 2) that scalar adaptive conformal methods can change the size of prediction sets but not their local shape. The axis ratios are determined by the metric tensor of the chart and therefore cannot be corrected by scalar thresholding.
- C3. A diagnostic and practical remedy.** We show that the Riemannian volume density provides a simple diagnostic for where chart-based systems are likely to undercover. Using coordinate-free geodesic scores eliminates the distortion entirely, requiring no retraining and adding negligible computation ($\leq 0.02 \mu\text{s}$ per sample).

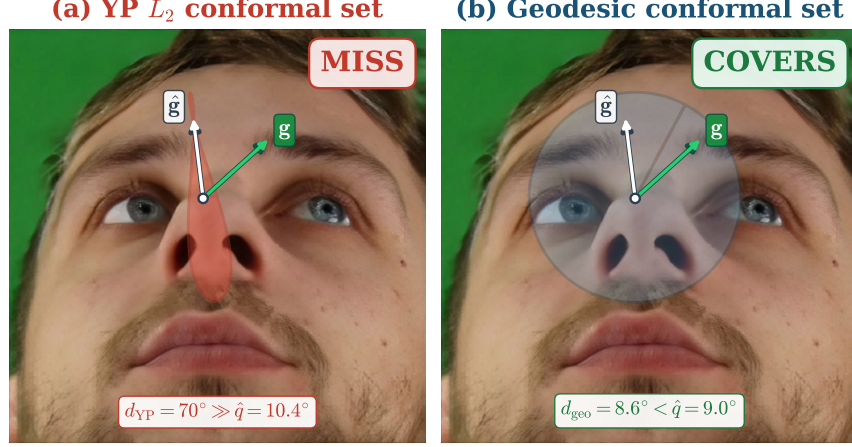


Fig. 1: Chart distortion breaks conformal coverage (schematic overlay, ETH-XGaze, $|\text{pitch}| = 82.5^\circ$). (a) The yaw–pitch (YP) L_2 conformal set ($\hat{q} = 10.4^\circ$) collapses into a narrow teardrop because $\cos(82.5^\circ) \approx 0.13$ compresses the yaw dimension; the ground truth $\mathbf{g}$ falls outside ($d_{YP} = 70^\circ \gg \hat{q}$). (b) The geodesic conformal set ($\hat{q} = 9.0^\circ$) forms a uniform circular cap on S^2 and covers $\mathbf{g}$ ($d_{geo} = 8.6^\circ < \hat{q}$). The prediction and calibration data are the same; only the scoring geometry differs. Chart coordinates inflate the true angular gap by 8.2 times.

2 Related Work

Conformal prediction. Split conformal prediction [22, 37] provides distribution-free marginal coverage under exchangeability, with extensions to adaptive settings [30, 33], group-conditional guarantees [37], and distribution shift [12, 36]; see [8, 35, 44] for surveys. Barber *et al.* [9] show that marginal guarantees do not imply conditional ones without further assumptions. Most conformal methods target Euclidean outputs; structured geometric outputs remain underexplored. Cholaquidis *et al.* [4] recently prove marginal validity for geodesic scores on Riemannian manifolds. Our work complements theirs. We show that commonly used chart-based scores in gaze and pose pipelines can produce systematic slice-conditional undercoverage. We further prove that scalar threshold adaptation cannot correct this geometric distortion.

Gaze and head pose estimation. Appearance-based gaze methods [1, 3, 17, 23, 40] and head pose estimators [14, 27, 34, 39] typically report angular or per-component MAE; uncertainty-aware variants [2, 24, 25, 38, 42, 43] provide confidence estimates but lack distribution-free guarantees. A persistent gap can arise between representation, training, calibration, and evaluation geometry. Gaze networks may predict yaw–pitch coordinates before angular-error evaluation [1, 23, 41], while head-pose systems may train with continuous rotation representations but still decode, report, or calibrate in Euler coordinates [14, 15, 45].

Our analysis concerns the geometry of the conformal score: if the score is computed in yaw–pitch or Euler coordinates, the conformal set inherits chart distortion. To our knowledge, no existing pipeline audits conformal reliability across pose or gaze ranges.

Uncertainty quantification for orientations. Deep ensembles [19], MC dropout [11], heteroscedastic regression [18], and distributional approaches on $\text{SO}(3)$ [28, 32] provide uncertainty estimates without finite-sample coverage guarantees. Johnstone *et al.* [16] study learned anisotropic conformal sets in Euclidean spaces; extending such methods to $\mathbb{S}^2$ or $\text{SO}(3)$ would require estimating local geometry. Directional distributions could give locally geometric sets but sacrifice distribution-free guarantees.

Positioning. Geometric distortion in Euler and yaw–pitch charts is well known [15, 45], but its consequences for conformal reliability have not been studied. To our knowledge, no prior work has shown that a nominally calibrated conformal system can silently undercover by 30–50 pp in safety-critical output regions, and no prior work has proved that standard adaptive remedies (normalised CP, CQR) are structurally unable to correct this. Our approach preserves the distribution-free guarantee of conformal prediction while eliminating chart-induced distortion via coordinate-free scoring.

3 Background and Formal Analysis

Conformal prediction primer. Given a predictor $\hat{y}(x)$ and calibration scores $s_i = s(y_i, \hat{y}_i)$, split conformal prediction chooses the empirical $(1 - \alpha)$ quantile $\hat{q}_\alpha$ and returns $\mathcal{C}(x) = \{y : s(y, \hat{y}(x)) \leq \hat{q}_\alpha\}$. Under exchangeability, $\Pr\{Y \in \mathcal{C}(X)\} \geq 1 - \alpha$, but marginal coverage need not hold on slices such as high gaze pitch or near-gimbal-lock poses. We study how the score s affects such slice-conditional coverage.

Notation. Let $\mathcal{M}$ be the output manifold, $\varphi : \mathcal{M} \supset U \rightarrow D \subset \mathbb{R}^d$ a chart, $\xi = \varphi(y)$ coordinates, $D\varphi$ the chart differential, $G(\xi)$ the metric tensor, and $\rho(\xi) = \sqrt{\det G(\xi)}$ the volume density.

Figure 2 summarises the protocol.

3.1 Setup

We analyse the common conformal-wrapper choice in which predictions and labels are represented in a coordinate chart and the nonconformity score is the Euclidean norm of the coordinate residual. For gaze, this chart score is $s_{\text{chart}} = \|(\Delta\psi, \Delta\theta)\|_2$ in yaw–pitch coordinates. For head pose, the analogous score uses ZYX Euler residuals. This is distinct from the final benchmark metric: a system may report angular error while a downstream conformal wrapper still calibrates

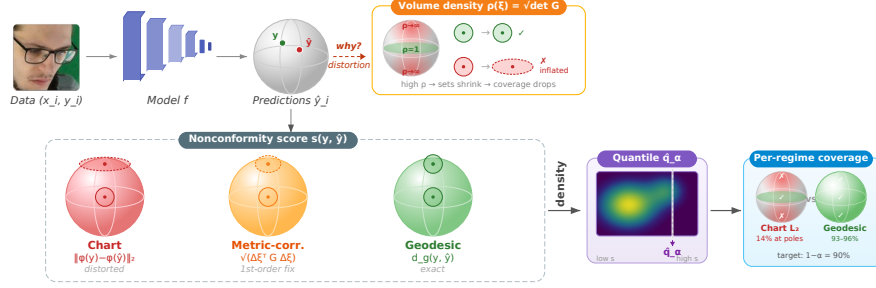


Fig. 2: Protocol overview. The key decision point is the nonconformity score. Chart-based scores inherit coordinate distortion; geodesic scores are coordinate-free. The volume density ρ indicates where chart scores are likely to fail.

coordinate residuals. (Section 3.4 defines the full score taxonomy; here we focus on s_{chart} and the geodesic score $s_{\text{geo}} = d_g(y, \hat{y})$.)

The discrepancy is captured by the *Riemannian metric tensor* $G(\xi)$, which describes how infinitesimal coordinate displacements relate to true distances and volumes on the manifold [5, 20]:

$$ds^2 = d\xi^\top G(\xi) d\xi, \quad d\text{Vol} = \sqrt{\det G(\xi)} d\xi. \quad (1)$$

When $G = I_d$, coordinates faithfully represent geometry and s_{chart} coincides with s_{geo} . When $G \neq I_d$, the chart score distorts intrinsic distances.

For yaw-pitch on $\mathbb{S}^2$, the metric is $G(\psi, \theta) = \text{diag}(\cos^2 \theta, 1)$. Yaw differences are scaled by $\cos \theta$, which vanishes at the poles. For ZYX Euler angles on $\text{SO}(3)$ (the convention used throughout; other sequences have analogous singularities at different angles), the metric (Supplementary Sec. A) is

$$G(\alpha, \beta, \gamma) = \begin{pmatrix} 1 & 0 & -\sin \beta \\ 0 & 1 & 0 \\ -\sin \beta & 0 & 1 \end{pmatrix}, \quad (2)$$

yielding $\det G = \cos^2 \beta$, which degenerates at gimbal lock ($|\beta| = 90^\circ$). In both cases, $\lambda_{\min}(G) \rightarrow 0$ near the singularity, so a fixed chart-score radius corresponds to a vanishing manifold volume.

Consequences for conformal prediction. Split conformal prediction [37] selects a threshold $\hat{q}_\alpha$ on calibration scores to guarantee marginal coverage $1 - \alpha$, where $\alpha \in (0, 1)$ is the target miscoverage level. Because most calibration samples lie where $G \approx I_d$, the threshold is set for those regions. Near singularities, the same threshold carves out a manifold region that is too small, producing systematic undercoverage. All experiments unwrap angles to $[-\pi, \pi]$; the effects documented below arise from metric distortion, not wrap-around.

3.2 Coverage Distortion from the Metric Tensor

We quantify how chart distortion affects the distribution of residuals and therefore conditional coverage. Fix a prediction $\hat{Y}(x) = p \in \mathcal{M}$ and model local error in the tangent space: $Y = \exp_p(\varepsilon)$, $\varepsilon \sim \mathcal{N}(0, \Sigma_p)$, where $\Sigma_p \succ 0$ is an arbitrary positive-definite covariance and ε is small enough that Y remains in the chart domain. The isotropic special case $\Sigma_p = \sigma^2 I_d$ isolates chart distortion from model anisotropy; the general case shows how the two interact.

Proposition 1 (Chart-induced coverage distortion). *Under the local error model above, expressed in a g -orthonormal basis of $T_p\mathcal{M}$, the chart residual $\Delta\xi_p = \varphi(Y) - \varphi(p)$ satisfies $\Delta\xi_p \approx \mathcal{N}(0, J_p \Sigma_p J_p^\top)$ where $J_p = D(\varphi \circ \exp_p)|_0$. Let $R_p = \|\Delta\xi_p\|_2$, $F_p(r) = \Pr(R_p \leq r)$, and $\mu_1(p) \leq \dots \leq \mu_d(p)$ be the eigenvalues of the combined matrix $M_p = J_p \Sigma_p J_p^\top$. Then for all $r \geq 0$,*

$$\Phi_d\left(\frac{r}{\sqrt{\mu_d(p)}}\right) \leq F_p(r) \leq \Phi_d\left(\frac{r}{\sqrt{\mu_1(p)}}\right), \quad (3)$$

where Φ_d is the CDF of the χ_d distribution. In the isotropic case ($\Sigma_p = \sigma^2 I_d$), letting $\xi_p = \varphi(p)$, the combined matrix is $M_p = \sigma^2 G(\xi_p)^{-1}$ and the set of eigenvalues reduces to $\{\sigma^2/\lambda_i^G\}_{i=1}^d$, recovering a bound in terms of the metric-tensor eigenvalues alone.

Proof (Proof sketch). Linearising $\varphi \circ \exp_p$ gives $\Delta\xi_p \approx \mathcal{N}(0, J_p \Sigma_p J_p^\top)$. The squared chart radius $R_p^2 = \|\Delta\xi_p\|^2$ is a quadratic form in a Gaussian vector; the Rayleigh quotient bounds $\mu_1\|w\|^2 \leq w^\top M_p w \leq \mu_d\|w\|^2$ (with $w \sim \mathcal{N}(0, I_d)$) yield (3). When $\Sigma_p = \sigma^2 I_d$, since $D\exp_p|_0 = I$, in a g -orthonormal basis of $T_p\mathcal{M}$ we have $J_p = D\varphi_p$. Thus, metric compatibility yields $J_p J_p^\top = G^{-1}$, so $M_p = \sigma^2 G^{-1}$, and the bounds depend only on the metric tensor. Full derivation in Supplementary Sec. A.

In the isotropic case, where $\lambda_{\min}^G(p)$ is small, near poles on $\mathbb{S}^2$ ($\lambda_{\min}^G = \cos^2\theta$) or gimbal lock on $\text{SO}(3)$ ($\lambda_{\min}^G = 1 - |\sin\beta|$), chart residuals become inflated in compressed directions. Thus a fixed chart-score threshold accepts a vanishing intrinsic volume near the singularity, producing undercoverage. When model error is anisotropic ($\Sigma_p \neq \sigma^2 I_d$), the coverage distortion depends on the alignment between Σ_p and the metric eigenvectors. If the model anisotropy opposes the chart distortion, the effect can be attenuated. If it is aligned with the distortion, the effect is amplified. Supplementary Sec. A.1 confirms both cases empirically. Setting $r = \hat{q}_\alpha$ in (3) gives explicit conditional coverage bounds (Supplementary Sec. A, Corollary 1). At singularities where $\lambda_{\min}^G \rightarrow 0$ and model error remains bounded, the lower bound tends to zero. Coverage is *redistributed*, not destroyed [37]. Regions near singularities undercover, while well-conditioned regions overcover. The residual ellipse eccentricity reaches 5.76:1 at $|\theta| = 80^\circ$ on $\mathbb{S}^2$ and 4.35:1 at $|\beta| = 60^\circ$ on $\text{SO}(3)$; both ETH-XGaze and BIWI contain substantial test samples at these angles. The propositions provide local, first-order predictions; the experiments in Sec. 4 confirm these predictions hold at finite radius across all settings tested.

Validity of the linearisation. Both propositions rely on a first-order approximation of the coordinate chart φ . Near singularities, the metric tensor varies rapidly. In these regions the approximation becomes less tight. The quantitative bounds may therefore be loose when a prediction set spans a region over which G changes substantially. Our experiments (Tables 2–5) confirm that the qualitative predictions hold well beyond the linearisation regime. These include directional collapse, redistribution, and the impossibility of scalar correction. However, the numerical bounds of Proposition 1 should be interpreted as indicative rather than exact at extreme angles.

Diagnostic: the volume density. The Riemannian volume density $\sqrt{\det G(\xi)}$ denoted as $\rho(\xi)$ measures how much manifold volume corresponds to one unit of coordinate volume. For both standard vision charts, ρ and $\lambda_{\min}$ are monotonically related ($\rho = |\cos \theta|$ on $\mathbb{S}^2$; $|\cos \beta|$ on $\text{SO}(3)$), making ρ an easily computed, closed-form proxy that strongly correlates with where coverage collapse is observed ($r > 0.90$ on controlled settings, $r = 0.81$ on AFLW2000; Sec. 4). For these standard charts, ρ reduces to $|\cos(\cdot)|$ and serves primarily as an interpretability aid that maps known coordinate geometry to expected coverage behaviour. On general manifolds with non-diagonal or higher-dimensional metric tensors, the eigenvalue-based bounds of Proposition 1 provide a more informative diagnostic (Supplementary Sec. C).

3.3 Why Scalar Adaptive Methods Cannot Fix This

Normalised CP [30] and CQR [33] rescale the acceptance threshold based on local difficulty. We show that this broad class of scalar threshold-modulating methods is structurally unable to correct chart distortion: it is a *shape* problem, not a *scale* problem.

Proposition 2 (Local impossibility of scalar chart correction). *Let φ be a coordinate chart on a d -dimensional Riemannian manifold $(\mathcal{M}, g)$, and assume the chart is nonsingular at the prediction $\hat{y}(x)$. Consider any conformal predictor whose acceptance sets take the form*

$$\mathcal{C}(x) = \{y \in \mathcal{M} : h(\|\varphi(y) - \varphi(\hat{y}(x))\|_2, x) \leq q\},$$

where $h(\cdot, x)$ is nondecreasing for each x . Then $\mathcal{C}(x)$ is the preimage of a chart-space ball of radius $r(x)$. In a g -orthonormal basis of the tangent space at $\hat{p} = \hat{y}(x)$, letting $J = D(\varphi^{-1})_{\hat{\xi}}$, the tangent-space approximation of $\mathcal{C}(x)$ is the ellipsoid $\{\delta \in T_{\hat{p}}\mathcal{M} : \delta^\top (JJ^\top)^{-1} \delta \leq r(x)^2\}$ with semi-axes $r(x)\sqrt{\lambda_i(\hat{\xi})}$, where $\lambda_i(\hat{\xi})$ are the eigenvalues of the metric tensor $G(\hat{\xi}) = J^\top J$. The local axis ratio

$$\frac{\text{major}}{\text{minor}} = \sqrt{\frac{\lambda_{\max}(\hat{\xi})}{\lambda_{\min}(\hat{\xi})}} \quad (4)$$

is independent of both h and $r(x)$. If $G(\hat{\xi}) \neq cI_d$ for any $c > 0$, no choice of h can make the acceptance set locally isotropic.

Proof (Proof sketch). Monotonicity gives $\{y : h(s, x) \leq q\} = \{y : s \leq r(x)\}$, a chart ball. Let $J = D(\varphi^{-1})_{\hat{\xi}}$. Expressing J in a g -orthonormal frame of $T_{\hat{p}}\mathcal{M}$, the pullback metric in chart coordinates is $G = J^\top J$. The chart constraint $\|\Delta\xi\|_2 \leq r$ maps to the tangent space via $\delta = J\Delta\xi$, yielding $\delta^\top (JJ^\top)^{-1} \delta \leq r^2$. Because the non-zero eigenvalues of $JJ^\top$ are identical to those of $J^\top J = G$, the ellipsoid’s semi-axes are exactly $r\sqrt{\lambda_i(G)}$. Full derivation in Supplementary Sec. A.

Instantiation. On $\mathbb{S}^2$, (4) gives $1/|\cos\theta|$ (diverges at poles). On $\text{SO}(3)$, $\sqrt{(1+|\sin\beta|)/(1-|\sin\beta|)}$ (diverges at gimbal lock).

Scope of the impossibility result. Proposition 2 applies to conformal predictors whose acceptance regions are radial in chart coordinates, *i.e.* sublevel sets of a scalar function of $\|\varphi(y) - \varphi(\hat{y}(x))\|_2$. Methods that explicitly learn anisotropic or non-radial acceptance shapes (e.g., full covariance or metric learning in the tangent space) fall outside this class. Our result therefore does not preclude shape correction via non-radial scores; it shows that scalar rescaling of a chart norm, including normalised conformal prediction and conformalized quantile regression, cannot eliminate chart-induced anisotropy.

Scale vs. shape correction. We distinguish two types of correction, formalised in the three-layer decomposition of Sec. 3.5. *Scale correction* (normalised CP, CQR) adjusts the radius $r(x)$ without changing axis ratios; it addresses heteroscedastic model error but not chart distortion. *Shape correction* changes the score geometry. The metric-corrected score $\sqrt{\Delta\xi^\top G \Delta\xi}$ (first-order) and the geodesic score $d_g(y, \hat{y})$ (exact) produce locally isotropic acceptance sets (Layer 3). Combining shape and scale correction substantially reduces conditional gaps across experiments (Sec. 4). Methods that learn anisotropic prediction sets [16] could achieve shape correction but do not yet exist for $\mathbb{S}^2$ or $\text{SO}(3)$; closed-form geodesics make this unnecessary on these manifolds.

3.4 Scoring Functions

Table 1 classifies scores by geometric properties and Layer-3 correction type.

Geodesic scores. On $\mathbb{S}^2$: $d_{\mathbb{S}^2}(v, \hat{v}) = \arccos(v^\top \hat{v})$. On $\text{SO}(3)$: $d_{\text{SO}(3)}(R_1, R_2) = \|\log(R_1^\top R_2)\|_F / \sqrt{2}$. Because conformal prediction depends only on score *ranking*, any strictly monotone surrogate of the geodesic distance yields identical conformal sets. For example, on $\mathbb{S}^2$ one may use $1 - v^\top \hat{v}$ (or chord distance), and on $\text{SO}(3)$ $1 - |q^\top \hat{q}|$, thereby avoiding trigonometric functions and matrix logarithms.

Metric-corrected scores. A first-order correction weights chart residuals by the local metric:

$$s_{\text{mc}} = \sqrt{\Delta\xi^\top G(\bar{\xi}) \Delta\xi} = \sqrt{(\Delta\psi)^2 \cos^2 \bar{\theta} + (\Delta\theta)^2}, \quad (5)$$

Table 1: Nonconformity score taxonomy. *Monotone-equivalent to geodesic; identical conformal sets.

Score	Manifold	Chart?	Intrinsic?	Shape
Yaw/pitch L_2	$\mathbb{S}^2$	Yes	No	None
Euler ZYX L_2	SO(3)	Yes	No	None
Metric-corr. YP	$\mathbb{S}^2$	Partial	No	1st order
Metric-corr. Euler	SO(3)	Partial	No	1st order
Geodesic (arc)	$\mathbb{S}^2$	No	Yes	Exact
$\mathbb{R}^3$ chord	$\mathbb{S}^2$	No	Yes*	Exact
Geodesic (angle)	SO(3)	No	Yes	Exact
Quaternion d_q	SO(3)	No	Yes*	Exact

evaluated at the midpoint $\bar{\xi}$. This removes first-order anisotropy but retains higher-order distortion.

Normalised scores. Dividing any base score by a difficulty estimate $\hat{\sigma}(x)$ [30] addresses heteroscedastic model error (Layer 2). In all experiments, $\hat{\sigma}(x)$ is a k -NN mean absolute residual in feature space (details in Supplementary Sec. F). By Proposition 2, normalisation cannot alter acceptance-set *shape*: normalised chart scores remain anisotropic; normalised geodesic scores address both layers simultaneously.

3.5 Three-Layer Decomposition

We decompose conditional coverage gaps into three layers: marginal validity (Layer 1, by construction), heteroscedastic model error (Layer 2, scale correction), and chart-induced distortion (Layer 3, shape correction only; Proposition 2). If scalar methods could absorb chart distortion, normalised yaw–pitch would match normalised geodesic; a persistent 40 pp gap on ETH-XGaze confirms they cannot (Sec. 4). Thus, plain geodesic scoring removes Layer 3 chart distortion, but residual scale variation may remain; normalised geodesic combines Layer 3 shape correction with Layer 2 scale correction. On ETH-XGaze, plain geodesic scoring raises near-pole coverage from 38.9% to 74.6% (Layer 3 removed), while normalised geodesic further closes the remaining gap to 89.3% (Layer 2 also addressed). A practical protocol is given in Sec. 6.

4 Experiments

We evaluate conformal coverage across two manifolds ($\mathbb{S}^2$, SO(3)), four datasets, and both controlled and real-model settings. All experiments use split conformal prediction [37] with $\alpha = 0.10$ (target: 90% marginal coverage). Per-regime coverage is reported as mean $\pm$ standard deviation over multiple random calibration/test splits (subject-disjoint). Throughout, “conditional coverage” means

Table 2: Controlled SO(3) experiment (BIWI, $\alpha = 0.10$, $\sigma = 5^\circ$, 50 splits). Near gimbal lock, Euler scoring collapses; geodesic maintains target coverage. All values in %. Bold marks the largest absolute deviation from the 90% target.

Region	Geo. [†]	Euler L_2	Met.-corr.	N. Euler	N. Geo.
Overall	90.0±0.5	89.9±0.5	89.9±0.5	90.0±0.4	90.0±0.5
$ \beta < 30^\circ$	89.7±0.6	95.6±0.4	90.2±0.6	92.9±0.4	89.5±0.6
$ \beta > 60^\circ$	89.5±1.4	57.5±2.1	87.8±1.5	72.7±1.7	89.0±1.2

ρ -coverage correlation: $r = 0.987$ ($p = 5 \times 10^{-6}$).

[†]Quaternion distance produces identical results (monotone equivalence).

slice-conditional coverage: empirical coverage in a specified output region (*e.g.*, $|\text{pitch}| > 70^\circ$). This is weaker than pointwise conditional coverage [9] but sufficient to reveal failures hidden by marginal metrics.

4.1 Controlled SO(3) Setting: Isolating the Mechanism

Setup. We use ground-truth rotations from BIWI [6] ($N = 15,678$, 24 subjects) and add calibrated isotropic Lie-algebra noise at $\sigma = 5^\circ$, producing geodesic MAE of 8.0° (comparable to modern methods [14, 39]). Because the noise is isotropic on the manifold by construction, any coverage variation is attributable solely to score geometry.

Results. Table 2: all methods achieve $\approx 90\%$ marginal coverage. Near gimbal lock ($|\beta| > 60^\circ$, 4.5% of samples), Euler L_2 collapses to $57.5 \pm 2.1\%$ while geodesic maintains $89.5 \pm 1.4\%$. At $|\beta| < 30^\circ$, Euler over-covers at 95.6%, compensating as Proposition 1 predicts.

Robustness. The pattern is robust across noise magnitudes ($\sigma \in \{3^\circ, 8^\circ, 12^\circ\}$; Euler stays at $\sim 61\%$, geodesic at $\approx 91\%$) and under anisotropic noise (ratios up to 4:1, CVF $r > 0.96$). Only at extreme anisotropy (5.7:1:1.3) does Euler’s failure attenuate (84.3%); geodesic remains at 90%. Full sweep in Supplementary Sec. A.1.

4.2 $\mathbb{S}^2$ Gaze Estimation: Real Models

Setup. We train ResNet-18 gaze estimators on ETH-XGaze ($N_{\text{test}} = 117,360$) and Gaze360 [17] ($N_{\text{test}} = 25,969$), comparing the non-redundant scoring functions in (Table 1). We define “near-pole” as $|\text{pitch}| > 70^\circ$, within the physiological range of combined eye and head rotation (Sec. 1) and relevant to safety-critical gaze targets including the instrument cluster, lap-mounted devices, and side mirrors [29].

Table 3: $\mathbb{S}^2$ gaze estimation (ETH-XGaze R18, $\alpha = 0.10$, 5 splits). Conditional coverage by output regime. Wilson 95% CIs in brackets. Bold marks the largest absolute deviation from the 90% target.

Method	Overall	Equator	Near Pole
Geodesic [†]	91.0 ± 0.7	94.0 ± 1.1	74.6 ± 3.9 [72.3, 76.8]
Yaw/Pitch L_2	90.8 ± 0.7	96.5 ± 1.0	38.9 ± 3.9 [36.4, 41.5]
Metric-corrected	91.0 ± 0.7	94.1 ± 1.1	74.1 ± 3.9 [71.7, 76.3]
Normalised YP	89.0 ± 3.3	94.2 ± 2.1	50.4 ± 5.9 [47.8, 53.0]
Norm. geodesic	89.1 ± 3.6	90.6 ± 3.0	89.3 ± 1.9 [87.6, 90.8]
ρ -coverage: $r = 0.957$ (ETH-XGaze), $r = 0.976$ (Gaze360).			

[†]Geodesic and $\mathbb{R}^3$ chord produce identical acceptance regions.

Table 4: Gaze360 conditional coverage (R18, $\alpha = 0.10$, 5 splits; means shown, full mean $\pm$ std in Supplementary Sec. B). Bold marks the largest absolute deviation from the 90% target.

Score	Overall	Equator	Near Pole
Geodesic	89.7	90.6	77.2
YP L_2	89.9	92.2	42.0
Metric-corr.	89.7	90.7	74.5
Norm. YP	89.9	91.6	50.8
Norm. geodesic	89.8	90.1	86.5

Results. Marginal coverage is indistinguishable across methods (90.8–91.0%, $\leq 9.18^\circ$ geodesic radius, ≤ 0.097 sr solid angle). Table 3 shows a sharp near-pole disparity: YP L_2 achieves $38.9 \pm 3.9\%$, while coordinate-free scoring reaches $74.6 \pm 3.9\%$. Table 4 replicates this on Gaze360: 42.0% for YP L_2 , 77.2% for geodesic, and 86.5% for normalised geodesic.

Scale vs. shape. Normalised YP improves near-pole coverage from 38.9% to 50.4%, but remains 40 pp below target; metric-corrected scoring reaches 74.1%, and normalised geodesic reaches 89.3%, confirming Proposition 2.

4.3 Real-Model $\text{SO}(3)$ Head Pose: AFLW2000-3D

Setup. We run SixDRepNet [14] (RepVGG-B1g2) on AFLW2000-3D [46] ($N = 1,969$), which includes substantial profile views ($|\beta| > 60^\circ$: $n = 278$, 14.1%). Normal cervical rotation reaches ± 70 – 80° (144° total [7]), placing profile-range orientations well within everyday head motion during conversation, shoulder checks, and blind-spot glances. SixDRepNet uses the continuous 6D rotation representation [45] internally; the failure arises *solely* from reverting to Euler coordinates at scoring time. No faces or poses are filtered; all 1,969 samples are used regardless of detection confidence or pose extremity.

Table 5: Real-model conformal on AFLW2000-3D (SixDRepNet, $\alpha = 0.10$, 50 splits). The 20 pp gap at $|\beta| > 60^\circ$ isolates Layer 3. All values in %. Bold marks the largest absolute deviation from the 90% target.

Region	Geo. [†]	Euler L_2	Met.-corr.	N. Geo.	Mondrian
Overall	89.8±1.4	89.8±1.2	89.8±1.3	89.5±1.5	—
$ \beta < 30$	95.6±0.8	98.2±0.5	95.0±0.9	93.1±1.0	—
$ \beta > 60$	75.1±3.7	55.2±4.2	78.1±2.9	80.3±3.1	90.1±3.3
ρ -coverage: $r = 0.807$ ($p = 1.5 \times 10^{-2}$).					

[†]Quaternion distance produces identical results.

Results. Table 5: at $|\beta| > 60^\circ$, geodesic scoring reaches $75.1 \pm 3.7\%$ (model weakness, Layer 2); Euler drops further to $55.2 \pm 4.2\%$, the 20 pp gap isolates chart distortion (Layer 3). Normalised geodesic reaches $80.3 \pm 3.1\%$, recovering part of the Layer-2 deficit; normalised Euler remains at $62.8 \pm 4.0\%$. Mondrian CP partitions samples into output regimes and calibrates a separate conformal threshold per bin. With a 3-bin geodesic base it restores $90.1 \pm 3.3\%$, showing that group-wise calibration can address Layer 2/model-regime effects once the base score geometry is intrinsic.

4.4 Separating the Layers: Cross-Setting Synthesis

The three-layer decomposition (Sec. 3.5) predicts that normalised chart scores should improve over unnormalised (Layer 2 addressed) but fall short of normalised geodesic (Layer 3 unresolved). This prediction holds across all datasets and backbone settings tested.

Layer 2 gains. Normalisation yields consistent but limited improvements: +15 pp on BIWI ($57.5 \rightarrow 72.7\%$), +12 pp on ETH-XGaze ($38.9 \rightarrow 50.4\%$), +8 pp on AFLW2000 ($55.2 \rightarrow 62.8\%$).

Layer 3 residuals. In every setting, a large gap persists between normalised chart and normalised geodesic scores. 39 pp on ETH-XGaze (50.4 vs. 89.3%), 16 pp on BIWI (72.7 vs. 89.0%), 18 pp on AFLW2000 (62.8 vs. 80.3%), and 43 pp on GazeTR-ViT (38.7 vs. 81.3%; trained on ETH-XGaze, $N_{\text{test}} = 117,360$). This residual is chart distortion in isolation (Proposition 2). We do not include a separate CQR baseline because CQR with chart-norm scores produces radial acceptance sets of the same form as normalised CP; Proposition 2 applies identically, and normalised YP already demonstrates the scalar-method ceiling. Notably, the strongest backbone (GazeTR-ViT) produces the *largest* gap. Better models concentrate predictions in well-conditioned regions, amplifying the relative penalty at singularities (Sec. 5).

Fig. 3 tracks each scoring function’s conditional coverage from equator to pole

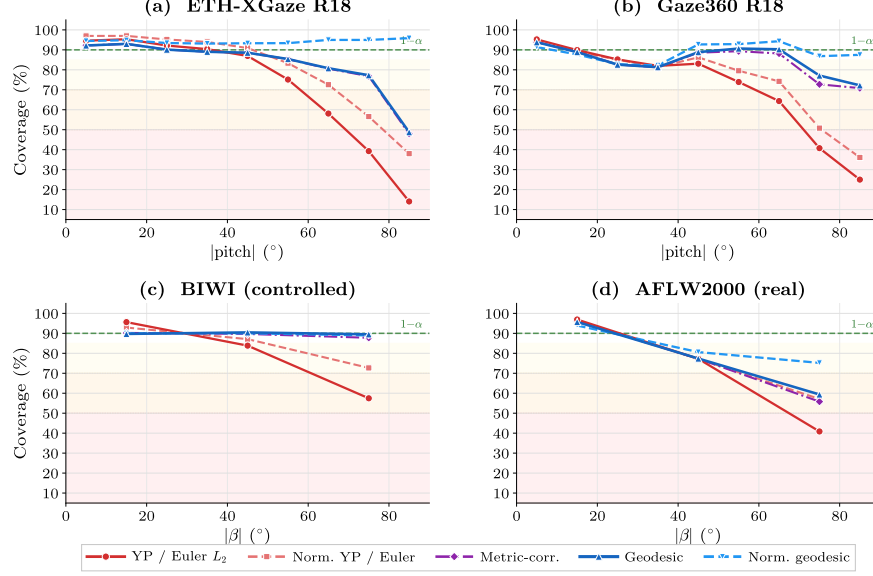


Fig. 3: Conditional coverage from equator to pole ($\alpha=0.10$). Shaded bands mark degraded (< 85%), severe (< 70%), and collapsed (< 50%) regimes. Chart-based scores (red) drop to 14–25% at extreme angles; normalised geodesic (blue dashed) remains closest to the target and substantially reduces the collapse. BIWI (c) isolates geometry as the sole factor; AFLW2000 (d) compounds model error, but intrinsic score-only methods retain the smallest gap to target among non-Mondrian scores.

across all four datasets. Chart-based scores (red) collapse sharply beyond moderate angles, while intrinsic scores (blue) substantially reduce the coverage gap.

5 Ablations

We summarise key ablations. Full tables appear in Supplementary Secs. B and F.

Stronger models amplify chart distortion. With a ViT-S backbone (GazeTR [3], ETH-XGaze), YP near-pole coverage drops to $30.3 \pm 7.8\%$, the lowest of all backbones tested. Normalised geodesic reaches $81.3 \pm 4.6\%$. Stronger models concentrate predictions in well-conditioned regions, so the conformal quantile is set more tightly there. This increases the relative penalty at singularities.

Chart-repair baselines. A multi-chart atlas reaches 59.7% near-pole on Gaze360 versus 77.2% for geodesic (Supplementary Sec. F). A learned Mahalanobis score partially compensates. It achieves $80.3 \pm 7.0\%$ on ETH-XGaze and $77.8 \pm 6.2\%$ on Gaze360 but with higher variance. On GazeTR it reaches $76.1 \pm 11.1\%$ versus

$81.3 \pm 4.6\%$ for normalised geodesic. Normalised geodesic requires no learned parameters and is up to 9 pp higher than the best Mahalanobis setting. Neither approach removes the geometric mismatch predicted by Proposition 2.

Robustness and Runtime. The effect persists across calibration sizes with < 0.5 pp variation from 1% to 100% of data. It persists across binning strategies. In the ETH-XGaze binning ablation, equal-mass Mondrian bins still yield 66.3% near-pole coverage, below the corresponding normalised-geodesic result; full details are in Supplementary Sec. F. It persists under cross-dataset recalibration. ETH-XGaze $\rightarrow$ Gaze360 still gives YP near 54%. Geodesic arc, $\mathbb{R}^3$ chord, and quaternion distance produce identical acceptance regions. Monotone-equivalent geodesic surrogates require only a dot product and cost $\leq 0.02 \mu\text{s}$ per sample on CPU. Chart-coordinate norms cost about the same. Metric-corrected scores are slightly slower. All scoring costs are $\ll 1\%$ of backbone inference time.

6 Discussion

A hidden failure mode under global summaries. The slice-conditional coverage collapse shown in Sections 4–5 is largely invisible under global summary metrics. Marginal coverage, MAE, and mean set size can appear normal (Table 3) even when reliability degrades in specific output regions. This is concerning in safety-relevant regimes such as extreme pitch and profile poses [27, 29, 31]. Chart coordinates persist because datasets, model heads, and downstream systems often store, decode, or calibrate yaw–pitch/Euler parameters, even when final point-estimation accuracy is reported using angular or geodesic metrics. A practical fix is to change only the nonconformity score. For manifold-valued outputs: convert predictions and labels to an intrinsic representation; compute a geodesic or monotone-equivalent score; optionally normalise for heteroscedasticity; calibrate with split conformal prediction; and report slice-conditional coverage alongside marginal coverage.

Scale adaptation cannot correct geometric distortion. The results distinguish scale and shape correction. Scalar adaptive methods adjust set size and can address heteroscedastic error. They do not change the local geometry induced by a chart. Proposition 2 formalises this limitation, and the experiments reflect it through consistent gaps between normalised chart scores and normalised geodesic scores. This complements known limits on conditional coverage [9]. Distribution-free pointwise guarantees are unattainable in general, but part of the conditional gap observed in geometric settings can be reduced by using intrinsic scores.

Implications beyond gaze and head pose. The choice of nonconformity score is important when outputs lie on a manifold. More generally, pipelines that (i) produce manifold-valued outputs, (ii) measure residuals in chart coordinates, and (iii) report only marginal conformal coverage may undercover in geometrically

ill-conditioned regions. The same diagnostic applies to other geometric tasks, including camera pose in $SE(3)$, angular prediction on products of circles, surface normal estimation on $\mathbb{S}^2$, and articulated human pose represented as products of joint rotations or directions. For $SE(3)$, translation is Euclidean while rotation inherits the $SO(3)$ analysis, with the translation–rotation scale chosen by the application. In such settings, chart singularities or poorly conditioned metric tensors indicate where reliability may degrade.

7 Limitations

Our approach inherits split conformal prediction’s standard assumptions (exchangeability, held-out calibration data). Propositions 1 and 2 are local, first-order results. The bounds may be loose when errors are large or strongly anisotropic, though robustness experiments (Sec. 4.1) confirm the qualitative predictions hold under moderate violations (anisotropy up to 4:1 on $SO(3)$, real model errors on AFLW2000). We evaluate on $\mathbb{S}^2$ and $SO(3)$; the theory applies to any Riemannian manifold with a non-isometric or nonconformal chart, but empirical validation on other spaces (*e.g.* $SE(3)$ or articulated pose) is future work. On manifolds without closed-form geodesics, intrinsic scoring may require numerical approximations or monotone surrogates, and the practical cost depends on the application. All conditional coverage results are *slice-conditional*, weaker than pointwise conditional coverage [9]. On non-singular but non-isometric charts (bounded distortion everywhere), the same redistribution mechanism applies but with smaller magnitude; geodesic scoring still eliminates it. Finally, geodesic scoring eliminates chart-induced distortion but does not address heteroscedastic model error, which requires complementary methods such as normalised conformal prediction.

8 Conclusion

We studied conformal prediction for manifold-valued vision outputs and showed that chart-coordinate nonconformity scores can silently redistribute coverage. Across four datasets, slice-conditional coverage drops by 30–50 pp near coordinate singularities despite nominal marginal coverage. We formalised this mechanism and proved that scalar adaptive conformal methods can adjust set size but cannot correct chart-induced shape distortion.

Replacing chart residuals with coordinate-free geodesic, chord, or quaternion-equivalent scores consistently reduces this failure without retraining and with negligible overhead. For geometric outputs, conformal reliability therefore depends not only on calibration, but also on measuring errors in a geometry-consistent way.

Acknowledgements

This work was supported by the UKRI BBSRC project EyeWarn (APP37953).

References

1. Abdelrahman, A.A., Hempel, T., Khalifa, A., Al-Hamadi, A., Dinges, L.: L2cs-net: Fine-grained gaze estimation in unconstrained environments. In: 2023 8th International Conference on Frontiers of Signal Processing (ICFSP). pp. 98–102. IEEE (2023)
2. Cantarini, G., Tomenotti, F.F., Noceti, N., Odone, F.: Hhp-net: A light heteroscedastic neural network for head pose estimation with uncertainty. In: Proceedings of the IEEE/CVF Winter Conference on applications of computer vision. pp. 3521–3530 (2022)
3. Cheng, Y., Lu, F.: Gaze estimation using transformer. In: 2022 26th International Conference on Pattern Recognition (ICPR). pp. 3341–3347. IEEE (2022)
4. Cholaquidis, A., Gamboa, F., Moreno, L.: Conformal inference for regression on riemannian manifolds. *Electronic Journal of Statistics* **19**(2), 6141–6166 (2025)
5. Do Carmo, M.P., Flaherty Francis, J.: *Riemannian geometry*, vol. 393. Springer (1992)
6. Fanelli, G., Dantone, M., Gall, J., Fossati, A., Van Gool, L.: Random forests for real time 3d face analysis. *International journal of computer vision* **101**(3), 437–458 (2013)
7. Feipel, V., Rondelet, B., Le Pallec, J.P., Rooze, M.: Normal global motion of the cervical spine:: an electrogoniometric study. *Clinical Biomechanics* **14**(7), 462–470 (1999)
8. Fontana, M., Zeni, G., Vantini, S.: Conformal prediction: a unified review of theory and new challenges. *Bernoulli* **29**(1), 1–23 (2023)
9. Foygel Barber, R., Candès, E.J., Ramdas, A., Tibshirani, R.J.: The limits of distribution-free conditional predictive inference. *Information and Inference: A Journal of the IMA* **10**(2), 455–482 (2021)
10. Fridman, L., Lee, J., Reimer, B., Victor, T.: ‘owl’and ‘lizard’: Patterns of head pose and eye pose in driver gaze classification. *IET Computer Vision* **10**(4), 308–314 (2016)
11. Gal, Y., Ghahramani, Z.: Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In: international conference on machine learning. pp. 1050–1059. PMLR (2016)
12. Gibbs, I., Candès, E.: Adaptive conformal inference under distribution shift. In: *Advances in Neural Information Processing Systems (NeurIPS)* (2021)
13. Hartley, R., Zisserman, A.: *Multiple view geometry in computer vision*. Cambridge university press (2003)
14. Hempel, T., Abdelrahman, A.A., Al-Hamadi, A.: 6d rotation representation for unconstrained head pose estimation. In: 2022 IEEE International Conference on image processing (ICIP). pp. 2496–2500. IEEE (2022)
15. Huynh, D.Q.: Metrics for 3d rotations: Comparison and analysis. *Journal of Mathematical Imaging and Vision* **35**(2), 155–164 (2009)
16. Johnstone, C., Cox, B.: Conformal uncertainty sets for robust optimization. In: *Conformal and Probabilistic Prediction and Applications*. pp. 72–90. PMLR (2021)
17. Kellnhofer, P., Recasens, A., Stent, S., Matusik, W., Torralba, A.: Gaze360: Physically unconstrained gaze estimation in the wild. In: *Proceedings of the IEEE/CVF international conference on computer vision*. pp. 6912–6921 (2019)
18. Kendall, A., Gal, Y.: What uncertainties do we need in bayesian deep learning for computer vision? In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) *Advances*

- in Neural Information Processing Systems. vol. 30. Curran Associates, Inc. (2017), https://proceedings.neurips.cc/paper_files/paper/2017/file/2650d6089a6d640c5e85b2b88265dc2b-Paper.pdf
19. Lakshminarayanan, B., Pritzel, A., Blundell, C.: Simple and scalable predictive uncertainty estimation using deep ensembles. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 30. Curran Associates, Inc. (2017), https://proceedings.neurips.cc/paper_files/paper/2017/file/9ef2ed4b7fd2c810847ffa5fa85bce38-Paper.pdf
 20. Lee, J.M.: Introduction to Riemannian manifolds, vol. 2. Springer (2018)
 21. Lee, W.J., Kim, J.H., Shin, Y.U., Hwang, S., Lim, H.W.: Correction: Differences in eye movement range based on age and gaze direction. *Eye* **33**(7), 1192 (2019)
 22. Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R.J., Wasserman, L.: Distribution-free predictive inference for regression. *Journal of the American Statistical Association* **113**(523), 1094–1111 (2018)
 23. Lei, Y., He, S., Khamis, M., Ye, J.: An end-to-end review of gaze estimation and its interactive applications on handheld mobile devices. *ACM Computing Surveys* **56**(2), 1–38 (2023)
 24. Lei, Y., Wang, Y., Buchanan, F., Zhao, M., Sugano, Y., He, S., Khamis, M., Ye, J.: Quantifying the impact of motion on 2d gaze estimation in real-world mobile interactions. *arXiv preprint arXiv:2502.10570* (2025)
 25. Lei, Y., Zhao, M., Wang, Y., He, S., Sugano, Y., Khamis, M., Ye, J.: Macgaze: Motion-aware continual calibration for mobile gaze tracking. *arXiv preprint arXiv:2505.22769* (2025)
 26. Ma, Y., Soatto, S., Kosecka, J., Sastry, S.S.: An invitation to 3-d vision: from images to geometric models, vol. 26. Springer Science & Business Media (2012)
 27. Martyniuk, T., Kupyn, O., Kurlyak, Y., Krasheniyi, I., Matas, J., Sharmanska, V.: Dad-3dheads: A large-scale dense, accurate and diverse dataset for 3d head alignment from a single image. In: Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition. pp. 20942–20952 (2022)
 28. Mohlin, D., Sullivan, J., Bianchi, G.: Probabilistic orientation estimation with matrix fisher distributions. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H. (eds.) Advances in Neural Information Processing Systems. vol. 33, pp. 4884–4893. Curran Associates, Inc. (2020), https://proceedings.neurips.cc/paper_files/paper/2020/file/33cc2b872dfe481abef0f61af181dfcf-Paper.pdf
 29. Nikan, S., Upadhyay, D.: Appearance-based gaze estimation for driver monitoring. In: NeuRIPS 2022 Workshop on Gaze Meets ML (2022)
 30. Papadopoulos, H., Proedrou, K., Vovk, V., Gammerman, A.: Inductive confidence machines for regression. In: European conference on machine learning. pp. 345–356. Springer (2002)
 31. Patney, A., Salvi, M., Kim, J., Kaplanyan, A., Wyman, C., Benty, N., Luebke, D., Lefohn, A.: Towards foveated rendering for gaze-tracked virtual reality. *ACM Transactions On Graphics (TOG)* **35**(6), 1–12 (2016)
 32. Prokudin, S., Gehler, P., Nowozin, S.: Deep directional statistics: Pose estimation with uncertainty quantification. In: Proceedings of the European conference on computer vision (ECCV). pp. 534–551 (2018)
 33. Romano, Y., Patterson, E., Candès, E.: Conformalized quantile regression. In: Advances in Neural Information Processing Systems (NeurIPS) (2019)

34. Ruiz, N., Chong, E., Rehg, J.M.: Fine-grained head pose estimation without key-points. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops. pp. 2074–2083 (2018)
35. Stephen, B., et al.: A gentle introduction to conformal prediction and distribution-free uncertainty quantification. arXiv preprint arXiv: 2107.07511 (2021)
36. Tibshirani, R.J., Foygel Barber, R., Candès, E., Ramdas, A.: Conformal prediction under covariate shift. *Advances in neural information processing systems* **32** (2019)
37. Vovk, V., Gammerman, A., Shafer, G.: *Algorithmic learning in a random world*. Springer (2005)
38. Wang, Y., Yan, R., Lei, Y., Fu, X.: Ptgaze: Cross-domain gaze estimation via proxy tuning. In: Proceedings of the 2025 Symposium on Eye Tracking Research and Applications. pp. 1–2 (2025)
39. Yang, T.Y., Chen, Y.T., Lin, Y.Y., Chuang, Y.Y.: Fsa-net: Learning fine-grained structure aggregation for head pose estimation from a single image. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1087–1096 (2019)
40. 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: European conference on computer vision. pp. 365–381. Springer (2020)
41. Zhang, X., Sugano, Y., Bulling, A., Hilliges, O.: Learning-based region selection for end-to-end gaze estimation. In: 31st British Machine Vision Conference (BMVC 2020). p. 86. British Machine Vision Association (2020)
42. Zheng, Q., Zhang, X.: Confidence-aware 3d gaze estimation and evaluation metric. arXiv preprint arXiv:2303.10062 (2023)
43. Zhong, W., Xia, C., Zhang, D., Han, J.: Uncertainty modeling for gaze estimation. *IEEE Transactions on Image Processing* **33**, 2851–2866 (2024)
44. Zhou, X., Chen, B., Gui, Y., Cheng, L.: Conformal prediction: A data perspective. *ACM computing surveys* **58**(2), 1–37 (2025)
45. Zhou, Y., Barnes, C., Lu, J., Yang, J., Li, H.: On the continuity of rotation representations in neural networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5745–5753 (2019)
46. Zhu, X., Lei, Z., Liu, X., Shi, H., Li, S.Z.: Face alignment across large poses: A 3d solution. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 146–155 (2016)