OUTPUT-AWARE CONFORMAL GEOMETRY FOR STRUCTURED NONLINEAR PARAMETRIC MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Structured parametric models generate complete meshes from compact parameters, but conformal coverage alone does not control the spatial extent of a decoded prediction set. We introduce output-aware conformal geometry (OACG), which shapes parameter-space ellipsoids using the matrix geometric mean (GM) of residual precision and decoder sensitivity. Independent calibration and exact decoding give finite-sample marginal coverage of representable references under exchangeability, regardless of linearization accuracy. For full-rank outputs, the unregularized GM core minimizes a pooled-moment surrogate for linearized RMS extent up to a common scale. Rank-deficient joint outputs require a ridge; we compare one isotropic in parameter coordinates (GM-I) with one isotropic in residual-normalized coordinates (GM-R). At nominal 90% coverage, both variants reduce decoded RMS extent against additive fusion in all five primary mesh settings. GM-R improves on additive fusion in all seven primary joint settings, whereas GM-I trails it on the three SMPL joint settings. From matched to participant-disjoint HOT3D evaluation, GM-I's extent increases by 1.6–9.9%, versus 22.8–58.2% for residual-only, additive, and GM-R geometries; population composition and participant overlap change together. A learned metric is smaller on EMDB but requires an additional geometry head. OACG makes fusion and ridge placement explicit choices for the spatial efficiency of structured prediction.

1 INTRODUCTION

Structured models such as SMPL, MANO, and FLAME decode compact parameters into 3D meshes (Loper et al., 2015; Romero et al., 2017; Li et al., 2017). An occluded limb can admit several plausible poses (Kolotouros et al., 2021). Conformal prediction turns a fixed predictor into a set with finite-sample marginal coverage under exchangeability (Vovk et al., 2005; Angelopoulos & Bates, 2023). Coverage alone, however, does not control how far the resulting meshes extend from the prediction.

The axes of a parameter region determine which mesh displacements it admits. Residual-only geometry captures prediction errors but ignores their spatial effect; sensitivity-only geometry penalizes decoded movement but ignores where the predictor errs. Additive precision fusion uses a relative weight, whose selected value varies across our tasks. This raises the question of how to combine the two sources of geometry without another mixing weight.

Output-aware conformal geometry (OACG; Figure 1) uses their matrix geometric mean to shape a parameter ellipsoid. A larger metric lowers linearized displacement per unit radius but raises residual scores and can enlarge the calibrated radius. In the full-rank setting, the geometric-mean core balances these effects by minimizing a pooled second-moment surrogate of linearized RMS extent, up to a common scale (Section 4.2). Development data select the geometry; independent calibration sets the radius; exact decoding defines the output set. The coverage guarantee holds for the complete decoded set without requiring the linearization to be accurate.

Joint outputs expose a second decision. Their decoder pullback is rank deficient, so the geometric-mean core alone does not bound every parameter direction. We compare an isotropic ridge in parameter coordinates (GM-I) with one in residual-normalized coordinates (GM-R). This choice is part of the geometry: the pooled optimum fixes the unregularized core, while calibration and nonlinear decoding determine whether either complete metric produces a smaller set.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

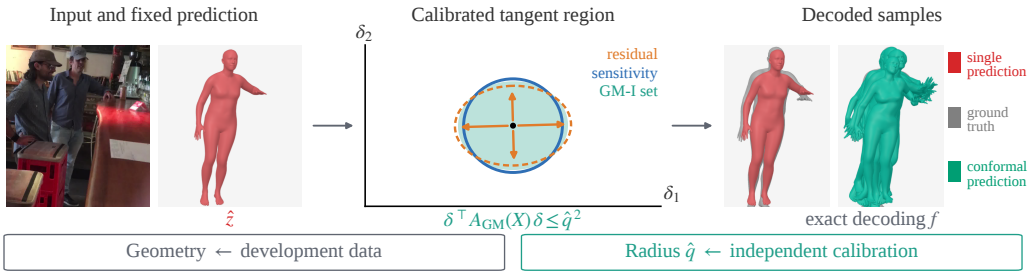

Figure 1: **Output-aware conformal geometry.** Development data determine the geometry; independent calibration sets the radius. The middle panel is a two-dimensional schematic slice; the right panel shows exact SMPL decodings of 16 draws from the GM-I region. Red denotes the point prediction and grey the representable reference. Membership concerns the full set, not the displayed draws.

The experiments show why both choices matter. Both GM variants reduce extent relative to additive fusion on all five primary mesh settings; GM-I's paired gain reaches 34.2%. GM-R improves on additive in all seven primary joint settings, while GM-I trails it on the three SMPL joint settings. Across the observed HOT3D population designs, GM-I's extent changes by 1.6–9.9%, compared with 22.8–58.2% for residual-only, additive, and GM-R geometries. A learned metric is smaller on EMDB but requires an additional geometry head. These results make fusion, ridge placement, and deployment population distinct design considerations.

We contribute (i) a geometric-mean fusion rule for shaping calibrated parameter regions according to decoded displacement, with exact whole-output coverage; (ii) a theory that identifies its full-rank surrogate optimum and separates that result from calibrated nonlinear extent; and (iii) comparisons across predictors, outputs, ridge coordinates, and HOT3D population designs that show where each design choice matters.

2 Related Work

Uncertainty in human pose and shape estimation. Human reconstruction methods produce point estimates such as HMR2.0 (Goel et al., 2023), or pose distributions like ProHMR (Kolotouros et al., 2021). CHAMP learns conformity scores to select pose hypotheses (Zhang & Carlone, 2025b); CUPS jointly trains a pose–shape estimator and uncertainty function before calibration (Zhang & Carlone, 2025a). OACG constructs continuous parameter regions around a fixed predictor without additional network training. Our experiments hold predictions fixed to isolate region geometry; pipelines that also change hypothesis generation or estimator training are outside this comparison.

Multivariate conformal regions and learned representations. Ellipsoidal conformal methods capture residual dependence (Messoudi et al., 2022). Our residual-based geometries share a regularized pooled second-moment estimator family, with hyperparameters selected separately. Whereas Feldman et al. (2023) construct regions using a learned response representation, OACG uses an existing parameterization and known decoder. Our learned-metric baseline C adapts ConfTr's simulated-calibration training (Stutz et al., 2022) to a linearized decoded-extent loss (Appendix B.3). Minimum-volume methods optimize norm-based regions (Braun et al., 2026); we measure RMS decoded displacement, which parameter volume does not determine.

Output-sensitive conformal prediction. COMPASS calibrates scalar intervals using directions sensitive to a segmentation metric (Cheung et al., 2026a); ConVOLT calibrates a volumetric scaling factor from deformation-field features (Cheung et al., 2026b). OACG uses decoder sensitivity to shape a multidimensional parameter region, incorporating output displacement into its geometry before scalar-score calibration.

Decoder geometry, precision fusion, and optimal transport. Our sensitivity-only control regularizes the pullback metric $G = J^\top W J$ (Shao et al., 2018). The additive control uses the precision-

addition form of Laplace–Gauss–Newton inference (Immer et al., 2021), combining decoder sensitivity and residual precision without fitting a posterior. The GM core follows a matrix variational principle (Bhatia et al., 2019). For positive-definite inputs, it has the algebraic form of the optimal-transport map between centered Gaussians (Peyré & Cuturi, 2019), although decoder sensitivity is not a covariance. Transport-based conformal methods construct multivariate ranks and quantile regions (Thurin et al., 2025); OACG instead uses the GM core to define a quadratic score with scalar calibration. Appendix A.4 (Table 4) compares the analytic cores and ridges.

3 Output-Aware Conformal Geometry

OACG selects a parameter-space metric on development data and freezes it before calibrating a shared radius. Exact decoding maps the resulting region to a set of complete outputs.

3.1 Structured Prediction Regions

For a fixed prediction $\hat{z}(x)$ and perturbation $\delta \in \mathbb{R}^d$, let $g_x(\delta) = f_T(\hat{z}(x) \oplus \delta)$, where f_T combines the known decoder, output selection, and alignment to each configuration's own root, requiring no reference labels. For SMPL and MANO, $\oplus$ applies rotation increments and additive shape updates (Appendix B.2). Coverage targets the reference-model output $y^\star = f_T(z^\star)$, with a measurable residual rule satisfying $\hat{z}(x) \oplus r^\star = z^\star$ and hence $g_x(r^\star) = y^\star$.

A metric field $A : \mathcal{X} \to \mathbb{S}^d_{++}$ and shared radius $q \geq 0$ define the parameter region and its exact decoded image,

$$\mathcal{R}_{A,q}(x) = \{\delta \in \mathbb{R}^d : \delta^\top A(x)\delta \leq q^2\}, \quad \mathcal{C}_{A,q}(x) = \{g_x(\delta) : \delta \in \mathcal{R}_{A,q}(x)\}. \tag{1}$$

The metric sets the axes and the relative scale across inputs; calibration sets q. We set $\mathcal{R}_{A,\infty}(x) = \mathbb{R}^d$ and $\mathcal{C}_{A,\infty}(x) = g_x(\mathbb{R}^d)$. Coverage applies to the full decoded set, not to its sampled approximation.

3.2 A GM Core for Decoded Extent

Output cost. For task weights $W \succeq 0$, write $\|v\|_W^2 = v^\top W v$. For $q < \infty$, define extent under uniform parameter-region sampling as

$$D(x; A, q) = \mathbb{E}_U \left\| g_x\Big(qA(x)^{-1/2}U\Big) - g_x(0) \right\|_W^2, \qquad U \sim \mathrm{Unif}(\mathbb{B}_d). \tag{2}$$

We report $E_{\mathrm{RMS}} = (\mathbb{E}_X D(X; A, q))^{1/2}$ over evaluation inputs. It measures typical decoded displacement under the specified uniform region measure, rather than parameter volume or the largest deformation (Proposition A.2). With $J(x) = Dg_x(0)$ computed through the full output map, the pullback metric and the linearized extent are

$$G(x) = J(x)^\top W J(x), \qquad D_{\mathrm{lin}}(x; A, q) = \frac{q^2}{d+2} \operatorname{tr}\big(G(x)A(x)^{-1}\big), \tag{3}$$

using $\mathbb{E}[UU^\top] = I/(d+2)$. Linearization guides the metric; the decoder defines the extent.

Moment proxy and GM core. Let $\widetilde{M} \succ 0$ be a regularized estimate of the pooled, uncentered residual second moment, retaining bias about the fixed prediction. Appendix A.4 specifies the estimator. Because A changes both output cost and the calibrated radius, we combine a pooled squared-score moment with linearized output cost:

$$\widetilde{\mathcal{U}}[A] = \mathbb{E}_X\Big[\operatorname{tr}(A(X)\widetilde{M})\Big]\, \mathbb{E}_X\big[\operatorname{tr}(G(X)A(X)^{-1})\big]. \tag{4}$$

Define the GM core

$$H(x) = \widetilde{M}^{1/2} G(x) \widetilde{M}^{1/2}, \qquad B(x) = \widetilde{M}^{-1} \# G(x) = \widetilde{M}^{-1/2} H(x)^{1/2} \widetilde{M}^{-1/2}, \tag{5}$$

where $\#$ denotes the matrix geometric mean. For $G(X) \succ 0$ almost surely and under the integrability conditions of Proposition A.3 in Appendix A.3, all minimizers of $\widetilde{\mathcal{U}}$ are $A(x) = cB(x)$ with one common $c > 0$ (Bhatia et al., 2019). We use $c = 1$; recalibration absorbs a common scaling of the complete metric. This optimality concerns the unregularized proxy, not calibrated nonlinear extent (Section 4). In residual-normalized coordinates $u = \widetilde{M}^{-1/2}\delta$, the GM core is $H^{1/2}$, whereas the additive core is $H + \gamma(x)I$. GM avoids the mixing coefficient $\gamma(x)$ for its core: the square root compresses the spread of whitened sensitivities, whereas additive fusion chooses a precision floor.

3.3 Regularization Coordinates

For singular G, $\ker B = \ker G$: directions with zero linearized cost can still incur finite-radius displacement (Section 4.3). This occurs for SMPL joint outputs, where $\operatorname{rank} G \leq 42 < 82$. We regularize the core as

$$A_{\mathrm{I}}(x) = B(x) + \lambda_{\mathrm{I}} s_B I, \qquad A_{\mathrm{R}}(x) = B(x) + \lambda_{\mathrm{R}} c_M \widetilde{M}^{-1}, \tag{6}$$

with positive strengths selected on development data. GM-I uses a ridge isotropic in parameter coordinates; GM-R uses one isotropic in residual-normalized coordinates, so estimated residual scales affect both its core and ridge. Both ridges act on all directions and supply the entire kernel penalty. Their calibrated extents depend on the chosen output and population design (Sections 5.3 and 5.4). Appendix A.4 defines the frozen scales s_B, c_M and compares all analytic geometries. We retain each metric's scale without per-input normalization; the normalization ablation is described in Appendix A.3. The pooled moment and scale constants are frozen, so for a fixed task each analytic metric depends on an image only through $G(x)$ at its predicted parameters. Two inputs with the same prediction therefore receive the same region even if their visual ambiguity differs.

3.4 Development Selection and Independent Calibration

All geometries, including the learned baseline, share five rounds of sequence-fold rotation. Within each round, three folds supply FIT, one supplies RADIUS, and one supplies RANK; the roles are sequence-disjoint. FIT estimates moments and scales or trains the learned metric, RADIUS sets a temporary threshold, and RANK measures exact-decoder extent. Selection uses prespecified grids at nominal level 0.90, averaging squared extent equally over all RANK sequences across rounds before taking the square root. Other coverage levels reuse the selected configuration and recalibrate its radius. After selection, we refit the configuration on all development data and freeze it before final calibration (Appendix B.3). For each frozen candidate h, define

$$S_h(x, z^\star) = \sqrt{r^{\star\top} A_h(x) r^\star}. \tag{7}$$

Using the same n calibration observations for all candidates, set

$$k = \lceil (n+1)(1-\alpha) \rceil, \qquad \widehat{q}_h = S_{h,(k)}, \qquad S_{h,(n+1)} = +\infty, \tag{8}$$

where $S_{h,(k)}$ is the kth ordered score. Substituting $(A_h, \widehat{q}_h)$ into Eq. (1) gives $\mathcal{R}_h$ and $\mathcal{C}_h$. Conditional on the frozen construction, exchangeability of calibration and test observations and reference reconstructibility yield

$$\Pr\{Y^\star \in \mathcal{C}_h(X)\} \geq \Pr\{r^\star \in \mathcal{R}_h(X)\} \geq 1 - \alpha. \tag{9}$$

Theorem 4.1 (Appendix A.1) establishes this guarantee without requiring accurate linearization. Sampling one observation per calibration sequence targets marginal observation coverage, not simultaneous sequence coverage. Algorithm 1 (Appendix B.1) summarizes the procedure.

4 Validity and Efficiency

We distinguish exact coverage from surrogate optimality and decoded extent, conditioning on the frozen construction. Write $S_A, \widehat{q}_A$ for the score and threshold of metric A in Eqs. (7)–(8). Proofs and auxiliary results are in Appendix A.

4.1 Coverage under Exact Decoding

Theorem 4.1 (Whole-output marginal coverage)**.** *Suppose calibration and test observations are exchangeable conditional on the frozen construction, a measurable reference residual satisfies $g_X(r^\star) = Y^\star$ almost surely, and scores and membership events are measurable. With Eq.* (8)*,*

$$\Pr\{Y^\star \in \mathcal{C}_{A,\widehat{q}_A}(X)\} \geq \Pr\{S_A(X, Z^\star) \leq \widehat{q}_A\} \geq 1 - \alpha. \tag{10}$$

The split-conformal rank guarantee (Vovk et al., 2005; Lei et al., 2018) transfers by exact decoding: score inclusion implies reference-output inclusion. Coverage concerns the full set and representable

outputs; identical independent sampling from exchangeable sequences targets a sampled observation, not simultaneous sequence coverage (Dunn et al., 2023).

Finite-pool randomization instead supports expected sequence-averaged coverage conditional on the pool, with a score fixed independently of uniform sequence allocation and observation sampling (Appendix A.1). It does not certify each realized split or a new population.

4.2 What the GM Core Optimizes

Let $M_x = \mathbb{E}[r^\star r^{\star\top} \mid X = x]$, $a_A(x) = \operatorname{tr}(A(x)M_x)$, and $b_A(x) = \operatorname{tr}(G(x)A(x)^{-1})$. The product

$$\mathcal{U}[A] = \mathbb{E}_X a_A(X)\, \mathbb{E}_X b_A(X) \tag{11}$$

bounds expected calibrated linearized extent up to a factor depending on n, α, and d under the assumptions of Proposition A.1, including i.i.d. sampling and finite moments. This motivates the objective without guaranteeing efficiency in our finite-pool evaluation.

Replacing M_x with $\widetilde{M}$ yields the pooled proxy $\widetilde{\mathcal{U}}$ from Section 3.2. For $\widetilde{M} \succ 0$, $G(X) \succ 0$ almost surely, and $0 < \mathbb{E}\tau < \infty$, where $\tau(x) = \operatorname{tr}(H(x)^{1/2})$, its minimum is $(\mathbb{E}\tau)^2$, attained exactly by $A(x) = cB(x)$ with one common $c > 0$ (Proposition A.3; Bhatia et al., 2019). Recalibration absorbs this common scale, whereas input-dependent scaling can change the regions. Pooling and regularization can separate the proxy from $\mathcal{U}$; neither objective determines score quantiles or nonlinear extent.

4.3 Regularization and Decoded Extent

When G is singular with positive probability and $0 < \mathbb{E}\tau < \infty$, the proxy has no positive-definite minimizer and selects no finite ridge (Appendix A.3). Although $\ker B = \ker G$, zero linearized cost need not imply zero finite-radius displacement: along a null direction, squared displacement can have a fourth-order leading term. Ridge placement matters, but each variant is recalibrated, so neither variant's section along a null direction is uniformly shorter (Appendices A.4–A.5).

For one calibration split and frozen model per candidate, bars denote common evaluation averages, with equal sequence weights and observations averaged within sequence. For finite extents with $\bar{D}_{\mathrm{lin},h} > 0$, let $\chi_h = \bar{D}_h / \bar{D}_{\mathrm{lin},h}$. The exact decomposition $D = D_{\mathrm{lin}} + \Delta_{\mathrm{odd}} + D_{\mathrm{even}}$ gives $\chi_h = 1 + (\bar{\Delta}_{\mathrm{odd},h} + \bar{D}_{\mathrm{even},h})/\bar{D}_{\mathrm{lin},h}$, where Δ_{odd} is the signed odd correction and $D_{\mathrm{even}} \geq 0$ is the even energy (Proposition A.4). Thus

$$E^2_{\mathrm{RMS},h} = \frac{\widehat{q}_h^2}{d+2}\, \bar{b}_h\, \chi_h. \tag{12}$$

This diagnostic separates calibrated radius, linearized cost, and nonlinear response; the factors remain coupled. Because the odd correction is signed, χ_h can exceed or fall below one. These identities apply before averaging across calibration splits or C's training replicates (Appendix A.5). Efficiency must therefore be evaluated after calibration and exact decoding; proxy optimality does not establish an ordering between GM-I and GM-R, even for matched populations.

5 Experiments

5.1 Setup and Evaluation

Settings and population designs. HMR2.0 with SMPL on 3DPW and EMDB provides a controlled comparison: fixed predictions are evaluated as 6890-vertex meshes and 14-joint outputs. Changing the output map changes its pullback G and the selected geometry without changing the predictor. ProHMR on 3DPW tests a second predictor; HaMeR with MANO on HOT3D tests a second decoder using mesh, R_{distal} (five joints), and R_{joints} (15 joints) outputs. HOT3D has a participant-disjoint design and a matched design in which every deployment participant is represented in development. A setting combines predictor, dataset, output, and population design. The reported comparison has five mesh and seven joint settings; development and deployment share participants on 3DPW and EMDB.

Protocol. All analytic geometries use the five-round FIT/RADIUS/RANK rotation of Section 3.4, with selection at nominal 0.90 over the frozen grids. After fitting and freezing each configuration, the evaluation uses 20 calibration splits per pool, each with $n = 40$ sequences and one observation per sequence, at nominal 0.80, 0.90, and 0.95. Candidate configurations and fitted statistics were frozen before evaluation (Appendix B.3); the reported study-wide GM-variant choice is a separate summary of evaluation outcomes.

Controls and metrics. We compare isotropic, residual-only, sensitivity-only, additive, GM-I, and GM-R geometries (Table 4). The learned-metric baseline C maps fixed image features to a positive-definite metric and has 20 frozen training-seed replicates in the four HMR2.0 settings (Appendix B.3). We report E_{RMS} in millimetres and score coverage. The paired gain of X over Y is the mean of $(E_Y - E_X)/E_Y$ across 20 calibration splits; C's split-level extent is first averaged over its replicates. The prespecified criterion (*) requires a mean gain of at least 5% and at least 16 of 20 gains agreeing in sign. It describes split-level consistency, not a hypothesis test.

Coverage. With $n = 40$, the finite-pool expected sequence-averaged score-coverage references are $33/41 = 0.805$, $37/41 = 0.902$, and $39/41 = 0.951$ at nominal 0.80, 0.90, and 0.95 for split-independent scores, with equality in the tie-free case (Appendix A.1). At nominal 0.90, observed mean score coverage spans 0.882–0.921 across the analytic settings and 0.890–0.897 for C. The appendix reports the SD and empirical fifth percentile across the 20 splits for every analytic family and nominal level (Tables 22– 24). Some means are below the $37/41$ reference, which is an expectation rather than a guarantee for each realized split.

Scope of claims. Coverage concerns complete decoded sets of representable references: neutral-model SMPL conversions and annotation-conditioned HOT3D crops. One observation per sequence gives marginal observation coverage, not simultaneous coverage of every frame in a sequence. Reported SDs describe calibration splits of fixed pools after one development selection. The HOT3D designs change participant overlap and population composition together, so their comparison describes a population-design contrast rather than an isolated overlap effect.

5.2 Full-Rank Outputs: Does Geometric-Mean Fusion Help?

Mesh outputs. Geometric-mean fusion reduces decoded extent relative to additive fusion in all five full-rank mesh settings (Tables 1, 2, and 10). GM-I gains 4.5–34.2% and meets the paired criterion in four of five; GM-R meets it in all five, with gains of 5.2–12.4%. The largest GM-I gain is on participant-disjoint HOT3D mesh: all 20 split-level extents are smaller, by a mean 13.9 mm (34.2% paired gain; Figure 5). GM-I almost satisfies the criterion on ProHMR/3DPW mesh ($+4.5\%$, 15 of 20 gains of the same sign).

5.3 Rank-Deficient Outputs: Where Should the Ridge Go?

Joint outputs. The ridge coordinate separates the two GM variants on joints. GM-R has smaller extent than additive fusion in all seven reported joint settings and meets the paired criterion in six. Its gains are 4.4–8.3% on SMPL and 12.1–26.5% on HOT3D; the sole criterion exception is ProHMR/3DPW ($+4.4\%$). GM-I instead loses on all three SMPL joint settings (-34.7 to -6.4%). On HOT3D it gains 30.1% on disjoint R_{joints}, its one joint criterion success, is near zero on disjoint R_{distal}, and loses 44.4% on matched R_{distal}. The joint results turn ridge placement into a substantive geometry choice rather than a numerical detail.

Table 1: **Decoded extent with frozen HMR2.0 at nominal 90%.** Mean score coverage and RMS decoded extent (mm; extent mean $\pm$ calibration-split SD over 20 splits of one pool). Bold marks the smallest mean in a column, not a verdict; paired gains and the prespecified criterion are reported in the text and Appendix B.3. Learned: mean over 20 training replicates.

	3DPW				EMDB			
	mesh		joint		mesh		joint	
Geometry	cov.	$E_{\mathrm{RMS}}\downarrow$	cov.	$E_{\mathrm{RMS}}\downarrow$	cov.	$E_{\mathrm{RMS}}\downarrow$	cov.	$E_{\mathrm{RMS}}\downarrow$
Isotropic	0.921	449.04 ± 17.31	0.921	397.11 ± 15.17	0.891	412.31 ± 11.42	0.891	355.86 ± 10.45
Residual-only	0.885	201.43 ± 14.59	0.885	170.09 ± 12.53	0.892	215.30 ± 11.54	0.892	182.48 ± 9.97
Sensitivity-only	0.899	149.57 ± 14.43	0.911	119.68 ± 10.81	0.888	163.10 ± 9.50	0.898	137.91 ± 6.34
Additive	0.894	105.74 ± 12.25	0.895	83.13 ± 8.00	0.897	123.44 ± 13.94	0.896	103.54 ± 10.29
GM-I	0.891	96.75 ± 12.90	0.911	102.58 ± 8.81	0.886	109.03 ± 8.84	0.893	109.64 ± 6.48
GM-R	0.890	$\mathbf{94.44} \pm 11.66$	0.893	$\mathbf{78.09} \pm 7.75$	0.887	107.57 ± 8.16	0.894	94.52 ± 6.19
Learned	0.897	99.38 ± 9.67	0.893	78.38 ± 7.57	0.890	$\mathbf{99.02} \pm 5.67$	0.890	$\mathbf{81.28} \pm 4.69$

Table 2: **HaMeR/MANO on HOT3D at nominal 90%.** RMS decoded extent (mm; mean $\pm$ calibration-split SD over 20 splits) under the primary participant-disjoint and matched population designs. Score coverage lies in $[0.882, 0.914]$ for these cells, compared with the finite-pool score reference $37/41 \simeq 0.902$; the pilot-excluded sensitivity check and all levels are in Appendix B.4. Bold marks the smallest mean in a column, not a verdict. Isotropic and sensitivity-only change by at most 2.5% between population designs; residual-only, additive fusion, and GM-R by 23–58%; GM-I by 2–10% (Table 3).

	mesh		R_{distal}		R_{joints}	
Geometry	disjoint	matched	disjoint	matched	disjoint	matched
Isotropic	110.19 ± 3.41	110.16 ± 1.68	124.25 ± 3.78	124.22 ± 1.86	109.83 ± 3.69	109.72 ± 1.82
Residual-only	66.96 ± 2.07	47.40 ± 2.77	75.97 ± 2.35	53.74 ± 3.14	65.43 ± 2.07	46.19 ± 2.71
Sensitivity-only	46.97 ± 1.07	47.61 ± 1.71	53.47 ± 1.48	54.83 ± 2.30	40.56 ± 1.08	41.13 ± 1.51
Additive	40.59 ± 1.65	27.54 ± 1.74	44.28 ± 1.88	30.20 ± 2.00	36.58 ± 1.44	25.20 ± 1.55
GM-I	$\mathbf{26.68} \pm 1.78$	$\mathbf{24.29} \pm 1.64$	44.21 ± 2.30	43.49 ± 2.06	$\mathbf{25.54} \pm 1.26$	24.87 ± 1.22
GM-R	38.45 ± 1.07	24.31 ± 1.57	$\mathbf{32.53} \pm 1.94$	$\mathbf{26.50} \pm 1.79$	31.12 ± 0.86	$\mathbf{21.81} \pm 1.59$

Rank and selected configurations. SMPL joint pullbacks have rank 42 of 82, while mesh pullbacks are full rank. HOT3D R_{distal} and R_{joints} have measured nullities 43 and 25. GM-I selects a larger outer ridge on SMPL joints and R_{distal} than on the corresponding meshes, but the R_{joints} selection combines the smallest ridge with a nonzero inner ridge (Table 5). These selected strengths act together with fitted scales and calibrated radius; their raw values alone do not order the resulting sets.

Ridge placement and strength. Sweep A varies the outer ridge with each variant's own selected moment. Sweep B gives both variants GM-R's moment and core, changing only the outer-ridge coordinate ($\lambda s_B I$ versus $\lambda c_M \widetilde{M}^{-1}$) between its arms; every sweep point is recalibrated (Figure 2; Appendix B.3). At $\lambda_h = 10^{-5}$, 3DPW joint extent is about $7\times$ and $18\times$ its selected GM-I and GM-R values. GM-I mesh changes by about 1% at the same strength. With a shared core there, both mesh arms give 98.49 mm, while the joint arms differ by nearly a factor of two. Strong regularization also exposes placement effects on mesh: at $\lambda_h = 10$, GM-I reaches 352 mm and GM-R 149 mm. Across these recalibrated sweeps mean score coverage remains around 0.88–0.92, while extent changes much more. Sweep B isolates placement within a shared core; the separately selected variants also differ in moments and strengths.

Exact decoding strengthens the case for measuring the output set directly. Across 1,440 split-level analytic records at nominal 0.90, the ratio χ_h of exact to linearized squared extent spans 0.52–2.13. Recalibrated trace normalization changes extent by at most 0.863% across 14 GM-I/GM-R arms on seven settings (Appendix B.8).

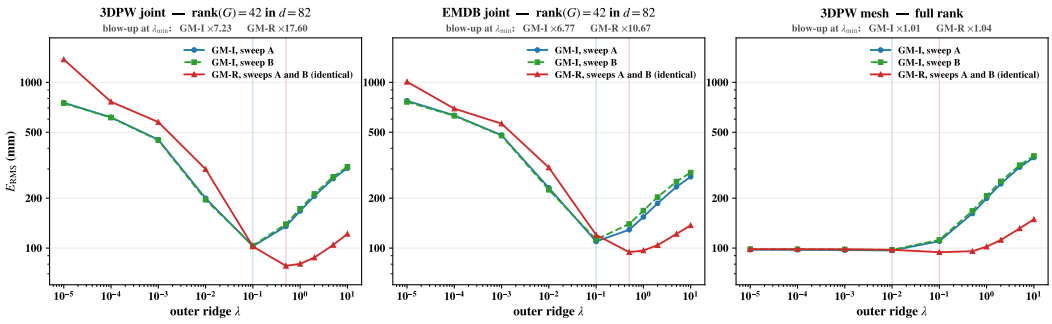

Figure 2: **Ridge sweeps at nominal** 0.90 **coverage.** The three panels show HMR2.0/SMPL 3DPW joints, EMDB joints, and 3DPW mesh. Curves give mean decoded RMS extent (mm) against outer ridge strength λ_h on a log scale. Sweep A uses each variant's selected moment; sweep B gives both variants GM-R's moment and core, so its two GM-R curves coincide. Every point is recalibrated. Vertical lines mark the development-selected strengths. Mean score coverage along the curves lies in $[0.88, 0.92]$. Sweep B isolates ridge coordinates under a shared core.

At 0.90, direct output-space controls methods B1–B3 have 1.8–$3.2\times$ GM-R's mean RMS extent on the tested SMPL outputs (Appendix B.7). Their ambient-space sets can contain off-model configurations; these ratios use each set's specified uniform measure.

5.4 Development-Population Shift

The HOT3D designs preserve calibration on each evaluation pool while changing who participates in development. From matched to disjoint, GM-I's mean extent rises by 1.6–9.9% across outputs, compared with 22.8–58.2% for GM-R and 41.3–47.4% for residual-only and additive fusion (Table 3(a)). Isotropic and sensitivity-only change still less but have larger absolute extent: on disjoint mesh, 110.19 and 46.97 mm versus GM-I's 26.68 mm. GM-R's post hoc gain over GM-I stays positive on R_{distal}, reverses on R_{joints}, and drops from parity to -44.6% on mesh (Table 3(b)). This is the observed tradeoff between spatial efficiency and change across the two population designs.

Choosing a ridge for the output and population. GM-R remains smaller on R_{distal} in both HOT3D designs, while R_{joints} changes winner (Table 2). GM-I's extent changes less across the observed population designs, but this gives no universal selection threshold. For a new decoder, compare exact decoded extent on development splits reflecting the intended output and population, including participant-held-out splits when shift is expected, then freeze the ridge choice before independent calibration. Rank deficiency motivates testing GM-R but does not settle the choice.

5.5 Comparison with a Learned Metric

In the four HMR2.0 settings with a trained C, no analytic family meets the paired criterion against it. GM-R nearly ties C on 3DPW joints, has 4.8% smaller extent on 3DPW mesh, and has 8.8% and 16.5% larger extent on EMDB mesh and joints. C's selected rank reaches the grid maximum in

Table 3: **HOT3D population-design comparison at nominal** 0.90**.** All entries are percentages. (a) Relative change in mean E_{RMS} from matched to participant-disjoint development, $(E_{\text{disjoint}} - E_{\text{matched}})/E_{\text{matched}}$. (b) Post hoc paired gain of GM-R over GM-I from frozen records.

(a) Change from matched to disjoint

Task	Iso.	Sens.	Resid.	Add.	GM-I	GM-R
R_{distal}	$+0.0$	-2.5	$+41.4$	$+46.6$	$+1.6$	$+22.8$
R_{joints}	$+0.1$	-1.4	$+41.7$	$+45.2$	$+2.7$	$+42.7$
mesh	$+0.0$	-1.3	$+41.3$	$+47.4$	$+9.9$	$+58.2$

(b) GM-R over GM-I (post hoc)

Task	matched	disjoint
R_{distal}	$+39.0$	$+26.4$
R_{joints}	$+12.4$	-22.1
mesh	-0.1	-44.6

three settings. Analytic methods fit task moments and select hyperparameters but train no geometry head. C uses predictor-specific features, was not adapted to ProHMR or HaMeR, and has untested fixed-head population transfer. OACG requires a per-input Jacobian and matrix square root; C's head was not timed alongside it (Appendix B.6).

5.6 Computational Cost

At batch size 128, GM's geometry-construction premium over additive fusion is 13.7% on 3DPW mesh and 36.0% on joints. On HOT3D mesh it rises from 31.4% at batch size 1 to 500.7% at 128 as Jacobians batch more efficiently than matrix square roots. Prediction and sampled mesh decoding are excluded (Appendix B.6).

6 Discussion and Limitations

What OACG contributes. A structured nonlinear decoder turns parameter regions into complete, model-generated outputs, but parameter-space volume does not determine their spatial extent. OACG designs a calibrated parameter ellipsoid for decoded displacement and evaluates it after exact decoding. Its GM core combines residual moments and decoder sensitivity without a mixing weight or an additional geometry network. The ridge makes explicit how directions weakly detected by the chosen output are constrained. Independent calibration provides marginal coverage of representable references under exchangeability without requiring accurate linearization.

What the results imply. Both GM variants have smaller mesh RMS extent than additive fusion in all five primary mesh settings and the pilot-excluded mesh sensitivity check. On joint outputs, GM-R has positive gains over additive throughout, whereas GM-I trails additive on SMPL joints. Mesh and joint maps assign different costs to the same parameter directions; joint maps are rank deficient, so the ridge determines penalties in their kernel. The shared-core sweep shows how this choice affects decoded extent, while separately selected variants also differ in fitted moments and ridge strengths. HOT3D adds a deployment consideration: GM-I has smaller relative extent change across its observed allocations, although GM-R has smaller extent on R_{distal} in each. Ridge choice should therefore be assessed for the specified output and deployment conditions, using exact decoded extent on representative development data when available.

Limits of inference. The full-rank GM result optimizes an unregularized, pooled linearized surrogate, so its optimum need not minimize calibrated nonlinear extent; we assess the latter by exact decoding. E_{RMS} summarizes typical displacement under the specified ellipsoid measure and does not bound the largest deformation. The learned baseline has smaller extent on EMDB and was evaluated only on HMR2.0; its population transfer was not measured. HOT3D allocations change participant composition and overlap together, so their comparison remains descriptive. Marginal coverage requires representable references and exchangeable calibration and future observations. Treating contiguous video frames as separate observations may violate exchangeability; the resulting nominal coverage is then not guaranteed, and one-observation coverage does not imply whole-sequence coverage.

7 Conclusion

OACG shapes calibrated parameter regions with residual moments and decoder sensitivity, then decodes them to complete model-generated outputs. Under exchangeability, calibration provides finite-sample marginal coverage of representable references without relying on linearization. Both GM variants have lower RMS extent than additive fusion in all five primary mesh settings; ridge placement changes joint efficiency, with GM-R giving more consistent gains over additive. GM-I has smaller relative extent changes than GM-R across the observed HOT3D allocations. In HMR2.0 comparisons, GM-R has comparable extent to the learned metric on 3DPW. While the learned metric is smaller on EMDB, OACG applies to frozen predictors without training an additional geometry network, making decoded spatial efficiency an explicit design target for structured nonlinear models.

REPRODUCIBILITY STATEMENT

An anonymous repository provides 18,480 frozen per-allocation records: 5,400 analytic, 5,040 learned-baseline, 840 normalisation-ablation, and 7,200 ridge-sweep rows, covering 308 units in those analyses. It also provides 72 frozen selected configurations, sequence-level split and rotation manifests, the learned baseline's selected settings and checkpoint hashes, and source code for geometry, protocol, reporting, and an asset-dependent experimental rerun. Per-allocation records and a table-generation script for the B1–B3 output-space controls are also included.

Two reproduction paths. Path 1 regenerates tables for those analyses, numerical sweep values, and the sweep figure from frozen records on a CPU. Its numerical artifacts can be compared byte for byte with released reference outputs; figure rendering is checked separately because it may vary across Matplotlib versions. This path checks declared reporting arithmetic while treating the recorded metrics as inputs. A separate script regenerates the B1–B3 table from its released per-allocation records. Path 2 supplies a four-stage analytic rerun: cache predictions, convert reference shapes, cache decoder pullbacks, and recompute the six analytic geometries, conformal radii, sampling, and exact decoded extent. It uses frozen selected configurations, without repeating development selection. Path 2 requires separately obtained datasets, body models, and predictor checkpoints; it is not a turnkey re-execution of the original evaluation program, which is not released. We ran Path 2 locally across the reported analytic task settings and reproduced their evaluation results within the documented numerical and Monte Carlo variation. Path 2 produces analytic records, not the frozen learned-baseline evaluation or the B1–B3 controls: C's training procedure and checkpoint hashes are provided, but its trained weights are not.

What the release verifies. The split manifests and tests check sequence-level development, calibration, evaluation, and rotation roles. The strict rerun check compares the reconstructed sequence layout with those manifests; it does not certify frame identity, because the manifests list sequences rather than sampled frames. All 72 analytic configurations record `test_rows_read: false`, but this is a recorded assertion. Released CSV rows omit record-to-split and record-to-code bindings, and the original evaluation program is absent. The reported single-read discipline and code-revision binding therefore cannot be independently established from the released records.

Numerical scope. Runs used NumPy 1.21.5, PyTorch 2.7.0, CUDA 12.8, driver 570.195.03, and an NVIDIA A100 80GB PCIe. Bitwise agreement of frozen inner-moment hashes requires the pinned environment and inputs. In one documented Path 2 check, coverage and radius matched exactly, while Monte Carlo extent differed by about 10^{-3} relative because CPU and CUDA use different sampling streams (Appendix B.3); that figure describes this check, not a uniform tolerance. Most asset file hashes were not recorded, limiting exact asset verification.

Assumptions and proofs are in Appendix A; geometry and evaluation procedures are in Appendices A.4 and B.3.

AI USE STATEMENT

Generative AI tools assisted with code development, manuscript organization and language editing, literature discovery, experimental planning, and interpretation of results. Numerical claims were checked against the underlying CSV and JSON records. The authors take responsibility for the final manuscript and its findings.

REFERENCES

Anastasios N Angelopoulos and Stephen Bates. Conformal prediction: A gentle introduction. *Foundations and Trends in Machine Learning*, 16(4):494–591, 2023.

Rajendra Bhatia, Tanvi Jain, and Yongdo Lim. On the bures–wasserstein distance between positive definite matrices. *Expositiones mathematicae*, 37(2):165–191, 2019.

Sacha Braun, Liviu Aolaritei, Michael I. Jordan, and Francis Bach. Minimum volume conformal sets for multivariate regression. *Journal of the American Statistical Association*, 2026. doi: 10.1080/01621459.2026.2692078.

Matt Y. Cheung, Ashok Veeraraghavan, and Guha Balakrishnan. COMPASS: Robust feature conformal prediction for medical segmentation metrics. In *International Conference on Learning Representations*, 2026a. URL https://openreview.net/forum?id=uBy4TCgGiT.

Matt Y. Cheung, Ashok Veeraraghavan, and Guha Balakrishnan. Efficient conformal volumetry for template-based segmentation. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2026*, 2026b. URL https://papers.miccai.org/miccai-2026/0339-Paper2641.html.

Robin Dunn, Larry Wasserman, and Aaditya Ramdas. Distribution-free prediction sets for two-layer hierarchical models. *Journal of the American Statistical Association*, 118(544):2491–2502, 2023.

Shai Feldman, Stephen Bates, and Yaniv Romano. Calibrated multiple-output quantile regression with representation learning. *Journal of Machine Learning Research*, 24(24):1–48, 2023.

Shubham Goel, Georgios Pavlakos, Jathushan Rajasegaran, Angjoo Kanazawa, and Jitendra Malik. Humans in 4D: Reconstructing and tracking humans with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14783–14794, 2023.

Alexander Immer, Maciej Korzepa, and Matthias Bauer. Improving predictions of bayesian neural nets via local linearization. In *International conference on artificial intelligence and statistics*, pp. 703–711. PMLR, 2021.

Nikos Kolotouros, Georgios Pavlakos, Dinesh Jayaraman, and Kostas Daniilidis. Probabilistic modeling for human mesh recovery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11605–11614, 2021.

Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523): 1094–1111, 2018.

Tianye Li, Timo Bolkart, Michael J. Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4D scans. *ACM Transactions on Graphics*, 36(6):194:1–194:17, 2017. doi: 10.1145/3130800.3130813.

Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Transactions on Graphics*, 34(6):248:1–248:16, 2015. doi: 10.1145/2816795.2818013.

Soundouss Messoudi, Sébastien Destercke, and Sylvain Rousseau. Ellipsoidal conformal inference for multi-target regression. In *Proceedings of the Eleventh Symposium on Conformal and Probabilistic Prediction with Applications*, volume 179 of *Proceedings of Machine Learning Research*, pp. 294–306. PMLR, 2022. URL https://proceedings.mlr.press/v179/messoudi22a.html.

Gabriel Peyré and Marco Cuturi. Computational optimal transport: With applications to data science. *Foundations and Trends in Machine Learning*, 11(5–6):355–607, 2019.

Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics*, 36(6):245:1–245:17, 2017. doi: 10.1145/3130800.3130883.

Hang Shao, Abhishek Kumar, and P. Thomas Fletcher. The riemannian geometry of deep generative models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 315–323, 2018.

David Stutz, Krishnamurthy Dvijotham, Ali Taylan Cemgil, and Arnaud Doucet. Learning optimal conformal classifiers. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=t8O-4LKFVx.

Gauthier Thurin, Kimia Nadjahi, and Claire Boyer. Optimal transport-based conformal prediction. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 59509–59527. PMLR, 2025. URL https://proceedings.mlr.press/v267/thurin25a.html.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*. Springer, 2005.

Harry Zhang and Luca Carlone. CUPS: Improving human pose-shape estimators with conformalized deep uncertainty. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 74583–74601. PMLR, 2025a. URL https://proceedings.mlr.press/v267/zhang25g.html.

Harry Zhang and Luca Carlone. CHAMP: Conformalized 3D human multi-hypothesis pose estimators. In *International Conference on Learning Representations*, 2025b. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/93bb6829b9a81c339d0b992372da2e95-Abstract-Conference.html.

A Proofs and Auxiliary Results

The proofs follow the construction from validity to efficiency. Exact decoding transfers score coverage to complete model outputs; a finite-calibration bound then motivates the pooled moment proxy. The remaining results distinguish the full-rank GM optimum from ridge placement and finite-radius nonlinear extent.

A.1 Coverage under Exact Decoding

Validity depends on set inclusion rather than local decoder accuracy: a reference residual inside the calibrated ellipsoid decodes to an output inside the prediction set.

Proof of Theorem 4.1. Condition on the frozen construction. Its measurable score rule maps exchangeable observations to exchangeable scores. Attach independent continuous tie breakers to the $n+1$ scores; the future score's rank is then uniform on $\{1,\dots,n+1\}$. For $k \le n$, rank at most k implies that the future score is no larger than the kth calibration score, so

$$\Pr\{S_{n+1} \le S_{(k)}\} \ge \frac{k}{n+1} \ge 1-\alpha.$$

For $k = n+1$ the threshold is infinite and score coverage is one. Tie breakers are a proof device; calibration uses the original scores. Representability and the definitions in Section 3.1 give

$$\{S_{n+1} \le \widehat{q}\} = \{r^\star_{n+1} \in \mathcal{R}_{A,\widehat{q}}(X_{n+1})\} \subseteq \{Y^\star_{n+1} \in \mathcal{C}_{A,\widehat{q}}(X_{n+1})\},$$

which proves the first inequality without a unique inverse parameterization or a local decoder approximation. The argument is the split-conformal rank proof (Vovk et al., 2005; Lei et al., 2018) followed by exact decoding. □

Sequence sampling and coverage targets. An identical randomized sampling kernel applied independently to each exchangeable sequence commutes with permutations, so the selected observations are exchangeable (Dunn et al., 2023) and Theorem 4.1 applies to the sampled observation. This does not make arbitrary dependent sequences exchangeable. Replacing each sequence's score by the maximum over its observations gives an exchangeable sequence statistic whose calibration covers all observations of a future finite sequence simultaneously; average framewise coverage under a sampled-observation threshold has no such simultaneous interpretation.

Finite-pool randomization. Fix $N > n$ nonempty finite sequences and a score rule fixed independently of the role allocation and within-sequence draws. Choose n calibration sequences uniformly without replacement and one observation uniformly from each. For sequence lengths L_j and $k \le n$,

$$\mathbb{E}_{\text{roles,draws}}\left[\frac{1}{N-n}\sum_{j\notin I_{\text{cal}}}\frac{1}{L_j}\sum_{t=1}^{L_j}\mathbf{1}\{S_{jt}\le \widehat{q}\}\right] \ge \frac{k}{n+1}.$$

For the proof, draw one observation independently from every sequence, including test sequences, as a coupling. Conditional on these draws, permute the sequence indices uniformly, using the first n for calibration and position $n+1$ as a distinguished test sequence. Its rank among these $n+1$ scores is uniform after tie breaking; averaging over permutations and, for the test sequence, over its uniform draw yields the displayed equal-sequence average. Decoded membership is at least as large by deterministic inclusion. This statement conditions on the pool and averages over fresh randomization; it does not guarantee coverage for one realized split, for a sequence outside the pool, or for a new participant. The score rule must be independent of the role allocation and within-sequence draws; population coverage additionally requires the exchangeability assumed in Theorem 4.1. Repeating allocations of the same pool does not create independent population samples.

A.2 Finite Calibration and the Moment Bound

Calibration sets the radius, while the metric affects both the scores used to determine it and the displacement that radius permits. The bound below couples these effects for linearized extent; it does not order the final nonlinear sets.

For a frozen metric A, let $M_x = \mathbb{E}[r^\star r^{\star\top} \mid X = x]$ and define

$$a_A(x) = \operatorname{tr}(A(x)M_x), \qquad b_A(x) = \operatorname{tr}(G(x)A(x)^{-1}). \tag{13}$$

Then $\mathbb{E}S_A^2 = \mathbb{E}_X a_A(X)$ and $D_{\mathrm{lin}}(x; A, q) = q^2 b_A(x)/(d+2)$, yielding the product objective in Eq. (11).

Proposition A.1 (Finite-calibration linearized extent bound)**.** *Let calibration and test observations be conditionally i.i.d. and let $k = \lceil (n+1)(1-\alpha) \rceil \le n$. For any frozen measurable $A(x) \succ 0$ with both factors of $\mathcal{U}[A]$ finite,*

$$\mathbb{E}_{\mathrm{cal,X}} D_{\mathrm{lin}}(X; A, \widehat{q}_A) \le \frac{n}{(d+2)(n-k+1)} \mathcal{U}[A]. \tag{14}$$

Proof of Proposition A.1. Condition on the frozen construction. For ordered nonnegative calibration scores and $k \le n$,

$$\widehat{q}_A^2 = S_{(k)}^2 \le \frac{1}{n-k+1} \sum_{j=k}^{n} S_{(j)}^2 \le \frac{1}{n-k+1} \sum_{i=1}^{n} S_i^2,$$

so $\mathbb{E}_{\mathrm{cal}} \widehat{q}_A^2 \le n\,\mathbb{E}S_A^2/(n-k+1)$, while conditioning on X gives $\mathbb{E}S_A^2 = \mathbb{E}_X \operatorname{tr}(A(X)M_X)$. Independence of the test input from the calibration sample implies

$$\mathbb{E}_{\mathrm{cal,X}} D_{\mathrm{lin}}(X; A, \widehat{q}_A) = \frac{\mathbb{E}_{\mathrm{cal}} \widehat{q}_A^2}{d+2}\, \mathbb{E}_X b_A(X) \le \frac{n}{(d+2)(n-k+1)} \mathcal{U}[A].$$

The order-statistic bound is deterministic; only the factorization uses independence. □

Proposition A.2 (Size criteria for fixed linearized outputs)**.** *Let $G, M \succ 0$ be input independent, fix a calibration sample with $k \le n$, and let $A \succ 0$ be input independent. Write $\mu_A = \lambda_{\max}(A^{-1/2} G A^{-1/2})$ and let $\widehat{q}_A$ be the calibrated score threshold. (i) The squared worst-case* linearized *output displacement over $\mathcal{R}_{A,\widehat{q}_A}$ is $\mu_A \widehat{q}_A^2$, and $A = G$ minimizes it over these metrics for this sample. (ii) The moment-based proxy $\operatorname{tr}(AM)\det(A)^{-1/d}$ for squared score times parameter-volume scale is minimized when $A \propto M^{-1}$. These statements do not rank the regularized, input-dependent metrics by their calibrated nonlinear extent or per-output-point width.*

Proof. For (i), $G \preceq \mu_A A$, so every calibration residual obeys $S_G(r_i) \le \sqrt{\mu_A} S_A(r_i)$. Order statistics preserve this inequality, giving $\widehat{q}_G^2 \le \mu_A \widehat{q}_A^2$; $\mu_G = 1$. For (ii), let $K = M^{1/2} A M^{1/2}$. The arithmetic–geometric mean inequality on its eigenvalues gives $\operatorname{tr} K \det(K)^{-1/d} \ge d$, with equality exactly when K is a scalar multiple of I. Hence the proxy is at least $d \det(M)^{1/d}$ and equality holds for $A \propto M^{-1}$. □

For fixed $\alpha \in (0,1)$, $(n-k+1)/n \to \alpha$. In our protocols ($n = 40$), nominal coverage levels 80%, 90%, and 95% give $k = 33, 37, 39$ and multipliers $5, 10, 20$. These quantify the looseness of the score-moment bound, not linearization accuracy or a method's efficiency gap. When $0 < \mathbb{E}S_A^2 < \infty$,

$$\mathbb{E}_{\mathrm{cal,X}} D_{\mathrm{lin}} = \frac{\kappa_A}{d+2} \mathcal{U}[A], \qquad \kappa_A = \frac{\mathbb{E}_{\mathrm{cal}} \widehat{q}_A^2}{\mathbb{E}S_A^2},$$

so quantile-to-moment ratios can differ between geometries. Neither statement certifies plug-in values for one realized calibration sample.

A.3 Global Moment Proxy: Optimality and Singular Case

Replacing M_x by $\widetilde{M}$ gives the pooled objective $\widetilde{\mathcal{U}}$. It equals $\mathcal{U}$ when $M_x \equiv M = \widetilde{M}$; otherwise, estimation, regularization, population differences, and $A(X)$-weighted conditional-moment mismatch can separate the objectives. Neither pooled optimality nor moment control determines calibrated nonlinear extent. Below we derive the proxy effect of input-dependent trace normalization and distinguish the unregularized core from the complete metrics used in the ablation.

Proposition A.3 (Global optimum of the pooled moment proxy)**.** *Fix $\widetilde{M} \succ 0$. If $G(X) \succ 0$ almost surely and $0 < \mathbb{E}_X \operatorname{tr}(H(X)^{1/2}) < \infty$, then all minimizers of Eq.* (4) *over measurable positive-definite fields with finite factors are, up to null sets,*

$$A(x) = cB(x), \qquad c > 0 \text{ common to all inputs}, \tag{15}$$

and the minimum value is $(\mathbb{E}_X \operatorname{tr}(H(X)^{1/2}))^2$.

Condition on development data and fix $\widetilde{M} \succ 0$. For a measurable field $A(x) \succ 0$, define

$$\widetilde{a}_A(x) = \operatorname{tr}(A(x)\widetilde{M}), \qquad b_A(x) = \operatorname{tr}(G(x)A(x)^{-1}), \qquad \tau(x) = \operatorname{tr}(H(x)^{1/2}),$$

so that $\widetilde{\mathcal{U}}[A] = \mathbb{E}\widetilde{a}_A\, \mathbb{E}b_A$ with both expectations over the same input distribution.

Lemma A.1 (Pointwise matrix inequality)**.** *For positive-definite $G(x)$ and $\widetilde{M}$,*

$$L_x(A) = \operatorname{tr}(A(x)\widetilde{M}) \operatorname{tr}(G(x)A(x)^{-1}) \geq \tau(x)^2, \tag{16}$$

with equality precisely when $A(x) = c(x)B(x)$ for some $c(x) > 0$ (Bhatia et al., 2019).

Proof. Suppress x and write $K = \widetilde{M}^{1/2} A \widetilde{M}^{1/2}$, so the objective is $\operatorname{tr} K \operatorname{tr}(HK^{-1})$. Hilbert–Schmidt Cauchy–Schwarz gives

$$(\operatorname{tr} H^{1/2})^2 = \langle K^{1/2}, K^{-1/2}H^{1/2}\rangle_F^2 \leq \|K^{1/2}\|_F^2 \, \|K^{-1/2}H^{1/2}\|_F^2 = \operatorname{tr} K \operatorname{tr}(HK^{-1}),$$

with equality iff $K^{-1/2}H^{1/2} = tK^{1/2}$ for some $t > 0$, i.e. $K = t^{-1}H^{1/2}$ and $A = t^{-1}B$. No commutation assumption is needed. □

Proof of Proposition A.3. Assume $G(X) \succ 0$ almost surely and $0 < \mathbb{E}\tau < \infty$. Lemma A.1 gives $\widetilde{a}_A b_A \geq \tau^2$ pointwise, hence over measurable positive-definite fields with finite factors

$$\widetilde{\mathcal{U}}[A] \geq \left(\mathbb{E}\sqrt{\widetilde{a}_A b_A}\right)^2 \geq (\mathbb{E}\tau)^2.$$

Equality in the second inequality requires $A(x) = c(x)B(x)$ almost surely. Using $\operatorname{tr}(B\widetilde{M}) = \operatorname{tr}(GB^{-1}) = \tau$, equality in the first inequality then requires $\mathbb{E}[c\tau]\,\mathbb{E}[\tau/c] = (\mathbb{E}\tau)^2$. The Cauchy–Schwarz equality condition forces c to be one positive constant almost surely. Hence the minimizers are exactly $A = cB$ up to null sets, and the minimum value is $(\mathbb{E}\tau)^2$. Averaging the pointwise objective L_x instead would give a different functional that does not fix relative scale across inputs. □

Singular case. Let $G(X) \succeq 0$ and $\mathbb{E}\tau < \infty$. In residual-normalized coordinates $K = \widetilde{M}^{1/2} A \widetilde{M}^{1/2}$, we still have $\widetilde{a}_A = \operatorname{tr} K$, $b_A = \operatorname{tr}(HK^{-1})$, and $\widetilde{a}_A b_A \geq \tau^2$. For $K_\epsilon = H^{1/2} + \epsilon I$ with eigenvalues h_j of H,

$$a_\epsilon = \tau + d\epsilon, \qquad b_\epsilon = \sum_j \frac{h_j}{\sqrt{h_j} + \epsilon} \leq \tau,$$

and dominated convergence gives $\widetilde{\mathcal{U}}[A_\epsilon] \to (\mathbb{E}\tau)^2$ as $\epsilon \downarrow 0$. If H is nonzero and singular on a set of positive probability, $\widetilde{a}_A b_A > \tau^2$ there for every $K \succ 0$, so the infimum is unattained. If $H = 0$ on a set of positive probability but $\mathbb{E}\tau > 0$, equality in the expectation-level Cauchy–Schwarz step fails there because $\widetilde{a}_A > 0$ and $b_A = 0$. Thus any positive-probability singular set precludes attainment when $\mathbb{E}\tau > 0$; if $G = 0$ almost surely, the proxy vanishes for every metric and expresses no preference.

The approximating family above is a witness for the infimum, not a ridge prescription. Note also that shrinking an arbitrary metric along $\ker G$ need not leave the inverse-trace term unchanged, since off-diagonal couplings enter that term. These facts motivate treating ridge coordinates and strength as explicit design choices selected on development data.

Scope of the proxy. The conditional squared-score moment is $\operatorname{tr}(A(x)M_x)$ with $M_x = \mathbb{E}[r^\star r^{\star\top} \mid X = x]$, which need not equal $\operatorname{tr}(A(x)\widetilde{M})$; even when they agree, the calibration quantile is not determined by the mean. Regularization, finite calibration, and nonlinear decoding therefore remain part of the empirical comparison. The normalization ablation replaces the complete metric by $A_h / \operatorname{tr}(A_h \widetilde{M})$ and recalibrates each variant with fitted statistics fixed, testing this scale choice without changing axis ratios or the core-to-ridge balance at a fixed input.

Trace normalization and the moment proxy. Fix $\widetilde{M} \succ 0$ and a measurable field $A(x) \succ 0$. Write $\widetilde{a}_A(x) = \mathrm{tr}(A(x)\widetilde{M})$, $b_A(x) = \mathrm{tr}(G(x)A(x)^{-1})$, and $A^\circ(x) = A(x)/\widetilde{a}_A(x)$. Assume $\mathbb{E}\widetilde{a}_A$, $\mathbb{E}b_A$, and $\mathbb{E}[\widetilde{a}_A b_A]$ are finite under the same input distribution. Since $\widetilde{a}_{A^\circ} = 1$ and $b_{A^\circ} = \widetilde{a}_A b_A$, trace normalization gives

$$\widetilde{\mathcal{U}}[A^\circ] - \widetilde{\mathcal{U}}[A] = \mathbb{E}[\widetilde{a}_A b_A] - \mathbb{E}\widetilde{a}_A\, \mathbb{E}b_A = \mathrm{Cov}(\widetilde{a}_A, b_A). \tag{17}$$

The change can have either sign. For the unregularized GM core with $G(X) \succ 0$ almost surely and $\mathbb{E}\tau^2 < \infty$, the identities $\widetilde{a}_B = b_B = \tau$ imply

$$\widetilde{\mathcal{U}}[B/\tau] - \widetilde{\mathcal{U}}[B] = \mathbb{E}\tau^2 - (\mathbb{E}\tau)^2 = \mathrm{Var}(\tau). \tag{18}$$

Thus the variance penalty applies to the unregularized core under this specific normalization. The ablation normalizes complete metrics, including their ridges, for which the covariance identity applies. Its empirical variability diagnostic therefore uses $t_h(x) = \mathrm{tr}(A_h(x)\widetilde{M})$ rather than the core-only $\tau(x) = \mathrm{tr}(H(x)^{1/2})$ (Appendix B.8). Neither identity determines the change in calibrated nonlinear extent, which also depends on score quantiles and decoder response.

Exactness and discrepancy of pooling. Let $M_{\mathrm{pool}} = \mathbb{E}_X M_X$ under the input distribution shared by both objectives. For fixed A and finite moments,

$$\begin{aligned} \widetilde{\mathcal{U}}[A] - \mathcal{U}[A] &= \mathbb{E}_X b_A(X)\,(\epsilon_{\mathrm{pool}} + \epsilon_{\mathrm{cond}}), \\ \epsilon_{\mathrm{pool}} &= \mathbb{E}_X\, \mathrm{tr}\Big[A(X)(\widetilde{M} - M_{\mathrm{pool}})\Big], \qquad \epsilon_{\mathrm{cond}} = \mathbb{E}_X\, \mathrm{tr}[A(X)(M_{\mathrm{pool}} - M_X)], \end{aligned} \tag{19}$$

by adding and subtracting M_{pool} in the first factor. The first term collects estimation, regularization, and any mismatch between the development moment and the population moment; the second is the conditional-moment mismatch weighted by $A(X)$. Neither is an unweighted measure of heteroscedasticity; they may cancel, and the second vanishes for an input-independent metric even when M_X varies. If $M_X \equiv M$ and $\widetilde{M} = M$, both vanish. Under the same constant-moment condition, for fixed A with $\mathbb{E}\,\mathrm{tr}\,A < \infty$ and $\mathbb{E}b_A < \infty$,

$$|\widetilde{\mathcal{U}}[A] - \mathcal{U}[A]| \le \|\widetilde{M} - M\|_{\mathrm{op}}\, \mathbb{E}\,\mathrm{tr}\,A\, \mathbb{E}b_A,$$

so operator-norm consistency of $\widetilde{M}$ implies convergence of the objectives for this fixed geometry; this is not uniform over refitted metrics. Since $M_x = \Sigma_x + \mu_x\mu_x^\top$, a constant conditional covariance alone does not give a constant second moment about the fixed prediction.

Conditional-moment oracle. If $M_X, G(X) \succ 0$ almost surely and $0 < \mathbb{E}\,\mathrm{tr}[(M_X^{1/2}G(X)M_X^{1/2})^{1/2}] < \infty$, the proof above with $\widetilde{M}$ replaced pointwise by M_X gives all minimizers of $\mathcal{U}[A]$ as $A(X) = c\,(M_X^{-1}\#G(X))$ with one common $c > 0$, and minimum equal to the square of the displayed expectation. This oracle statement does not identify a pooled estimate with M_X.

A.4 Geometry Details

Residual-moment estimation. From N_{fit} development residuals, estimate $\widehat{M} = N_{\mathrm{fit}}^{-1}\sum_i r_i^\star r_i^{\star\top}$ and set

$$M_\rho = (1-\rho)\widehat{M} + \rho\,\mathrm{diag}(\widehat{M}), \qquad \widetilde{M} = M_\rho + \eta s_M I, \tag{20}$$

where $s_M = \mathrm{tr}(M_\rho)/d$, $\rho \in [0,1]$, and $\eta \ge 0$; only $\widetilde{M} \succ 0$ is retained. The uncentered moment retains squared bias about the fixed prediction. Shrinkage and the inner ridge stabilize estimation; residual-based geometries share this estimator family but select (ρ, η) separately.

As the outer ridge strengths vanish, GM-I and GM-R approach B. The development grid includes small strengths for full-rank settings.

Table 4 compares the complete analytic geometries and separates their cores from their outer ridges.

Table 4: **Analytic geometries in residual-normalized coordinates.** The last two columns sum to $\bar{A} = \widetilde{M}^{1/2} A \widetilde{M}^{1/2}$, separating the fusion rule from the coordinates of the outer ridge.

Geometry	Complete A	Core of $\bar{A}$	Outer ridge of $\bar{A}$
Residual-only	$\widetilde{M}^{-1}$	I	0
Sensitivity-only	$G + \lambda_h s_G I$	H	$\lambda_h s_G \widetilde{M}$
Additive	$G + \gamma \widetilde{M}^{-1} + \lambda_h s_G I$	$H + \gamma I$	$\lambda_h s_G \widetilde{M}$
GM-I	$B + \lambda_{\mathrm{I}} s_B I$	$H^{1/2}$	$\lambda_{\mathrm{I}} s_B \widetilde{M}$
GM-R	$B + \lambda_{\mathrm{R}} c_M \widetilde{M}^{-1}$	$H^{1/2}$	$\lambda_{\mathrm{R}} c_M I$

Each family uses its own frozen configuration. Scale definitions and trace matching are given in this subsection. Additive uses the scale of its G component. Residual-only is regularized through $\widetilde{M}$ (Eq. 20); the isotropic reference $A = I$ gives $\bar{A} = \widetilde{M}$.

Additive fusion. The additive coefficient is

$$\gamma(x) = w \, \frac{\operatorname{tr} G(x) + \varepsilon_g}{\operatorname{tr}(\widetilde{M}^{-1})}, \tag{21}$$

with w selected on development data and $\varepsilon_g = 10^{-6}$ a numerical offset expressed in the stored Gram-matrix scale of G; it must be transformed with G when that scale changes. The complete additive geometry and its outer ridge appear in Table 4.

Frozen scales and trace matching. The averaging set is FIT during each selection round and all development rows after selection; RADIUS and RANK never supply scale constants. In the table, $\langle \cdot \rangle_{\mathrm{dev}}$ denotes the applicable fitting set, and s_B corresponds to `s_gm`.

Constant	Definition	Role
s_M	$\operatorname{tr}(M_\rho)/d$	Inner moment ridge
s_G	$\langle \operatorname{tr} G \rangle_{\mathrm{dev}}/d$	Sensitivity and additive ridge
s_B	$\langle \operatorname{tr} B \rangle_{\mathrm{dev}}/d$	GM-I ridge
c_M	$d\, s_B / \operatorname{tr}(\widetilde{M}^{-1})$	GM-R ridge

Development averages are computed after fitting the selected moment estimator on all development data and are reused, not recomputed, at calibration and evaluation. For the same B, $\widetilde{M}$, and coefficient λ, $\operatorname{tr}(\lambda s_B I) = \lambda d s_B = \operatorname{tr}(\lambda c_M \widetilde{M}^{-1})$; this matches total trace, not eigenvalues, and makes no claim across independently tuned families.

Coordinate changes. SMPL parameter coordinates combine rotation increments in radians with shape-coefficient units. For an invertible linear change $\delta' = C\delta$, transform

$$\widetilde{M}' = C \widetilde{M} C^\top, \qquad G' = C^{-\top} G C^{-1}, \qquad A' = C^{-\top} A C^{-1},$$

which preserves scores and the represented region. The GM core obeys $B' = C^{-\top} B C^{-1}$, as does $B + \kappa \widetilde{M}^{-1}$ for a fixed absolute coefficient κ, whereas an identity ridge becomes $\kappa C^{-\top} C^{-1}$, generally anisotropic. Re-estimating diagonal shrinkage, recomputing trace normalizers, or searching the same finite grid in new coordinates need not preserve this relation; the property concerns the matrix construction with consistently transformed inputs, not the full fitting procedure. Uniform measure on a fixed ellipsoid also transforms consistently, preserving exact decoded extent.

Kernel and directional sections. Along a unit $v \in \ker G$, the GM-I and GM-R central sections have respective half-lengths

$$\frac{\widehat{q}_{\mathrm{I}}}{\sqrt{\lambda_{\mathrm{I}} s_B}}, \qquad \frac{\widehat{q}_{\mathrm{R}}}{\sqrt{\lambda_{\mathrm{R}} c_M \, v^\top \widetilde{M}^{-1} v}}. \tag{22}$$

Each uses its own fitted statistics and calibrated radius; neither is uniformly shorter. Both ridges also act outside the kernel.

The GM formula gives $B\widetilde{M}B = G$, hence for every v,

$$v^\top G v = \|\widetilde{M}^{1/2} B v\|_2^2 = \|W^{1/2} J v\|_2^2,$$

and since $G, B \succeq 0$ and $\widetilde{M} \succ 0$, $\ker B = \ker G = \ker(W^{1/2}J)$. For a unit v in this kernel, substituting $\delta = tv$ into each complete metric gives $t^2 \lambda_{\mathrm{I}} s_B \le \widehat{q}_{\mathrm{I}}^2$ and $t^2 \lambda_{\mathrm{R}} c_M v^\top \widetilde{M}^{-1} v \le \widehat{q}_{\mathrm{R}}^2$, proving Eq. (22). The kernel does not depend on the residual moment, although the penalty and radius do. These are intersections with a line; projected widths depend on other directions.

A.5 Finite-Radius Identities and Local Expansions

For fixed x, let $\delta = qA(x)^{-1/2}U$ with $U \sim \mathrm{Unif}(\mathbb{B}_d)$, and define

$$e_x(\delta) = \tfrac{1}{2}\{g_x(\delta) + g_x(-\delta)\} - g_x(0), \tag{23}$$

$$o_x(\delta) = \tfrac{1}{2}\{g_x(\delta) - g_x(-\delta)\}. \tag{24}$$

Proposition A.4 (Exact finite-radius decomposition)**.** *Suppose $J(x) = Dg_x(0)$ exists, $q < \infty$, $A(x) \succ 0$, and the decoded displacement is square-integrable. Define*

$$\Delta_{\mathrm{odd}}(x; A, q) = \mathbb{E}_U\left[\|o_x(\delta)\|_W^2 - \|J(x)\delta\|_W^2\right], \tag{25}$$

$$D_{\mathrm{even}}(x; A, q) = \mathbb{E}_U\|e_x(\delta)\|_W^2. \tag{26}$$

Then

$$D(x; A, q) = D_{\mathrm{lin}}(x; A, q) + \Delta_{\mathrm{odd}}(x; A, q) + D_{\mathrm{even}}(x; A, q). \tag{27}$$

Proof of Proposition A.4. Suppress x. The centered outputs at $\pm\delta$ are $e(\delta) \pm o(\delta)$, so

$$\tfrac{1}{2}\left(\|g(\delta) - g(0)\|_W^2 + \|g(-\delta) - g(0)\|_W^2\right) = \|e(\delta)\|_W^2 + \|o(\delta)\|_W^2.$$

Under the symmetric perturbation law the expected left side equals D; adding and subtracting $\mathbb{E}\|J\delta\|_W^2 = D_{\mathrm{lin}}$ gives Eq. (27). Equivalently, the cross term $\langle e, o\rangle_W$ is odd with zero expectation. Square integrability justifies the expectations. □

Null-direction consequence. For a unit $v \in \ker G$ and g_x twice differentiable at zero, as $t \to 0$,

$$\|g_x(tv) - g_x(0)\|_W^2 = \tfrac{1}{4}t^4\|D^2 g_x(0)[v, v]\|_W^2 + o(t^4). \tag{28}$$

If g is twice differentiable at zero, then along fixed v, $g(tv) - g(0) = tJv + \frac{1}{2}t^2 D^2 g(0)[v, v] + o(t^2)$. For $v \in \ker G$, $W^{1/2}Jv = 0$, which yields Eq. (28). If g is locally C^3, cancellation of quadratic terms gives $W^{1/2}o(tv) = O(t^3)$; without $W \succ 0$ the unweighted Jv need not vanish. Under the second-order expansion alone the weighted odd part is $o(t^2)$, which suffices for the fourth-order squared response. A zero Hessian coefficient does not rule out higher-order effects or exact invariance.

Proposition A.5 (Fourth-order extent expansion)**.** *Fix $A \succ 0$ and let $P = A^{-1}$. Suppose g has a third-order Taylor expansion at zero and $h(\delta) = \|g(\delta) - g(0)\|_W^2$ is C^6 on a neighborhood of the ball of radius $r_A = q/\sqrt{\lambda_{\min}(A)}$ with $\|D^6 h\|_{\mathrm{op}} \le K_h$ there. For output coordinates a, b, write $Q_a = D^2 g_a(0)$ and $T_{a,ijk} = \partial_{ijk} g_a(0)$. Then*

$$D(A, q) = \frac{q^2}{d+2}\,\mathrm{tr}(GP) + \frac{q^4}{(d+2)(d+4)}\left(\frac{\mathcal{H}_A}{4} + \mathcal{T}_A\right) + R_6, \tag{29}$$

$$\mathcal{H}_A = \sum_{a,b} W_{ab}\left\{\mathrm{tr}(PQ_a)\,\mathrm{tr}(PQ_b) + 2\,\mathrm{tr}(PQ_aPQ_b)\right\} \ge 0, \tag{30}$$

$$\mathcal{T}_A = \sum_{a,b,i,j,k,l} W_{ab} J_{ai} T_{b,jkl} P_{ij} P_{kl}, \tag{31}$$

$$|R_6| \le \frac{K_h}{720}\,\mathbb{E}\|\delta\|_2^6 \le \frac{K_h}{720}\,\frac{d}{d+6}\,r_A^6. \tag{32}$$

With only local C^3 smoothness of g, the same coefficients hold with an $o(q^4)$ remainder as $q \to 0$ for fixed A.

Proof. Write a uniform unit-ball draw as $U = RV$ with V uniform on the sphere and $\Pr(R \le r) = r^d$, so $\mathbb{E}R^4 = d/(d+4)$ and

$$\mathbb{E}[\delta_i\delta_j\delta_k\delta_l] = \frac{q^4}{(d+2)(d+4)}(P_{ij}P_{kl} + P_{ik}P_{jl} + P_{il}P_{jk}).$$

The degree-four terms of the squared Taylor displacement are $\frac{1}{4}\|D^2g(0)[\delta,\delta]\|_W^2$ and $\frac{1}{3}\langle J\delta, D^3g(0)[\delta,\delta,\delta]\rangle_W$; the three pairings give $\mathcal{H}_A$ and $\mathcal{T}_A$, symmetry of the third derivative cancelling the factor $1/3$. The Hessian contraction is a positive multiple of an expected squared weighted norm, hence nonnegative. Applying the fifth-degree Taylor formula to h, the constant and linear terms vanish, odd degrees vanish by symmetry, and the remainder is bounded by $K_h\|\delta\|^6/6!$ with $\mathbb{E}\|\delta\|^6 \le r_A^6\,\mathbb{E}\|U\|^6 = r_A^6 d/(d+6)$. Under local C^3 regularity only, the displacement remainder is $o(\|\delta\|^3)$ and its product with the linear term is $o(\|\delta\|^4)$. □

For a quadratic decoder, $\mathcal{T}_A = R_6 = 0$ and the expansion is exact at every radius. In general the Jacobian–third-derivative contraction is signed and cannot be discarded; it vanishes for perturbations supported in $\ker(W^{1/2}J)$, but full-ellipsoid draws retain mixed terms. An $O(q^6)$ statement for fixed A gives no uniform control as $\lambda_{\min}(A) \downarrow 0$, and we do not certify a uniform remainder constant on the calibration regions.

Coarse uniform remainder bound. If $\|g(\delta) - g(0) - J\delta\|_W \le L\|\delta\|_2^2/2$ on the ellipsoid, the reverse triangle inequality in L^2 gives

$$|\sqrt{D} - \sqrt{D_{\mathrm{lin}}}| \le \big(\mathbb{E}\|g(\delta) - g(0) - J\delta\|_W^2\big)^{1/2} \le \frac{Lq^2}{2\lambda_{\min}(A)}.$$

This requires a uniform bound, not directional probes, and is independent of the coverage argument.

Calibration, averages, and metric scale. For finite averaged quantities with $\bar{D}_{\mathrm{lin},h} > 0$, define

$$\chi_h = \frac{\bar{D}_h}{\bar{D}_{\mathrm{lin},h}} = 1 + \frac{\bar{\Delta}_{\mathrm{odd},h} + \bar{D}_{\mathrm{even},h}}{\bar{D}_{\mathrm{lin},h}}. \tag{33}$$

This is a ratio of averages under the common evaluation weights. Fix the realized calibration and use the same input weights for all averages. Averaging Eq. (27) gives $\bar{D}_h = \bar{D}_{\mathrm{lin},h} + \bar{\Delta}_{\mathrm{odd},h} + \bar{D}_{\mathrm{even},h}$; for $\bar{D}_{\mathrm{lin},h} > 0$, division gives Eq. (33), and substituting $\bar{D}_{\mathrm{lin},h} = \widehat{q}_h^2\bar{b}_h/(d+2)$ proves Eq. (12). No calibration–test independence is needed for this conditional identity. Replacing A_h by c_hA_h for a common $c_h > 0$ scales scores and threshold by $\sqrt{c_h}$ and $\bar{b}_h$ by $1/c_h$, leaving the ellipsoid, its sampling law, χ_h, and the decoded extent unchanged; this does not extend to input-dependent rescaling. Ratios of calibrated squared RMS extents factor into ratios of $\widehat{q}_h^2$, $\bar{b}_h$, and χ_h as descriptive, not causal, factors. Across repeated splits, the identity is formed within each split before aggregation; the product of across-split means is not the mean of the products.

For m evaluation sequences of lengths L_j, the protocol uses

$$\bar{D}_h = \frac{1}{m}\sum_{j=1}^{m}\frac{1}{L_j}\sum_{t=1}^{L_j} D_h(x_{jt}),$$

with the same weights for $\bar{b}_h$, $\bar{\Delta}_{\mathrm{odd},h}$, and $\bar{D}_{\mathrm{even},h}$. Thus the square root follows averaging squared displacement across observations and equally weighted sequences. The development objective in Eq. (37) uses the corresponding sequence weights across all RANK groups, with one square root after pooling; unequal group sizes do not receive equal group weights.

For C, let $\bar{D}_{C,r,s}$ denote the mean squared extent of frozen training replicate r on calibration split s. Its main-table summary for that allocation is

$$\widehat{E}_{C,s} = \frac{1}{R}\sum_{r=1}^{R}\sqrt{\bar{D}_{C,r,s}}, \qquad R = 20. \tag{34}$$

Equation (12) applies to each (r,s) separately, before this average. In general, $\widehat{E}_{C,s}^2 \neq R^{-1}\sum_r \bar{D}_{C,r,s}$; the reported average does not define a single calibrated ellipsoid or ensemble

prediction set. Paired comparisons use $\widehat{E}_{C,s}$ and the analytic method's RMS extent on the same allocation. For every method, the table SD is the descriptive spread over the 20 calibration-split-level results, computed with divisor 20 rather than 19. It is not divided by $\sqrt{20}$ and is not a standard error. For C, it describes variation of $\widehat{E}_{C,s}$ across calibration splits, not variation across training seeds.

B Instantiations and Protocol

The same calibrated-region construction applies to body and hand models. The output map changes sensitivity and fitted geometry, while the shared development and calibration roles keep comparisons aligned.

B.1 Algorithm Overview

Algorithm 1 summarizes development selection, independent calibration, and prediction using the frozen analytic geometry.

Algorithm 1 OACG for an analytic geometry family h

Require: Fixed predictor and decoder; development and calibration data; miscoverage level α

1: **Preparation**
2: Select settings using the shared five-round FIT/RADIUS/RANK protocol and Eq. (37).
3: Refit moments and scales on all development data; freeze and hash statistics, hyperparameters, and scale constants before calibration.
4: Compute S_h on calibration observations; obtain $\widehat{q}_h$ from Eq. (8).
5: **Prediction for input** x
6: Obtain $\hat{z}(x)$, compute $G(x)$, and form $A_h(x)$ from the frozen configuration.
7: **return** $\{g_x(\delta) : \delta^\top A_h(x)\delta \le \widehat{q}_h^2\}$.

B.2 Body and Hand Instantiations

Both models use right-multiplicative rotation increments and additive shape updates:

$$R_j(\delta_{R,j}) = \hat{R}_j \operatorname{Exp}(\delta_{R,j}), \qquad \beta = \hat{\beta} + \delta_\beta, \qquad r^\star_{R,j} = \operatorname{Log}(\hat{R}_j^\top R_j^\star), \tag{35}$$

where the principal logarithm selects reference rotation residuals and the exponential defines updates for arbitrary increments.

Neutral SMPL uses $\mathcal{Z} = SO(3)^{24} \times \mathbb{R}^{10}$ with $d = 82$, retaining global orientation and fixing translation to zero. Rotations use Eq. (35); shape coefficients use additive updates. The exponential defines updates on all of $\mathbb{R}^{82}$, so ellipsoids are decoded without chart truncation: the principal logarithm selects the reference residual, but sampled increments with $\|\delta_{R,j}\| > \pi$ are neither rejected nor folded, so distinct increments can decode to the same rotation; folding would change the specified perturbation measure. The joint map uses the 4DHumans asset `SMPL_to_J19.pkl`: 25 joints reordered by `SMPL_TO_OPENPOSE` are concatenated with 19 regressed joints to form $j_{44}(z)$. All indices below are zero-based. The pelvis is $p(z) = j_{44}(z)_{39}$, the regressor output `extra_joints[14]`. The evaluated subset is $\{25, \dots, 37, 43\}$ (`KEYPOINT_LIST_3DPW`); it excludes the pelvis (`USE_HIPS=False`). Let $\mathcal{V}_T(z)$ denote vertices or this 14-joint subset. In metres,

$$f_T(z) = \operatorname{vec}\big(\mathcal{V}_T(z) - \mathbf{1}_{N_T} p(z)^\top\big), \qquad W_T = I_{3N_T}/N_T, \tag{36}$$

with $N_{\text{mesh}} = 6890$ and $N_{\text{joint}} = 14$; RMS values are reported in millimetres. Each candidate and reference configuration uses its own pelvis for both mesh and joint outputs. The joint implementation subtracts $j_{44}(z)_{39}$ before selecting the 14 joints; the mesh implementation subtracts the same regressed pelvis from every vertex. Jacobians differentiate this complete output map, including root subtraction, so the displacement measure and G use identical alignment. No rotational or scale alignment is applied. Each task has its own geometry and calibration. The joint pullback is singular because $\operatorname{rank}(J) \le 42 < 82$; this dimensional bound is not an exact rank calculation. Coverage targets the neutral-model references obtained under the fixed conversion protocol, not the original gender-specific surfaces (Section B.3).

Right-hand MANO v1.2 (Romero et al., 2017) uses $SO(3)^{16} \times \mathbb{R}^{10}$ with $d = 58$ and the same updates. Its mesh task uses Eq. (36) with $N_T = 778$ and each configuration's own wrist as root; full ellipsoids are decoded without chart truncation. Section B.3.1 specifies the HaMeR/HOT3D protocol.

B.3 Development and Calibration Protocol

Each predictor remains fixed. Development selects and fits the geometry, independent calibration sets its radius, and evaluation measures coverage and exact decoded extent. The learned comparator shares the development roles but also trains a metric head.

Computational environment and reproducibility. Experiments use NumPy 1.21.5, PyTorch 2.7.0, CUDA 12.8, and an NVIDIA A100 GPU. Bitwise reproducibility is scoped to this pinned software and hardware environment with the same frozen inputs, random seeds, numerical precision, batching, and execution settings. It is not guaranteed across software versions, GPU architectures, or execution settings; fixed random seeds alone do not ensure it. The supplementary records and manifests identify frozen configurations and sequence assignments, and the regeneration scripts reconstruct their tables and numerical sweep values from these records. A separate script regenerates Table 26 from released B1–B3 per-allocation records. The four-stage analytic path caches predictions, converts reference shapes, caches pullbacks, and recomputes analytic geometry, calibration, and decoded extent from externally supplied assets. It uses the frozen configurations rather than repeating their development selection; the original evaluation program and learned-baseline weights are not released. We ran this path locally across the reported analytic task settings and reproduced their results within the documented numerical and Monte Carlo variation. The strict split check verifies sequence assignments, not frame identity. CPU and CUDA generators draw different samples from the same seed, so a rerun of Monte Carlo extent need not match the released value bitwise. In the documented HMR2.0/3DPW joint check (allocation 0, GM-R, 0.90), sequence-weighted coverage and radius match the released record, $\bar{b}_h$ and D_{lin} agree within 1.2×10^{-8} relative, and E_{RMS} differs by about 10^{-3} relative. Across eight streams on the same 240 frames, the latter has a $0.09\,\mathrm{mm}$ standard deviation. This single check illustrates sampling variability rather than a universal tolerance.

Data roles and sampling units. Development data fix the estimated moments, hyperparameters, scale constants, and learned checkpoints; final calibration sets only the threshold of each frozen geometry; evaluation data measure coverage and extent. In the SMPL sequence protocol, 3DPW uses 10 development, 40 calibration, and 10 evaluation sequences, and EMDB uses 20, 40, and 21; development allocations use seeds 12345 and 54321, and development and deployment participants overlap on both datasets. HOT3D uses 29 development sequences under three allocations described in Section B.3.1. Calibration draws one uniformly sampled eligible observation per sequence. Evaluation averages observation-level indicators within each sequence and then weights sequences equally. Repeated allocations of one pool are randomizations of that pool, not independent datasets. The same reference conversion, output map, and eligibility rule apply on both sides of calibration; the body experiments use the fixed neutral-model conversion and joint regressor. Eligibility uses principal-log residuals under the predictors sharing an allocation and does not truncate the forward perturbations that define or integrate a region.

Shared development roles. All methods use the same five sequence groups, generated by `sequence_folds(dev, 5, seed=20260914)`. Distinct sequence names are sorted, permuted with NumPy `RandomState(20260914)`, and assigned by taking every fifth element of the permutation from offsets $0, \dots, 4$. In round $f \in \{0, \dots, 4\}$, group f supplies RANK observations, group $(f + 1) \bmod 5$ supplies RADIUS observations, and the remaining three groups supply FIT observations, so every development sequence serves as RANK once and as RADIUS once. The three roles are sequence-disjoint within each round, and every sequence supplies FIT in three rounds. In particular, RADIUS in round f is RANK in round $(f + 1) \bmod 5$; the two roles are not interchangeable. The FIT/RADIUS/RANK counts are $6/2/2$ for 3DPW and $12/4/4$ for EMDB. HOT3D has group sizes $6, 6, 6, 6, 5$, giving $17/6/6$ in rounds 0–2, $18/5/6$ in round 3, and $18/6/5$ in round 4. Memberships are retained in the split manifest. ProHMR uses the 3DPW sequences and groups. Analytic methods estimate $\widetilde{M}$ and all scale constants from FIT alone; the learned range baseline uses FIT for gradient updates. RADIUS determines only a temporary threshold, the empirical 0.90 quantile with equal-sequence, uniform-within-sequence weights; RANK determines only exact-decoder extent,

from at most 24 sampled observations per sequence and 64 uniform-ball draws per observation, with seeds shared by all methods. Frame selection uses `stable_seed(seed, seq, 'extent')`; ball sampling uses `stable_seed(seed, seq, global_idx, 'ball')`. The evaluation code records both `disjoint_radius`, using RADIUS scores, and `fit_radius`, using FIT scores. Only `disjoint_radius` enters configuration selection; `fit_radius` is a diagnostic control.

For candidate h, let $\widehat{D}_{h,s}$ denote its mean squared decoded displacement on ranking sequence s, computed with the FIT assets and RADIUS threshold of the round in which s is ranked. Selection minimizes

$$\widehat{E}_{\mathrm{dev},h} = \left(\frac{1}{N_{\mathrm{dev}}} \sum_{f=0}^{4} \sum_{s \in \mathcal{S}_f} \widehat{D}_{h,s} \right)^{1/2}, \tag{37}$$

where $\mathcal{S}_f$ is the RANK group of round f; sequences are weighted equally regardless of group size, and the square root is taken once after averaging; metre-valued RMS is multiplied by 1000 for reporting. Selection is performed at nominal level 0.90 only; the 0.80 and 0.95 results reuse the selected configuration and are recalibrated at their own levels. Exact ties favor the smaller grid index. The temporary threshold is an empirical proxy for every method with no finite-sample guarantee; its sequence support is smallest on 3DPW, with two RADIUS sequences per round. For GM-I and GM-R the input-independent factor $\widetilde{M}^{\pm 1/2}$ of Eq. (5) is precomputed once per fitted moment; this is an algebraically identical evaluation order whose maximum absolute deviation from the direct computation was 3.7×10^{-10} on the verification set.

Common analytic search grids. Every instance uses the same grids, restricted to the parameters a family uses:

Parameter	Values
ρ	$0, 0.1, 0.3, 0.5, 0.7, 0.9, 1$
η	$0, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 0.3, 1$
w	$0.1, 0.25, 0.5, 1, 2, 4, 8, 16, 32$
λ_h	$10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 0.1, 0.5, 1, 2, 5, 10$

Nominal grid sizes are 1 (isotropic), 49 (residual-only), 10 (sensitivity-only), 4,410 (additive), and 490 each (GM-I, GM-R); positive-definiteness remains an admissibility requirement. Selected values at grid endpoints are marked † in Table 5; the grids are not expanded in response to a selection. $\rho = 0$ and $\eta = 0$ respectively mean no diagonal shrinkage and no inner ridge; these natural lower bounds are not marked.

Learned range baseline. The learned range baseline (C in tables) uses development sequences only. From each sequence it samples at most 200 frames without replacement, seeded by `stable_seed(20260914, seq, 'sub')`; 3DPW supplies 2,000 rows from ten sequences. Predictor image features are frozen, with no gradient propagated to the predictor. The head maps these features through two fully connected layers of width 256, each followed by ReLU, to outputs $a(x) \in \mathbb{R}^d$ and $L(x) \in \mathbb{R}^{d \times r}$, defining

$$A_\theta(x) = \mathrm{diag}(\exp(2\,\mathrm{clip}(a(x), -6, 6))) + L(x)L(x)^\top + \lambda I, \qquad d = 82. \tag{38}$$

The diagonal head has zero weights and biases, so its initial diagonal term is I. Low-rank head weights are drawn from $\mathcal{N}(0, \sigma_L^2)$ with standard deviation $\sigma_L = 10^{-3}$ and zero biases. The initial metric is therefore $(1 + \lambda)I + L_0(x)L_0(x)^\top$, retaining a nonzero, input-dependent low-rank term. These initializations are specified explicitly rather than inherited from the linear-layer defaults. Clipping the diagonal log-scale to $[-6, 6]$ bounds its exponential by e^{12}; the metric is positive definite by construction. This head is named `EllipsoidHead` upstream and `SigmaHead` in the anonymous release.

For a training mini-batch $\mathcal{B}$, let $s_i = \sqrt{r_i^\top A_\theta(x_i) r_i}$. The training objective is

$$\mathcal{L}_{\mathcal{B}}(\theta) = \frac{q_{\mathrm{soft}}^2}{d+2} \frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} \mathrm{tr}\big(G(x_i) A_\theta(x_i)^{-1}\big). \tag{39}$$

For $m = |\mathcal{B}|$ sorted scores $s_{(0)} \le \cdots \le s_{(m-1)}$, define $p_i = i/\max(m-1, 1)$ and

$$q_{\text{soft}} = \sum_{i=0}^{m-1} w_i s_{(i)}, \qquad w_i = \frac{\exp[-(p_i - \ell)^2/(2\tau_q^2)]}{\sum_{j=0}^{m-1} \exp[-(p_j - \ell)^2/(2\tau_q^2)]}, \qquad \ell = 0.90, \quad \tau_q = 0.05. \quad (40)$$

The Gaussian kernel acts on normalized ranks, with standard deviation τ_q in rank units. For $m > 1$, the rank grid includes both 0 and 1; for $m = 1$, the sole score is returned. This L-statistic distributes gradients over multiple ordered scores, whereas a hard order statistic selects one score away from ties. It is used only in the linearized training loss. Configuration selection instead uses exact decoded extent in Eq. (37), and final calibration always uses the exact order statistic in Eq. (8). This distinction also matters for joint outputs: their linearized cost cannot measure displacement along the kernel of G_T, whose dimension is at least $82 - 42 = 40$.

The search uses $r \in \{4, 16, 32, 64\}$, $\lambda \in \{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1\}$, and learning rates $\{10^{-4}, 3 \cdot 10^{-4}, 10^{-3}, 3 \cdot 10^{-3}\}$, giving 80 configurations, each assessed at the 11 checkpoints $2, 4, 8, 12, 20, 30, 40, 60, 100, 150, 200$ epochs; configuration and epoch are selected jointly, with ties resolved by candidate index then epoch. Training uses AdamW, batch size 128, weight decay 10^{-4}, and gradient-norm clipping at 5; a nonfinite loss raises `FloatingPointError`. PyTorch deterministic algorithms are enabled, `CUBLAS_WORKSPACE_CONFIG` is set to `:4096:8`, and the thread count is two. Round f uses `torch.manual_seed(20260914+`f`)`. After selecting candidate and epoch, the network is retrained from scratch on all retained development rows with seed 20260914 for that epoch count; this gives more gradient steps per epoch than the FIT groups. The selected checkpoint is frozen only after a second run of the unchanged harness reproduces every tensor bitwise. Its record sets `final_calibration_or_test_read: False`.

The selected rank is 64, the grid maximum, for 3DPW mesh, 3DPW joint, and EMDB mesh, and 32 for EMDB joint. No selected epoch reaches the 200-epoch budget. These comparisons therefore concern the specified rank grid and training budget.

In addition to the selected checkpoint, 20 training-seed replicates of the selected configuration are trained and frozen in each of the four HMR2.0 settings before evaluation. Replicate k uses seed $20260914 + 100000k$, with all other settings fixed; none is selected on evaluation results. The main tables report, per calibration split, the mean over the 20 replicates, compared with the analytic families on the same split. The original selected checkpoint is distinct from these 20 replicates, giving 21 frozen learned models evaluated per setting without selection on evaluation results. The reported standard deviation across calibration splits describes calibration-repartition variation of the replicate mean; it does not measure training-seed variation. The learned baseline is available for the four HMR2.0 settings only; it was not trained for ProHMR or HOT3D and was not evaluated under a population shift.

Its input features are predictor-specific. Applying the HMR2.0 head to HaMeR would require changing its input layer and training a new head. The analytic families also refit any required moments and scales for a new predictor, but do not train a geometry network. This cross-predictor cost is distinct from population transfer of a fixed head, which was not tested for C. HOT3D does not support inferring a transfer loss from dependence on $\widetilde{M}$ alone: GM-I uses that moment but changes by only 1.6–9.9% between the reported designs, versus 22.8–58.2% for residual-only, additive, and GM-R geometries (Table 3). C's transfer behaviour remains unmeasured.

Refitting, freezing, and evaluation scope. In the original evaluation campaign, after selection, analytic moments and scale constants are re-estimated on all development observations and frozen; the scoring batch never supplies replacement constants. Every configuration, fitted statistic, scale constant, and checkpoint is recorded in one hashed manifest, and the evaluation program refuses an incomplete manifest. The six analytic families are evaluated on every reported setting: HMR2.0 (3DPW and EMDB, mesh and joint), ProHMR (3DPW, mesh and joint), and HOT3D (mesh, R_{distal}, R_{joints}) under the participant-disjoint and matched designs; the learned baseline is evaluated on the four HMR2.0 settings. The 14-cell normalization ablation (four HMR2.0 settings and three HOT3D tasks, each for GM-I and GM-R) and the ridge sweeps are regenerated under this protocol; the three HOT3D cells use the `main` allocation.

Frozen evaluation and decision rule. In that campaign, after every asset is frozen, one evaluation program uses the 20-seed 3DPW `seq60` and EMDB `seq81` allocations and the two reported HOT3D

designs, calibrating every candidate separately on the same $n = 40$ observations at nominal 0.80, 0.90, and 0.95 ($k = 33, 37, 39$). A comparison is reported as an improvement if the mean paired relative gain over the 20 allocations is at least 5% and at least 16 of the 20 per-allocation gains share the sign of the mean; the gain of allocation-averaged extents is reported alongside and never substituted. This is a prespecified consistency criterion, not a hypothesis test; we report no p-values. The allocations randomize one fixed pool. The evaluation pool was withheld throughout configuration fitting and grid selection for every method. All configurations, fitted statistics, scale constants, and checkpoints were frozen and hashed before calibration and evaluation. The reported results come from this blind evaluation campaign and the frozen learned-model replicates; no configuration was selected using evaluation results. All 72 frozen analytic configurations record `test_rows_read: false`; the learned baseline was frozen under the same protocol. The evaluation pool was read once, after the complete manifest had been frozen and hashed. Source records retain their generating code revisions; a revision mismatch invalidates the evaluation run to prevent mixing records from different revisions. Released CSVs contain metric rows but omit these per-record code hashes, which remain in the source records. The released files therefore document but do not independently establish the single-read discipline or each record's code-revision binding. Tables and empirical claims in this manuscript summarize the reported frozen evaluation protocol without selecting configurations on evaluation results.

Recorded default rule and post hoc comparisons. The default-selection rule was specified before evaluation: GM-R is selected if and only if (A) it improves on GM-I under the criterion above in all four HMR2.0 settings and (B) its relative degradation from matched to participant-disjoint HOT3D evaluation is no larger than GM-I's on every task. The rule names no nominal level; we evaluate it at 0.90 and report that both conditions fail identically at 0.80 and 0.95.

Condition (B) fails (Table 3), which suffices to retain GM-I without a verdict on (A). Applying this rule to the frozen evaluation results is a summary decision, distinct from the development-only selection of each candidate configuration. It is not an independent prospective validation of a task-wide default; a variant used for future prediction must be fixed before independent calibration.

The prespecified analysis script compares methods with the isotropic reference, additive fusion, and C, and compares normalized with unnormalized metrics. It omits the GM-R versus GM-I comparison needed for (A). A separate post hoc script imports the frozen paired-gain routine and rule unchanged and applies them to the frozen records. At nominal 90% coverage, this analysis gives GM-R gains of 1.3% and 2.2% on EMDB and 3DPW mesh, below the 5% threshold. These are post hoc results, not verdicts from the frozen comparison set; their status is retained wherever they are reported. Condition (B) compares the matched and participant-disjoint designs reported in this paper.

Selected grid endpoints. On the four SMPL instances where additive fusion is closest to GM (3DPW and ProHMR, mesh and joint), additive fusion selected the smallest positive outer-ridge strength $\lambda = 10^{-5}$. The development objective is monotone over the evaluated λ grid and increases by 0.0004–0.008% from 10^{-5} to 10^{-4}. Computed post hoc at $\lambda = 0$ on development data with all other parameters frozen, removing the outer ridge changes additive fusion's development objective by -0.00003% to -0.00087% across the four instances, so additive fusion is underestimated by at most 0.001% at this endpoint; the four $\lambda = 10^{-5}$ values reproduced the recorded grid surface bitwise. The runner-up gaps at these endpoints ($< 10^{-5}$) are 20–400 times smaller than for interior additive selections. On HOT3D additive fusion is interior in every parameter ($w = 16$ of $\{\ldots, 8, 16, 32\}$). The zero bounds $\rho = 0$ (no shrinkage) and $\eta = 0$ (no inner ridge) are natural model choices and are not marked. Other selected analytic grid endpoints are marked with † in Table 5.

Ridge sweeps, directional widths, and record verification. Two ridge sweeps are evaluated with the main program, descriptively and without altering any selected configuration, on 3DPW joint, EMDB joint, and 3DPW mesh with all ten λ_h values: sweep A uses each variant's own selected (ρ, η), and sweep B fixes both coordinate choices to GM-R's selected (ρ, η), so that the two arms share $\widetilde{M}$ and the core B and differ only in ridge coordinates. Every record stores $\widehat{q}$, $\bar{b}_h$, $\bar{D}_{\mathrm{lin},h}$, and, for main-table units and the 14 normalization arms at level 0.90, the odd and even components of Eq. (27) from paired $\pm\delta$ decodes of the same draws, with the identity residual recorded as a self-check (maximum 9.2×10^{-9}). Directional extent is reported as the linearized width $2\widehat{q}\sqrt{u^\top J A^{-1} J^\top u}$ along prescribed directions u, the linearized maximal width $2\widehat{q}\sqrt{\lambda_{\max}(J_v A^{-1} J_v^\top)}$ per output point v (exact under

the linearized forward map, aggregated as an RMS over points with weight W), and the sampled nonlinear lower bound from the 64 draws; no nonlinear extremal search is performed, because the maximum over 64 sampled directions is only a lower bound and increased as the number of sampled directions grew. A separately instrumented evaluation path invokes the same geometry routine and checks records bitwise; the program halts on any discrepancy. This checks record consistency but is not independent validation of the shared geometry implementation. The sweep units recompute the λ-independent core at all ten grid points, performing nine redundant core evaluations per sweep instead of reusing one; this does not imply a ninefold increase in total runtime.

Table 5: **Selected configurations under the shared protocol.** All 72 analytic configurations are taken from the frozen selection files. ρ and η define $\widetilde{M} = M_\rho + \eta s_M I$, with $s_M = \mathrm{tr}(M_\rho)/d$; w is the additive mixing weight and λ_h the outer ridge. A dash denotes an unused parameter. For HOT3D, M denotes `matched`; unmarked settings use the participant-disjoint development allocation.

Setting	Family	ρ	η	w	λ_h
HMR2.0: 3DPW mesh	isotropic	—	—	—	—
	residual-only	0.5	0.01	—	—
	sensitivity-only	—	—	—	0.1
	additive	0.7	0.1	0.5	10^{-5}
	GM-I	0.1	0.01	—	0.01
	GM-R	0.3	0.01	—	0.1
HMR2.0: 3DPW joint	isotropic	—	—	—	—
	residual-only	0.5	0.01	—	—
	sensitivity-only	—	—	—	0.1
	additive	0.5	0.1	0.5	10^{-5}
	GM-I	0.3	0	—	0.1
	GM-R	0.5	0.001	—	0.5
HMR2.0: EMDB mesh	isotropic	—	—	—	—
	residual-only	0.3	0.1	—	—
	sensitivity-only	—	—	—	0.5
	additive	0.5	0.01	1	0.01
	GM-I	0	0	—	0.01
	GM-R	0.1	0	—	0.5
HMR2.0: EMDB joint	isotropic	—	—	—	—
	residual-only	0.3	0.1	—	—
	sensitivity-only	—	—	—	0.5
	additive	0.5	0.01	1	0.01
	GM-I	0	0	—	0.1
	GM-R	0.3	0.01	—	0.5
ProHMR mesh	isotropic	—	—	—	—
	residual-only	0.5	0.01	—	—
	sensitivity-only	—	—	—	0.1
	additive	0.7	0.1	0.5	10^{-5}
	GM-I	0.3	0	—	0.01
	GM-R	0.3	0.01	—	0.1
ProHMR joint	isotropic	—	—	—	—
	residual-only	0.5	0.01	—	—
	sensitivity-only	—	—	—	0.1
	additive	0.5	0.1	0.5	10^{-5}
	GM-I	0.3	0	—	0.1
	GM-R	0.3	0.01	—	0.5

Table 5: **Selected configurations (continued).** M: `matched`

Setting	Family	ρ	η	w	λ_h
HOT3D mesh	isotropic	—	—	—	—
	residual-only	0.1	0.3	—	—
	sensitivity-only	—	—	—	0.5
	additive	0.1	0.01	16	0.01
	GM-I	0	0	—	0.001
	GM-R	0	0.001	—	0.5
HOT3D R_{distal}	isotropic	—	—	—	—
	residual-only	0.1	0.3	—	—
	sensitivity-only	—	—	—	0.5
	additive	0.1	0.01	16	0.01
	GM-I	0	0	—	0.1
	GM-R	0.1	0.01	—	2
HOT3D R_{joints}	isotropic	—	—	—	—
	residual-only	0.1	0.3	—	—
	sensitivity-only	—	—	—	0.5
	additive	0.1	0.01	16	0.001
	GM-I	0	0.1	—	10^{-5}
	GM-R	0	0.001	—	0.5
HOT3D mesh, M	isotropic	—	—	—	—
	residual-only	0.3	0.3	—	—
	sensitivity-only	—	—	—	0.5
	additive	0.1	0.01	16	0.01
	GM-I	0	0	—	0.001
	GM-R	0.1	10^{-4}	—	0.5
HOT3D R_{distal}, M	isotropic	—	—	—	—
	residual-only	0.3	0.3	—	—
	sensitivity-only	—	—	—	0.5
	additive	0.1	0.01	16	0.01
	GM-I	0	0	—	0.1
	GM-R	0.1	0.01	—	2
HOT3D R_{joints}, M	isotropic	—	—	—	—
	residual-only	0.3	0.3	—	—
	sensitivity-only	—	—	—	0.5
	additive	0.1	0.01	16	0.01
	GM-I	0	0.1	—	10^{-5}
	GM-R	0.1	0	—	0.5

B.3.1 HAND PROTOCOL

The hand tasks use right-hand MANO v1.2 and fixed HaMeR predictions on HOT3D. Reference-mesh projections supply the input crop boxes, so the tasks exclude uncertainty from image-only hand detection (stated with the results in Section 5). Outputs are in metres and aligned to each configuration's own wrist, with $W = I_{3N_T}/N_T$; RMS values are reported in millimetres. The mesh task uses 778 vertices, R_{distal} uses joints 3,6,9,12,15 ($3N_T = 15$, nullity 43), and R_{joints} uses joints 1–15 ($3N_T = 45$, measured nullity 25). Rank is measured in float64 rather than inferred from $3N_T$: for the mesh over 232 development inputs (rank 58 at every input); for the joint tasks over 16 development base points, with the ratio of consecutive singular values at index $d - \mathrm{rank}$ at least 1.3×10^9 at every point. For R_{joints} the dimension bound gives a kernel of at least 13; the 15 rotation coordinates of the five distal joints lie in the kernel because those joints are leaves of the kinematic tree and move no joint, raising the bound to 15; the measured kernel is 25. All hand tasks use the shared three-role selection and grids above, and all six analytic families are run under both reported designs.

Participant-disjoint design. In the participant-disjoint design, all sequences of participants P0001, P0002 and P0003 (29 sequences) form the development set, and the calibration/evaluation pool contains only sequences of P0009, P0010, P0011, P0012, P0014 and P0015 (106 sequences). Development and deployment participants are disjoint, while calibration and evaluation may share participants. Each of 20 seeds uses $\mathrm{RandomState}(3000 + s)$ to allocate 40 calibration sequences and one eligible observation per calibration sequence; the sequence roles are recorded in the released manifests.

Matched-population allocation. The `matched` allocation keeps 29 development sequences but stratifies them by participant so that all nine participants included in this study (P0001, P0002, P0003, P0009, P0010, P0011, P0012, P0014, P0015) are represented, contributing 3/1/2/4/3/5/5/2/4 development sequences in that order. The remaining 106 sequences (13/4/6/16/10/16/17/9/15 in the same order) form the pool, from which each of 20 seeds allocates 40 calibration and 66 evaluation sequences with the same sampling rule and pool size as `main`. A split is matched in the sense that every calibration and every evaluation participant appears in the development set; this holds for all 20 seeds. Because P0002 and P0003 contribute only four and six pool sequences, the calibration sets of seeds 10 and 12 contain no P0002 sequence and that of seed 11 no P0003 sequence; all 20 evaluation sets contain all nine participants. The allocation was generated once and hashed, not redrawn. Relative to `main`, participant overlap and population composition change together: development includes nine rather than three participants (`main` develops on P0001–P0003 and deploys on the six remaining participants, which are disjoint), and deployment includes participants absent from the `main` pool. This design does not isolate the effect of participant overlap. Section 5.4 therefore reports each family's extent change between allocations as a descriptive comparison; all six analytic families are evaluated on every allocation.

Table 6: **3DPW mesh: all nominal levels.** Score-coverage reference $k/(n+1) = 33/41, 37/41, 39/41$ at nominal $80/90/95\%$. Score coverage / RMS extent in mm (mean $\pm$ calibration-split SD over 20 splits of one pool). See Table 1.

Method	80%	90%	95%
Isotropic	0.827 / 425.33 $\pm$ 13.79	0.921 / 449.04 $\pm$ 17.31	0.960 / 465.58 $\pm$ 15.26
Residual-only	0.778 / 175.99 $\pm$ 9.76	0.885 / 201.43 $\pm$ 14.59	0.941 / 222.39 $\pm$ 14.27
Sensitivity-only	0.815 / 129.56 $\pm$ 7.44	0.899 / 149.57 $\pm$ 14.43	0.954 / 190.74 $\pm$ 36.51
Additive	0.781 / 88.57 $\pm$ 6.60	0.894 / 105.74 $\pm$ 12.25	0.955 / 125.28 $\pm$ 18.61
GM-I	0.787 / 80.36 $\pm$ 5.61	0.891 / 96.75 $\pm$ 12.90	0.952 / 120.70 $\pm$ 24.37
GM-R	0.788 / **78.85** $\pm$ 5.53	0.890 / **94.44** $\pm$ 11.66	0.950 / 117.05 $\pm$ 22.85
Learned	0.776 / 84.41 $\pm$ 8.24	0.897 / 99.38 $\pm$ 9.67	0.951 / **114.81** $\pm$ 16.05

Table 7: **3DPW joint: all nominal levels.** Score-coverage reference $k/(n+1) = 33/41, 37/41, 39/41$ at nominal $80/90/95\%$. Score coverage / RMS extent in mm (mean $\pm$ calibration-split SD over 20 splits of one pool). See Table 1.

Method	80%	90%	95%
Isotropic	0.827 / 374.59 $\pm$ 11.94	0.921 / 397.11 $\pm$ 15.17	0.960 / 413.04 $\pm$ 13.32
Residual-only	0.778 / 148.19 $\pm$ 8.18	0.885 / 170.09 $\pm$ 12.53	0.941 / 188.28 $\pm$ 12.49
Sensitivity-only	0.818 / 103.63 $\pm$ 7.20	0.911 / 119.68 $\pm$ 10.81	0.946 / 135.58 $\pm$ 21.19
Additive	0.797 / 72.00 $\pm$ 4.65	0.895 / 83.13 $\pm$ 8.00	0.943 / 93.67 $\pm$ 11.54
GM-I	0.810 / 89.01 $\pm$ 5.03	0.911 / 102.58 $\pm$ 8.81	0.951 / 116.52 $\pm$ 16.40
GM-R	0.792 / **66.44** $\pm$ 4.72	0.893 / **78.09** $\pm$ 7.75	0.946 / 90.53 $\pm$ 12.26
Learned	0.774 / 66.76 $\pm$ 6.04	0.893 / 78.38 $\pm$ 7.57	0.943 / **90.11** $\pm$ 14.13

B.4 Full Result Tables

The following tables report score coverage and RMS decoded extent at nominal 0.80, 0.90, and 0.95 for every setting and family; the main text reports the 0.90 level. The finite-pool reference levels for expected sequence-averaged score coverage are $33/41$, $37/41$, and $39/41$ under the conditions of Appendix A.1; individual splits need not reach those levels. HOT3D pilot-excluded tables are a sensitivity check with substantial pool overlap. Rows are the six analytic families and, where trained, the learned-metric baseline (mean over 20 training replicates).

Table 8: **EMDB mesh: all nominal levels.** Score-coverage reference $k/(n+1) = 33/41, 37/41, 39/41$ at nominal $80/90/95\%$. Score coverage / RMS extent in mm (mean $\pm$ calibration-split SD over 20 splits of one pool). † marks a selection at a grid endpoint other than the natural zero bounds; see Table 1.

Method	80%	90%	95%
Isotropic	0.797 / 391.91 $\pm$ 11.23	0.891 / 412.31 $\pm$ 11.42	0.951 / 437.88 $\pm$ 18.91
Residual-only	0.783 / 197.27 $\pm$ 6.92	0.892 / 215.30 $\pm$ 11.54	0.952 / 235.73 $\pm$ 13.83
Sensitivity-only	0.790 / 148.76 $\pm$ 5.01	0.888 / 163.10 $\pm$ 9.50	0.951 / 183.88 $\pm$ 18.21
Additive	0.784 / 104.66 $\pm$ 8.06	0.897 / 123.44 $\pm$ 13.94	0.951 / 143.85 $\pm$ 20.75
GM-I	0.788 / 95.52 $\pm$ 7.29	0.886 / 109.03 $\pm$ 8.84	0.952 / 133.19 $\pm$ 20.50
GM-R	0.778 / 93.96 $\pm$ 6.15	0.887 / 107.57 $\pm$ 8.16	0.949 / 127.69 $\pm$ 17.76
Learned	0.789 / **89.38** $\pm$ 2.66	0.890 / **99.02** $\pm$ 5.67	0.939 / **111.74** $\pm$ 13.79

Table 9: **EMDB joint: all nominal levels.** Score-coverage reference $k/(n+1) = 33/41, 37/41, 39/41$ at nominal $80/90/95\%$. Score coverage / RMS extent in mm (mean $\pm$ calibration-split SD over 20 splits of one pool). See Table 1.

Method	80%	90%	95%
Isotropic	0.797 / 337.26 $\pm$ 9.96	0.891 / 355.86 $\pm$ 10.45	0.951 / 379.57 $\pm$ 17.81
Residual-only	0.783 / 166.93 $\pm$ 5.85	0.892 / 182.48 $\pm$ 9.97	0.952 / 200.20 $\pm$ 12.02
Sensitivity-only	0.794 / 125.77 $\pm$ 4.73	0.898 / 137.91 $\pm$ 6.34	0.950 / 153.68 $\pm$ 17.32
Additive	0.801 / 90.84 $\pm$ 6.37	0.896 / 103.54 $\pm$ 10.29	0.948 / 118.76 $\pm$ 17.61
GM-I	0.801 / 98.33 $\pm$ 4.52	0.893 / 109.64 $\pm$ 6.48	0.950 / 128.64 $\pm$ 20.16
GM-R	0.801 / 83.98 $\pm$ 4.33	0.894 / 94.52 $\pm$ 6.19	0.952 / 109.76 $\pm$ 14.44
Learned	0.788 / **72.89** $\pm$ 2.56	0.890 / **81.28** $\pm$ 4.69	0.942 / **93.12** $\pm$ 11.83

Table 10: **A second predictor: ProHMR on 3DPW.** Same sequence partitions and calibration observations, nominal 90% (finite-pool score-coverage reference $37/41 \simeq 0.902$). No learned baseline was trained for ProHMR. See Table 1.

	ProHMR mesh		ProHMR joint	
Geometry	Score cov.	$E_{\mathrm{RMS}}\downarrow$	Score cov.	$E_{\mathrm{RMS}}\downarrow$
Isotropic	0.921	523.51 $\pm$ 23.12	0.921	468.56 $\pm$ 21.91
Residual-only	0.894	226.40 $\pm$ 16.75	0.894	190.54 $\pm$ 14.44
Sensitivity-only	0.911	197.27 $\pm$ 25.98	0.904	147.14 $\pm$ 12.34
Additive	0.898	121.14 $\pm$ 13.92	0.896	93.24 $\pm$ 8.09
GM-I	0.907	115.52 $\pm$ 13.42	0.909	125.14 $\pm$ 9.67
GM-R	0.907	**112.50** $\pm$ 12.74	0.901	**89.13** $\pm$ 8.54

Table 11: **ProHMR 3DPW mesh: all nominal levels.** Score-coverage reference $k/(n+1) = 33/41, 37/41, 39/41$ at nominal $80/90/95\%$. Score coverage / RMS extent in mm (mean $\pm$ calibration-split SD over 20 splits of one pool). See Table 1.

Method	80%	90%	95%
Isotropic	0.823 / 466.18 $\pm$ 14.35	0.921 / 523.51 $\pm$ 23.12	0.974 / 555.67 $\pm$ 15.53
Residual-only	0.784 / 194.98 $\pm$ 12.32	0.894 / 226.40 $\pm$ 16.75	0.946 / 248.78 $\pm$ 17.13
Sensitivity-only	0.809 / 162.31 $\pm$ 10.20	0.911 / 197.27 $\pm$ 25.98	0.954 / 228.82 $\pm$ 39.61
Additive	0.798 / 101.16 $\pm$ 7.74	0.898 / 121.14 $\pm$ 13.92	0.957 / 145.65 $\pm$ 18.84
GM-I	0.800 / 95.32 $\pm$ 7.79	0.907 / 115.52 $\pm$ 13.42	0.948 / 136.46 $\pm$ 22.51
GM-R	0.801 / **92.91** $\pm$ 8.35	0.907 / **112.50** $\pm$ 12.74	0.949 / **133.01** $\pm$ 21.34

Table 12: **ProHMR 3DPW joint: all nominal levels.** Score-coverage reference $k/(n+1) = 33/41, 37/41, 39/41$ at nominal $80/90/95\%$. Score coverage / RMS extent in mm (mean $\pm$ calibration-split SD over 20 splits of one pool). See Table 1.

Method	80%	90%	95%
Isotropic	0.823 / 411.75 $\pm$ 12.52	0.921 / 468.56 $\pm$ 21.91	0.974 / 502.04 $\pm$ 15.25
Residual-only	0.784 / 163.50 $\pm$ 10.41	0.894 / 190.54 $\pm$ 14.44	0.946 / 209.98 $\pm$ 15.01
Sensitivity-only	0.823 / 130.32 $\pm$ 8.41	0.904 / 147.14 $\pm$ 12.34	0.958 / 172.34 $\pm$ 25.81
Additive	0.780 / 78.62 $\pm$ 6.84	0.896 / 93.24 $\pm$ 8.09	0.955 / 108.65 $\pm$ 10.47
GM-I	0.819 / 109.93 $\pm$ 5.58	0.909 / 125.14 $\pm$ 9.67	0.967 / 146.48 $\pm$ 15.20
GM-R	0.790 / **74.70** $\pm$ 6.34	0.901 / **89.13** $\pm$ 8.54	0.953 / **103.46** $\pm$ 11.72

Table 13: **MANO: all nominal levels.** Score-coverage reference $k/(n+1) = 33/41, 37/41, 39/41$ at nominal $80/90/95\%$. Score coverage / RMS decoded extent in mm (mean $\pm$ calibration-split SD over 20 splits of one pool). The SD excludes new-participant and development-selection variation. † marks a selection at a grid endpoint other than the natural zero bounds; see Table 1.

Method	80%	90%	95%
Isotropic	0.792 / 99.29 $\pm$ 7.38	0.890 / 110.19 $\pm$ 3.41	0.946 / 111.75 $\pm$ 0.70
Residual-only	0.785 / 62.53 $\pm$ 2.93	0.883 / 66.96 $\pm$ 2.07	0.940 / 70.15 $\pm$ 2.71
Sensitivity-only	0.771 / 43.93 $\pm$ 1.04	0.891 / 46.97 $\pm$ 1.07	0.953 / 50.40 $\pm$ 2.45
Additive	0.779 / 34.33 $\pm$ 2.92	0.898 / 40.59 $\pm$ 1.65	0.950 / 43.32 $\pm$ 2.50
GM-I	0.777 / **23.83** $\pm$ 0.80	0.893 / **26.68** $\pm$ 1.78	0.953 / **29.74** $\pm$ 2.05
GM-R	0.776 / 35.62 $\pm$ 1.43	0.892 / 38.45 $\pm$ 1.07	0.952 / 40.91 $\pm$ 1.76

Table 14: **HOT3D R_{distal}: all nominal levels.** Score-coverage reference $k/(n+1) = 33/41, 37/41, 39/41$ at nominal $80/90/95\%$. Score coverage / RMS extent in mm (mean $\pm$ calibration-split SD over 20 splits of one pool). † marks a selection at a grid endpoint other than the natural zero bounds; see Table 1.

Method	80%	90%	95%
Isotropic	0.792 / 112.12 $\pm$ 8.21	0.890 / 124.25 $\pm$ 3.78	0.946 / 125.96 $\pm$ 0.76
Residual-only	0.785 / 70.94 $\pm$ 3.32	0.883 / 75.97 $\pm$ 2.35	0.940 / 79.58 $\pm$ 3.07
Sensitivity-only	0.773 / 49.77 $\pm$ 1.28	0.885 / 53.47 $\pm$ 1.48	0.952 / 58.33 $\pm$ 3.37
Additive	0.779 / 37.31 $\pm$ 3.23	0.900 / 44.28 $\pm$ 1.88	0.950 / 47.19 $\pm$ 2.76
GM-I	0.782 / 40.29 $\pm$ 2.07	0.882 / 44.21 $\pm$ 2.30	0.951 / 49.50 $\pm$ 3.46
GM-R	0.781 / **29.16** $\pm$ 1.50	0.893 / **32.53** $\pm$ 1.94	0.949 / **35.74** $\pm$ 2.61

Table 15: **HOT3D R_{joints}: all nominal levels.** Score-coverage reference $k/(n+1) = 33/41, 37/41, 39/41$ at nominal $80/90/95\%$. Score coverage / RMS extent in mm (mean $\pm$ calibration-split SD over 20 splits of one pool). † marks a selection at a grid endpoint other than the natural zero bounds; see Table 1.

Method	80%	90%	95%
Isotropic	0.792 / 98.04 $\pm$ 7.98	0.890 / 109.83 $\pm$ 3.69	0.946 / 111.53 $\pm$ 0.80
Residual-only	0.785 / 61.03 $\pm$ 2.93	0.883 / 65.43 $\pm$ 2.07	0.940 / 68.61 $\pm$ 2.71
Sensitivity-only	0.768 / 37.64 $\pm$ 0.82	0.894 / 40.56 $\pm$ 1.08	0.954 / 43.71 $\pm$ 2.24
Additive	0.781 / 31.11 $\pm$ 2.61	0.903 / 36.58 $\pm$ 1.44	0.954 / 39.01 $\pm$ 2.15
GM-I	0.771 / **23.25** $\pm$ 0.60†	0.900 / **25.54** $\pm$ 1.26†	0.954 / **27.83** $\pm$ 1.68†
GM-R	0.774 / 28.86 $\pm$ 0.97	0.891 / 31.12 $\pm$ 0.86	0.953 / 33.83 $\pm$ 1.82

Table 16: **HOT3D mesh, pilot-excluded sensitivity check: all nominal levels.** Score coverage / RMS extent (mm; mean ± calibration-split SD over 20 splits of one pool). The score-coverage reference is $k/(n+1) = 33/41, 37/41, 39/41$ at nominal $80/90/95\%$. The code name `pool_minus_pilot` is defined in Appendix B.3.1. Bold marks the smallest mean extent at each nominal level.

Method	80%	90%	95%
Isotropic	0.782 / 97.99 ± 6.78	0.902 / 111.08 ± 0.39	0.949 / 111.71 ± 0.46
Residual-only	0.798 / 63.15 ± 2.62	0.910 / 67.97 ± 1.59	0.954 / 71.35 ± 3.18
Sensitivity-only	0.791 / 44.52 ± 1.40	0.891 / 47.25 ± 1.58	0.954 / 50.50 ± 2.20
Additive	0.796 / 35.24 ± 3.30	0.902 / 40.67 ± 1.41	0.955 / 43.71 ± 2.32
GM-I	0.805 / **24.43** ± 1.00	0.892 / **26.97** ± 2.29	0.950 / **30.27** ± 3.39
GM-R	0.789 / 36.03 ± 1.59	0.900 / 38.85 ± 1.54	0.955 / 41.54 ± 2.36

Table 17: **HOT3D R_{distal}, pilot-excluded sensitivity check: all nominal levels.** Score-coverage reference $k/(n+1) = 33/41, 37/41, 39/41$ at nominal $80/90/95\%$. Score coverage / RMS extent in mm (mean ± calibration-split SD over 20 splits of one pool). † marks a selection at a grid endpoint other than the natural zero bounds; see Table 1.

Method	80%	90%	95%
Isotropic	0.782 / 110.69 ± 7.56	0.902 / 125.25 ± 0.43	0.949 / 125.93 ± 0.50
Residual-only	0.798 / 71.65 ± 2.99	0.910 / 77.11 ± 1.80	0.954 / 80.95 ± 3.61
Sensitivity-only	0.791 / 50.49 ± 1.80	0.893 / 54.11 ± 2.04	0.955 / 58.52 ± 2.85
Additive	0.795 / 38.27 ± 3.57	0.901 / 44.23 ± 1.51	0.958 / 47.72 ± 2.42
GM-I	0.790 / 40.72 ± 2.36	0.902 / 45.26 ± 2.63	0.952 / 50.45 ± 4.76
GM-R	0.796 / **29.60** ± 1.54	0.904 / **32.93** ± 1.88	0.959 / **37.01** ± 3.52

Table 18: **HOT3D R_{joints}, pilot-excluded sensitivity check: all nominal levels.** Score-coverage reference $k/(n+1) = 33/41, 37/41, 39/41$ at nominal $80/90/95\%$. Score coverage / RMS extent in mm (mean ± calibration-split SD over 20 splits of one pool). See Table 1.

Method	80%	90%	95%
Isotropic	0.782 / 96.60 ± 7.34	0.902 / 110.73 ± 0.44	0.949 / 111.42 ± 0.50
Residual-only	0.798 / 61.61 ± 2.61	0.910 / 66.39 ± 1.59	0.954 / 69.77 ± 3.20
Sensitivity-only	0.793 / 38.28 ± 1.14	0.888 / 40.61 ± 1.53	0.953 / 43.62 ± 1.96
Additive	0.793 / 31.72 ± 2.80	0.907 / 36.62 ± 1.07	0.958 / 39.13 ± 1.92
GM-I	0.786 / **23.51** ± 0.77	0.899 / **25.68** ± 1.58	0.944 / **27.66** ± 2.21
GM-R	0.791 / 29.26 ± 1.10	0.899 / 31.48 ± 1.35	0.952 / 34.06 ± 2.37

Table 19: **MANO matched-population allocation.** Development is stratified by participant over the matched universe, and the pool is drawn from the same participants. See Table 1.

Method	80%	90%	95%
Isotropic	0.787 / 104.05 ± 2.87	0.889 / 110.16 ± 1.68	0.948 / 111.58 ± 0.81
Residual-only	0.805 / 43.51 ± 1.96	0.899 / 47.40 ± 2.77	0.947 / 50.92 ± 3.49
Sensitivity-only	0.790 / 44.82 ± 1.21	0.898 / 47.61 ± 1.71	0.945 / 50.21 ± 2.51
Additive	0.817 / 24.57 ± 1.22	0.914 / 27.54 ± 1.74	0.957 / 31.27 ± 5.31
GM-I	0.817 / 21.39 ± 0.92	0.909 / **24.29** ± 1.64	0.949 / 27.64 ± 5.12
GM-R	0.814 / **21.33** ± 0.74	0.911 / 24.31 ± 1.57	0.949 / **27.53** ± 4.84

Table 20: **HOT3D R_{distal}, matched: all nominal levels.** Score-coverage reference $k/(n+1) = 33/41, 37/41, 39/41$ at nominal $80/90/95\%$. Score coverage / RMS extent in mm (mean $\pm$ calibration-split SD over 20 splits of one pool). See Table 1.

Method	80%	90%	95%
Isotropic	0.787 / 117.46 $\pm$ 3.18	0.889 / 124.22 $\pm$ 1.86	0.948 / 125.78 $\pm$ 0.90
Residual-only	0.805 / 49.33 $\pm$ 2.22	0.899 / 53.74 $\pm$ 3.14	0.947 / 57.73 $\pm$ 3.95
Sensitivity-only	0.790 / 50.91 $\pm$ 1.72	0.902 / 54.83 $\pm$ 2.30	0.944 / 58.05 $\pm$ 3.45
Additive	0.820 / 26.92 $\pm$ 1.37	0.914 / 30.20 $\pm$ 2.00	0.956 / 34.21 $\pm$ 5.98
GM-I	0.787 / 39.12 $\pm$ 1.72	0.909 / 43.49 $\pm$ 2.06	0.949 / 47.63 $\pm$ 6.08
GM-R	0.817 / **23.24** $\pm$ 0.90	0.914 / **26.50** $\pm$ 1.79	0.950 / **29.81** $\pm$ 5.28

Table 21: **HOT3D R_{joints}, matched: all nominal levels.** Score-coverage reference $k/(n+1) = 33/41, 37/41, 39/41$ at nominal $80/90/95\%$. Score coverage / RMS extent in mm (mean $\pm$ calibration-split SD over 20 splits of one pool). See Table 1.

Method	80%	90%	95%
Isotropic	0.787 / 103.05 $\pm$ 3.13	0.889 / 109.72 $\pm$ 1.82	0.948 / 111.28 $\pm$ 0.89
Residual-only	0.805 / 42.37 $\pm$ 1.91	0.899 / 46.19 $\pm$ 2.71	0.947 / 49.65 $\pm$ 3.44
Sensitivity-only	0.787 / 38.45 $\pm$ 1.09	0.900 / 41.13 $\pm$ 1.51	0.946 / 43.55 $\pm$ 2.39
Additive	0.820 / 22.59 $\pm$ 0.97	0.912 / 25.20 $\pm$ 1.55	0.956 / 28.61 $\pm$ 5.01
GM-I	0.809 / 22.70 $\pm$ 0.76	0.910 / 24.87 $\pm$ 1.22	0.948 / 26.82 $\pm$ 2.08
GM-R	0.816 / **19.29** $\pm$ 0.83	0.904 / **21.81** $\pm$ 1.59	0.950 / **24.72** $\pm$ 3.53

Setting	Iso.	Residual	Sens.	Additive	GM-I	GM-R
HMR2.0 / 3DPW mesh	9.7/69.5	11.6/53.9	6.8/73.8	8.4/63.5	7.3/65.6	7.1/65.6
HMR2.0 / 3DPW joints	9.7/69.5	11.6/53.9	7.9/72.8	9.3/62.4	7.9/71.5	8.2/65.2
HMR2.0 / EMDB mesh	7.0/70.1	5.9/68.2	5.5/69.7	8.3/62.4	6.9/69.2	7.1/69.9
HMR2.0 / EMDB joints	7.0/70.1	5.9/68.2	6.3/67.5	8.4/64.2	6.0/68.9	7.0/69.4
ProHMR / 3DPW mesh	8.1/71.0	11.8/54.9	7.5/70.1	8.1/68.1	7.9/66.8	8.2/65.4
ProHMR / 3DPW joints	8.1/71.0	11.8/54.9	8.1/67.3	10.2/63.5	7.7/68.3	9.3/59.4
HOT3D disjoint / mesh	7.0/69.5	8.2/64.2	6.0/66.2	7.6/66.2	6.1/68.5	7.3/66.5
HOT3D disjoint / distal	7.0/69.5	8.2/64.2	5.9/67.4	7.7/66.0	7.4/66.3	7.7/67.5
HOT3D disjoint / joints	7.0/69.5	8.2/64.2	5.9/67.2	7.9/66.5	5.9/68.9	7.0/66.4
HOT3D matched / mesh	7.4/67.4	6.9/67.7	7.1/66.5	5.7/72.1	4.3/75.1	3.7/75.4
HOT3D matched / distal	7.4/67.4	6.9/67.7	7.4/65.9	5.9/69.7	7.2/69.1	4.5/72.2
HOT3D matched / joints	7.4/67.4	6.9/67.7	7.0/69.8	5.0/75.9	5.2/72.5	4.4/74.8

Table 22: **Across-split score-coverage variation at nominal** 0.80**.** Each cell is SD in percentage points / empirical fifth percentile in percent, computed from 20 released sequence-weighted coverage values for a fixed selected configuration and evaluation pool. SD divides by 20; the fifth percentile linearly interpolates the two lowest sorted values. These descriptive fixed-pool split statistics are not uncertainty intervals for new participants or new development selections.

Setting	Iso.	Residual	Sens.	Additive	GM-I	GM-R
HMR2.0 / 3DPW mesh	6.8/78.0	10.0/69.0	5.6/84.8	6.4/75.9	6.7/80.5	6.3/80.8
HMR2.0 / 3DPW joints	6.8/78.0	10.0/69.0	5.1/82.1	6.3/78.4	5.2/85.1	6.6/75.7
HMR2.0 / EMDB mesh	4.8/80.2	5.3/81.3	5.5/78.3	5.0/82.0	4.1/81.9	4.3/82.6
HMR2.0 / EMDB joints	4.8/80.2	5.3/81.3	4.9/83.8	6.4/80.7	4.6/85.0	5.1/81.8
ProHMR / 3DPW mesh	5.8/78.6	9.1/69.4	6.8/80.0	5.3/79.6	4.7/84.5	4.5/84.2
ProHMR / 3DPW joints	5.8/78.6	9.1/69.4	6.4/81.5	6.8/78.7	5.8/81.4	5.9/79.1
HOT3D disjoint / mesh	5.2/82.3	5.7/80.5	4.3/81.3	4.8/83.0	5.4/80.2	4.8/83.0
HOT3D disjoint / distal	5.2/82.3	5.7/80.5	4.6/79.0	4.9/81.9	5.4/79.2	5.4/80.7
HOT3D disjoint / joints	5.2/82.3	5.7/80.5	4.3/81.5	5.1/81.0	5.1/81.8	4.4/82.3
HOT3D matched / mesh	4.3/80.4	5.4/79.2	4.7/84.2	3.8/85.3	3.6/84.3	3.5/84.8
HOT3D matched / distal	4.3/80.4	5.4/79.2	4.6/83.8	4.1/85.7	3.9/82.9	3.6/84.4
HOT3D matched / joints	4.3/80.4	5.4/79.2	4.8/81.0	4.4/82.8	4.1/81.9	4.5/80.9

Table 23: **Across-split score-coverage variation at nominal** 0.90**.** Each cell is SD in percentage points / empirical fifth percentile in percent, computed from 20 released sequence-weighted coverage values for a fixed selected configuration and evaluation pool. SD divides by 20; the fifth percentile linearly interpolates the two lowest sorted values. These descriptive fixed-pool split statistics are not uncertainty intervals for new participants or new development selections.

Setting	Iso.	Residual	Sens.	Additive	GM-I	GM-R
HMR2.0 / 3DPW mesh	4.1/87.6	6.4/82.8	6.0/92.9	3.5/91.8	4.5/92.4	4.8/91.9
HMR2.0 / 3DPW joints	4.1/87.6	6.4/82.8	4.4/90.6	4.7/86.3	4.3/87.8	4.3/86.7
HMR2.0 / EMDB mesh	3.3/90.0	3.1/90.6	3.8/89.2	2.8/91.2	3.0/89.0	2.9/89.9
HMR2.0 / EMDB joints	3.3/90.0	3.1/90.6	3.3/89.6	4.5/87.5	3.2/89.7	3.5/89.0
ProHMR / 3DPW mesh	2.8/92.1	6.0/84.0	3.2/89.5	2.6/90.4	4.2/87.8	4.1/88.3
ProHMR / 3DPW joints	2.8/92.1	6.0/84.0	2.6/92.9	2.6/92.0	1.8/93.5	3.2/88.3
HOT3D disjoint / mesh	3.8/87.6	4.3/87.7	2.9/91.5	3.3/89.1	2.6/90.3	3.3/89.5
HOT3D disjoint / distal	3.8/87.6	4.3/87.7	2.9/91.1	3.3/89.8	3.4/89.2	3.8/89.1
HOT3D disjoint / joints	3.8/87.6	4.3/87.7	2.9/90.3	3.4/88.8	3.0/88.6	3.5/89.0
HOT3D matched / mesh	3.0/90.4	3.0/88.8	4.1/87.6	2.6/90.4	2.8/90.9	2.7/90.8
HOT3D matched / distal	3.0/90.4	3.0/88.8	4.4/88.1	2.7/90.2	3.1/90.2	2.8/90.4
HOT3D matched / joints	3.0/90.4	3.0/88.8	4.2/89.3	3.1/88.9	3.0/90.6	2.8/90.4

Table 24: **Across-split score-coverage variation at nominal** 0.95**.** Each cell is SD in percentage points / empirical fifth percentile in percent, computed from 20 released sequence-weighted coverage values for a fixed selected configuration and evaluation pool. SD divides by 20; the fifth percentile linearly interpolates the two lowest sorted values. These descriptive fixed-pool split statistics are not uncertainty intervals for new participants or new development selections.

B.5 Additional Visualizations

Decoded body and hand sets. Figures 3 and 4 show samples from calibrated regions. Finite samples illustrate geometry; the coverage guarantee concerns the full decoded set.

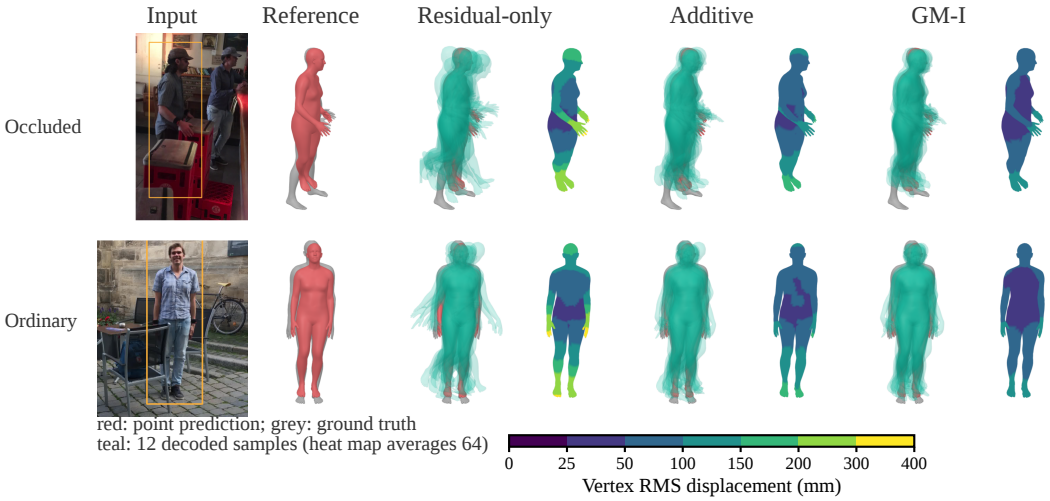

Figure 3: **Decoded prediction sets on 3DPW (HMR2.0/SMPL).** The two seed-0 evaluation inputs have point-prediction errors nearest the median within their respective truncation-proxy groups (≥ 0.32 and ≤ 0.08); selection does not use method gains. Grey is the neutral-model reference, red the point prediction, and teal 12 decoded samples at nominal 0.90. Heatmaps show per-vertex RMS displacement from the prediction over 64 draws, with one discrete colour scale in millimetres. View and scale are shared within each row. The samples are neither the complete prediction set nor a simultaneous vertex band. SMPL realizability does not ensure collision-free or anatomical poses. Both references satisfy $S_h/\widehat{q}_h < 1$ for the three methods; this is checked from their scores, not inferred from image overlap.

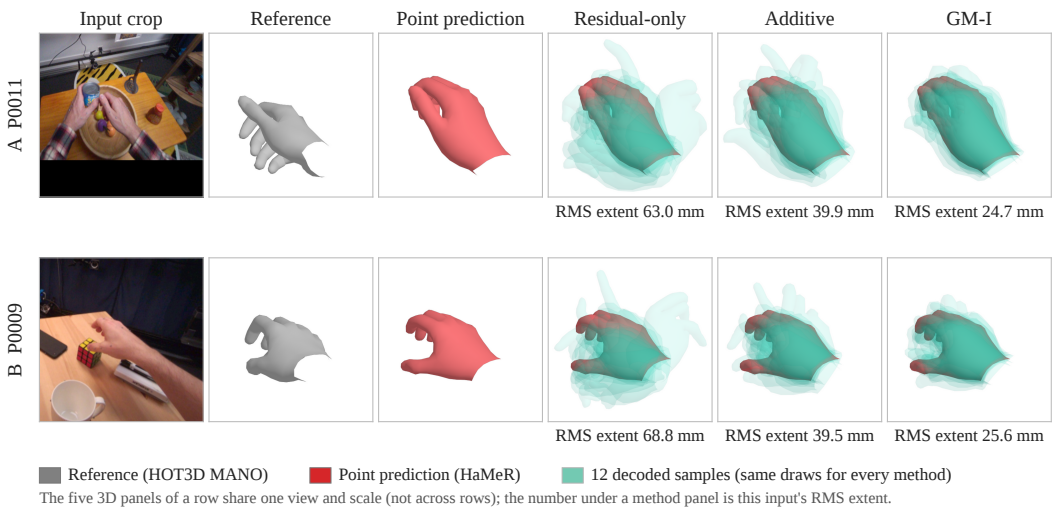

Figure 4: **Decoded hand sets on HOT3D (HaMeR/MANO).** Examples under `main` at nominal 0.90 come from the two held-out participants with the most seed-0 evaluation rows. Each example is nearest its participant's median point-prediction mesh error; selection does not use region extent or appearance. Input crops are annotation-conditioned. Grey is the reference, red the prediction, and teal 12 decoded draws, using the same unit-ball draws for every method. Each row shares one view and scale; scales differ between rows. Numbers beneath panels are this input's RMS extent from 1,024 shared draws, not a dataset mean. The rendered samples do not constitute the full prediction set, and visual occlusion does not establish reference coverage.

Paired extent and geometric diagnostics. Figure 5 displays calibration-split-level differences under the frozen HOT3D comparison. Figures 6 and 7 provide geometric illustrations of the distinction between parameter displacement, local output sensitivity, and exact decoded displacement.

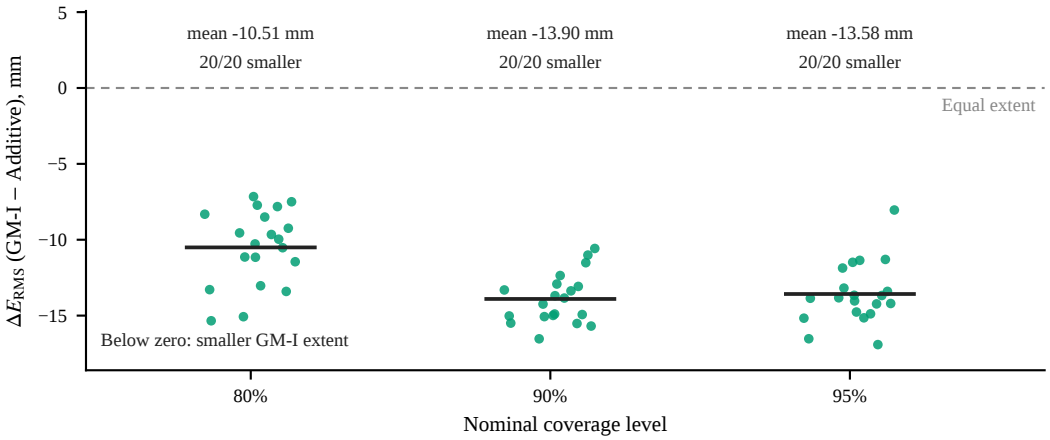

Figure 5: **Paired HOT3D mesh extent differences.** Each point is $E_{\mathrm{RMS}}^{\mathrm{GM\text{-}I}} - E_{\mathrm{RMS}}^{\mathrm{additive}}$ for one of 20 split seeds under `main`, using the analytic block. Negative values indicate smaller GM-I extent. Horizontal bars are means: -10.51, -13.90, and -13.58 mm at nominal coverage 0.80, 0.90, and 0.95, respectively; all 20 differences are negative at each level. These absolute differences complement the paired relative gains used by the prespecified comparison rule.

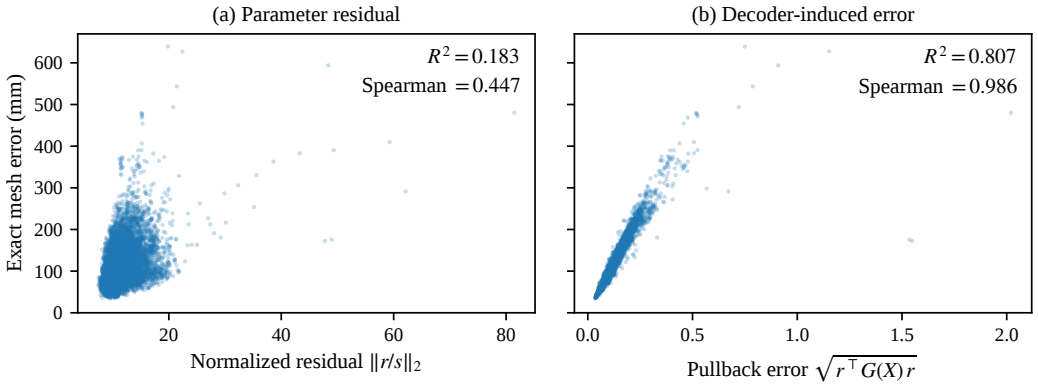

Figure 6: **Parameter residuals and decoder-induced error on 3DPW.** The diagnostic uses 10,000 frames from 50 sequences: the union of TEST indices across all 20 seeds, capped at 200 frames per sequence using `RandomState(0)`. For the 82-dimensional tangent residual r, s is the development residual standard deviation in each coordinate, floored at 10^{-8}. The horizontal axes show (a) dimensionless $\|r/s\|_2$ and (b) $\sqrt{r^\top G(X) r}$ in metres, where $G = J^\top J/V$ is the root-aligned mesh pullback. Multiply (b) by 10^3 to compare with the vertical axis in millimetres. The displayed R^2 values are 0.183 and 0.807, and Spearman correlations are 0.447 and 0.986, respectively. This descriptive analysis uses root-aligned decoding and does not enter any prespecified comparison. The stronger association for (b) does not establish optimal calibrated region efficiency or coverage.

Construction of the diagnostic direction. The input in Figure 7 is seed-0 TEST row 40821 of 3DPW, reused from the ordinary example in Figure 3. In the 82-dimensional tangent coordinates, we compute the root-aligned joint and mesh Jacobians by float64 forward AD, giving $G_T = J_T^\top J_T/14$ and $G_M = J_M^\top J_M/V$. We restrict perturbations to local rotations (coordinates 3:72 in zero-based indexing), fixing global rotation and shape. Within this subspace, eigenvectors of the restricted joint pullback with eigenvalues below $10^{-8}\lambda_{\max}(G_T)$ define a numerical near-null subspace. We choose its unit vector u maximizing $u^\top G_M u$, then scale u to target linearized mesh RMS displacements of approximately 25 and 100 mm. Reported perturbation norms are rounded to two significant figures. The norms use the raw, unnormalized implementation coordinates and have no direct physical-unit interpretation. All displayed displacements are evaluated at these perturbations; none is an expectation over directions.

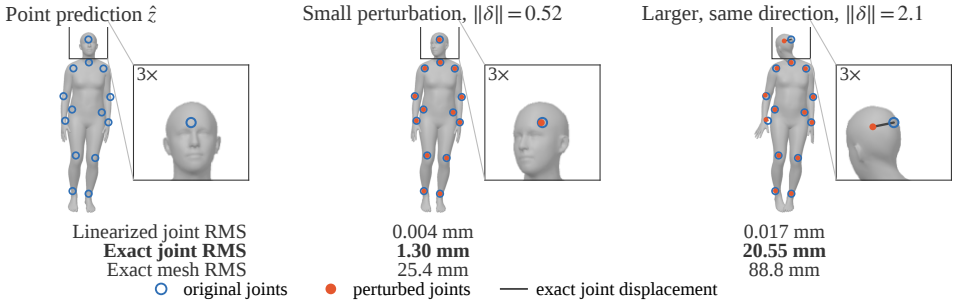

Figure 7: **Finite displacement along a selected joint near-null direction.** For one input, the direction is deliberately chosen to maximize local mesh sensitivity within a numerical joint near-null subspace; it is not a typical direction. Near-null membership uses relative eigenvalue tolerance 10^{-8} and is not evidence of an exact null vector. At raw coordinate norms $\|\delta\| = 0.52$ and 2.1, linearized joint RMS is 0.004 and 0.017 mm, whereas exact root-aligned joint RMS is 1.30 and 20.55 mm and mesh RMS is 25.4 and 88.8 mm. Insets enlarge the head threefold. Neither perturbation was checked against a calibrated set. The example illustrates differences between local and finite responses, and between joint and mesh sensitivity; it makes no coverage claim.

Table 25: **Geometry-construction cost at matched batch sizes.** Median of five runs using the same input rows and batch size B for GM and additive fusion, with checks before and after measurement. Component times are milliseconds per input; the premium is the measured relative runtime increase of GM over additive fusion. Prediction and sampled mesh generation are excluded.

Setting	Component	$B = 1$	$B = 8$	$B = 32$	$B = 128$
3DPW mesh	Jacobian	17.37	17.41	17.25	17.18
	GM square root	7.98	3.00	2.49	2.35
	GM premium	+42.6%	+17.0%	+14.4%	+13.7%
3DPW joints	Jacobian	19.52	18.80	19.13	19.12
	GM square root	10.88	7.55	7.01	6.90
	GM premium	+51.2%	+39.7%	+36.6%	+36.0%
HOT3D mesh	Jacobian	15.44	2.03	0.59	0.26
	GM square root	5.41	1.82	1.43	1.34
	GM premium	+31.4%	+82.0%	+224.4%	+500.7%

B.6 Computational Cost Details

Development selection dominates the reported analytic workloads. On one GPU, additive fusion (4,410 candidates) takes 26.5 GPU-h on 3DPW mesh and 6–6.5 GPU-h per HOT3D task; GM-I and GM-R together (980 candidates) take 11.6 and 2.5 GPU-h, respectively. C uses 80 configurations and 11 checkpoints, taking 35–50 GPU-h across its four settings, in addition to 20 training replicates per setting.

Table 25 reports medians of five runs using the same input rows and batch size for GM and additive fusion, with checks before and after measurement. These geometry-construction timings exclude prediction and sampled mesh generation. For SMPL, the Jacobian remains the larger component at every tested batch; the GM premium falls from 42.6% to 13.7% on 3DPW mesh and from 51.2% to 36.0% on 3DPW joints as B increases from 1 to 128.

For HOT3D mesh, the Jacobian cost falls from 15.44 to 0.26 ms per input (about $60\times$), whereas the square-root cost falls from 5.41 to 1.34 ms (about $4\times$). The Jacobian is the larger component at $B = 1$ and 8; the square root is larger at $B = 32$ and 128. Consequently, the relative GM premium grows from 31.4% to 500.7% even as both components become cheaper per input. The premium is measured from the compared pipelines, not reconstructed from the two displayed component timings.

Decoding 64 perturbations takes approximately 10.4, 10.7, and 7.0 ms per frame on 3DPW mesh, 3DPW joints, and HOT3D mesh, respectively, varying by less than 3% over $B = 1$ to 128. This cost applies to extent evaluation and requested mesh samples; returning region parameters does not require these decodes.

B.7 Direct Output-Space Conformal Controls

We compare three direct output-space sets on the SMPL tasks. Let $d_j \in \mathbb{R}^3$ and $d_v \in \mathbb{R}^3$ be root-aligned joint and vertex displacements, with $J = 14$ and $V = 6890$. The scales s_j and s_v are the mean Euclidean error for each joint and vertex on development data. Each score is calibrated from the same 20 sequence-level allocations as the main comparison, with $n = 40$ calibration observations per allocation. Because each score covers a complete output vector, split conformal gives finite-sample marginal coverage under exchangeability.

B1: jointwise balls. The score $S_{\mathrm{B1}} = \max_j \|d_j\|_2/s_j$ defines the product of 14 three-dimensional balls, each of radius $\hat{q}_{\mathrm{B1}} s_j$. Under the standard deterministic split-conformal quantile, separate Bonferroni calibration of the 14 balls yields an infinite radius with $n = 40$: for all three reported levels, its order-statistic index is $\lceil 41(1 - \alpha/14) \rceil = 41 > 40$. For the uniform measure on this product set, $E_{\mathrm{RMS}} = \hat{q}_{\mathrm{B1}} \sqrt{3 \sum_j s_j^2/(5J)}$.

B2: aggregate vertex-RMS ball. The score $S_{\mathrm{B2}} = \sqrt{V^{-1}\sum_v \|d_v\|_2^2}$ defines one ball in $\mathbb{R}^{3V}$ of radius $\hat{q}_{\mathrm{B2}}\sqrt{V}$. Its uniform-measure extent is $E_{\mathrm{RMS}} = \hat{q}_{\mathrm{B2}}\sqrt{3V/(3V+2)}$; replacing this ball with a product of per-vertex balls would define a different set.

B3: per-vertex balls. The score $S_{\mathrm{B3}} = \max_v \|d_v\|_2/s_v$ defines a product of V three-dimensional balls of radii $\hat{q}_{\mathrm{B3}}s_v$. Its uniform-measure extent is $E_{\mathrm{RMS}} = \hat{q}_{\mathrm{B3}}\sqrt{3\sum_v s_v^2/(5V)}$. B3 imposes pointwise vertex bounds, whereas B2 constrains aggregate RMS; their different scores and sets make this a comparison of two output-space constructions, not a single-factor ablation of scaling.

Setting	GM-R	B1: joints	B2: RMS	B3: vertices
HMR2.0 / 3DPW joints	78.1	214.5	–	–
ProHMR / 3DPW joints	89.1	233.7	–	–
HMR2.0 / EMDB joints	94.5	274.5	–	–
HMR2.0 / 3DPW mesh	94.4	–	181.3	273.5
ProHMR / 3DPW mesh	112.5	–	205.3	303.6
HMR2.0 / EMDB mesh	107.6	–	219.6	343.7

Table 26: **Direct output-space controls at nominal** 0.90**.** Mean E_{RMS} in mm across the same 20 calibration allocations as the main tables. GM-R is included for scale; dashes mark inapplicable output types. Each control uses the uniform measure on its own calibrated output-space set, whereas GM-R decodes the uniform measure on its parameter ellipsoid.

Table 26 shows that B2 has smaller extent than B3 on each tested mesh task, yet its mean remains 1.8–$2.0\times$ that of GM-R. All three controls are calibrated sets in ambient output space and can contain configurations that the decoder cannot generate. Their extent ratios compare the specified uniform measures, not a measure-independent ordering of set sizes. The table reports mean extent only; we make no paired effect-rule or observed-coverage claim for these controls.

B.8 Additional Extent Diagnostics

Finite-radius response on evaluation data. Across 1,440 calibration-split-level records in the 12 reported settings at nominal 0.90 (72 setting–family combinations $\times$ 20 splits), χ_h ranges from 0.52 to 2.13, with median 0.98. The odd correction is negative in 97.6% of records, while the nonnegative even energy has median 28% of $\bar{D}_{\mathrm{lin}}$ and range 7–139% (Eq. (27)). These descriptive summaries show that nonlinear response can either reduce or increase squared extent relative to its linearization.

Normalization on evaluation data. Normalizing each complete metric by $\mathrm{tr}(A_h(x)\widetilde{M})$ and recalibrating changes E_{RMS} by at most 0.863% across 14 GM-I/GM-R arms on seven settings, with seven increases and seven decreases. These are evaluation-side results; the development diagnostics below use a different set of inputs.

Trace variability on development data. For each complete, unnormalized metric, define the trace factor $t_h(x) = \mathrm{tr}(A_h(x)\widetilde{M})$, computed in float64. The ablation replaces A_h by A_h/t_h and recalibrates. If t_h is constant across inputs, calibration absorbs this global rescaling; its variability therefore measures departure from that invariant case. This empirical factor includes the ridge and is distinct from the theoretical $\tau(x) = \mathrm{tr}(H(x)^{1/2})$ for the GM core.

For cell h and rotation f, $\widetilde{M}_{h,f}$ and all moment inverses and scale constants are fitted on that rotation's FIT rows. The complete metric $A_{h,f}$ is then evaluated on *all* n_h retained development rows, not just FIT, RADIUS, or RANK; 3DPW contributes 2,000 rows after sequence-capped subsampling. Thus $t_{h,f,i} = \mathrm{tr}(A_{h,f}(x_i)\widetilde{M}_{h,f})$ uses the FIT moment on both sides. A consistency check requires the largest absolute entry of the inverse–moment product minus I to be below 10^{-4}, using the moment inverse stored for A. With $\bar{t}_{h,f} = n_h^{-1}\sum_i t_{h,f,i}$, the empirical coefficient of variation and its cell

summary are

$$c_{h,f} = \frac{\sqrt{\frac{1}{n_h-1}\sum_{i=1}^{n_h}(t_{h,f,i} - \bar{t}_{h,f})^2}}{\bar{t}_{h,f}}, \qquad \bar{c}_h = \frac{1}{5}\sum_{f=0}^{4} c_{h,f}. \tag{41}$$

These values come from the unnormalized arm. The within-rotation standard deviation uses Bessel correction, whereas the descriptive calibration-split SD in the result tables uses divisor 20 without that correction.

Normalization on development data. The 14 development cells comprise seven instances and two geometries, GM-I and GM-R. For cell h, let $d_h = 100(E^{\mathrm{norm}}_{\mathrm{dev},h} - E^{\mathrm{unnorm}}_{\mathrm{dev},h})/E^{\mathrm{unnorm}}_{\mathrm{dev},h}$, so a negative value means normalization reduces RMS extent. We use `dE_RMS_pct_cell` for d_h and `cv_tau_measured_here` for $\bar{c}_h$ from `cells.csv`. These are cell-level changes, not averages of fold-paired or sequence-paired percentage changes. All correlations below are computed from full-precision values before rounding.

Development extent decreases in nine cells and increases in five, with changes bounded in magnitude by 0.747%. The temporary threshold decreases in 12 of 14 cells, and threshold and extent changes have opposite signs in five cells. Across the 14 development cells, the two associations are

Association on development data	Pearson	Spearman
$\bar{c}_h$ versus $\lvert d_h \rvert$ (effect magnitude)	-0.2566	-0.0462
$\bar{c}_h$ versus d_h (signed change)	$+0.2554$	$+0.1121$

The magnitude association does not distinguish improvement from deterioration. The positive signed Pearson coefficient points toward less benefit at greater trace variability in this development summary, but the weak rank associations do not establish a monotonic relation. These are descriptive correlations, not tests of a prespecified success criterion.

Development and evaluation extent changes have the same sign in only seven of the 14 cells. These descriptive development results do not establish an association between trace variability and evaluation extent changes; trace variability has not been recomputed on evaluation inputs. The unregularized core's variance identity motivates the diagnostic but does not determine the effect on regularized metrics or calibrated nonlinear extent (Appendix A.3).