SPARC: Single-Pass Scaling for Motion Forecasting with Conformal Bayesian Last Layers

Sakif Hossain[0000−0003−0800−8392] [1⋆], Julian Teusch[0000−0002−4103−8430] [1⋆⋆], and Jörg P. Müller[0000−0001−7533−3852] [1]

Clausthal University of Technology (TU Clausthal)
sakif.hossain@tu-clausthal.de, julian.teusch@tu-clausthal.de
joerg.mueller@tu-clausthal.de

Abstract. Human motion forecasters are increasingly accurate and fast, but reliable deployment requires uncertainty estimates that are structured, calibrated, and efficient. Bayesian and ensemble-based uncertainty estimates often require repeated stochastic inference [15, 26], while conformal calibration alone does not provide an epistemic signal or preserve trajectory covariance structure [14, 50]. We introduce *SPARC* (Single-Pass Adaptive Risk Calibration), a Bayesian–conformal uncertainty layer for motion forecasting. A deterministic MLP backbone predicts the future mean, and a conjugate Bayesian last layer converts time-domain feature leverage into an analytic horizon-wise epistemic scale $\kappa_t(x)$. This scale inflates a graph-temporal Gaussian covariance without changing its correlation structure, and split conformal calibration produces 95% marginal prediction tubes with finite-sample validity under exchangeability. The key interface is the structured factorization $\kappa_t(x)\Sigma_{\mathrm{str},t}(x)$, which injects feature-space epistemic uncertainty into trajectory densities without Monte Carlo sampling. Across nine dataset/protocol blocks and deterministic, multimodal, and calibration baselines, *SPARC* ranks first on NLL and on the combined MPJPE+NLL criterion while retaining competitive point accuracy and efficient calibrated tubes. Ranking windows by κ separates high-error cases, making the scale usable as a lightweight risk monitor.

Keywords: Human motion forecasting · uncertainty quantification · Bayesian last layer · conformal prediction

1 Introduction

Human motion forecasting—predicting future 3D pose trajectories from a short observed prefix—is a core primitive for human–robot interaction, animation, and anticipatory perception. Recent forecasters achieve low mean per-joint position error (MPJPE) with fast inference [18,37], but point forecasts alone are insufficient for downstream decision-making. Reliable systems must reason about *uncertainty*:

⋆ Equal contribution. Corresponding author.
⋆⋆ Equal contribution.

multiple futures may be plausible due to intrinsic ambiguity (aleatoric), and models may become unreliable when extrapolating beyond their training support (epistemic), especially under distribution shift [23, 40].

Bayesian deep learning models epistemic uncertainty through a posterior over model parameters, yet practical approaches often rely on approximate inference and repeated stochastic forward passes (e.g., Monte Carlo (MC) dropout [15] or ensembles [26]) and can degrade under shift [40]. Moreover, modern neural networks are frequently miscalibrated even in independent and identically distributed (i.i.d.) settings, motivating calibration layers with explicit guarantees [17]. Neural-linear and Bayesian last-layer methods reduce this cost by keeping a deterministic feature extractor and placing a conjugate Bayesian model on the final linear layer, yielding closed-form posterior predictive uncertainty at near single-pass cost [12, 33, 41, 47, 51]. However, neural-linear uncertainty is typically paired with factorized (diagonal or isotropic) observation models [12, 24, 33], making it nontrivial to preserve structured spatio-temporal correlations for trajectory outputs.

In parallel, conformal prediction provides distribution-free predictive inference with finite-sample marginal validity under exchangeability [1, 29, 50]. Conformal methods have also been extended to functional and trajectory-like outputs, yielding finite-sample valid prediction bands around a base forecaster [13, 30]. For regression, conformalization can wrap any base predictor (including heteroscedastic quantile regressors [42]) to guarantee target coverage, but it does not by itself separate epistemic from aleatoric uncertainty, and achieving strong conditional guarantees remains challenging in general [16].

We introduce *SPARC*, a Bayesian–conformal interface that adds structured calibrated uncertainty to a fast deterministic motion forecaster in a single pass. For deployment, uncertainty must attach to future joint trajectories with temporal and skeletal dependencies, not only to a scalar confidence score. Starting from a strong deterministic MLP backbone, SPARC places a conjugate Bayesian model on the final linear layer and derives an analytic, input-dependent *epistemic scale* $\kappa(x)$ (or timewise $\kappa_t(x)$) with a leverage-style interpretation. This scale multiplicatively inflates a structured Gaussian aleatoric covariance that models spatio-temporal and skeletal correlations, and split conformal calibration converts the resulting distribution into 95% marginal prediction tubes under exchangeability [13, 30], complementing Bayesian credible regions under model misspecification [14, 58]. The same κ signal also enables κ-conditioned (Mondrian) tube scaling and optional trajectory-level conformal ellipsoid scores (Sec. 3.2).

Contributions. We contribute:

- **Structured Bayesian last-layer scaling.** Under a multi-output conjugate model, the predictive covariance factorizes exactly as $\kappa_t(x)\,\Sigma_{\mathrm{str},t}(x)$ (Prop. 1), enabling analytic epistemic inflation while preserving arbitrary positive-definite trajectory covariance structure.
- **Single-pass Bayesian–conformal motion uncertainty.** We combine this scale with a graph-temporal Gaussian covariance head and split conformal

calibration, producing calibrated marginal prediction tubes without Monte Carlo sampling or ensembles.

– **Cross-dataset probabilistic gains.** Across nine dataset/protocol blocks, *SPARC* ranks first on average NLL and on overall MPJPE+NLL, with efficient calibrated tubes and competitive point accuracy.
– **Deployment-oriented epistemic monitoring.** On Human3.6M, ranking windows by κ separates high-error cases: the highest-κ decile has $1.79\times$ MPJPE compared to the average, and filtering that decile reduces MPJPE by $\approx 9\%$ (supplementary Sec. *Epistemic scale diagnostics*; thresholds tuned on held-out data, with no formal guarantee implied).

2 Related Work

Human motion forecasting has long been dominated by deterministic point predictors, ranging from early recurrent/seq2seq models [37] to modern feed-forward backbones that offer strong runtime–accuracy trade-offs [18]. Recent deterministic improvements add structure via graph convolutions [10], auxiliary supervision [53], stability-inducing integration [6], or deviation-feedback refinement [46]. To represent the inherent ambiguity of future motion, multimodal predictors pursue latent-variable and generative formulations, including normalizing flows [55], diverse candidate generation [35], and diffusion-style denoising models [7]. While these approaches can capture multiple plausible futures, their inference often involves sampling or multiple hypotheses, which motivates lightweight uncertainty layers that preserve real-time point-forecasting backbones [7, 35, 55].

Uncertainty quantification for deep regression typically distinguishes aleatoric ambiguity from epistemic uncertainty due to limited training support [23]. Common epistemic estimators such as MC dropout and deep ensembles [15, 26] can improve robustness but require repeated stochastic forward passes and can still fail under distribution shift [40]. Neural-linear and last-layer Bayesian approximations keep the feature extractor deterministic and place Bayesian inference on the final linear layer, yielding analytic predictive uncertainty at near single-pass cost [12, 33, 41, 47, 51]. However, these epistemic estimators are frequently used with factorized (diagonal or isotropic) likelihoods in regression [12, 24, 33], which does not capture spatio-temporal or skeletal correlations needed for coherent trajectory tubes and likelihood.

Structured Gaussian models such as matrix-normal/Kronecker factorizations capture separable spatio-temporal correlations, while graph/Gaussian Markov random field (GMRF) constructions encode skeletal coupling via sparse precisions [3, 11, 43]. Separately, conformal prediction offers distribution-free predictive inference with finite-sample *marginal* validity under exchangeability [1, 29, 50]. It has been extended to functional and trajectory bands [13, 30]; for regression, conformalized quantile regression is a strong calibration baseline [42], while dependent time series require additional assumptions or online adaptation [52]. Recent work adapts conformal calibration to structured prediction, scene graphs, and multi-dimensional time series, illustrating that for non-scalar outputs the task-

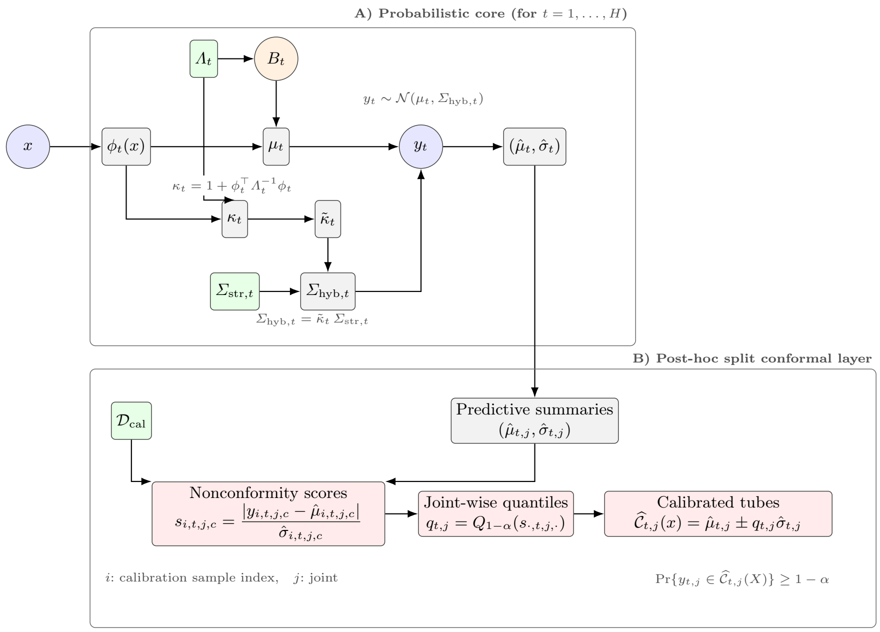

Fig. 1: SPARC dependency view. (A) Core: features yield $\kappa_t = 1 + \phi_t^\top \Lambda_{n,t}^{-1} \phi_t$ and inflate the structured covariance. (B) Split CP: calibration scores produce per-joint quantiles $q_{t,j}$ for marginal tubes.

specific score and output interface are central [28,39,56]. Conformal methods have also been used for trajectory uncertainty under distribution shift and safety-aware planning [21, 31]; in 3D human motion forecasting, latent conformal prediction calibrates multidimensional trajectories [32]. Approximate Bayesian posteriors have likewise been paired with conformal layers, including aleatoric–epistemic scaling via last-layer Laplace for conformal regression [24] and posterior conformal prediction [58]. SPARC jointly integrates analytic conjugate Bayesian κ_t, graph-temporal covariance, marginal conformal tubes, and optional trajectory-level conformal set scores in a single-pass motion forecasting pipeline.

3 Method

SPARC turns a deterministic forecaster into calibrated marginal prediction tubes through a Bayesian last-layer scale and a post-hoc conformal layer. Figure 1 shows the dependency structure. We then specify the calibrated deployment recipe used in experiments. Algorithm 1 summarizes the recipe for adding this uncertainty layer to deterministic forecasters. We additionally describe two conformal variants enabled by the analytic κ and structured covariance: κ-conditioned (Mondrian) tubes and a trajectory-level conformal ellipsoid (Sec. 3.2). Simpler variants in our ablation are recovered by dropping the graph coupling, the timewise shrinkage, or the hybrid inflation.

Method overview. SPARC combines a deterministic mean backbone $\hat{\mu}$ with a structured Gaussian head for aleatoric temporal/skeletal correlations Σ_{str}. A conjugate Bayesian last layer maps time-domain feature leverage to $\kappa_t(x)$, and split conformal calibration turns standardized residuals into marginal tubes; inference uses one forward pass plus closed-form quadratic forms.

3.1 Hybrid structured Gaussian predictor (ConjGraph)

Let $x_{1:T} \in \mathbb{R}^{T\times C}$ be an observed pose prefix and $y_{1:H} \in \mathbb{R}^{H\times C}$ the future trajectory, with $C = 3J$ coordinates (flattened joints), where J is the number of joints. As in our implementation, we train and evaluate on *displacements* relative to the last observed pose x_T (i.e., $y_t \leftarrow y_t - x_T$); absolute poses are recovered by adding x_T back at inference.

Goal: calibrated marginal tubes. Our primary reported object is a collection of axis-aligned marginal intervals, written compactly as a prediction tube $\mathcal{C}_\alpha(x)$ with miscoverage $\alpha \in (0, 1)$:

$$\widehat{\mathcal{C}}_\alpha(x) \;=\; \prod_{t=1}^{H}\prod_{j=1}^{J}\prod_{d=1}^{3}\big[\hat{\mu}_{t,j,d}(x) \;\pm\; q_{t,j}\,\hat{\sigma}^{\text{CP}}_{t,j,d}(x)\big], \tag{1}$$

where $(\hat{\mu}, \hat{\sigma}^{\text{CP}})$ come from a probabilistic predictor and $\{q_{t,j}\}$ are split-conformal calibration factors (Sec. 3.2). The product notation denotes the reported rectangle of marginal intervals; the guarantee below is marginal for the induced score distribution, while stricter trajectory-level set scores are considered separately (Sec. 3.2).

Backbone and time-domain design vector. We follow the siMLPe/DCT-bridge backbone [18]: the input sequence is transformed along the time axis into a discrete cosine transform (DCT) coefficient representation and passed through a deterministic trunk f_θ that outputs coefficient-wise features $h_n(x) \in \mathbb{R}^C$ for $n \in \{1, \ldots, N\}$. Let $A \in \mathbb{R}^{N\times N}$ denote the inverse DCT (iDCT) matrix that maps coefficient outputs back to the time domain. In the fully-connected output (FC-Out) parameterization used throughout this work, the time-domain mean at horizon step t is linear in an augmented design vector $\phi_t(x) \in \mathbb{R}^P$:

$$\hat{\mu}_t(x) = B_t^\top \phi_t(x), \tag{2}$$

$$\phi_t(x) = \begin{bmatrix} g_t(x) \\ s_t \end{bmatrix} \in \mathbb{R}^P, \qquad P = C + 1,$$

$$g_t(x) = \sum_{n=1}^{N} A_{t,n}\, h_n(x) \in \mathbb{R}^C, \qquad s_t = \sum_{n=1}^{N} A_{t,n} \in \mathbb{R}. \tag{3}$$

The scalar s_t accounts for the coefficient-space bias after mapping back through the iDCT.

Conjugate Bayesian last-layer scale. We place a conjugate Bayesian model on the final linear map in (2) while keeping the trunk deterministic. For each horizon step t, consider the multi-output linear-Gaussian model

$$y_t \;=\; B_t^\top \phi_t(x) + \varepsilon_t, \qquad \varepsilon_t \sim \mathcal{N}(0, \Sigma_{\mathrm{str},t}(x)), \tag{4}$$

with a matrix-normal prior [11]

$$B_t \mid \Sigma_{\mathrm{str},t} \;\sim\; \mathcal{MN}(B_{0,t}, \Lambda_0^{-1}, \Sigma_{\mathrm{str},t}), \qquad \Lambda_0 = \lambda_0 I_P. \tag{5}$$

Proposition 1 (Exact structured-covariance scaling). *Conditioning on any positive-definite $\Sigma_{\mathrm{str},t}(x)$, integrating out B_t under the conjugate matrix-normal model in (4)–(5) preserves the structured covariance up to a* scalar *multiplicative inflation:*

$$y_t \mid x, \mathcal{D} \;\sim\; \mathcal{N}\big(\mu_t^{\mathrm{post}}(x),\; \kappa_t(x)\, \Sigma_{\mathrm{str},t}(x)\big), \qquad \kappa_t(x) = 1 + \phi_t(x)^\top \Lambda_{n,t}^{-1} \phi_t(x),$$

where $\Lambda_{n,t} = \Lambda_0 + \sum_{i\in\mathcal{D}} \phi_t(x_i)\phi_t(x_i)^\top$. In particular, κ_t rescales only *the covariance magnitude and preserves the correlation/structure encoded by $\Sigma_{\mathrm{str},t}$.*

The intuition is that integrating over the last-layer weights adds uncertainty in the output direction induced by the feature leverage $\phi_t(x)^\top \Lambda_{n,t}^{-1} \phi_t(x)$. Because the matrix-normal prior shares the same output covariance as the structured residual model, this additional uncertainty has the same correlation structure as $\Sigma_{\mathrm{str},t}$. Thus epistemic uncertainty can be injected by the scalar multiplier $\kappa_t(x)$ instead of replacing the structured covariance with a diagonal approximation.

Proof. By conjugacy, $B_t \mid \Sigma_{\mathrm{str},t}, \mathcal{D} \sim \mathcal{MN}(B_{n,t}, \Lambda_{n,t}^{-1}, \Sigma_{\mathrm{str},t})$, hence $\mathrm{vec}(B_t) \mid \Sigma_{\mathrm{str},t}, \mathcal{D}$ is Gaussian with covariance $\Sigma_{\mathrm{str},t} \otimes \Lambda_{n,t}^{-1}$. Therefore $B_t^\top \phi_t(x)$ has covariance $(\phi_t(x)^\top \Lambda_{n,t}^{-1} \phi_t(x))\, \Sigma_{\mathrm{str},t}(x)$, and adding $\varepsilon_t \sim \mathcal{N}(0, \Sigma_{\mathrm{str},t}(x))$ yields the stated inflation.

Remark 1 (Arbitrary structured covariance). The factorization in Prop. 1 holds for any positive-definite $\Sigma_{\mathrm{str},t}(x)$ (dense, Kronecker-factored, or defined via a sparse precision), enabling epistemic inflation without simplifying the structured aleatoric head.

We use $\kappa_t(x)$ as an analytic epistemic inflation factor (no MC sampling) that increases when the test feature vector $\phi_t(x)$ has large leverage under the training design. This is a feature-space signal: its informativeness depends on the learned representation ϕ_t and should not be interpreted as a universal out-of-distribution detector independent of the backbone. In SPARC we keep the backbone mean predictor $\hat{\mu}_t(x)$ and use only $\kappa_t(x)$ (and its stabilized version $\tilde{\kappa}_t(x)$ below) to rescale the structured covariance head.

Closed-form fitting and stabilization. We only require $\Lambda_{n,t}^{-1}$ to compute κ_t. With $\Lambda_0 = \lambda_0 I_P$, we fit $\Lambda_{n,t}^{-1}$ in closed form by accumulating the sufficient statistic $S_{xx,t} = \sum_i \phi_t(x_i)\phi_t(x_i)^\top$ on the training set and computing $(S_{xx,t} + \lambda_0 I_P)^{-1}$ once via a Cholesky factorization. At inference, we evaluate the quadratic form $\kappa_t(x) = 1 + \phi_t(x)^\top \Lambda_{n,t}^{-1}\phi_t(x)$ directly, avoiding MC sampling. To stabilize horizon-wise scaling, we optionally shrink κ_t toward a time-homogeneous sequence-level scale derived from pooled statistics. Let $\Lambda_{\text{glob}} = \lambda_0 I_P + \sum_{t,i} \phi_t(x_i)\phi_t(x_i)^\top$ and define

$$\bar{\kappa}(x) \;=\; \frac{1}{H}\sum_{t=1}^{H}\Big(1 + \phi_t(x)^\top \Lambda_{\text{glob}}^{-1}\phi_t(x)\Big). \tag{6}$$

Our stabilized scale is

$$\tilde{\kappa}_t(x) \;=\; (1-\rho)\,\kappa_t(x) + \rho\,\bar{\kappa}(x), \qquad \rho \in [0,1]. \tag{7}$$

The global-κ variant is recovered by setting $\rho = 1$. In experiments we also allow a monotone "epistemic temperature" transform $\tilde{\kappa}_t \leftarrow \tilde{\kappa}_t^{\gamma}$ with $\gamma > 0$ (with $\gamma = 1$ corresponding to the default construction) to tune the NLL/interval-width (mean marginal interval width; reported as W95) trade-off.

Per-joint leverage scale for CP tubes. The global timewise scale $\tilde{\kappa}_t(x)$ is shared across all joints at horizon t. To allow per-joint epistemic inflation in the CP layer without changing the likelihood, we compute an additional per-joint scale $\kappa_j(x)$ from reduced per-joint design vectors. Let $\phi_{t,j}(x) \in \mathbb{R}^4$ contain the 3 coordinates of joint j in $g_t(x)$ plus the shared bias feature s_t. We fit a per-joint conjugate statistic $\Lambda_{n,j}^{-1} = (\lambda_{0,\text{joint}} I + \sum_{i,t} \phi_{t,j}(x_i)\phi_{t,j}(x_i)^\top)^{-1}$ on $\mathcal{D}_{\text{train}}$ and define $\kappa_{t,j}(x) = 1 + \phi_{t,j}(x)^\top \Lambda_{n,j}^{-1}\phi_{t,j}(x)$, aggregated as $\kappa_j(x) = \frac{1}{H}\sum_{t=1}^{H}\kappa_{t,j}(x)$. For conformal tubes we use

$$\hat{\sigma}_{t,j,d}^{\text{CP}}(x) \;=\; \sqrt{\kappa_j(x)}\,\hat{\sigma}_{t,j,d}(x), \tag{8}$$

while the probabilistic core (and NLL) remains unchanged.

Structured Gaussian aleatoric covariance (MN-GraphJ). We refer to this matrix-normal + joint-graph construction as *MN-GraphJ* (MatrixNormal-GraphJ). The backbone predicts a structured *aleatoric* covariance for the residual matrix $R(x) = y_{1:H} - \hat{\mu}_{1:H}(x) \in \mathbb{R}^{H\times C}$. We use a separable matrix-normal parameterization,

$$\begin{aligned} \operatorname{vec}(R(x)) &\sim \mathcal{N}\big(0,\ \Sigma_T(x)\otimes \Sigma_C(x)\big),\\ R(x) &\sim \mathcal{MN}(0, \Sigma_T(x), \Sigma_C(x)), \end{aligned} \tag{9}$$

where $\Sigma_T(x) \in \mathbb{R}^{H\times H}$ captures temporal correlations and $\Sigma_C(x) \in \mathbb{R}^{C\times C}$ captures spatial/skeletal coupling. We parameterize $\Sigma_T(x) = L_T(x)L_T(x)^\top$ with a learned lower-triangular $L_T(x)$ (packed Cholesky representation), enabling dense temporal correlations with stable optimization. To encode skeletal structure, we model Σ_C via a joint-graph precision (a GMRF construction [3, 43]). Let

$L_{\text{joint}} \in \mathbb{R}^{J\times J}$ be the unweighted graph Laplacian of a chosen joint graph (e.g., the kinematic tree, or a data-driven k-nearest-neighbor (kNN) graph estimated from training motions). The network predicts positive scalars $\tau(x)$ and $\epsilon(x)$, and we define

$$Q_J(x) \;=\; \tau(x)\,L_{\text{joint}} + \epsilon(x)\,I_J, \qquad \Sigma_C(x) \;=\; (Q_J(x)\otimes I_3)^{-1}. \tag{10}$$

Hybridization. We combine epistemic and aleatoric components by inflating the temporal covariance factor with $\tilde{\kappa}_t(x)$. Define

$$D_\kappa(x) = \operatorname{diag}\bigl(\sqrt{\tilde{\kappa}_t(x)}\bigr)_{t=1}^{H}, \tag{11}$$

and set

$$\Sigma_T^{\text{hyb}}(x) \;=\; D_\kappa(x)\,\Sigma_T(x)\,D_\kappa(x)^\top, \qquad \Sigma^{\text{hyb}}(x) = \Sigma_T^{\text{hyb}}(x)\otimes\Sigma_C(x). \tag{12}$$

If $\Sigma_T(x) = L_T(x)L_T(x)^\top$, we implement (12) by scaling the t-th row of $L_T(x)$ by $\sqrt{\tilde{\kappa}_t(x)}$, preserving correlation structure while inflating scale. The resulting marginal variance factorizes as

$$\hat{\sigma}^2_{t,c}(x) \;=\; \operatorname{Var}(y_{t,c}\mid x) \;=\; \tilde{\kappa}_t(x)\,[\Sigma_T(x)]_{tt}\,[\Sigma_C(x)]_{cc}. \tag{13}$$

Scope of the exact result and stabilizers. Prop. 1 establishes the exact last-layer covariance factorization under the stated conjugate Gaussian model, conditional on the structured covariance. The shrinkage coefficient ρ and epistemic temperature γ are held-out stabilizers for the NLL–width trade-off, not additional assumptions in the theorem. Likewise, the per-joint scale in (8) is used only for conformal tube scoring and construction; it does not change the Gaussian NLL core. The split-conformal guarantee below therefore attaches to the final calibrated score under exchangeability, while the Bayesian factorization explains the probabilistic core that supplies the scale.

3.2 Split conformal calibration for prediction tubes

Bayesian credible regions can be miscalibrated under model misspecification or dataset shift [14, 58]. We therefore conformalize the hybrid Gaussian predictor to obtain calibrated marginal tubes with finite-sample validity under exchangeability [1, 29, 50].

Following split conformal calibration [1, 29, 50], on a held-out calibration set $\mathcal{D}_{\text{cal}} = \{(x_i, y_i)\}_{i=1}^n$, we compute standardized absolute residual scores

$$s_{i,t,j,d} \;=\; \frac{|y_{i,t,j,d} - \hat{\mu}_{t,j,d}(x_i)|}{\hat{\sigma}^{\text{CP}}_{t,j,d}(x_i)}, \tag{14}$$

where $\hat{\sigma}^{\text{CP}}$ is the scale used for tube construction (in SPARC this includes the hybrid timewise inflation $\tilde{\kappa}_t$ and the CP-only joint factor in (8)). For each horizon step t and joint j, we set $q_{t,j}$ using the split-conformal "higher" quantile rule,

$$q_{t,j} \;=\; \operatorname{Quantile}\bigl(\{s_{i,t,j,d}\}_{i,d},\; \lceil (N_{t,j}+1)(1-\alpha)\rceil / N_{t,j}\bigr), \tag{15}$$

where $N_{t,j}$ is the number of pooled scores for (t, j) (in our implementation $N_{t,j} = n\cdot 3$). We then output the tube in (1).

Algorithm 1 SPARC: single-pass structured uncertainty with analytic κ and split conformal calibration

Require: Training data $\mathcal{D}_{\text{train}}$, calibration data $\mathcal{D}_{\text{cal}}$, miscoverage α, prior precisions λ_0 and $\lambda_{0,\text{joint}}$

1: Train a deterministic mean forecaster (mean fine-tune; MeanFT) to obtain time-domain features $\phi_t(x)$ and mean predictions $\hat{\mu}_t(x)$.

2: Fit a structured Gaussian covariance head (e.g., MN-GraphJ) on residuals with frozen mean, yielding $\Sigma_{\text{str},t}(x)$.

3: Accumulate $S_{xx,t} \leftarrow \sum_{(x_i,\cdot)\in\mathcal{D}_{\text{train}}} \phi_t(x_i)\phi_t(x_i)^\top$ and compute $\Lambda_{n,t}^{-1} = (\lambda_0 I + S_{xx,t})^{-1}$ (e.g., via Cholesky factorization).

4: Fit per-joint leverage statistics: accumulate $S_{xx,j} \leftarrow \sum_{(x_i,\cdot)\in\mathcal{D}_{\text{train}},t} \phi_{t,j}(x_i)\phi_{t,j}(x_i)^\top$ and compute $\Lambda_{n,j}^{-1} = (\lambda_{0,\text{joint}} I + S_{xx,j})^{-1}$. Use it to compute $\kappa_j(x)$ and $\hat{\sigma}^{\text{CP}}$ via (8).

5: **Split conformal calibration:** on $\mathcal{D}_{\text{cal}}$, compute
$s_{i,t,j,d} = |y_{i,t,j,d} - \hat{\mu}_{t,j,d}(x_i)|/\hat{\sigma}^{\text{CP}}_{t,j,d}(x_i)$ and set
$q_{t,j} \leftarrow \text{Quantile}(\{s_{i,t,j,d}\}_{i,d}, \lceil (N_{t,j}+1)(1-\alpha)\rceil / N_{t,j})$ (higher), where $N_{t,j}$ is the number of pooled scores for (t,j).

6: **(Optional)** κ**-Mondrian CP:** bin sequences by $r(x)$ (e.g., mean $\sqrt{\kappa_t(x)}$) and compute $q_{t,j,b}$ per bin b; at test time use $q_{t,j,b(x)}$.

7: **(Optional) trajectory CP:** compute $s_i = \| L_T^{\text{hyb}}(x_i)^{-1}(y_i - \hat{\mu}(x_i)) L_C(x_i)^{-\top}\|_F$ and conformalize to get q_{traj}, yielding an ellipsoidal set $\{y : \; s(y) \le q_{\text{traj}}\}$.

8: **Prediction for new** x**:** compute $\kappa_t(x) = 1 + \phi_t(x)^\top \Lambda_{n,t}^{-1}\phi_t(x)$, optionally shrink to $\tilde{\kappa}_t(x)$ via (7), and form $\Sigma_{\text{hyb},t}(x) = \tilde{\kappa}_t(x)\Sigma_{\text{str},t}(x)$.

9: Output calibrated tube $\widehat{\mathcal{C}}_\alpha(x)$ in (1) (or $q_{t,j,b(x)}$), and optionally the trajectory ellipsoid.

Guarantee and practical notes. Split conformal provides finite-sample *marginal* validity for the induced score distribution under exchangeability between calibration and evaluation windows. Calibration is post-hoc: it computes quantiles on $\mathcal{D}_{\text{cal}}$ and does not update network weights, covariance parameters, or the conjugate statistics. Under temporal dependence or distribution shift that violates exchangeability, the conformal layer should be read as empirical calibration rather than an unconditional coverage guarantee. Pooling over the 3 coordinates per joint targets the mean marginal coverage metric we report; stricter max-score variants are more conservative.

Risk-adaptive and set-valued variants. The supplement reports two extensions: κ-conditioned (Mondrian) conformal tubes that calibrate bin-wise $q_{t,j,b}$ [4,49], and a trajectory-level conformal ellipsoid based on a structured Mahalanobis/whitened-residual score under $\Sigma_T^{\text{hyb}} \otimes \Sigma_C$.

4 Experiments

4.1 Experimental setup

Datasets. We treat Human3.6M (H36M) [22] as the primary benchmark (indoor motion capture (MoCap)) and additionally evaluate cross-dataset generalization on AMASS [34], LaFAN1 [19], CMU-MoCap [5], 3DPW [36], and three human–robot interaction (HRI)-style datasets with different capture / interaction setups: CHICO [44], HA4M [8], and AnDy [38]. Unless explicitly stated otherwise, non-H36M datasets use T=50 observed frames and we forecast H=25 future frames.

Evaluation protocol. We use a random-but-deterministic partition of the official evaluation windows. We fix the random number generator (RNG) seed to 304 and permute the windows. We use the first n_{cal}=512 windows for split conformal calibration and report all metrics on the next n_{eval}=1024 windows. We use α=0.05 throughout. We keep $(n_{\text{cal}}, n_{\text{eval}}, \alpha)$ fixed across models. Data usage is separated by operation: the training split fits the backbone, covariance heads, and conjugate sufficient statistics; held-out training windows select the five uncertainty/calibration hyperparameters: λ_0 ('lambda0'), ρ ('time_shrink_rho'), γ ('kappa_power'), $\lambda_{0,\text{joint}}$ ('lambda0_joint'), and γ_{joint} ('joint_kappa_power'). The calibration subset computes only split-conformal quantiles $\{q_{t,j}\}$, and the disjoint evaluation subset is used only once for the reported metrics. Thus the conformal step is post-hoc calibration rather than gradient training, and no model weights, covariance parameters, conjugate statistics, or hyperparameters are updated on the final evaluation windows.

Metrics. We report mean per-joint position error (MPJPE), final displacement error (FDE; here implemented as final-horizon MPJPE), and per-element Gaussian negative log-likelihood (NLL) (averaged over horizons and coordinates). For space, we present MPJPE/FDE as "MPJPE / FDE" in the main tables. Because absolute metric scales and units vary across datasets, we additionally report per-dataset ranks in cross-dataset summaries; MPJPE+NLL rank denotes the mean of MPJPE and NLL ranks (Table 1). For uncertainty, we evaluate nominal two-sided 95% marginal intervals $\hat{\mu} \pm Z_{0.95}\hat{\sigma}$ (**Cov95/W95**) and split-conformalized intervals $\hat{\mu} \pm q_{t,j}\,\hat{\sigma}^{\text{CP}}$ (**Cov95(CP)/W95(CP)**). Here $Z_{0.95} = \Phi^{-1}(0.975) \approx 1.960$. Cov95 is empirical marginal coverage (fraction of scalar targets (t, c) covered), and W95 is the corresponding mean marginal interval width (averaged over (t, c)). Conformal scales $\{q_{t,j}\}$ are calibrated *per horizon and joint* by pooling standardized residuals $|y_{t,j,d} - \hat{\mu}_{t,j,d}|/\hat{\sigma}^{\text{CP}}_{t,j,d}$ across the 3 coordinates d of each joint on the calibration split. Deterministic models are paired with a fixed isotropic observation noise σ_{obs}=0.017 m to compute NLL and nominal intervals; they do not output $\hat{\sigma}_{t,c}(x)$ and thus do not report conformal tube metrics. The supplementary material additionally reports κ-binned Mondrian tubes and trajectory-level conformal ellipsoid metrics.

Training and calibration. All methods share the siMLPe/DCT backbone [18] (MLP in DCT coefficient space, mapped to the time domain via iDCT; we forecast displacements relative to x_T and add x_T back at inference, as described in Sec. 3.1). All methods are trained with Adam. For the siMLPe backbone we use 2000 steps for base training and an 800-step mean fine-tune (MeanFT). Probabilistic heads are trained for 800 steps with the mean frozen, using batch sizes 256/128/64 respectively. For structured Gaussian heads, we initialize the mean from a trained MeanFT model and (when applicable) freeze the mean while optimizing only covariance parameters via NLL. For **SPARC**, we fit the conjugate Bayesian last-layer statistics $\Lambda_{n,t}^{-1}$ in closed form on the training split, compute $\kappa_t(x)$, and tune the uncertainty/calibration hyperparameters per dataset on held-out training windows for the NLL–W95 trade-off. These values are fixed for all reported evaluations. Finally, we compute $\{q_{t,j}\}$ on the calibration split and evaluate calibrated tubes on the evaluation split.

4.2 Baselines

We compare against baseline families evaluated under the same protocol: (i) deterministic point forecasters (siMLPe/MeanFT and recent variants such as AuxTasks [53], Symplectic [6], DeFeeNet [46], and HumanMAC [7], plus graph-based SeSGCN [44]); (ii) sampling/ensemble uncertainty baselines (deep ensembles [26] and multi-head predictors such as GSPS [35]); (iii) calibration baselines (Bridge-CQR: conformalized quantile regression [42]); and (iv) Gaussian uncertainty heads (diagonal DCT-Bridge; separable matrix-normal [11]; and graph/GMRF-style structured covariances [3, 43]); and (v) external stochastic/diffusion forecasters (SkeletonDiffusion [9], BeLFusion [2], TransFusion [48], CoMusion [45], SPARD [57], MotionMap [20], and SLD-HMP [54]). When external methods provide public checkpoints, we initialize from the released weights and fine-tune under our protocol. If checkpoints exist only for a subset of datasets, we initialize from the closest available checkpoint (e.g., Human3.6M/AMASS) and fine-tune. Otherwise, we train the method from scratch.

4.3 Cross-dataset comparison

Because units differ across datasets, Table 1 reports mean ranks (MR; ↓ better) across the cross-dataset benchmark. Bold (underline) indicates best (second-best) among the shown methods (ties included); for coverage, best/second-best is defined by closeness to the 0.95 target. Full per-dataset metrics are reported in the supplement. The rank table compares point accuracy (MPJPE/FDE), density quality (NLL), and calibrated tube efficiency (W95(CP)) on a common scale. SPARC has the leading density-quality and MPJPE+NLL ranks while retaining competitive point accuracy. MeanFT remains a strong deterministic point predictor, but it does not provide calibrated uncertainty. Sampling/generative baselines use multiple hypotheses or stochastic inference, while calibration-only baselines improve coverage without separating epistemic risk from structured aleatoric covariance. SPARC combines competitive point accuracy ($\mathrm{MR}_{\mathrm{MPJPE}} = 4.39$), rank-1 density quality ($\mathrm{MR}_{\mathrm{NLL}} = 1.00$), rank-1 combined MPJPE+NLL trade-off (2.69), and efficient calibrated tubes ($\mathrm{MR}_{W95(CP)} = 2.50$) in one forward pass.

Practical benefits over existing methods. SPARC is not a uniform MPJPE winner, but it has the lowest NLL among shown methods on every dataset/protocol block while retaining competitive MPJPE; the supplementary detailed tables report the full metrics. Its calibrated tubes remain efficient, so coverage is not obtained

Table 1: Mean rank (MR; ↓ better) across datasets. Ranks are computed per dataset and averaged over available datasets. $\mathrm{MR}_{\mathrm{MPJPE+NLL}}$ is the mean of MPJPE and NLL ranks (available for all models).

Model	$\mathrm{MR}_{\mathrm{MPJPE}}\downarrow$	$\mathrm{MR}_{\mathrm{NLL}}\downarrow$	$\mathrm{MR}_{W95(\mathrm{CP})}\downarrow$	$\mathrm{MR}_{\mathrm{MPJPE+NLL}}\downarrow$
siMLPe (base)	7.89	10.67	–	9.28
MeanFT (siMLPe fine-tune)	**2.22**	6.11	–	4.17
AuxTasks (50/50)	9.06	11.78	–	10.42
Symplectic (50/50)	16.22	15.78	–	16.00
DeFeeNet (50/50)	7.61	10.44	–	9.03
GSPS (50/50)	7.67	9.00	5.44	8.33
HumanMAC (50/50)	10.22	10.00	–	10.11
SeSGCN (teacher)	13.39	11.78	–	12.58
Deep ensemble	8.22	5.56	5.39	6.89
Bridge-CQR	4.61	<u>3.00</u>	9.11	<u>3.81</u>
SkeletonDiffusion (CVPR 2025) [9]	15.00	13.67	9.78	14.33
BeLFusion (ICCV 2023) [2]	9.56	8.00	5.89	8.78
TransFusion (RA-L 2024) [48]	12.44	9.78	4.78	11.11
CoMusion (ECCV 2024) [45]	6.06	6.44	3.78	6.25
SPARD (AAAI 2026) [57]	7.11	6.56	<u>3.39</u>	6.83
MotionMap (CVPR 2025) [20]	16.11	16.22	9.44	16.17
SLD-HMP (ECCV 2024) [54]	13.22	15.22	6.50	14.22
SPARC†	<u>4.39</u>	**1.00**	**2.50**	**2.69**

Legend: bold = best; <u>underline</u> = second-best (ties included).

by indiscriminately widening intervals. Beyond average metrics, the supplement shows that κ stratifies error monotonically and supports selective deployment via abstention/fallback on high-κ windows (thresholds tuned on held-out data, with no formal guarantee implied). For robustness, the supplement re-evaluates fixed SPARC checkpoints over five split seeds, re-fits selected uncertainty checkpoints with three fit seeds, and repeats held-out Optuna-TPE HPO with three sampler seeds; coverage and tube width remain stable.

To isolate the contributions of structured covariance, epistemic scaling, and time-coupled κ_t shrinkage, Fig. 2 summarizes an FC-Out component ablation across datasets; the supplement gives the per-dataset results. Fig. 2 shows both the MPJPE/NLL component effects and a compact calibrated tube-efficiency rank; the supplementary tables give CP coverage/width for all FC-Out variants under the same fixed protocol.

4.4 FC-Out ablation: component contributions

Ablation findings (all nine dataset/protocol blocks). Within this family, *hybrid MN-GraphJ* gives the lowest MPJPE+NLL mean rank ($\mathrm{MR}_{\mathrm{MPJPE+NLL}} = 2.47$), followed by uncoupled timewise κ_t (2.53) and *SPARC* (time-coupled κ_t; 3.17) (Fig. 2). The difference reflects the metric being optimized. Hybrid MN-GraphJ favors MPJPE+NLL, whereas the supplementary complete-CP comparison gives SPARC the lower tube-width rank ($\mathrm{MR}_{W95(CP)} = 2.56$ vs. 4.78). We therefore use *SPARC* (time-coupled κ_t) as the default because it gives the best tube-efficiency trade-off for the calibrated deployment recipe, while the full benchmark against all baselines still ranks it first on global $\mathrm{MR}_{\mathrm{MPJPE+NLL}}$ (Table 1). The main

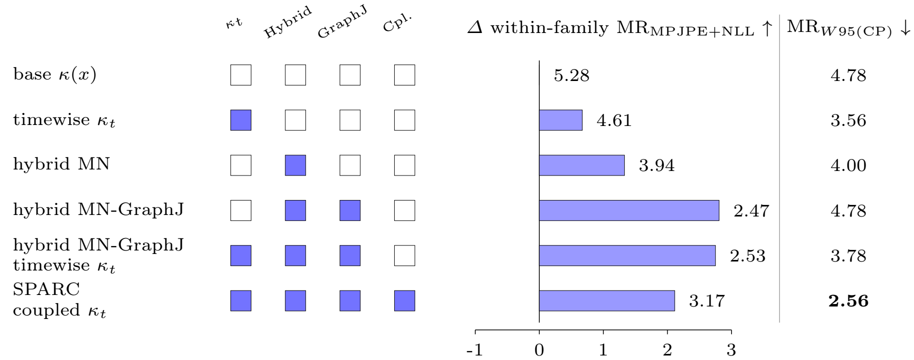

Fig. 2: FC-Out trade-off ablation. Left: active components. Center: within-family MPJPE+NLL-rank gain over base $\kappa(x)$; labels give absolute rank. Right: fixed-protocol $\mathrm{MR}_{W95(\mathrm{CP})}$ (lower is better). The center panel is within-family only and not comparable to Table 1.

gains come from hybrid structured covariance and GraphJ coupling: relative to basic FC-Out $\kappa(x)$ (within-family $\mathrm{MR}_{\mathrm{MPJPE+NLL}} = 5.28$), hybrid MN improves to 3.94, and GraphJ improves further to 2.47. Time-coupled shrinkage should be read as a calibration-efficiency knob rather than a within-family NLL win. Within this ablation it sacrifices some NLL rank ($\mathrm{MR}_{\mathrm{NLL}} = 3.11$ vs. 2.44 for hybrid MN-GraphJ without coupling) for the conformalized default used in the full benchmark (Table 1). The DCT-pinv variant is omitted from Fig. 2 for space; the supplementary tables report it with overall rank 6.00, so it is not part of the recommended recipe.

Compute. All structured heads (including SPARC) require one forward pass at inference. SPARC adds only horizon-wise quadratic forms for κ_t and does not require MC sampling. The closed-form fit of $\Lambda_{n,t}^{-1}$ is an offline pass over training features, and split conformal calibration uses only the 512-window held-out calibration subset from the official evaluation protocol. The supplement reports compute/memory overhead and latency measurements.

Controlled synthetic benchmark (supplementary). The controlled synthetic benchmark in the supplement tests component recovery under a known ground-truth uncertainty decomposition.

4.5 Qualitative uncertainty visualization

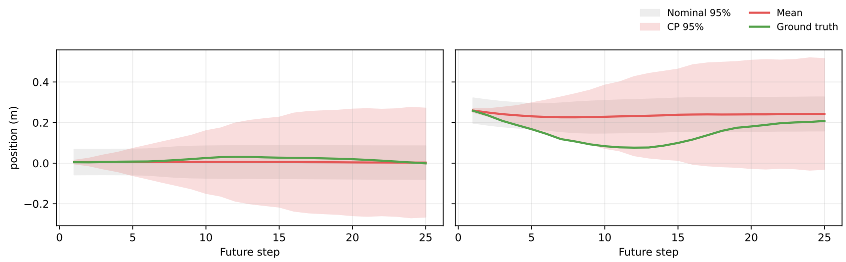

Fig. 3: Calibrated intervals for low vs. high mean epistemic scale $\bar{\kappa}$. Gray: nominal 95% intervals ($\pm 1.960\,\sigma$). Red: split-conformal marginal intervals ($\pm q_{t,j}\,\sigma$), matching the target 95% marginal coverage.

5 Discussion and limitations

Risk-sensitive HRI, human-aware navigation, and safety filters need predictive envelopes for collision avoidance and personal-space constraints [25, 27, 44]. Overconfident tubes can leave insufficient safety margin under novel motions or shift, even when the mean forecast is accurate [23, 40]. SPARC addresses this with a single-pass epistemic risk score that inflates structured spatio-temporal covariance, while split conformal calibration restores target marginal coverage under exchangeability. Practically, κ can trigger conservative fallbacks using thresholds tuned on held-out calibration/validation data.

Limitations. The approach depends on the learned representation; poorly aligned features can weaken κ. Conformal guarantees are marginal, and strong conditional guarantees remain open without additional assumptions [16]. We report split-seed, fit-seed, and repeated-HPO diagnostics; hyperparameters are selected on held-out training windows and fixed before final calibration/evaluation. Richer dependencies may require denser coordinate models at additional computational cost.

6 Conclusion

We introduced *SPARC*, a hybrid uncertainty model combining conjugate Bayesian last-layer scaling, structured Gaussian aleatoric covariance, and conformal calibration. It provides interpretable epistemic scaling, efficient inference, calibrated 95% marginal prediction tubes, and low NLL across diverse motion forecasting datasets. The full ablation supports *SPARC* (hybrid MN-GraphJ with time-coupled κ_t shrink) as the default, while FC-Out and DCT-pinv are analysis baselines.

Acknowledgments This work is funded by the federal state of Lower Saxony, Germany, as part of the project 76251-1337/2022 Kognitiv und Empathisch Intelligente Kobots – "KEIKO" of the SPRUNG funding instrument. We thank the anonymous reviewers for their constructive feedback.

References

1. Angelopoulos, A.N., Bates, S.: Conformal prediction: A gentle introduction. Foundations and Trends® in Machine Learning (2023). https://doi.org/10.1561/2200000101
2. Barquero, G., Escalera, S., Palmero, C.: BeLFusion: Latent diffusion for behavior-driven human motion prediction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 2317–2327 (2023). https://doi.org/10.48550/arXiv.2211.14304, https://arxiv.org/abs/2211.14304
3. Besag, J.: Spatial interaction and the statistical analysis of lattice systems. Journal of the Royal Statistical Society: Series B (Methodological) **36**(2), 192–225 (1974). https://doi.org/10.1111/j.2517-6161.1974.tb00999.x, https://doi.org/10.1111/j.2517-6161.1974.tb00999.x
4. Boström, H., Johansson, U.: Mondrian conformal regressors. In: Proceedings of the Ninth Symposium on Conformal and Probabilistic Prediction and Applications. Proceedings of Machine Learning Research, vol. 128, pp. 114–133. PMLR (2020), https://proceedings.mlr.press/v128/bostrom20a.html
5. Carnegie Mellon University: CMU graphics lab motion capture database. Online (2026), https://mocap.cs.cmu.edu/, accessed: 2026-02-09
6. Chen, H., Lyu, K., Liu, Z., Yin, Y., Yang, X., Lyu, Y.: Rethinking human motion prediction with symplectic integral. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2134–2143 (2024). https://doi.org/10.1109/CVPR52733.2024.00208, https://openaccess.thecvf.com/content/CVPR2024/html/Chen_Rethinking_Human_Motion_Prediction_with_Symplectic_Integral_CVPR_2024_paper.html
7. Chen, L.H., Zhang, J., Li, Y., Pang, Y., Xia, X., Liu, T.: HumanMAC: Masked motion completion for human motion prediction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 9510–9521 (2023). https://doi.org/10.1109/ICCV51070.2023.00875
8. Cicirelli, G., Marani, R., Romeo, L., García Domínguez, M., Heras, J., Perri, A.G., D'Orazio, T.: The HA4M dataset: Multi-modal monitoring of an assembly task for human action recognition in manufacturing. Scientific Data **9**, 745 (2022). https://doi.org/10.1038/s41597-022-01843-z, https://doi.org/10.1038/s41597-022-01843-z
9. Curreli, C., Muhle, D., Saroha, A., Ye, Z., Marin, R., Cremers, D.: Nonisotropic gaussian diffusion for realistic 3D human motion prediction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1871–1882 (2025). https://doi.org/10.48550/arXiv.2501.06035, https://arxiv.org/abs/2501.06035
10. Dang, L., Nie, Y., Long, C., Zhang, Q., Li, G.: MSR-GCN: Multi-scale residual graph convolution networks for human motion prediction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 11467–11476 (2021). https://doi.org/10.1109/ICCV48922.2021.01127
11. Dawid, A.P.: Some matrix-variate distribution theory: Notational considerations and a Bayesian application. Biometrika **68**(1), 265–274 (1981). https://doi.org/10.1093/biomet/68.1.265
12. Daxberger, E., Kristiadi, A., Immer, A., Eschenhagen, R., Bauer, M., Hennig, P.: Laplace redux — effortless bayesian deep learning. In: Advances in Neural Information Processing Systems (NeurIPS). pp. 20089–20103 (2021), https://proceedings.neurips.cc/paper/2021/file/a7c9585703d275249f30a088cebba0ad-Paper.pdf

13. Diquigiovanni, J., Fontana, M., Vantini, S.: Conformal prediction bands for multivariate functional data (2021). https://doi.org/10.48550/arXiv.2106.01792, https://arxiv.org/abs/2106.01792
14. Fong, E., Holmes, C.: Conformal Bayesian computation (2021). https://doi.org/10.48550/arXiv.2106.06137, https://arxiv.org/abs/2106.06137
15. Gal, Y., Ghahramani, Z.: Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In: Proceedings of the 33rd International Conference on Machine Learning (ICML). Proceedings of Machine Learning Research, vol. 48, pp. 1050–1059. PMLR (2016), https://proceedings.mlr.press/v48/gal16.html
16. Gibbs, I., Cherian, J.J., Candès, E.J.: Conformal prediction with conditional guarantees (2023). https://doi.org/10.48550/arXiv.2305.12616, https://arxiv.org/abs/2305.12616
17. Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. In: Proceedings of the 34th International Conference on Machine Learning (ICML). Proceedings of Machine Learning Research, vol. 70, pp. 1321–1330. PMLR (2017), https://proceedings.mlr.press/v70/guo17a.html
18. Guo, W., Du, Y., Shen, X., Lepetit, V., Alameda-Pineda, X., Moreno-Noguer, F.: Back to MLP: A simple baseline for human motion prediction. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) (2023). https://doi.org/10.48550/arXiv.2207.01567, https://openaccess.thecvf.com/content/WACV2023/html/Guo_Back_to_MLP_A_Simple_Baseline_for_Human_Motion_Prediction_WACV_2023_paper.html
19. Harvey, F.G., Yurick, M., Nowrouzezahrai, D., Pal, C.J.: Robust motion inbetweening. ACM Transactions on Graphics **39**(4), 60 (2020). https://doi.org/10.1145/3386569.3392480, https://doi.org/10.1145/3386569.3392480
20. Hosseininejad, R., Shukla, M., Saadatnejad, S., Salzmann, M., Alahi, A.: MotionMap: Representing multimodality in human pose forecasting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2025). https://doi.org/10.48550/arXiv.2412.18883, https://arxiv.org/abs/2412.18883
21. Huang, H., He, S., Miao, F.: CUQDS: Conformal uncertainty quantification under distribution shift for trajectory prediction (2024). https://doi.org/10.48550/arXiv.2406.12100, https://arxiv.org/abs/2406.12100
22. Ionescu, C., Papava, D., Olaru, V., Sminchisescu, C.: Human3.6M: Large scale datasets and predictive methods for 3D human sensing in natural environments. IEEE Transactions on Pattern Analysis and Machine Intelligence **36**(7), 1325–1339 (2014). https://doi.org/10.1109/TPAMI.2013.248
23. Kendall, A., Gal, Y.: What uncertainties do we need in Bayesian deep learning for computer vision? In: Advances in Neural Information Processing Systems (NeurIPS). vol. 30 (2017), https://proceedings.neurips.cc/paper/2017/hash/2650d6089a6d640c5e85b2b88265dc2b-Abstract.html
24. Kim, D., Choi, H., Rasool, M.J.A., Oh, G.: CLAPS: Aleatoric-epistemic scaling via last-layer laplace for conformal regression (2026). https://doi.org/10.48550/arXiv.2512.01384, https://arxiv.org/abs/2512.01384v4, version 4, updated 2026-05-05
25. Kruse, T., Pandey, A.K., Alami, R., Kirsch, A.C.: Human-aware robot navigation: A survey. Robotics and Autonomous Systems **61**(12), 1726–1743 (2013). https://doi.org/10.1016/j.robot.2013.05.007
26. Lakshminarayanan, B., Pritzel, A., Blundell, C.: Simple and scalable predictive uncertainty estimation using deep ensembles. In: Advances in Neural Information

Processing Systems (NeurIPS) (2017). https://doi.org/10.48550/arXiv.1612.01474, https://arxiv.org/abs/1612.01474
27. Lasota, P.A., Fong, T., Shah, J.A.: A survey of methods for safe human-robot interaction. Foundations and Trends® in Robotics **5**(4), 261–349 (2017). https://doi.org/10.1561/2300000052
28. Lee, J., Xu, C., Xie, Y.: Flow-based conformal prediction for multi-dimensional time series. In: The Fourteenth International Conference on Learning Representations (ICLR) (2026), https://openreview.net/forum?id=Uv3efQiPBZ
29. Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R.J., Wasserman, L.: Distribution-free predictive inference for regression. Journal of the American Statistical Association **113**(523), 1094–1111 (2018). https://doi.org/10.1080/01621459.2017.1307116, https://doi.org/10.1080/01621459.2017.1307116
30. Lei, J., Rinaldo, A., Wasserman, L.: A conformal prediction approach to explore functional data (2013). https://doi.org/10.48550/arXiv.1302.6452, https://arxiv.org/abs/1302.6452
31. Lindemann, L., Cleaveland, M., Shim, G., Pappas, G.J.: Safe planning in dynamic environments using conformal prediction (2022). https://doi.org/10.48550/arXiv.2210.10254, https://arxiv.org/abs/2210.10254
32. Ma, Y., Li, F.W.B., Liang, X.: Uncertainty-aware calibrated 3D human motion forecasting with latent conformal prediction. Pattern Recognition **175**, 113144 (2026). https://doi.org/10.1016/j.patcog.2026.113144, https://doi.org/10.1016/j.patcog.2026.113144
33. MacKay, D.J.C.: Bayesian interpolation. Neural Computation **4**(3), 415–447 (1992). https://doi.org/10.1162/neco.1992.4.3.415
34. Mahmood, N., Ghorbani, N., Troje, N.F., Pons-Moll, G., Black, M.J.: AMASS: Archive of motion capture as surface shapes. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 5442–5451 (2019). https://doi.org/10.1109/ICCV.2019.00554, https://doi.org/10.1109/ICCV.2019.00554
35. Mao, W., Liu, M., Salzmann, M.: Generating smooth pose sequences for diverse human motion prediction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 13289–13298 (2021). https://doi.org/10.1109/ICCV48922.2021.01306
36. von Marcard, T., Henschel, R., Black, M.J., Rosenhahn, B., Pons-Moll, G.: Recovering accurate 3D human pose in the wild using IMUs and a moving camera. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 614–631 (2018). https://doi.org/10.1007/978-3-030-01249-6_37, https://doi.org/10.1007/978-3-030-01249-6_37
37. Martinez, J., Black, M.J., Romero, J.: On human motion prediction using recurrent neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017). https://doi.org/10.48550/arXiv.1705.02445, https://arxiv.org/abs/1705.02445
38. Maurice, P., Malaisé, A., Ivaldi, S., Rochel, O., Amiot, C., Paris, N., Richard, G.J., Fritzsche, L.: Andydata-lab-oneperson. Zenodo (2018). https://doi.org/10.5281/zenodo.1471975, https://doi.org/10.5281/zenodo.1471975
39. Nag, S., Ghosh, U., Ta, C.K., Bose, S., Li, J., Roy-Chowdhury, A.K.: Conformal prediction and MLLM-aided uncertainty quantification in scene graph generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 11676–11686 (2025), https://openaccess.thecvf.com/content/CVPR2025/html/Nag_Conformal_Prediction_and_MLLM_aided_Uncertainty_Quantification_in_Scene_Graph_CVPR_2025_paper.html

40. Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Nowozin, S., Dillon, J., Lakshminarayanan, B., Snoek, J.: Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. In: Advances in Neural Information Processing Systems (NeurIPS). vol. 32. Curran Associates, Inc. (2019), https://proceedings.neurips.cc/paper_files/paper/2019/file/8558cb408c1d76621371888657d2eb1d-Paper.pdf
41. Riquelme, C., Tucker, G., Snoek, J.: Deep Bayesian bandits showdown: An empirical comparison of bayesian deep networks for thompson sampling. In: International Conference on Learning Representations (ICLR) (2018). https://doi.org/10.48550/arXiv.1802.09127, https://arxiv.org/abs/1802.09127
42. Romano, Y., Patterson, E., Candès, E.: Conformalized quantile regression. In: Advances in Neural Information Processing Systems (NeurIPS) (2019). https://doi.org/10.48550/arXiv.1905.03222, https://arxiv.org/abs/1905.03222
43. Rue, H., Held, L.: Gaussian Markov Random Fields: Theory and Applications. Chapman and Hall/CRC (2005). https://doi.org/10.1201/9780203492024, https://doi.org/10.1201/9780203492024
44. Sampieri, P., Alameda-Pineda, X., Solhugot, J., Lepetit, V., Moreno-Noguer, F.: Pose forecasting in industrial human-robot collaboration. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 51–69 (2022). https://doi.org/10.1007/978-3-031-19839-7_4, https://doi.org/10.1007/978-3-031-19839-7_4
45. Sun, J., Chowdhary, G.: CoMusion: Towards consistent stochastic human motion prediction via motion diffusion. In: Proceedings of the European Conference on Computer Vision (ECCV) (2024). https://doi.org/10.48550/arXiv.2305.12554, https://arxiv.org/abs/2305.12554
46. Sun, X., Sun, H., Li, B., Wei, D., Li, W., Lu, J.: DeFeeNet: Consecutive 3D human motion prediction with deviation feedback. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5527–5536 (2023). https://doi.org/10.1109/CVPR52729.2023.00535
47. Thakur, S., Lorsung, C., Yacoby, Y., Doshi-Velez, F., Pan, W.: Uncertainty-aware (UNA) bases for deep bayesian regression using multi-headed auxiliary networks (2020). https://doi.org/10.48550/arXiv.2006.11695, https://arxiv.org/abs/2006.11695
48. Tian, S., Zheng, M., Liang, X.: TransFusion: A practical and effective transformer-based diffusion model for 3D human motion prediction. IEEE Robotics and Automation Letters **9**(7), 6232–6239 (2024). https://doi.org/10.1109/LRA.2024.3401116
49. Vovk, V.: Conditional validity of inductive conformal predictors. In: Proceedings of the Asian Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 25, pp. 475–490. PMLR (2012), https://proceedings.mlr.press/v25/vovk12.html
50. Vovk, V., Gammerman, A., Shafer, G.: Algorithmic Learning in a Random World. Springer (2005). https://doi.org/10.1007/b106715, https://doi.org/10.1007/b106715
51. Watson, J., Lin, J.A., Klink, P., Pajarinen, J., Peters, J.: Latent derivative bayesian last layer networks. In: Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS). Proceedings of Machine Learning Research, vol. 130, pp. 1198–1206. PMLR (2021), https://proceedings.mlr.press/v130/watson21a.html

52. Xu, C., Xie, Y.: Sequential predictive conformal inference for time series. In: Proceedings of the 40th International Conference on Machine Learning (ICML). Proceedings of Machine Learning Research, vol. 202, pp. 38707–38727. PMLR (2023), https://proceedings.mlr.press/v202/xu23r.html
53. Xu, C., Tan, R.T., Tan, Y., Chen, S., Wang, X., Wang, Y.: Auxiliary tasks benefit 3D skeleton-based human motion prediction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 9475–9486 (2023). https://doi.org/10.1109/ICCV51070.2023.00872
54. Xu, G., Tao, J., Li, W., Duan, L.: Learning semantic latent directions for accurate and controllable human motion prediction (2024). https://doi.org/10.48550/arXiv.2407.11494, https://arxiv.org/abs/2407.11494
55. Yuan, Y., Kitani, K.M.: DLow: Diversifying latent flows for diverse human motion prediction. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 346–364 (2020). https://doi.org/10.1007/978-3-030-58545-7_20
56. Zhang, B., Li, S., Bastani, O.: Conformal structured prediction. In: The Thirteenth International Conference on Learning Representations (ICLR). pp. 1–21 (2025), https://openreview.net/forum?id=2ATD8a8P3C
57. Zhang, Y., Han, B., Li, X., Pang, W., Giunchiglia, F., Feng, X., Guan, R.: SPARD: Single-step inference with adaptive sampling in residual diffusion for human motion prediction. In: Proceedings of the AAAI Conference on Artificial Intelligence (2026), https://github.com/ZYM-JLU/SPARD, accepted
58. Zhang, Y., Candès, E.J.: Posterior conformal prediction (2024). https://doi.org/10.48550/arXiv.2409.19712, https://arxiv.org/abs/2409.19712