You Only Flow Once: Calibrated and Real-Time Radar Pose Estimation with Multi-Hypothesis Normalizing Flows

Jonas Leo Mueller[1,2,3] Sebastian Hoefler[1,2] Dario Zanca[1,2]
Naga Venkata Sai Jitin Jami[1,2] Thomas Altstidl[1,2,3] Bjoern M. Eskofier[1,2,3,4]

[1]Department Artificial Intelligence in Biomedical Engineering (AIBE), Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany
[2]Munich Center for Machine Learning (MCML), Munich, Germany [3]Chair of AI-supported Therapy Decisions, LMU München, Germany
[4]Institute of AI for Health, Helmholtz Zentrum München, Germany jonas.leo.mueller@fau.de

Abstract

Sparse and noisy millimeter-wave radar point cloud observations often correspond to multiple plausible human poses, making deterministic pose estimation fundamentally ill-posed. Yet existing radar methods remain deterministic, collapsing this ambiguity into a single estimate. Diffusion-based alternatives can model multi-hypothesis distributions but require costly sequential denoising for each distribution sample and lack calibrated uncertainty. We propose Multi-Hypothesis Normalizing Flow Pose Generator (MH-NFPG), which models pose distributions from radar point clouds using a conditional normalizing flow. Specifically, we combine a spatiotemporal transformer backbone with a normalizing flow that transforms a Laplace base distribution into an expressive posterior, generated in parallel through a single forward pass. Leveraging this efficiency, we outperform diffusion-based alternatives in calibration across three radar benchmarks (MM-Fi, mmRadPose, mRI), improve pose accuracy on two, and match it on the third, while achieving over 20× faster inference for applications and reducing calibration error by up to 85%. We find that calibration degrades substantially for diffusion models, whereas our flow-based approach maintains reliable coverage, also in cross-environment settings. These results demonstrate normalizing flows as a practical alternative to diffusion models for real-time, uncertainty-aware radar pose estimation. Our code will be made publicly available.

1. Introduction

Uncertainty quantification underpins trustworthy machine learning, from medical decisions [5] to computer vision [1, 21]. It is no less vital for radar human pose estimation, a privacy-preserving, lighting-robust modality [7, 12, 18, 23, 44, 49] whose safety-critical use makes confident errors costly. An overconfident pose can corrupt a clinician's assessment in rehabilitation monitoring [32], or drive a robot into the worker it should avoid in human-robot collaboration [39]. How strongly a body part scatters the incident waves is set by its radar cross section (RCS) σ_{rcs}, which is small for distal parts such as the hands and feet, so they produce faint echoes, and the received power falls off with range as R^{-4} [37]. Physically, the RCS is the effective reflecting area a body part presents to the radar, set by its size, shape, and orientation relative to the sensor, so the small, curved, fast-moving extremities present the least area and are exactly the joints whose returns are weakest. For real-time use, a Constant False Alarm Rate (CFAR) detector thresholds the dense raw radar tensor into a sparse point cloud [11, 37] and can drop these weak returns, while multipath and scattering can further influence the points that remain [3].

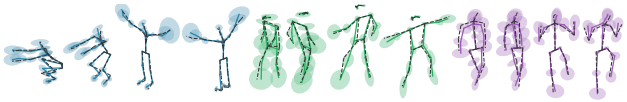

Figure 1. **MH-NFPG** produces calibrated 3D pose distributions from radar point clouds in a single forward pass, with predictions (colored), ground truth (dashed), and 50% confidence ellipsoids on mmRadPose (blue), mRI (green), and MM-Fi (purple).

Recovering 3D pose from such a point cloud is therefore an ill-posed inverse problem that we learn to solve. The forward map from pose to measurement is lossy and many-to-one, so one point cloud matches many plausible poses. The resulting uncertainty is shaped by the sensor physics, and changes with the recording environment, the subject, and the radar aspect angle. A deployable estimator should therefore report not a single guess but a calibrated distribution over the plausible poses, one that stays calibrated even

when the test environment was not seen during training.

Despite these sensor characteristics, radar pose estimation from point clouds is still almost entirely deterministic. Standard regressors collapse the ambiguous signal into one compromise estimate [2, 3, 7, 11, 30, 40, 48] and discard uncertainty, and the uncertainty quantification work that exists operates on full five-dimensional complex radar tensors rather than the point cloud [31], a far heavier input to transmit on device, and requires recalibration after training. Diffusion models can in principle represent ambiguity by drawing many hypotheses [13, 14, 19], yet producing N hypotheses needs $N \times M$ sequential denoising steps, which is too slow for real-time use. Recent radar diffusion work [12] avoids this cost with deterministic inference, but at the cost of stochasticity and uncertainty quantification (Section 2).

No prior radar method provides both at once, a calibrated multi-hypothesis posterior over the ambiguous point-cloud-to-pose mapping and real-time inference. Normalizing flows (NFs) can close this gap. Their bijective structure gives an exact likelihood for training, and batching can draw every hypothesis in one forward pass. NFs have been applied to RGB pose estimation [16, 46], where the NF is conditioned on structured 2D keypoints. Radar offers no comparable keypoint detector, so we instead derive the conditioning signal from a prior learned on the radar itself, namely a heteroscedastic multivariate Gaussian [31]. Our model, the Multi-Hypothesis Normalizing Flow Pose Generator (MH-NFPG), combines this prior with a conditional NF. A permutation-invariant spatiotemporal transformer backbone ingests the unordered, variable-size point set and parameterizes the prior, whose full covariance captures the anisotropic, input-dependent uncertainty of radar detections. A temporal graph-convolutional conditioning branch then propagates prior motion context across frames, so low-radar-cross-section joints such as hands and feet stay constrained on the frames where they go undetected. The conditional Real NVP NF [9] starts from a Laplace base distribution, whose heavier tails than a Gaussian accommodate multipath and ghost-target outliers, and transforms it through affine coupling layers into the full pose distribution. Our contributions are as follows.

- **Intrinsic calibration.** Exact-likelihood training yields empirically calibrated distributions with no post-hoc recalibration, on all three datasets, with 34% to 85% lower expected calibration error than the diffusion baselines, and holds even when an entire environment is held out.
- **Real-time multi-hypothesis inference.** A full posterior is produced in one forward pass at 80 frames per second, 14–21$\times$ faster than the stochastic diffusion baselines.
- **No accuracy trade-off.** Across all three datasets our accuracy matches or improves on the strongest baseline on MPJPE and PA-MPJPE.

Our code will be made publicly available.

2. Related Work

2.1. Multi-Hypothesis Pose Estimation

Multi-hypothesis modelling (MHM) was introduced for inherently ambiguous prediction tasks, where one input admits several valid outputs, by predicting a set of N candidates instead of a single answer [38]. Applied to human pose estimation it represents the conditional distribution $p(\mathbf{y}|\mathbf{x})$ over poses $\mathbf{y} \in \mathbb{R}^{K\times 3}$ with K keypoints. It has been used mostly for monocular RGB pose estimation, where lifting ordered 2D keypoints to 3D leaves an irreducible depth ambiguity [53] that a deterministic model cannot resolve and that motivated multi-hypothesis RGB methods [13, 14, 19, 46]. Early methods picked among hypotheses with best-of-N losses and the minimum Mean Per Joint Position Error [38, 46], which needs an oracle at inference and is impractical [36, 42], so recent methods model the posterior directly and aggregate its samples into a final estimate [13, 14, 19, 20, 25, 28]. These RGB methods fall into three classes, namely Gaussian models [24], diffusion models [19], and normalizing flows [46].

Radar point clouds carry ambiguity that arises from sparsity, dropped detections, and multipath rather than depth, yet MHMs for radar remain largely under-explored. The few diffusion-based attempts, by Li *et al*. [27, 28] and Fan *et al*. [12], can sample multiple pose hypotheses but report only point predictions and quantify no uncertainty.

2.2. Uncertainty for Radar Pose Estimation

Calibration has received almost no attention in radar pose estimation. Chiang *et al*. [8] reduce uncertainty during training but do not quantify uncertainty, and only Mueller *et al*. [31] report calibrated uncertainty, fitting a heteroscedastic multivariate Gaussian on raw complex radar tensors and recalibrating it post-hoc on a held-out set. We compare against the radar Gaussian of Mueller *et al*., the radar diffusion model mmDiff of Fan *et al*. [12] evaluated in its stochastic multi-hypothesis form, and the canonical RGB diffusion model DiffPose [19] adapted to radar.

2.3. Normalizing Flows

NFs learn invertible maps between a simple base distribution and a complex data distribution [35, 51], which yields exact likelihoods and efficient sampling. Conditional NFs have been used for probabilistic regression [36, 47] and for pose estimation, including multi-hypothesis 2D-to-3D lifting [16, 46], weakly supervised reconstruction [50], and anomaly detection [17]. Unlike continuous NFs that integrate an ODE over many steps [29], discrete NFs use a fixed set of layers [9]. To our knowledge, we are the first to apply conditional NFs to calibrated radar point-cloud pose estimation.

3. Methods

We propose a two-phase training pipeline (Figure 2) that combines a multivariate Gaussian prior with a conditional NF to produce calibrated posteriors over 3D poses. In the first phase, we train the prior distribution, a spatiotemporal transformer that encodes an input sequence of radar point clouds into a latent distribution, from which samples are decoded into 3D pose hypotheses by a graph convolutional network (GCN) reflecting the kinematic structure of the human body.

We then use a conditional bijective NF in the second phase to predict the final output distribution. The NF starts from a standard Laplace distribution and is conditioned on two complementary signals, the prior latent features providing per-frame spatial context, and a temporal context vector from a second GCN encoding the past n_t prior predictions to enforce temporal consistency.

3.1. Multi-Hypothesis Normalizing Flow Pose Generator

Spatiotemporal Transformer Backbone A radar sequence consists of T frames $\mathcal{P} = \{\mathbf{P}_1, \ldots, \mathbf{P}_T\}$, where each frame $\mathbf{P}_t = \{\mathbf{x}_{t,1}, \ldots, \mathbf{x}_{t,N_t}\}$ is an unordered point cloud whose size N_t varies across frames as CFAR retains a different number of detections. Each point is first embedded into a d-dimensional feature space via an MLP $f_{\text{emb}} : \mathbb{R}^{d_{\text{in}}} \to \mathbb{R}^d$. A learnable temporal embedding $\mathbf{e}_t$ is then added to the embedded points of each frame t, yielding $\mathbf{P}_t + \mathbf{e}_t$. Finally, all T frames are concatenated into a single point set $\mathbf{X} \in \mathbb{R}^{(T \cdot N_t) \times d}$, which serves as input to the transformer. Our backbone is a transformer encoder [45] $\Psi(\cdot)$ that processes $\mathbf{X}$ using multi-head self-attention (MHSA). Since we do not use positional encodings for individual points, the spatial ordering within each frame is irrelevant, exploiting the permutation-equivariance of MHSA. The temporal embeddings, however, enable the model to distinguish points from different timesteps, preserving temporal structure. A learnable class token aggregates the sequence into a fixed-dimensional representation $\mathbf{F} \in \mathbb{R}^{d_F}$, achieving permutation invariance over the input point cloud.

Prior Distribution Radar signals are inherently ambiguous due to sensor noise, multipath propagation and scattering [37]. We model this aleatoric uncertainty with a heteroscedastic multivariate Gaussian (MG) prior distribution [8, 15, 31], which the NF uses as a conditioning signal to predict a more expressive posterior (Figure 2). The MG prior captures inter-keypoint correlations through the multivariate Gaussian likelihood, which optimizes the full covariance structure of the decoded pose hypotheses.

To achieve this, the backbone representation $\mathbf{F}$ is fed into two branches predicting mean $\boldsymbol{\mu}_l$ and log-variance $\log \boldsymbol{\sigma}_l^2$ for a latent Gaussian, regularized with a KL divergence term

$$\mathcal{L}_{\text{KL}} = -\frac{1}{2}\mathbb{E}\left[\sum_{j=1}^{d_l}\left(1 + \log \boldsymbol{\sigma}_{l,j}^2 - \boldsymbol{\mu}_{l,j}^2 - \boldsymbol{\sigma}_{l,j}^2\right)\right]. \quad (1)$$

We sample N hypotheses and decode each with a spectral GCN [12, 14] to predict the pose hypothesis distribution $\mathbf{E} \in \mathbb{R}^{3K \times N}$. The empirical mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$ of the decoded hypotheses are used in the multivariate negative log likelihood (NLL) loss. We use a variance regularizer $\mathcal{R} = \frac{\text{tr}(\boldsymbol{\Sigma})}{d_{\text{pose}}}$ scaled by a hyperparameter γ penalizing the mean diagonal variance of the covariance matrix to stabilize training.

$$\mathcal{L}_{\text{NLL}}^{\text{prior}} = \mathbb{E}\left[\frac{1}{2}\left(\log|\boldsymbol{\Sigma}| + (\mathbf{y} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{y} - \boldsymbol{\mu})\right) + \gamma\,\mathcal{R}\right]. \quad (2)$$

The final training loss for the prior is

$$\mathcal{L}_{\text{backbone}} = \mathcal{L}_{\text{NLL}}^{\text{prior}} + \lambda_{\text{KL}}\mathcal{L}_{\text{KL}}. \quad (3)$$

We additionally test Laplace prior distribution formulations (see Appendix A) but achieve superior performance with the MG formulation. We tune hyperparameters γ and λ_{KL} with targeted grid search on a validation set (see Appendix B). The final prior MG is then evaluated on its own in subsequent experiments and its features and temporal predictions are used for conditioning the NF.

3.1.1. Conditional Normalizing Flow

While Gaussians provide tractable uncertainty, their parametric assumptions may not capture the true uncertainty structure [16, 46]. We therefore use a conditional NF to transform a Laplace base distribution [26] into a more expressive posterior conditioned on our prior. Formally, the NF models pose $\mathbf{y} \in \mathbb{R}^{d_{\text{pose}}}$, where $d_{\text{pose}} = 3K$ represents the flattened 3D coordinates of K keypoints. The NF transforms samples from a base distribution of independent Laplace$(0,\ 1/\sqrt{2})$ components, chosen such that each marginal has unit variance, into an expressive pose distribution through a series of invertible transformations. We ablate further base distribution assumptions to validate this choice in Section 4.4.

Affine Coupling Layers Following prior multi-hypothesis RGB pose estimation work [26, 46], we use Real NVP affine coupling layers [9], which allow efficient parallel sampling. The NF consists of L such layers. Each layer ℓ applies a bijective transformation $f_\ell : \mathbb{R}^{d_{\text{pose}}} \to \mathbb{R}^{d_{\text{pose}}}$ using a binary mask $\mathbf{m}_\ell \in \{0,1\}^{d_{\text{pose}}}$ that alternates between even and odd dimensions across layers. The forward transformation is defined as

$$\mathbf{h}_\ell = \mathbf{m}_\ell \odot \mathbf{h}_{\ell-1} + (1 - \mathbf{m}_\ell) \odot (\mathbf{h}_{\ell-1} \odot \exp(\mathbf{s}_\ell) + \mathbf{t}_\ell), \quad (4)$$

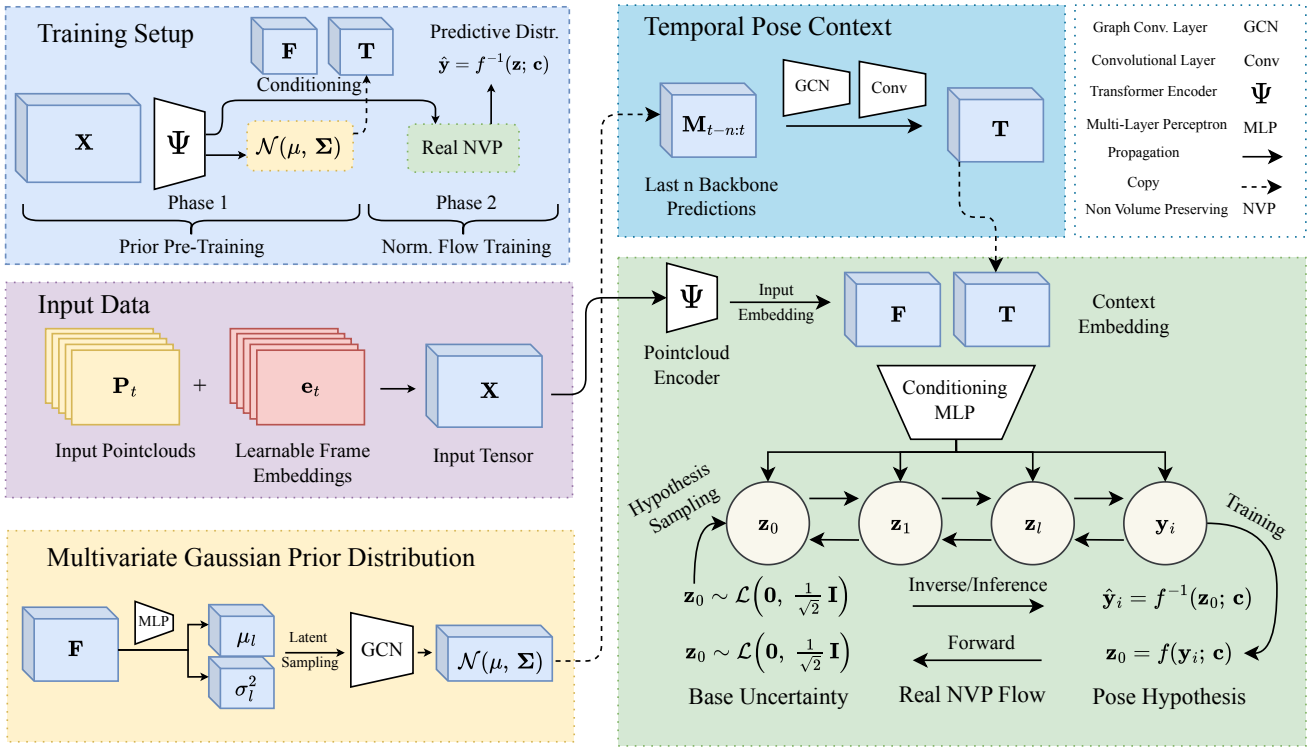

Figure 2. Overview of MH-NFPG. Phase 1 trains a transformer encoder on radar point cloud sequences to parameterize a heteroscedastic Gaussian prior over 3D poses. Phase 2 freezes the backbone and prior and trains a conditional Real NVP NF that transforms a Laplace base distribution into expressive pose hypotheses, conditioned on backbone features and temporal predictions from Phase 1. At inference, hypotheses are generated by sampling from the base distribution and applying the inverse NF.

where $\mathbf{h}_0 = \mathbf{y}$ is the input pose and $\odot$ denotes element-wise multiplication. The scale $\mathbf{s}_\ell \in \mathbb{R}^{d_{\text{pose}}}$ and translation $\mathbf{t}_\ell \in \mathbb{R}^{d_{\text{pose}}}$ parameters are computed by a neural network g_ℓ that takes both the masked input and the conditioning vector:

$$[\tilde{\mathbf{s}}_\ell, \mathbf{t}_\ell] = g_\ell([\mathbf{m}_\ell \odot \mathbf{h}_{\ell-1};\ \mathbf{c}]), \quad \mathbf{s}_\ell = \tanh(\tilde{\mathbf{s}}_\ell). \tag{5}$$

Each g_ℓ is a three-layer MLP with dropout after each layer.

Conditioning Mechanism We condition the model on two components, namely the backbone features $\mathbf{F} \in \mathbb{R}^{d_F}$ and the past temporal pose predictions $\mathbf{M} \in \mathbb{R}^{n_t \times K \times 3}$ of the prior distribution, where d_F is the backbone feature dimension and n_t the number of past frames. These temporal predictions supply motion context that compensates for frames where low-cross-section joints such as hands and feet drop out of the radar point cloud. To capture biomechanical dependencies between skeletal keypoints, we employ a GCN for conditioning. Backbone features $\mathbf{F}$ are first projected to per-keypoint embeddings $\mathbf{P} \in \mathbb{R}^{K \times d_g}$ via a two-layer MLP, where d_g is the per-keypoint embedding dimension. These are then processed through a Chebyshev graph convolution on the skeleton adjacency matrix to obtain a global embedding $\mathbf{G} \in \mathbb{R}^{K \times d_h}$, where d_h is the GCN hidden dimension [12, 14]. Similarly, each of the n_t past pose frames in $\mathbf{M}$ is independently processed through a graph convolution with shared weights, producing per-frame embeddings in $\mathbb{R}^{K \times d_h}$. The resulting temporal sequence is aggregated via 1D convolution and max-pooling to form $\mathbf{T} \in \mathbb{R}^{K \times d_h}$. The global and temporal embeddings are summed element-wise, flattened to $\mathbb{R}^{K \cdot d_h}$, and projected through an MLP to produce the final conditioning vector $\mathbf{c} \in \mathbb{R}^{d_c}$, where d_c is the conditioning dimension. In each affine coupling layer of the NF, $\mathbf{c}$ is concatenated with the masked input before passing through the scale-translation network.

NF Composition The complete NF ϕ composes all L affine coupling layers:

$$\mathbf{z} = \phi(\mathbf{y}; \mathbf{c}) = f_L \circ f_{L-1} \circ \cdots \circ f_1(\mathbf{y}; \mathbf{c}). \tag{6}$$

The inverse transformation $\mathbf{y} = \phi^{-1}(\mathbf{z}; \mathbf{c})$ maps from the base space $\mathbf{z} \in \mathbb{R}^{d_{\text{pose}}}$ back to pose space by applying the inverse of each coupling layer in reverse order. The log-likelihood of a pose $\mathbf{y}$ under the model is obtained via the change-of-variables formula [4, 9]

$$\log p(\mathbf{y}|\mathbf{c}) = \log p_0(\mathbf{z}) + \log \left| \det \frac{\partial \mathbf{z}}{\partial \mathbf{y}} \right|, \quad \text{where } \mathbf{z} = \phi(\mathbf{y}; \mathbf{c}). \tag{7}$$

These components together form our **M**ulti-**H**ypothesis **N**ormalizing **F**low **P**ose **G**enerator model (MH-NFPG), which we report as MG-Prior + NFPG in our tables to make the conditioning prior explicit alongside the baselines.

NF Training Training follows a two-phase design. We first train the prior distribution model, then freeze it, and train the NF on its features and temporal predictions. In the forward direction, ground-truth poses $\mathbf{y}^*$ are mapped to the base distribution via $\mathbf{z} = \phi(\mathbf{y}^*; \mathbf{c})$, and we minimize the negative log-likelihood

$$\mathcal{L}_{\text{NLL}}^{\text{NF}} = -\mathbb{E}\left[\log p_0(\mathbf{z}) + \log\left|\det \frac{\partial \mathbf{z}}{\partial \mathbf{y}^*}\right|\right]. \tag{8}$$

The joint log-density of the base distribution p_0 with scale $b = 1/\sqrt{2}$ factorizes as

$$\log p_0(\mathbf{z}) = d_{\text{pose}} \log \frac{1}{2b} - \frac{1}{b}\|\mathbf{z}\|_1. \tag{9}$$

Since ϕ is a composition of L affine coupling layers, each with a triangular Jacobian, the log-determinant decomposes as

$$\log\left|\det \frac{\partial \mathbf{z}}{\partial \mathbf{y}^*}\right| = \sum_{\ell=1}^{L} \sum_{q \in \mathcal{U}_\ell} s_{\ell,q}, \tag{10}$$

where $s_{\ell,q}$ is the log-scale output of layer ℓ for dimension q and $\mathcal{U}_\ell = \{q : m_{\ell,q} = 0\}$ denotes the set of unmasked dimensions in layer ℓ. For inference, we sample $\mathbf{z}^{(s)} \sim p_0(\mathbf{z})$ and generate pose predictions via the inverse NF $\hat{\mathbf{y}}^{(s)} = \phi^{-1}(\mathbf{z}^{(s)}; \mathbf{c})$.

3.2. Datasets

We focus on compact frequency-modulated continuous-wave (FMCW) radars with 3Tx/4Rx antenna arrays, which enable fast on-device CFAR-based point cloud extraction [3, 11, 22]. While large-aperture arrays [7, 12] achieve higher angular resolution, their substantially larger data volume [22] prohibits real-time on-device processing, practical data transfer and hardware costs for applications. In line with prior radar pose estimation work [3, 11, 12, 18, 23, 41], we evaluate on three publicly available single-person datasets. To ensure point cloud density, we concatenate the points of the last five frames into one frame for each dataset and use temporal sequences of five frames for our models as inputs [12].

MM-Fi Dataset [49]: This dataset comprises over 320k frames from 40 subjects (11 female, 29 male) performing 27 daily and rehabilitation activities across four environments. Ground-truth 17-keypoint poses are obtained via HRNet-w48 2D detection from two infrared cameras, triangulation, and optimization-based refinement. We follow the splits of Fan *et al*. [12] but use all available activities to increase dataset size and generalization.

mmRadPose Dataset [11]: This dataset comprises recordings of 12 participants performing 11 rehabilitation exercises with 26-keypoint annotations. Data were captured from three distinct radar aspect angles (0°, 45°, and 90° relative to the subject's frontal plane), providing viewpoint diversity for evaluating pose estimation robustness. The held-out test set contains one male and one female participant, evaluated across all exercises. Ground-truth poses are recorded with an optical motion capture system, yielding the highest-fidelity annotations among the three datasets.

mRI Dataset [3]: This dataset includes 20 participants performing 12 distinct activities with 17-keypoint annotations. Ten activities consist of structured rehabilitation exercises, while the remaining two capture free-form stretching and relaxation movements, as well as straight-line walking. This broader activity set enables assessment of generalization across varying motion patterns. We use all exercises and 16 participants for training and 4 for testing. Ground-truth poses are obtained using 2D keypoint detection with HRNet from two RGB cameras, followed by triangulation and optimization-based refinement. The ground truth annotations contain occasional glitches, where poses jump around 2 meters between frames. We filter out those confounding frames for training (1%) and testing (0.9%).

3.3. Evaluation Metrics

We evaluate pose accuracy with Mean Per Joint Position Error (MPJPE), the average Euclidean distance between predicted and ground-truth keypoints, MPJPE $= \frac{1}{K}\sum_{k=1}^{K} \|\hat{\mathbf{y}}_k - \mathbf{y}_k^*\|_2$. We also report PA-MPJPE, which adds Procrustes alignment.

We assess uncertainty directly from the pose hypotheses. For each pose coordinate they form an empirical CDF $\hat{F}$, and evaluating it at the ground truth yields the probability integral transform $u_i = \hat{F}(y_i^*)$, the fraction of hypotheses below the true value, which is uniform on $[0, 1]$ under perfect calibration. The Expected Calibration Error (ECE) [33, 36] measures its deviation from uniformity across C=100 levels $p_j \in [0.01, 0.99]$,

$$\text{ECE} = \sqrt{\frac{1}{C}\sum_{j=1}^{C}\left(\hat{P}(p_j) - p_j\right)^2}, \tag{11}$$

where $\hat{P}(p_j) = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}[u_i \le p_j]$ and N is the number of test samples. We use the L2 variant, which penalizes large deviations more heavily than the original L1 formulation [33]. We also report interval coverage at the 50%, 90%, and 95% levels, the fraction of ground truths inside the central quantile interval $[\hat{F}^{-1}(\frac{1-c}{2}), \hat{F}^{-1}(\frac{1+c}{2})]$ for level c, and sharpness, the average predictive standard deviation,

$$\text{Sharpness} = \frac{1}{NK}\sum_{i=1}^{N}\sum_{k=1}^{K} \sigma_{i,k}. \tag{12}$$

Table 1. Pose estimation accuracy and uncertainty calibration on MM-Fi with standard testing protocols. Position errors in cm. @$X\%$ denotes the proportion of ground-truth joints within the $X\%$ credible interval.

Model	MPJPE↓	PA-MPJPE↓	@50%	@90%	@95%	ECE↓	Sharp.
Random							
PointTransformer [52]	6.392	4.967	–	–	–	–	–
mmDiff (η=0) [12]	6.242	4.791	–	–	–	–	–
mmDiff (η=1) [12]	6.272	4.794	0.253	0.572	0.646	0.123	2.176
MG-Prior (Ours)	6.533	4.937	0.447	0.763	0.823	0.040	2.901
MG-Prior + DiffPose [19]	6.506	4.872	0.354	0.698	0.760	0.071	3.014
MG-Prior + NFPG (Ours)	**6.206**	**4.687**	**0.583**	**0.902**	**0.943**	**0.038**	5.457
Cross-Subject							
PointTransformer [52]	6.573	5.071	–	–	–	–	–
mmDiff (η=0) [12]	6.365	4.880	–	–	–	–	–
mmDiff (η=1) [12]	6.379	4.877	0.305	0.671	0.735	0.098	2.464
MG-Prior (Ours)	6.313	4.841	0.428	0.762	0.824	0.050	2.927
MG-Prior + DiffPose [19]	6.359	4.787	0.342	0.684	0.746	0.078	2.932
MG-Prior + NFPG (Ours)	**6.090**	**4.660**	**0.604**	**0.909**	**0.946**	**0.046**	5.738
Cross-Environment							
PointTransformer [52]	9.046	7.020	–	–	–	–	–
mmDiff (η=0) [12]	8.979	6.926	–	–	–	–	–
mmDiff (η=1) [12]	9.034	6.970	0.188	0.450	0.515	0.167	2.319
MG-Prior (Ours)	8.780	6.770	0.392	0.687	0.747	0.065	3.028
MG-Prior + DiffPose [19]	8.693	6.682	0.281	0.588	0.650	0.109	2.971
MG-Prior + NFPG (Ours)	**8.187**	**6.312**	**0.533**	**0.852**	**0.901**	**0.025**	5.841

Table 2. Leave-one-environment-out cross-validation on MM-Fi. Position errors in cm. Each block holds out one environment for testing and trains on the rest. Holding out Env 4 corresponds to the Cross-Environment split of Tab. 1.

Left-out	Model	MPJPE↓	PA-MPJPE↓	@50%	@90%	@95%	ECE↓	Sharp.
Env 1	PointTransformer [52]	6.989	5.516	–	–	–	–	–
	mmDiff (η=0) [12]	6.785	5.350	–	–	–	–	–
	mmDiff (η=1) [12]	6.794	5.349	0.225	0.528	0.601	0.135	2.215
	MG-Prior (Ours)	**6.699**	**5.238**	0.437	0.747	0.807	0.044	2.818
	MG-Prior + DiffPose [19]	6.894	5.339	0.366	0.718	0.782	0.065	3.367
	MG-Prior + NFPG (Ours)	6.744	5.329	**0.601**	**0.911**	**0.947**	**0.043**	6.294
Env 2	PointTransformer [52]	7.064	5.507	–	–	–	–	–
	mmDiff (η=0) [12]	6.914	5.352	–	–	–	–	–
	mmDiff (η=1) [12]	6.930	5.350	0.220	0.501	0.569	0.143	2.037
	MG-Prior (Ours)	7.021	5.440	0.390	0.706	0.771	0.063	2.755
	MG-Prior + DiffPose [19]	7.276	5.643	0.359	0.698	0.761	0.071	3.364
	MG-Prior + NFPG (Ours)	**6.813**	**5.326**	**0.599**	**0.908**	**0.943**	**0.042**	6.091
Env 3	PointTransformer [52]	6.779	5.265	–	–	–	–	–
	mmDiff (η=0) [12]	6.570	5.028	–	–	–	–	–
	mmDiff (η=1) [12]	6.590	5.036	0.236	0.545	0.618	0.129	2.303
	MG-Prior (Ours)	6.716	5.220	0.420	0.736	0.797	0.050	2.744
	MG-Prior + DiffPose [19]	6.914	5.245	0.365	0.712	0.776	0.068	3.300
	MG-Prior + NFPG (Ours)	**6.382**	**4.899**	**0.576**	**0.890**	**0.931**	**0.037**	5.382
Env 4	PointTransformer [52]	9.046	7.020	–	–	–	–	–
	mmDiff (η=0) [12]	8.979	6.926	–	–	–	–	–
	mmDiff (η=1) [12]	9.034	6.970	0.188	0.450	0.515	0.167	2.319
	MG-Prior (Ours)	8.821	6.821	0.373	0.658	0.718	0.075	2.790
	MG-Prior + DiffPose [19]	8.693	6.682	0.281	0.588	0.650	0.109	2.971
	MG-Prior + NFPG (Ours)	**8.187**	**6.312**	**0.533**	**0.852**	**0.901**	**0.025**	5.841

3.4. Baseline Models

We evaluate our approach against baselines spanning deterministic and probabilistic paradigms using state-of-the-art models from the RGB and radar domain adapted to our use case.

PointTransformer [52]**:** A deterministic point cloud processing baseline that directly regresses 3D poses without probabilistic modeling.

mmDiff [12]**:** We implement the deterministic DDIM framework of Fan *et al*., which treats the backbone's coarse prediction as a noisy pose estimate and iteratively refines it through a deterministic diffusion process. While effective for accuracy, this formulation cannot quantify predictive ambiguity. Since a DDIM can equivalently be formulated as a DDPM by setting $\eta = 1$ in the denoising equation [43], we additionally evaluate this stochastic variant to enable multi-hypothesis generation and uncertainty quantification.

MG-Prior: Our multi-hypothesis multivariate Gaussian prior model, trained in Phase 1, serves as both a standalone baseline and the conditioning source for downstream models.

MG-Prior + DiffPose [19]**:** We adapt the MHM DiffPose from the RGB domain to radar-based pose estimation by replacing its original Gaussian mixture model conditioning with samples from our MG prior distribution $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. In line with the original formulation, the N sampled poses are embedded, weighted by their Gaussian likelihood, and aggregated via a joint-wise transformer. The output is concatenated with projected backbone features to form the conditioning vector. This baseline generates diverse pose hypotheses from standard Gaussian noise instead of a backbone prediction, enabling calibration and coverage analysis.

3.5. Implementation Details

We follow a two-phase training procedure. In Phase 1, we train the backbone and MG prior distribution end-to-end. In Phase 2, we freeze the backbone and prior and train the NF on its features and temporal predictions. Hyperparameters for the prior are tuned on the MM-Fi dataset using cross-subject evaluation with three held-out validation participants (see Appendix B for full grid search results). We select hyperparameters that balance accuracy and calibration, yielding λ_{KL}=15 and γ=1 for the Gaussian prior distribution, which we use for all datasets. We use 5 consecutive radar frames as inputs to the model and 6 consecutive backbone predictions for the conditioning of the NF. The NF is implemented with 8 affine coupling layers [9]. The prior and MH-NFPG are trained with the Adam optimizer, a batch size of 32 and a learning rate of 0.0001 for both phases on NVIDIA A40 GPUs. We train the model using early stopping, resulting in 4 epochs of prior training and 10 epochs of NF training for the three MM-Fi conditions, 15 and 25 epochs on mmRadPose, and 15 and 6 epochs on the mRI dataset, respectively.

Baseline models are trained using the optimized parameters reported in their respective implementations [12, 19]. For mmDiff, we retain the limb loss weight of 10 from the original code repository. All evaluations of probabilistic models are done with 200 hypotheses [19, 46]. The final pose estimate is obtained by averaging across all hypotheses.

4. Experiments

We evaluate MH-NFPG on three radar pose estimation benchmarks, MM-Fi [49], mmRadPose [11], and mRI [3]. Tables 1–4 report pose estimation accuracy and uncertainty calibration jointly. We first analyze calibration, the primary contribution of this work, and then confirm that calibrated uncertainty does not sacrifice prediction accuracy. We subsequently evaluate inference efficiency and ablate key design choices.

4.1. Uncertainty Calibration and Pose Estimation

Table 1 reports MM-Fi under its official testing protocol, the Random, Cross-Subject, and Cross-Environment splits, and Table 2 additionally evaluates a leave-one-environment-out cross-validation that holds out each environment in turn. MH-NFPG consistently produces well-calibrated posteriors across all three benchmarks. Across the two diffusion baselines, MH-NFPG reduces ECE by 34–85%, with the largest reduction on the Cross-Environment split of MM-Fi (ECE 0.025 against 0.167 for mmDiff and 0.109 for MG-Prior + DiffPose), where distribution shift causes the diffusion baselines to degrade severely while our posterior stays calibrated. Coverage at 95% closely tracks the nominal rate across all settings (88–95%), whereas mmDiff (η=1) exhibits severe undercoverage (51–74%). MG-Prior + DiffPose also falls short of nominal coverage, so neither diffusion baseline matches the calibration of our NF.

Under this protocol, MH-NFPG stays well-calibrated across all four held-out environments (ECE 0.025–0.043) with the highest coverage at every credible level, whereas mmDiff (η=1) ranges from 0.129 to 0.167, a 3× to nearly 7× gap. MH-NFPG also attains the best position accuracy in three of the four splits, with the standalone MG prior marginally ahead on MPJPE and PA-MPJPE only when Env 1 is held out. This confirms that the learned posterior adapts to unseen environments rather than reflecting a single favorable split.

While MH-NFPG produces broader predictive intervals than baselines (reflected in higher sharpness values), the near-nominal coverage rates demonstrate that these intervals are appropriately sized rather than overconfident, which is critical for safety-relevant applications. In contrast, diffusion baselines exhibit low sharpness but severely undercover, indicating overconfident uncertainty estimates. Notably, MH-NFPG produces its tightest predictive intervals on mmRadPose, which uses optical motion capture ground truth, suggesting that the model's uncertainty appropriately reflects ground truth precision.

MH-NFPG matches or improves on the strongest baseline in pose accuracy across all three datasets. It attains the lowest MPJPE on every MM-Fi split and on mmRadPose (6.030 cm), with the largest gains on the most challeng-

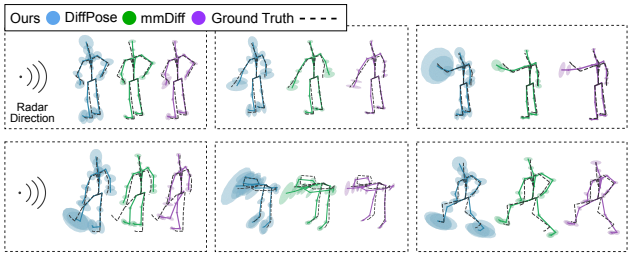

Figure 3. Qualitative comparison on mmRadPose. Predicted skeletons (solid, colored) and ground truth (dotted) are shown with per-joint 90% confidence ellipsoids. **Top**: Easy examples. Joints with smaller radar cross section show higher uncertainties. **Bottom**: Challenging examples with occlusions and more complicated multipath effects. MH-NFPG achieves lower pose error with well-calibrated uncertainty estimates, where ground-truth joints fall consistently within the 90% confidence regions. Baseline methods produce poorly localized ellipsoids under ambiguous conditions.

Table 5. Computational comparison of multi-hypothesis pose estimation models on mmRadPose using 200 hypotheses for all probabilistic models. The bottom section shows the MH-NFPG component breakdown.

Model	Params (M)	GFLOPs	Time (ms)
MG-Prior	10.606	39.165	8.127 ± 0.174
MG-Prior + NFPG	19.709	39.664	12.460 ± 0.130
MG-Prior + DiffPose [19]	26.222	489.613	173.232 ± 1.017
mmDiff [12] η=0 (det., 1 hyp.)	48.344	103.550	48.090 ± 0.102
mmDiff [12] η=1	48.344	318.133	257.184 ± 1.281
MG-Prior + NFPG Component Breakdown			
Spatiotemporal Transformer Backbone	8,945,152	39.039	7.683 ± 0.093
NF Conditioning Module	9,230,713	0.013	2.535 ± 0.024
Standard Laplace Sampling	—	0.000	0.101 ± 0.021
Real NVP NF	1,533,152	0.612	2.140 ± 0.021
Total	**19,709,017**	**39.664**	**12.460 ± 0.130**

Table 3. Pose estimation accuracy and uncertainty calibration on mmRadPose. Position errors in cm. Per-angle columns report MPJPE only.

Model	Overall		MPJPE by Angle			Calibration (Overall)				
	MPJPE↓	PA-MPJPE↓	0°	45°	90°	@50%	@90%	@95%	ECE↓	Sharp.
PointTransformer [52]	6.467	5.116	6.183	6.239	6.991	–	–	–	–	–
mmDiff (η=0) [12]	6.313	4.693	6.095	6.058	6.797	–	–	–	–	–
mmDiff (η=1) [12]	6.450	4.724	6.245	6.225	6.890	0.192	0.451	0.514	0.163	2.108
MG-Prior (Ours)	6.645	4.933	6.565	6.252	7.121	0.346	0.698	0.773	0.109	2.755
MG-Prior + DiffPose [19]	6.539	4.812	6.331	6.176	7.120	0.254	0.545	0.613	0.145	2.497
MG-Prior + NFPG (Ours)	**6.030**	**4.597**	**5.895**	**5.587**	**6.613**	**0.480**	**0.839**	**0.891**	**0.056**	4.739

Table 4. Pose estimation accuracy and uncertainty calibration on mRI. Position errors in cm.

Model	MPJPE↓	PA-MPJPE↓	@50%	@90%	@95%	ECE↓	Sharp.
PointTransformer [52]	8.385	6.049	–	–	–	–	–
mmDiff (η=0) [12]	**8.225**	5.650	–	–	–	–	–
mmDiff (η=1) [12]	8.300	5.659	0.247	0.566	0.643	0.138	2.778
MG-Prior (Ours)	8.371	5.808	0.303	0.635	0.709	0.099	2.796
MG-Prior + DiffPose [19]	8.802	6.043	0.245	0.528	0.589	0.132	3.351
MG-Prior + NFPG (Ours)	8.269	**5.616**	**0.477**	**0.829**	**0.888**	**0.038**	6.175

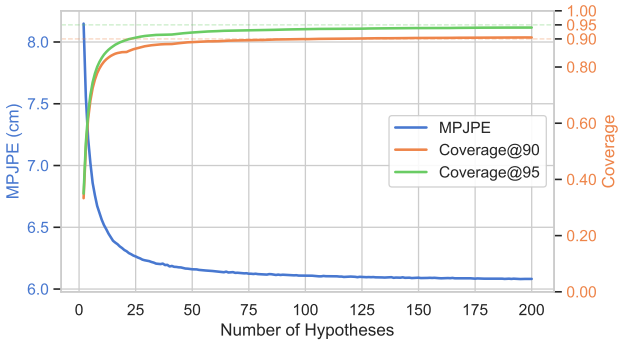

Figure 4. Effect of the number of hypotheses on MPJPE and calibration coverage on MM-Fi. Both improve with more hypotheses, with diminishing returns beyond approximately 100.

ing conditions, the Cross-Environment split of MM-Fi and the 90° view on mmRadPose. On mRI it achieves the lowest PA-MPJPE (5.616 cm) and essentially matches the best MPJPE (8.269 versus 8.225 cm), showing that calibrated uncertainty comes at no cost to point estimation accuracy.

4.2. Qualitative Analysis

Figure 1 shows MH-NFPG predictions across all three datasets, with ground truth consistently falling within the predicted confidence intervals, confirming that calibrated uncertainty generalizes across diverse actions, subjects, and radar configurations. Figure 3 compares MH-NFPG against DiffPose and mmDiff on mmRadPose. Our model produces well-localized estimates with confidence ellipsoids that reflect radar scattering properties [37], especially under more complicated multipath effects and occlusions. In contrast, ground-truth joints frequently fall outside the predicted confidence ellipsoids for both MG-Prior + DiffPose and mmDiff.

4.3. Inference Efficiency

We evaluate all models in a causal, online setting with batch size 1 (see Appendix C for details) and high-fidelity sampling from the posterior. For temporal models, we cache backbone predictions in a circular buffer, requiring only a single forward pass per frame. All 200 hypotheses [19] are drawn in parallel and stored in the batch dimension of the input tensor for all models, whereas diffusion baselines require additional sequential denoising steps (25 for both DiffPose [19] and mmDiff [12]).

Table 5 reports computational cost when generating high-fidelity multi-hypothesis distributions with 200 samples per frame on an NVIDIA RTX 3090 GPU. MH-NFPG runs at 12.5 ms per frame (80 FPS), over 20× faster than mmDiff (η=1) and 14× faster than MG-Prior + DiffPose, with 12× fewer FLOPs than DiffPose. The efficiency gain stems from replacing sequential denoising with a single NF forward pass. The Real NVP NF requires only 2.1 ms.

4.4. Ablation Study

Table 6 ablates the major design choices across all three datasets. The default configuration (L=8 coupling layers, GCN conditioning, Standard Laplace base, affine coupling, temporal context) gives the best overall accuracy/calibration trade-off on every dataset, attaining the lowest MPJPE on MM-Fi (6.090 cm) and mmRadPose (6.030 cm) and remaining within 0.04 cm of the best on mRI at a substantially lower ECE. Varying the coupling depth shows that deeper NFs ($L\in\{10, 12\}$) can shave a little MPJPE but degrade calibration, whereas replacing the GCN conditioning with an MLP consistently worsens accuracy (e.g., 6.030 to 6.333 cm on mmRadPose), confirming the value of encoding skeletal structure. Among base distributions, the heavy-tailed Standard Laplace yields the best accuracy on MM-Fi and mmRadPose, supporting our choice. Substituting rational quadratic spline NFs [10] for affine coupling worsens accuracy and greatly inflates sharpness (up to 11.7 cm), and we omit autoregressive NFs due to their higher computational complexity [34]. Finally, removing the temporal conditioning increases MPJPE on all three datasets, confirming that temporal context disambiguates sparse radar observations. Figure 4 further shows that both MPJPE and calibration coverage improve with increasing hypothesis count, with diminishing returns beyond approximately 100 hypotheses, validating our choice of 200.

5. Conclusion

We presented MH-NFPG, a normalizing flow pose generator for radar-based human pose estimation that combines probabilistic expressiveness with single forward-pass inference. A conditional NF with a heteroscedastic Gaussian prior and Laplace base distribution produces calibrated uncertainties while enabling parallel hypothesis sampling. On MM-Fi, mmRadPose, and mRI, MH-NFPG achieves the

Table 6. Ablation across all three datasets. Cells: MPJPE (cm) / ECE / Sharpness (cm). Base dist.: StdN=Std. Normal; ScG/ScL=Gaussian/Laplace scaled by the prior uncertainty. MM-Fi uses the cross-subject protocol. Default (bold): L=8, GCN, fixed-scale Laplace, affine, temporal.

	Coupling layers					Cond.	Base distribution			NF	Temp.	
	L=2	L=4	L=6	L=10	L=12	MLP	StdN	ScG	ScL	Spline	None	**Ours**
MM-Fi	6.372/.056/5.582	6.260/.080/7.644	6.159/.060/6.341	6.253/.033/4.739	6.213/.026/4.400	6.186/.057/6.388	6.354/.032/4.849	6.369/.028/4.690	6.314/.036/5.349	6.361/.042/11.73	6.281/.049/5.829	**6.090/.046/5.738**
mmRadPose	6.150/.061/3.284	6.294/.061/5.794	6.282/.054/5.815	6.164/.083/3.913	6.215/.113/3.434	6.333/.062/4.941	6.109/.074/3.719	6.285/.072/4.171	6.248/.059/5.220	6.481/.128/10.52	6.408/.061/4.672	**6.030/.056/4.739**
mRI	8.979/.072/3.311	8.493/.024/7.038	8.442/.041/7.456	8.236/.060/4.205	8.246/.069/4.026	8.588/.033/7.028	8.288/.039/4.956	8.259/.074/4.421	8.275/.055/4.734	8.448/.088/11.70	8.397/.017/6.782	**8.269/.038/6.175**

best uncertainty calibration and matches or exceeds the accuracy of all baselines, while enabling real-time deployment. Operating on radar point clouds at 80 frames per second, it delivers the calibrated uncertainty that a safety-critical system needs to know when its prediction can be trusted.

References

[1] Moloud Abdar, Farhad Pourpanah, Sadiq Hussain, Dana Rezazadegan, Li Liu, Mohammad Ghavamzadeh, Paul Fieguth, Xiaochun Cao, Abbas Khosravi, U Rajendra Acharya, et al. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information fusion*, 76:243–297, 2021.

[2] Sizhe An and Umit Y. Ogras. Fast and scalable human pose estimation using mmWave point cloud. In *Proceedings of the 59th ACM/IEEE Design Automation Conference*, pages 889–894, 2022.

[3] Sizhe An, Yin Li, and Umit Ogras. mRI: Multi-modal 3D human pose estimation dataset using mmWave, RGB-D, and inertial sensors. *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[4] L Ardizzone, J Kruse, C Rother, and U Köthe. Analyzing inverse problems with invertible neural networks. 7th int. In *Conf. on Learning Representations, ICLR 2019 (New Orleans, LA)*, 2019.

[5] Edmon Begoli, Tanmoy Bhattacharya, and Dimitri Kusnezov. The need for uncertainty quantification in machine-assisted medical decision making. *Nature Machine Intelligence*, 1(1):20–23, 2019.

[6] Lennart Bramlage, Michelle Karg, and Cristóbal Curio. Plausible uncertainties for human pose regression. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15133–15142, 2023.

[7] Anjun Chen, Xiangyu Wang, Shaohao Zhu, Yanxu Li, Jiming Chen, and Qi Ye. mmBody benchmark: 3D body reconstruction dataset and analysis for millimeter wave radar. *ACM Multimedia*, 2022.

[8] Hsin-Che Chiang, Guan-Hua Li, Fan Wang, Shervin Shirmohammadi, and Cheng-Hsin Hsu. Enhancing skeletal pose estimation from mmWave point clouds through uncertainty reduction. In *Proceedings of the 5th International Workshop on Human-centric Multimedia Analysis*, pages 45–53, 2024.

[9] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using Real NVP. In *International Conference on Learning Representations (ICLR)*, 2017.

[10] Conor Durkan, Artur Bekasov, Iain Murray, and George Papamakarios. Neural spline flows. *Advances in Neural Information Processing Systems*, 32, 2019.

[11] Lukas Engel, Jonas Mueller, Eduardo Javier Feria Rendon, Eva Dorschky, Daniel Krauss, Ingrid Ullmann, Bjoern M. Eskofier, and Martin Vossiek. Advanced millimeter wave radar-based human pose estimation enabled by a deep learning neural network trained with optical motion capture ground truth data. *IEEE Journal of Microwaves*, 2025.

[12] Junqiao Fan, Jianfei Yang, Yuecong Xu, and Lihua Xie. Diffusion model is a good pose estimator from 3D RF-vision. In *European Conference on Computer Vision (ECCV)*, pages 1–18. Springer, 2024.

[13] Runyang Feng, Yixing Gao, Tze Ho Elden Tse, Xueqing Ma, and Hyung Jin Chang. DiffPose: Spatiotemporal diffusion model for video-based human pose estimation. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14861–14872, 2023.

[14] Jia Gong, Lin Geng Foo, Zhipeng Fan, Qiuhong Ke, Hossein Rahmani, and Jun Liu. DiffPose: Toward more reliable 3D pose estimation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13041–13051, 2023.

[15] Nitesh B. Gundavarapu, Divyansh Srivastava, Rahul Mitra, Abhishek Sharma, and Arjun Jain. Structured aleatoric uncertainty in human pose estimation. In *CVPR Workshops*, 2019.

[16] Ju-Min Han, Jun-Hee Kim, and Seong-Whan Lee. Propose: Probabilistic 3d human pose estimation with instance-level distribution and normalizing flow. In *AAAI Conference on Artificial Intelligence*, 2025.

[17] Or Hirschorn and Shai Avidan. Normalizing flows for human pose anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13545–13554, 2023.

[18] Yuan-Hao Ho, Jen-Hao Cheng, Sheng Yao Kuan, Zhongyu Jiang, Wenhao Chai, Hsiang-Wei Huang, Chih-Lung Lin, and Jenq-Neng Hwang. Rt-pose: A 4d radar tensor-based 3d human pose estimation and localization benchmark. In *European Conference on Computer Vision*, pages 107–125. Springer, 2024.

[19] Karl Holmquist and Bastian Wandt. DiffPose: Multi-hypothesis human pose estimation using diffusion models. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15977–15987, 2023.

[20] Liu Jinwei, Zhang Feng, and Chen Lei. DiffPose: Reliable 2D pose estimation through denoising diffusion. In *China Automation Congress (CAC)*, pages 4041–4046. IEEE, 2023.

[21] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in neural information processing systems*, 30, 2017.

[22] Daniel Krauss, Lukas Engel, Tabea Ott, Johanna Bräunig, Robert Richer, Markus Gambietz, Nils Albrecht, Eva M. Hille, Ingrid Ullmann, Matthias Braun, et al. A review and tutorial on machine learning-enabled radar-based biomedical monitoring. *IEEE Open Journal of Engineering in Medicine and Biology*, 5:680–699, 2024.

[23] Shih-Po Lee, Niraj Prakash Kini, Wen-Hsiao Peng, Ching-Wen Ma, and Jenq-Neng Hwang. HuPR: A benchmark for human pose estimation using millimeter wave radar. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5715–5724, 2023.

[24] Chen Li and Gim Hee Lee. Generating multiple hypotheses for 3d human pose estimation with mixture density network. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9887–9895, 2019.

[25] Jiefeng Li, Siyuan Bian, Ailing Zeng, Can Wang, Bo Pang, Wentao Liu, and Cewu Lu. Human pose regression with

residual log-likelihood estimation. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11005–11014, 2021.

[26] Jiefeng Li, Siyuan Bian, Ailing Zeng, Can Wang, Bo Pang, Wentao Liu, and Cewu Lu. Human pose regression with residual log-likelihood estimation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11025–11034, 2021.

[27] Lanxin Li, Che Liu, and Tie Jun Cui. Rpdiff: Multi-hypothesis human pose estimation based on mmwave radar. In *2024 IEEE MTT-S International Microwave Workshop Series on Advanced Materials and Processes for RF and THz Applications (IMWS-AMP)*, pages 1–3. IEEE, 2024.

[28] Lanxin Li, Che Liu, Wenming Yu, and Tie Jun Cui. Diff-HPE: Multi-hypothesis human pose estimation based on mmWave radar and diffusion framework. *IEEE Sensors Journal*, 2025.

[29] Shipeng Liu, Ziliang Xiong, Bastian Wandt, and Per-Erik Forssén. Continuous normalizing flows for uncertainty-aware human pose estimation. In *Image Analysis*, pages 276–291, Cham, 2025. Springer Nature Switzerland.

[30] Chris Xiaoxuan Lu, Muhamad Risqi U. Saputra, Peijun Zhao, Yasin Almalioglu, Pedro P. B. De Gusmao, Changhao Chen, Ke Sun, Niki Trigoni, and Andrew Markham. milliEgo: Single-chip mmWave radar aided egomotion estimation via deep sensor fusion. In *Proceedings of the 18th Conference on Embedded Networked Sensor Systems*, pages 109–122, 2020.

[31] Jonas Leo Mueller, Lukas Engel, Eva Dorschky, Daniel Krauss, Ingrid Ullmann, Martin Vossiek, and Bjoern M. Eskofier. RadProPoser: A framework for human pose estimation with uncertainty quantification from raw radar data. *arXiv preprint arXiv:2508.03578*, 2025.

[32] Jonas Leo Mueller, Alexander Weiss, and Bjoern M Eskofier. Adaptive biofeedback for digital physiotherapy using sakoe-chiba constrained pose matching. In *International Conference on AI in Healthcare*, pages 227–241. Springer, 2025.

[33] Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using bayesian binning. In *Proceedings of the AAAI conference on artificial intelligence*, 2015.

[34] George Papamakarios, Theo Pavlakou, and Iain Murray. Masked autoregressive flow for density estimation. *Advances in Neural Information Processing Systems*, 30, 2017.

[35] George Papamakarios, Eric Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan. Normalizing flows for probabilistic modeling and inference. *Journal of Machine Learning Research*, 22(57):1–64, 2021.

[36] Paweł A Pierzchlewicz, R James Cotton, Mohammad Bashiri, and Fabian H Sinz. Multi-hypothesis 3d human pose estimation metrics favor miscalibrated distributions. *arXiv preprint arXiv:2210.11179*, 2022.

[37] Mark A. Richards. *Fundamentals of Radar Signal Processing*. McGraw-Hill, 2005.

[38] Christian Rupprecht, Iro Laina, Robert DiPietro, Maximilian Baust, Federico Tombari, Nassir Navab, and Gregory D. Hager. Learning in an uncertain world: Representing ambiguity through multiple hypotheses. In *IEEE International Conference on Computer Vision (ICCV)*, pages 3591–3600, 2017.

[39] Constantin Scholz, Hoang-Long Cao, Emil Imrith, Nima Roshandel, Hamed Firouzipouyaei, Aleksander Burkiewicz, Milan Amighi, Sebastien Menet, Dylan Warawout Sisavath, Antonio Paolillo, et al. Sensor-enabled safety systems for human-robot collaboration: A review. *IEEE Sensors Journal*, 2024.

[40] Arindam Sengupta, Feng Jin, Renyuan Zhang, and Siyang Cao. mm-Pose: Real-time human skeletal posture estimation using mmWave radars and CNNs. *IEEE Sensors Journal*, 2019.

[41] Arindam Sengupta, Feng Jin, Renyuan Zhang, and Siyang Cao. mm-Pose: Real-time human skeletal posture estimation using mmWave radars and CNNs. *IEEE Sensors Journal*, 20 (17):10032–10044, 2020.

[42] Wenkang Shan, Zhenhua Liu, Xinfeng Zhang, Zhao Wang, Kai Han, Shanshe Wang, Siwei Ma, and Wen Gao. Diffusion-based 3D human pose estimation with multi-hypothesis aggregation. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14761–14771, 2023.

[43] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021.

[44] Yangfan Sun, Renlong Hang, Zhu Li, Mouqing Jin, and Kelvin Xu. Privacy-preserving fall detection with deep learning on mmwave radar signal. In *2019 IEEE Visual Communications and Image Processing (VCIP)*, pages 1–4. IEEE, 2019.

[45] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017.

[46] Tom Wehrbein, Marco Rudolph, Bodo Rosenhahn, and Bastian Wandt. Probabilistic monocular 3D human pose estimation with normalizing flows. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11199–11208, 2021.

[47] Christina Winkler, Daniel Worrall, Emiel Hoogeboom, and Max Welling. Learning likelihoods with conditional normalizing flows. *arXiv preprint arXiv:1912.00042*, 2019.

[48] Hongfei Xue, Yan Ju, Chenglin Miao, Yijiang Wang, Shiyang Wang, Aidong Zhang, and Lu Su. mmMesh: Towards 3D real-time dynamic human mesh construction using millimeter-wave. In *ACM SIGMOBILE International Conference on Mobile Systems, Applications, and Services*, 2021.

[49] Jianfei Yang, He Huang, Yunjiao Zhou, Xinyan Chen, Yuecong Xu, Shenghai Yuan, Han Zou, Chris Xiaoxuan Lu, and Lihua Xie. MM-Fi: Multi-modal non-intrusive 4D human dataset for versatile wireless sensing. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2023.

[50] Andrei Zanfir, Eduard Gabriel Bazavan, Hongyi Xu, William T. Freeman, Rahul Sukthankar, and Cristian Sminchisescu. Weakly supervised 3d human pose and shape reconstruction with normalizing flows. In *Computer Vision –*

ECCV 2020, pages 465–481, Cham, 2020. Springer International Publishing.

[51] Shuangfei Zhai, Ruixiang Zhang, Preetum Nakkiran, David Berthelot, Jiatao Gu, Huangjie Zheng, Tianrong Chen, Miguel Angel Bautista, Navdeep Jaitly, and Josh Susskind. Normalizing flows are capable generative models. *arXiv preprint arXiv:2412.06329*, 2024.

[52] Hengshuang Zhao, Li Jiang, Jiaya Jia, Philip HS Torr, and Vladlen Koltun. Point transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 16259–16268, 2021.

[53] Ce Zheng, Wenhan Wu, Chen Chen, Taojiannan Yang, Sijie Zhu, Ju Shen, Nasser Kehtarnavaz, and Mubarak Shah. Deep learning-based human pose estimation: A survey. *ACM Computing Surveys*, 56(1):1–37, 2023.

A. Prior Loss Derivation

We test two formalizations of the prior loss.

Gaussian Prior NLL. Starting from the multivariate Gaussian likelihood [6]

$$p(\mathbf{y}) = \frac{1}{(2\pi)^{d/2}|\mathbf{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{y}-\boldsymbol{\mu})^\top \mathbf{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu})\right), \tag{13}$$

and dropping the constant normalization term, we obtain the negative log-likelihood. Adding a variance regularizer $\gamma\,\mathcal{R}$ with $\mathcal{R} = \frac{\mathrm{tr}(\mathbf{\Sigma})}{d_{\text{pose}}}$ to stabilize training yields

$$\mathcal{L}_{\text{NLL}}^{\text{prior}} = \mathbb{E}\left[\frac{1}{2}\left(\log|\mathbf{\Sigma}| + (\mathbf{y}-\boldsymbol{\mu})^\top \mathbf{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu})\right) + \gamma\,\mathcal{R}\right]. \tag{14}$$

In practice, the empirical mean $\boldsymbol{\mu}$ and covariance $\mathbf{\Sigma}$ are computed from the N decoded pose hypotheses. To avoid explicit matrix inversion, we factorize $\mathbf{\Sigma} = \mathbf{L}\mathbf{L}^\top$ via Cholesky decomposition. The Mahalanobis distance is then obtained by solving the triangular system $\mathbf{L}\mathbf{v} = (\mathbf{y}-\boldsymbol{\mu})$ and computing $\|\mathbf{v}\|^2$, while the log-determinant reduces to $\log|\mathbf{\Sigma}| = 2\sum_i \log L_{ii}$. Adaptive diagonal jitter ensures numerical stability when $\mathbf{\Sigma}$ is near-singular.

Laplace Prior NLL. Starting from the Laplace log-likelihood

$$\log p(\mathbf{y}) = \sum_{i=1}^{d_{\text{pose}}} \left(-\log(2b_i) - \frac{|y_i-\mu_i|}{b_i}\right), \tag{15}$$

and dropping constants, we obtain the negative log-likelihood with scale regularizer $\mathcal{R} = \frac{\sum_{i=1}^{d_{\text{pose}}} b_i}{d_{\text{pose}}}$

$$\mathcal{L}_{\text{L1}} = \mathbb{E}\left[\sum_{i=1}^{d_{\text{pose}}} \left(\log b_i + \frac{|y_i-\mu_i|}{b_i}\right) + \gamma\,\mathcal{R}\right]. \tag{16}$$

KL Divergence. The KL divergence between the learned latent distribution $q = \mathcal{N}(\boldsymbol{\mu}_l, \mathrm{diag}(\boldsymbol{\sigma}_l^2))$ and the standard normal prior $p = \mathcal{N}(\mathbf{0}, \mathbf{I})$ has the closed-form solution

$$\mathcal{L}_{\text{KL}} = -\frac{1}{2}\mathbb{E}\left[\sum_{j=1}^{d_l} \left(1 + \log \boldsymbol{\sigma}_{l,j}^2 - \boldsymbol{\mu}_{l,j}^2 - \boldsymbol{\sigma}_{l,j}^2\right)\right]. \tag{17}$$

The final training loss for the prior combines the reconstruction and regularization terms

$$\mathcal{L}_{\text{backbone}} = \mathcal{L}_{\text{NLL}}^{\text{prior}} + \lambda_{\text{KL}}\mathcal{L}_{\text{KL}}. \tag{18}$$

B. Hyperparameter Tuning

We tune hyperparameters of the prior distribution on the MM-Fi dataset with cross-subject evaluation using three random held-out train-set participants as validation. Tables 7 and 8 present the full grid search results.

We also tested the Laplace likelihood formulation (Appendix A).

In summary, we therefore chose the Gaussian likelihood formulation for the prior because of our empirical results.

C. Detailed Computational Analysis

We design our model for efficient inference without sacrificing the benefits of probabilistic uncertainty quantification. To ensure a fair comparison, we implement optimized inference pipelines for all evaluated models under their best-performing configurations. All models are evaluated in a causal, online setting with batch size 1, reflecting a realistic deployment scenario.

For temporally-aware models such as MH-NFPG and mmDiff, we employ a circular buffer that maintains the last n_t backbone predictions as a sliding window of temporal context. This allows the model to perform a single forward pass per frame, using cached backbone predictions until the buffer is filled and continuously updated thereafter.

We exploit batched computation for parallelism wherever possible and present the resulting inference procedures in Algorithms 1, 2, and 3. For the normalizing flow model, we draw a batch of N hypotheses in the backbone's latent space and propagate them through the flow in parallel, producing the full output distribution in a single forward pass. For the diffusion baselines, we parallelize hypothesis sampling across the batch dimension, while computing the conditioning only once (Table 9). For MG-Prior + DiffPose [19], we use 25 denoising steps, matching the original paper's configuration. For mmDiff [12], the full diffusion schedule comprises 51 timesteps, but because the backbone already provides a partially denoised pose estimate, the first 25 steps are skipped at inference. The deterministic DDIM variant (η=0) further reduces this to 2 steps as in the original formulation, while the probabilistic DDPM variant (η=1) uses the remaining 25 steps. All models are benchmarked on an NVIDIA RTX 3090 GPU.

In Algorithm 1, all operations on N hypotheses are batched along a single tensor dimension and executed in parallel, with the conditioning computed once and shared across all hypotheses. In contrast, both diffusion baselines require M=25 sequential denoising steps. The two diffusion models differ in their conditioning. DiffPose (Algorithm 2) uses stochastic conditioning where each denoising step and hypothesis receives a unique conditioning vector derived from fresh Gaussian samples, making the conditioning per-hypothesis and per-step. While this conditioning computation can be batched across all $M \times N$ pairs, the denoising loop remains sequential. mmDiff (Algorithm 3) instead uses deterministic conditioning computed once from per-joint backbone features and temporal context, shared across all hypotheses. Here, stochasticity arises only from the noise initialization, which is then refined

Table 7. Tuning of base distribution hyperparameters λ_{KL} (KL weight) and γ (variance regularization) with Gaussian likelihood. MPJPE in cm, Sharpness in cm.

	λ_{KL}=1					λ_{KL}=5					λ_{KL}=10					λ_{KL}=15					λ_{KL}=20				
γ	1	3	5	7	9	1	3	5	7	9	1	3	5	7	9	1	3	5	7	9	1	3	5	7	9
MPJPE↓	6.71	6.86	**6.60**	6.78	6.68	6.73	6.67	6.77	6.68	6.64	6.74	6.72	6.78	6.73	6.73	<u>6.67</u>	6.74	6.70	6.73	6.84	6.87	6.76	7.01	6.73	6.66
ECE↓	.044	.069	.079	.094	.095	.052	.072	.085	.090	.092	.051	.067	.085	.090	.094	<u>.042</u>	.073	.083	.100	.093	.047	.070	**.040**	.085	.098
Sharpness	8.26	5.26	4.06	3.59	3.24	7.77	5.03	3.81	3.46	3.21	7.74	4.95	4.05	3.40	3.12	8.41	4.90	4.06	3.38	3.10	8.11	4.93	7.93	3.44	3.25

Best values in **bold**, <u>underlined</u> denotes best accuracy-calibration trade-off (λ_{KL}=15, γ=1). Evaluated on MM-Fi cross-subject split with three validation participants.

Table 8. Tuning of loss weights λ_{KL} (KL divergence) and γ (variance regularization) with Laplace likelihood. MPJPE in cm, Sharpness in cm.

	λ_{KL}=1					λ_{KL}=5					λ_{KL}=10					λ_{KL}=15					λ_{KL}=20				
γ	1	3	5	7	9	1	3	5	7	9	1	3	5	7	9	1	3	5	7	9	1	3	5	7	9
MPJPE↓	6.85	6.78	6.81	**6.68**	6.69	6.92	6.78	6.70	6.80	7.15	7.57	7.35	7.49	7.37	7.40	7.53	6.84	6.87	6.87	6.79	11.05	10.98	10.74	—	—
ECE↓	.148	.165	.180	.188	.187	.148	.175	.183	.190	.182	.146	.163	.182	.176	.183	**.138**	.175	.182	.191	.200	.124	.155	.164	—	—
Sharpness	0.94	0.47	0.31	0.27	0.24	0.89	0.43	0.32	0.28	0.31	1.23	0.62	0.45	0.37	0.32	1.32	0.49	0.34	0.29	0.24	1.83	0.87	0.68	—	—

Best values in **bold**. Evaluated on MM-Fi cross-subject validation split.

Algorithm 1 MH-NFPG Inference (Single Forward Pass)

Require: Point cloud sequence $\mathbf{X}$, temporal buffer $\mathcal{B}$, number of hypotheses N
Ensure: N pose hypotheses $\{\hat{\mathbf{y}}^{(s)}\}_{s=1}^{N}$

$\mathbf{F} \leftarrow \text{SpatiotemporalTransformer}(\mathbf{X})$ ▷ Backbone: single forward pass
$\boldsymbol{\mu}_l, \log \boldsymbol{\sigma}_l^2 \leftarrow \text{Encoder}(\mathbf{F})$ ▷ Latent parameters
$\{\mathbf{z}_l^{(s)}\}_{s=1}^{N} \leftarrow \boldsymbol{\mu}_l + \boldsymbol{\sigma}_l \odot \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ ▷ Batched latent sampling
$\{\hat{\mathbf{y}}_{\text{prior}}^{(s)}\}_{s=1}^{N} \leftarrow \text{GCN}_{\text{dec}}(\{\mathbf{z}_l^{(s)}\}_{s=1}^{N})$ ▷ Batched decoding
$\mathbf{c} \leftarrow \text{CondGCN}(\mathbf{F},\ \mathcal{B})$ ▷ Features + temporal buffer → conditioning
$\{\mathbf{z}^{(s)}\}_{s=1}^{N} \sim \text{Laplace}(\mathbf{0},\ 1/\sqrt{2})$ ▷ Batched base distribution sampling
$\{\hat{\mathbf{y}}^{(s)}\}_{s=1}^{N} \leftarrow f^{-1}(\{\mathbf{z}^{(s)}\}_{s=1}^{N};\ \mathbf{c})$ ▷ Batched inverse flow (Real NVP)
$\mathcal{B}.\text{update}(\bar{\mathbf{y}}_{\text{prior}})$ ▷ Update temporal buffer with prior mean
return $\{\hat{\mathbf{y}}^{(s)}\}_{s=1}^{N}$

Algorithm 2 MG-Prior + DiffPose Inference (M Sequential Steps)

Require: Point cloud $\mathbf{X}$, hypotheses N, denoising steps M=25, Gaussian samples S=32
Ensure: N pose hypotheses $\{\hat{\mathbf{y}}^{(s)}\}_{s=1}^{N}$

$\mathbf{F} \leftarrow \text{SpatiotemporalTransformer}(\mathbf{X})$ ▷ Backbone (shared with MH-NFPG)
$\boldsymbol{\mu}_l, \log \boldsymbol{\sigma}_l^2 \leftarrow \text{Encoder}(\mathbf{F})$ ▷ Latent parameters (shared with MH-NFPG)
for each step $t \in \{1, \ldots, M\}$ and hypothesis $s \in \{1, \ldots, N\}$ **do** ▷ Batched
  $\{\mathbf{z}_i\}_{i=1}^{S} \leftarrow \boldsymbol{\mu}_l + \boldsymbol{\sigma}_l \odot \boldsymbol{\epsilon}_i, \quad \boldsymbol{\epsilon}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ ▷ Sample S latent codes
  $\{\hat{\mathbf{y}}_i\}_{i=1}^{S} \leftarrow \text{GCN}_{\text{dec}}(\{\mathbf{z}_i\}_{i=1}^{S})$ ▷ Decode to S pose samples
  $\mathbf{c}_{t,s} \leftarrow \text{Transformer}(\{\hat{\mathbf{y}}_i\}_{i=1}^{S})$ ▷ Stochastic conditioning per (t, s)
end for
$\{\mathbf{x}_M^{(s)}\}_{s=1}^{N} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ ▷ Initialize N noise samples
for $t = M, \ldots, 1$ **do** ▷ **Sequential** denoising loop
  $\boldsymbol{\epsilon}_\theta \leftarrow \text{Denoiser}(\{\mathbf{x}_t^{(s)}\}_{s=1}^{N},\ t,\ \{\mathbf{c}_{t,s}\}_{s=1}^{N})$ ▷ Batched over N
  $\{\mathbf{x}_{t-1}^{(s)}\} \leftarrow \text{DDPM}(\{\mathbf{x}_t^{(s)}\},\ \boldsymbol{\epsilon}_\theta,\ t)$
end for
return $\{\mathbf{x}_0^{(s)}\}_{s=1}^{N}$

Algorithm 3 mmDiff Inference (M Sequential Steps)

Require: Point cloud $\mathbf{X}$, temporal buffer $\mathcal{B}$, hypotheses N, steps M, stochasticity η
Ensure: N pose hypotheses $\{\hat{\mathbf{y}}^{(s)}\}_{s=1}^{N}$

$\hat{\mathbf{y}}_{\text{coarse}},\ \mathbf{F}_{\text{joint}} \leftarrow \text{PointTransformer}(\mathbf{X})$ ▷ Backbone + coarse pose
$\mathbf{c} \leftarrow \text{GlobalCond}(\mathbf{F}_{\text{joint}}) + \text{TemporalCond}(\mathcal{B})$ ▷ Deterministic conditioning
$\mathbf{e}_{\text{limb}} \leftarrow \text{LimbEmbed}(\mathbf{F}_{\text{joint}})$ ▷ Predicted limb lengths
Expand $\mathbf{c}$, $\mathbf{e}_{\text{limb}}$ to N hypotheses ▷ Shared across all hypotheses
$\{\mathbf{x}_M^{(s)}\}_{s=1}^{N} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ ▷ Initialize N noise samples
for $t = M, \ldots, 1$ **do** ▷ **Sequential** denoising loop
  $\boldsymbol{\epsilon}_\theta \leftarrow \text{Denoiser}(\{\mathbf{x}_t^{(s)}\}_{s=1}^{N},\ t,\ \mathbf{c},\ \mathbf{e}_{\text{limb}})$ ▷ Batched over N
  $\{\mathbf{x}_{t-1}^{(s)}\} \leftarrow \text{DDIM/DDPM}(\{\mathbf{x}_t^{(s)}\},\ \boldsymbol{\epsilon}_\theta,\ t,\ \eta)$ ▷ η=0: det., η=1: stoch.
end for
$\mathcal{B}.\text{update}(\bar{\mathbf{x}}_0)$ ▷ Update temporal buffer
return $\{\mathbf{x}_0^{(s)}\}_{s=1}^{N}$

Table 9. Component-wise breakdown of model architectures (200 hypotheses).

Component	Parameters	GFLOPs	Time (ms)
MG-Prior			
Spatiotemporal Transformer Backbone	8,945,152	39.039	7.707 ± 0.142
Gaussian Head - Encoder	1,575,424	0.091	0.169 ± 0.022
Gaussian Head - Latent Sampling	—	—	0.088 ± 0.008
Gaussian Head - Decoder	85,838	0.034	0.081 ± 0.004
Aggregation	—	—	0.083 ± 0.008
Total	**10,606,415**	**39.165**	**8.127 ± 0.174**
MG-Prior + NFPG			
Spatiotemporal Transformer Backbone	8,945,152	39.039	7.683 ± 0.093
Flow Conditioning Module	9,230,713	0.013	2.535 ± 0.024
Standard Laplace Sampling	—	0.000	0.101 ± 0.021
Normalizing Flow	1,533,152	0.612	2.140 ± 0.021
Total	**19,709,017**	**39.664**	**12.460 ± 0.130**
MG-Prior + DiffPose			
Spatiotemporal Transformer Backbone	8,945,152	39.039	7.604 ± 0.044
Gaussian Head - Encoder	1,575,424	0.091	0.156 ± 0.008
Gaussian Head - Latent Sampling	—	—	0.950 ± 0.003
Gaussian Head - Decoder	85,838	27.361	2.034 ± 0.008
Conditioning Embedding	13,605,760	403.056	144.242 ± 0.912
Diffusion Denoiser	2,009,678	20.065	17.430 ± 0.071
Total	**26,221,853**	**489.613**	**173.232 ± 1.017**
mmDiff η=0 (deterministic)			
PointTransformer Backbone	8,009,123	103.456	19.352 ± 0.048
Global Conditioning	43,424	0.003	0.413 ± 0.002
Temporal Conditioning	37,386,048	0.001	2.814 ± 0.007
Limb Conditioning	1,879,056	0.004	0.194 ± 0.002
GCNDiff Denoising Layers	1,026,455	0.086	25.318 ± 0.074
Total	**48,344,106**	**103.550**	**48.090 ± 0.102**
mmDiff η=1			
PointTransformer Backbone	8,009,123	103.456	19.336 ± 0.039
Global Conditioning	43,424	0.003	0.413 ± 0.003
Temporal Conditioning	37,386,048	0.001	2.815 ± 0.009
Limb Conditioning	1,879,056	0.004	0.198 ± 0.004
GCNDiff Denoising Layers	1,026,455	214.669	234.421 ± 1.273
Total	**48,344,106**	**318.133**	**257.184 ± 1.281**

through the sequential denoising loop.

D. Qualitative Results

Figure 5 shows MH-NFPG predictions across all three datasets, demonstrating that calibrated uncertainty estimates generalize across diverse poses, subjects, and radar configurations, with ground-truth joints consistently falling within the predicted confidence ellipsoids. Figure 6 provides an extended comparison against DiffPose and mmDiff on mmRadPose. The first three rows show easy examples, while the last three rows show ambiguous examples with occlusions and more complex multipath effects. MH-NFPG produces uncertainty estimates appropriate for the ambiguity of each sample and recognizes which joints have smaller radar cross-sections and are therefore subject to noisier observations, assigning them larger uncertainties accordingly.

E. Additional Uncertainty Analysis

Figure 7 shows the reliability diagram on mmRadPose, plotting empirical coverage against expected coverage for all evaluated probabilistic models. A perfectly calibrated model follows the diagonal. MH-NFPG closely tracks the diagonal across all confidence levels, confirming well-calibrated uncertainty estimates. In contrast, MG-Prior + DiffPose and mmDiff (η=1) fall far below the diagonal, indicating severe overconfidence where predicted confidence intervals consistently fail to cover the ground truth.

Tables 10 and 11 report the per-joint sharpness for MH-NFPG across all three datasets, computed from 200 hypotheses averaged over all test frames. Across all datasets, a consistent pattern emerges. Joints with smaller radar cross-sections (hands, feet, head) exhibit higher per-joint sharpness values, while joints with larger reflective surface area show lower uncertainty. This aligns with physical expectations, as smaller radar cross-sections yield weaker and less consistent reflections and are more susceptible to self-occlusion and multipath effects. Crucially, MH-NFPG dy-

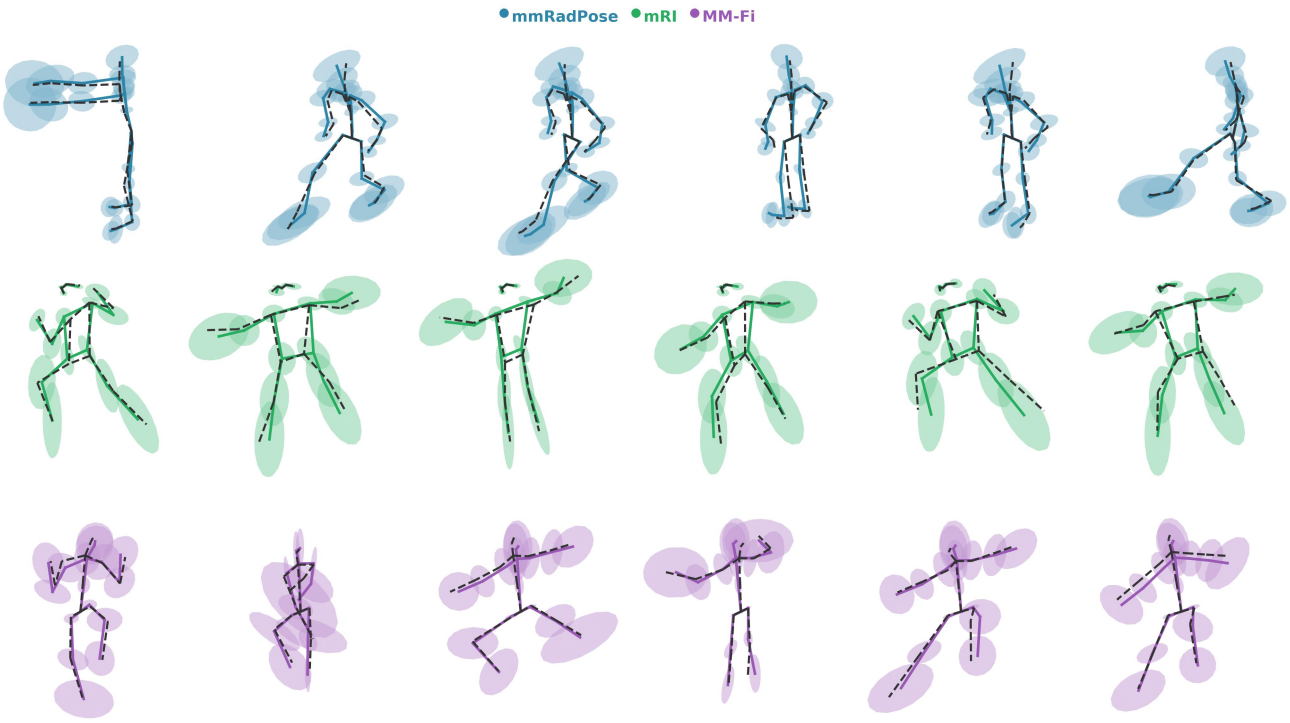

Figure 5. Qualitative results of MH-NFPG across mmRadPose, mRI, and MM-Fi. Predicted poses with per-joint 90% confidence ellipsoids consistently contain the ground truth, confirming calibrated uncertainty across diverse poses and radar configurations.

namically expresses this variation in its per-joint uncertainties, capturing meaningful physical structure of radar observability rather than arbitrary noise, as additionally visualized by the confidence ellipsoids in Figures 1 and 3 of the main paper. Furthermore, the overall per-joint sharpness is notably lower on mmRadPose, which uses clean optical motion capture ground truth, compared to mRI and MM-Fi, where ground truth is obtained from a vision-based pose estimator. This indicates that the model additionally adapts its uncertainty to reflect the precision of the supervision signal.

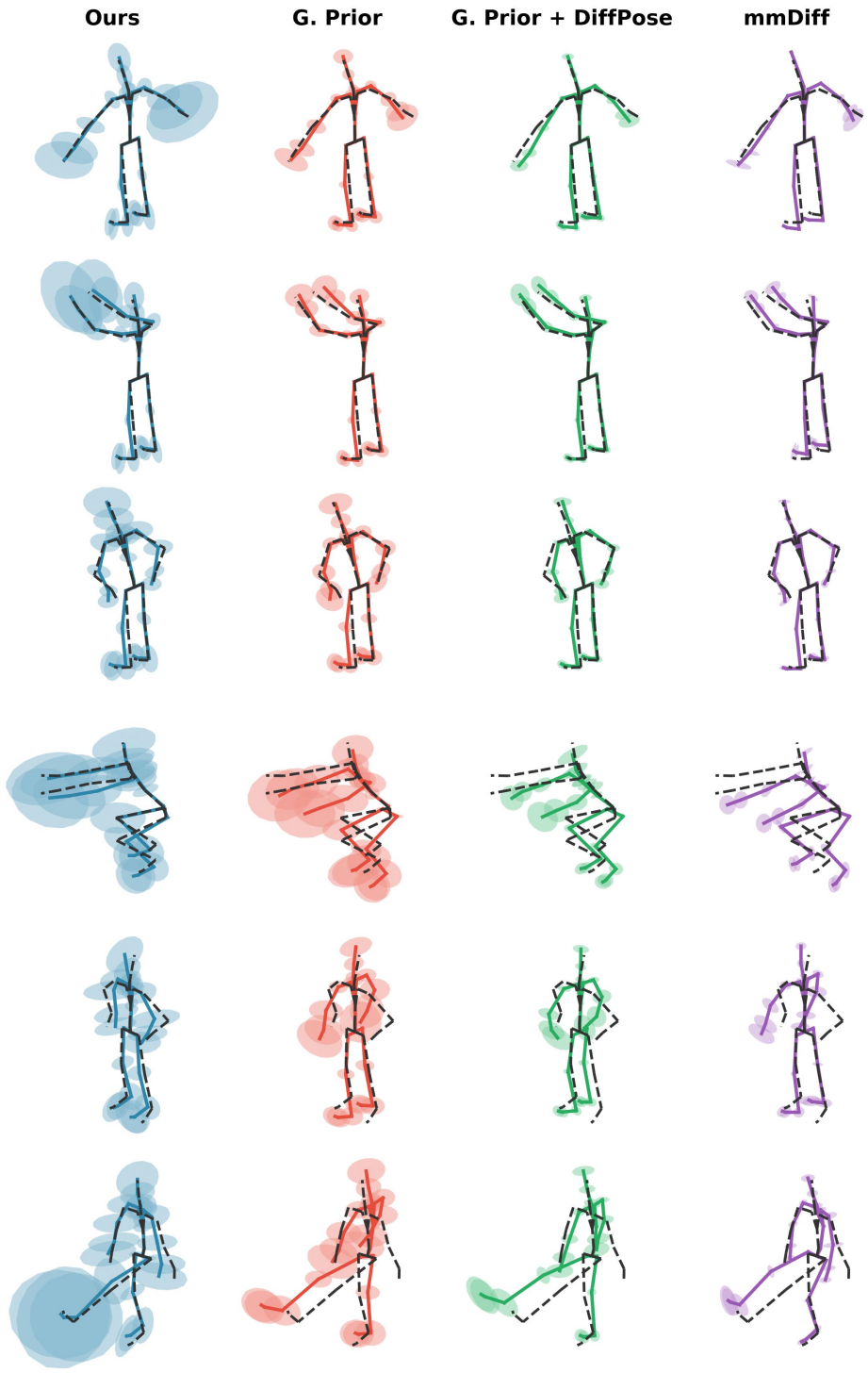

Figure 6. Extended qualitative comparison on mmRadPose. Predictions (black) and ground truth (red) with 90% confidence ellipsoids for MH-NFPG, DiffPose, and mmDiff. The first three rows show easy examples, while the last three rows show ambiguous examples with occlusions and more complex multipath effects. MH-NFPG consistently produces well-calibrated confidence regions.

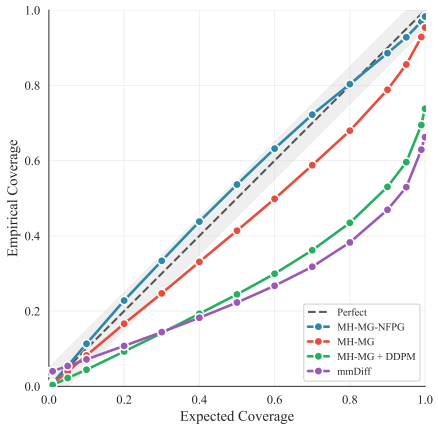

Figure 7. Reliability diagram on mmRadPose. MH-NFPG (blue) closely follows the perfect calibration diagonal (dashed), while diffusion baselines exhibit severe undercoverage.

Table 10. Per-joint sharpness – mmRadPose (31 835 test frames).

Joint	Sharpness (cm)
RightUpLeg	0.699
RightLeg	3.067
RightFoot	4.842
RightToeBase	5.050
RightToeEnd	5.243
LeftUpLeg	0.657
LeftLeg	3.190
LeftFoot	4.930
LeftToeBase	5.691
LeftToeEnd	5.368
Spine	0.409
Spine1	2.045
RightShoulder	3.656
RightArm	3.793
RightForeArm	4.468
RightHand	5.975
RightHandEnd	7.386
LeftShoulder	3.447
LeftArm	3.889
LeftForeArm	4.492
LeftHand	6.848
LeftHandEnd	9.409
Neck	4.043
Head	4.574
HeadEnd	6.166

Table 11. Per-joint sharpness – mRI (26 511 test frames) and MM-Fi (62 424 test frames).

	mRI		MM-Fi
Joint	Sharpness (cm)	Joint	Sharpness (cm)
L_Eye	1.006	R_Hip	1.661
R_Eye	1.058	R_Knee	3.771
L_Ear	2.562	R_Ankle	6.283
R_Ear	2.329	L_Hip	1.600
L_Shoulder	3.596	L_Knee	3.694
R_Shoulder	3.604	L_Ankle	5.795
L_Elbow	5.331	Spine	2.840
R_Elbow	5.335	Thorax	5.679
L_Wrist	8.019	Neck	6.645
R_Wrist	8.050	Head	6.836
L_Hip	5.204	L_Shoulder	5.907
R_Hip	5.173	L_Elbow	6.775
L_Knee	8.548	L_Wrist	8.960
R_Knee	8.121	R_Shoulder	5.898
L_Ankle	10.493	R_Elbow	6.585
R_Ankle	10.837	R_Wrist	8.343