BEYOND TOKEN ALIGNMENT: HUMAN-CENTERED POSTERIOR LEARNING FOR MULTIMODAL HUMAN SENSING

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal human sensing integrates heterogeneous sensors such as RGB cameras, depth sensors, LiDAR, mmWave radar, WiFi CSI, and RFID for robust and privacy-preserving human perception. Existing fusion methods typically convert sensor observations into modality tokens and then align or aggregate them in a shared embedding space. We argue that this token-centered view overlooks the central structure of human sensing: different sensors observe the same human state through distinct physical mechanisms, while their measurements are entangled with environment factors, sensing noise, and modality-dependent uncertainty. The shared object across modalities is therefore not token correspondence, but the underlying human state. Based on this insight, we propose **HUPER**, a **Hu**man-centered uncertainty-aware **P**osterior **E**stimation and **R**efinement framework. HUPER first uses modality-aware encoders to preserve sensor-specific inductive biases, estimates a probabilistic human posterior for each modality, separates human content from environment factors, and aggregates available modalities through reliability-calibrated posterior precision. Since the fused posterior may still contain residual uncertainty, HUPER further refines it in the human state space with an uncertainty-conditioned vector field that transports the consensus posterior toward a cleaner human representation. Experiments on MM-Fi and XRF55 demonstrate strong performance across pose estimation, activity recognition, missing-modality, and heterogeneous RF settings. On MM-Fi HPE, HUPER-Flow reduces full-modality MPJPE by **22.7%** over X-Fi and reduces partial-modality average MPJPE by **15.9%**. On MM-Fi HAR, it improves full-modality accuracy over X-Fi by **13.3** percentage points. On XRF55 HAR, HUPER-Flow achieves **97.1%** full-modality accuracy, improving over X-Fi by **7.3** percentage points, and improves the dual-modality average by **13.7** percentage points. These results show that aligning heterogeneous sensors through human-centered posterior learning provides a more reliable foundation than direct token alignment for multimodal human sensing.

1 INTRODUCTION

Human sensing aims to infer human activities, poses, and behaviors from sensor observations, supporting applications in smart homes, health monitoring, rehabilitation, human computer interaction, robotics, and augmented reality (Vrigkas et al., 2015; Nanzer, 2017; Nirmal et al., 2021). While vision based methods provide rich appearance and geometry cues, they are sensitive to illumination, occlusion, field of view, and privacy concerns (Nguyen et al., 2016; Elharrouss et al., 2021; Toshev & Szegedy, 2014; Sun et al., 2019). Non-intrusive sensors such as LiDAR, mmWave radar, WiFi CSI, and RFID offer complementary sensing properties, making multimodal sensing a promising direction for robust human perception (Li et al., 2022; Ren et al., 2024; An & Ogras, 2022; Yang et al., 2023a; Wang et al., 2024).

Yet multimodal human sensing faces a fundamental question often obscured by model design: ***what should be aligned across modalities?*** Existing methods commonly extract modality-specific tokens and combine them through concatenation, attention, cross-attention, transformer fusion, decision aggregation, or modality-invariant learning (Vaswani et al., 2017; Lu et al., 2019; Xue & Marculescu,

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

2023; Bruce et al., 2021; Chen & Yang, 2025). These approaches implicitly assume comparable token semantics. In heterogeneous human sensing, however, RGB pixels, LiDAR points, mmWave reflections, WiFi CSI variations, and RFID phase responses are distinct physical observations of the same latent human state, each affected by environment propagation, reflection, occlusion, multipath effects, and hardware noise (Nanzer, 2017; Nirmal et al., 2021; Yang et al., 2023a; Wang et al., 2024).

As illustrated in Fig. 1, directly aligning such tokens may amplify modality gaps, environment entanglement, and noise sensitivity. This leads to our central insight: **the common space of multimodal human sensing should be the human state space, not the token space**. Each modality should first infer a human-centered posterior over the underlying state, followed by posterior aggregation across available modalities.

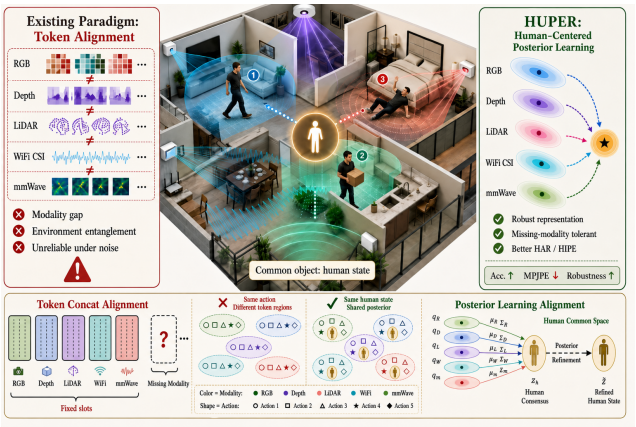


Figure 1: The human state is the common space. RGB, depth, LiDAR, WiFi CSI, and mmWave observe the same human event through different physical mechanisms, making direct token alignment unreliable. HUPER maps each modality to a human-centered posterior and performs fusion in the human state space, reducing modality gaps and environment entanglement for robust activity recognition and pose estimation.

This view introduces three challenges: preserving modality-specific inductive biases across images, point clouds, and RF signals (He et al., 2016; Dosovitskiy, 2020; Qi et al., 2017; Zhao et al., 2021); separating human content from environment and sensor-dependent factors; and aggregating modality estimates according to sample-level reliability and uncertainty before refining the uncertain consensus.

We address these challenges with **HUPER**, a human-centered uncertainty-aware posterior estimation and refinement framework. Modality-aware encoders map heterogeneous observations to a unified feature interface while preserving sensor-specific structure. HUPER then decomposes each modality into a human content posterior and an environment latent, estimates uncertainty and reliability, and aggregates available posteriors through reliability-calibrated precision. This formulation makes refinement a continuation of posterior estimation rather than an independent generative step. Instead of generating from pure noise, Human Latent Flow learns an uncertainty-conditioned vector field that transports the consensus posterior toward a cleaner human representation. We also consider a DDPM residual variant that removes uncertainty-scaled corruption in the human latent space. The vector-field formulation better matches our posterior-correction view and serves as the default model.

HUPER differs from existing multimodal fusion in both motivation and mechanism. Token fusion combines heterogeneous features, whereas HUPER infers the shared human state they imply (Ma et al., 2022; Dai et al., 2024; Han et al., 2024; Zhang et al., 2023; Girdhar et al., 2023). Modality-balancing methods treat unreliable inputs mainly as an optimization issue, whereas HUPER models reliability as a posterior property for probabilistic consensus (Xue et al., 2022; Wang et al., 2020). Finally, rather than inheriting a full generative sampling pipeline, HUPER uses refinement only to correct uncertain posteriors in the human state space.

Our contributions are summarized as follows:

- We revisit multimodal human sensing from an alignment perspective and argue that heterogeneous sensors should be aligned through the underlying human state rather than raw modality tokens.

- We propose HUPER, a human-centered posterior learning framework with modality-aware encoders that preserve sensor-specific inductive biases while projecting heterogeneous observations into a unified posterior estimation interface.
- We introduce a reliability-calibrated posterior consensus mechanism that separates human content from environment factors, estimates sample-level uncertainty and reliability, and aggregates modality posteriors through precision weighting.
- We introduce uncertainty-aware posterior refinement in the human state space. The default realization estimates an uncertainty-conditioned vector field for posterior transport, while a residual perturbation variant provides a complementary noise-removal view of the same correction problem.

2 Related Work

2.1 Multimodal Human Sensing

Human sensing has been studied with visual, wearable, and radio-frequency sensors (Vrigkas et al., 2015; Nanzer, 2017; Nirmal et al., 2021; Das et al., 2021; Marcard et al., 2016). Vision provides rich appearance and geometry but remains sensitive to illumination, occlusion, and privacy concerns (Nguyen et al., 2016; Elharrouss et al., 2021; Shahroudy et al., 2016; Pavlakos et al., 2018; Goel et al., 2023; Cai et al., 2023). WiFi CSI, mmWave radar, RFID, and LiDAR provide complementary cues but suffer from low resolution, multipath propagation, environment sensitivity, or device noise (Zhao et al., 2018; Zhou et al., 2023; 2022; D. Gian et al., 2024; An & Ogras, 2022; Wang et al., 2021; Li et al., 2022; Ren et al., 2024; Wang et al., 2024). Existing methods combine extracted features through feature, decision, attention, or modality-invariant fusion (Zheng et al., 2022; An et al., 2022; Chen et al., 2023; Man et al., 2023; Xue & Marculescu, 2023; Ma et al., 2022; Chen & Yang, 2025). HUPER instead asks what should be aligned, arguing that heterogeneous sensor tokens should be connected through the underlying human state.

2.2 Wireless Human Perception

WiFi and RF sensing offer privacy-preserving, contact-free, and occlusion-resilient human perception (Nanzer, 2017; Nirmal et al., 2021; Yang et al., 2023a). Prior work studies activity and gesture recognition, localization, vital-sign monitoring, and pose estimation (Zhao et al., 2018; Zhou et al., 2023; 2022; D. Gian et al., 2024; Wang et al., 2021; An & Ogras, 2021; 2022; Wang et al., 2024). However, wireless signals are indirect body observations entangled with environment layout, reflection, multipath propagation, attenuation, and device noise. HUPER therefore maps sensor observations to a human-centered probabilistic representation before multimodal fusion.

2.3 Uncertainty and Posterior Refinement

A key challenge in multimodal learning is that modalities differ in informativeness across samples (Ma et al., 2022; Xue et al., 2022; Wang et al., 2020). Existing methods address this imbalance through contribution estimation, gradient modulation, auxiliary unimodal heads, or optimization regularization (Xue & Marculescu, 2023; Bruce et al., 2021; Dai et al., 2024). HUPER instead models reliability as a sample-level posterior property and aggregates modalities through reliability-calibrated precision. It further casts refinement as uncertainty-conditioned correction in the human state space, using either vector-field transport or uncertainty-scaled residual removal. Thus, refinement extends posterior learning rather than acting as an independent generative process.

3 Method

3.1 Overview

HUPER formulates multimodal fusion as human-centered posterior estimation rather than token alignment. For available modalities $\mathcal{A}$, it encodes and decomposes each observation into a human posterior and an environment latent, forms a reliability-calibrated consensus, and refines it for downstream tasks. The pipeline is shown in Fig. 2.

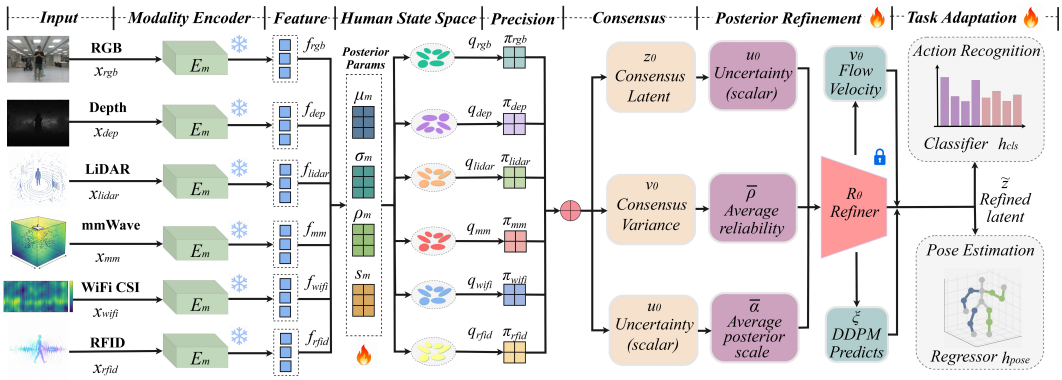


Figure 2: Overview of HUPER. The consensus module aggregates modality posteriors into a human latent $\mathbf{z}_0$, consensus variance $\mathbf{v}_0$, and scalar uncertainty u_0. The posterior refinement module R_θ uses u_0, average reliability $\bar{\rho}$, and average posterior scale $\bar{a}$ as conditions to refine the uncertain consensus into $\tilde{\mathbf{z}}$. R_θ can be instantiated as DDPM-style noise prediction or flow-matching posterior transport. In our final model, HUPER-Flow directly predicts a velocity field toward a cleaner human-state representation.

3.2 Human Posterior Alignment

Let $\mathcal{M}$ denote the set of candidate modalities, such as RGB, depth, LiDAR, mmWave radar, WiFi CSI, and RFID. For a sample, let $\mathcal{A} \subseteq \mathcal{M}$ be the available modality subset. The multimodal observation is $\mathbf{X} = \{\mathbf{x}_m\}_{m\in\mathcal{A}}$, where $\mathbf{x}_m$ is the raw observation from modality m. Our goal is to learn a human-centered latent representation that is robust to heterogeneous sensing mechanisms, unreliable observations, and missing modalities.

Instead of directly fusing modality tokens, HUPER models each modality as a probabilistic estimator of a shared human content variable $\mathbf{c} \in \mathbb{R}^{d_c}$:

$$q_m(\mathbf{c} \mid \mathbf{x}_m) = \mathcal{N}\left(\boldsymbol{\mu}_m, \mathrm{diag}(\boldsymbol{\sigma}_m^2)\right), \tag{1}$$

where q_m is the modality posterior, $\boldsymbol{\mu}_m \in \mathbb{R}^{d_c}$ is the posterior mean, and $\boldsymbol{\sigma}_m \in \mathbb{R}_+^{d_c}$ is the posterior standard deviation. Each modality also predicts an environment latent $\mathbf{s}_m \in \mathbb{R}^{d_s}$ and a scalar reliability score $\rho_m \in (0, 1)$. The posterior represents human-related semantics, while $\mathbf{s}_m$ captures modality and environment factors.

For human activity recognition, the final representation is mapped to a class distribution over K categories. For human pose estimation, it is mapped to a 3D pose $\mathbf{P} \in \mathbb{R}^{3\times J}$, where J is the number of body joints.

3.3 Modality-Aware Encoding and Decomposition

For each modality $m \in \mathcal{A}$, a modality encoder E_m maps the raw observation to a feature vector:

$$\mathbf{f}_m = E_m(\mathbf{x}_m), \tag{2}$$

where $\mathbf{f}_m \in \mathbb{R}^{d_f}$ is the encoded modality feature. HUPER uses lightweight modality-aware encoders: convolutional encoders process image-like RGB, depth, WiFi CSI, and dense RF inputs, while point-wise MLPs with masked mean and max pooling process LiDAR and mmWave point sets. All encoders map their inputs to the common dimension d_f, preserving sensor-specific inductive biases while providing a unified interface for posterior decomposition. Architectural details are provided in the Appendix.

The feature $\mathbf{f}_m$ is decomposed into human posterior parameters and an environment latent:

$$\boldsymbol{\mu}_m = g_m^{\mu}(\mathbf{f}_m), \quad \boldsymbol{\ell}_m = g_m^{\ell}(\mathbf{f}_m), \quad \mathbf{s}_m = g_m^{s}(\mathbf{f}_m), \tag{3}$$

where g_m^{μ}, g_m^{ℓ}, and g_m^{s} are prediction heads, and $\boldsymbol{\ell}_m \in \mathbb{R}^{d_c}$ is the log variance. The posterior standard deviation is

$$\boldsymbol{\sigma}_m = \exp\left(\frac{1}{2}\,\mathrm{clip}(\boldsymbol{\ell}_m, \ell_{\min}, \ell_{\max})\right), \tag{4}$$

where $\ell_{\min}$ and $\ell_{\max}$ are constants used for numerical stability. During training, the human content sample is obtained by reparameterization:

$$\mathbf{c}_m = \boldsymbol{\mu}_m + \boldsymbol{\sigma}_m \odot \boldsymbol{\eta}_m, \quad \boldsymbol{\eta}_m \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{d_c}), \tag{5}$$

where $\boldsymbol{\eta}_m$ is standard Gaussian noise and $\odot$ denotes element-wise multiplication. At inference, $\boldsymbol{\mu}_m$ is used as the modality estimate of human content.

The reliability score is predicted as

$$\rho_m = \mathrm{sigmoid}\left(-a_m + g_m^{\rho}([\mathbf{f}_m; \boldsymbol{\mu}_m; \mathbf{s}_m])\right), \tag{6}$$

where g_m^{ρ} is a reliability head, $[\cdot;\cdot]$ denotes concatenation, and

$$a_m = \frac{1}{d_c}\sum_{j=1}^{d_c} \sigma_{m,j} \tag{7}$$

is the average posterior scale of modality m. This form gives a useful inductive bias: a modality with high posterior uncertainty tends to receive lower reliability unless the learned residual predicts otherwise.

To discourage degenerate decomposition, we reconstruct the modality feature from human content and environment latent:

$$\hat{\mathbf{f}}_m = g_m^{\mathrm{rec}}([\mathbf{c}_m; \mathbf{s}_m]), \tag{8}$$

where g_m^{rec} is a reconstruction head. This path encourages $\mathbf{s}_m$ to retain modality and environment factors, while $\boldsymbol{\mu}_m$ focuses on human semantics shared across modalities.

3.4 Posterior Calibration

The modality posteriors are trained with a human-oriented calibration objective. In the main paper, we group the objective into four terms:

$$\begin{aligned}\mathcal{L}_{\mathrm{HUPER}} =& \lambda_{\mathrm{align}}\mathcal{L}_{\mathrm{align}} + \lambda_{\mathrm{dist}}\mathcal{L}_{\mathrm{dist}} \\ &+ \lambda_{\mathrm{rec}}\mathcal{L}_{\mathrm{rec}} + \lambda_{\mathrm{cal}}\mathcal{L}_{\mathrm{cal}},\end{aligned} \tag{9}$$

where each λ is nonnegative. $\mathcal{L}_{\mathrm{align}}$ aligns same-sample posterior means while separating different samples, and $\mathcal{L}_{\mathrm{dist}}$ aligns posterior means and scales to preserve uncertainty. $\mathcal{L}_{\mathrm{rec}}$ and an orthogonality regularizer encourage complementary content and environment factors, while $\mathcal{L}_{\mathrm{cal}}$ calibrates reliability and regularizes posterior variance. Full definitions are provided in the Appendix.

3.5 Reliability-Calibrated Human Consensus

After each modality predicts a posterior over the human state, HUPER aggregates available modalities by posterior precision. For modality m, the reliability-calibrated precision vector is

$$\boldsymbol{\pi}_m = \frac{\rho_m}{\boldsymbol{\sigma}_m^2 + \varepsilon}, \tag{10}$$

where $\boldsymbol{\pi}_m \in \mathbb{R}_+^{d_c}$ and $\varepsilon > 0$ is a numerical stabilizer. The consensus human latent is

$$\mathbf{z}_0 = \frac{\sum_{m\in\mathcal{A}} \boldsymbol{\pi}_m \odot \boldsymbol{\mu}_m}{\sum_{m\in\mathcal{A}} \boldsymbol{\pi}_m + \varepsilon}. \tag{11}$$

The consensus variance and scalar uncertainty are

$$\mathbf{v}_0 = \frac{1}{\sum_{m\in\mathcal{A}} \boldsymbol{\pi}_m + \varepsilon}, \quad u_0 = \frac{1}{d_c}\sum_{j=1}^{d_c}\sqrt{v_{0,j}}. \tag{12}$$

Table 1: MM-Fi HPE results under different modality subsets. Errors are reported in millimeters. "Baseline1" denotes the decision-level fusion results, while "Baseline2" denotes the feature-level fusion results. Imp.$_{\mathrm{X}}$, Imp.$_{\mathrm{B1}}$, and Imp.$_{\mathrm{B2}}$ denote relative error reductions over X-Fi, Baseline1, and Baseline2, respectively.

Modality	MPJPE ↓								PA-MPJPE ↓							
	Flow	**DDPM**	X-Fi	Imp.$_{\mathrm{X}}$	Baseline1	Imp.$_{\mathrm{B1}}$	Baseline2	Imp.$_{\mathrm{B2}}$	**Flow**	**DDPM**	X-Fi	Imp.$_{\mathrm{X}}$	Baseline1	Imp.$_{\mathrm{B1}}$	Baseline2	Imp.$_{\mathrm{B2}}$
I+D	**82.4**	88.0	86.1	+4.3%	96.7	+14.8%	109.2	+24.6%	**39.1**	41.4	48.1	+18.8%	53.9	+27.5%	57.1	+31.6%
I+L	**67.7**	70.4	93.0	+27.2%	122.1	+44.5%	121.2	+44.1%	**37.9**	40.7	59.7	+36.5%	75.6	+49.9%	73.0	+48.1%
I+R	**86.5**	112.8	88.8	+2.6%	101.0	+14.4%	140.6	+38.5%	**38.1**	40.7	57.3	+33.5%	57.3	+33.5%	68.2	+44.1%
I+W	**92.9**	101.3	93.0	+0.1%	146.4	+36.5%	156.7	+40.7%	**39.1**	41.4	59.5	+34.2%	77.3	+49.4%	86.3	+54.7%
D+L	**75.5**	87.8	102.5	+26.4%	111.7	+32.4%	108.0	+30.1%	**47.3**	58.6	48.4	+2.3%	68.8	+31.2%	55.8	+15.2%
D+R	**91.8**	132.0	98.0	+6.3%	94.1	+2.4%	116.4	+21.1%	**46.2**	74.7	47.3	+2.3%	51.8	+10.8%	57.0	+18.9%
D+W	**98.9**	142.5	101.8	+2.8%	141.7	+30.2%	155.5	+36.4%	**46.9**	87.3	48.1	+2.5%	71.4	+34.3%	81.5	+42.5%
L+R	**77.8**	91.3	109.8	+29.1%	116.4	+33.1%	137.0	+43.2%	**45.7**	57.2	63.4	+28.0%	71.5	+36.1%	71.9	+36.5%
L+W	**79.8**	92.5	159.5	+50.0%	167.1	+52.2%	206.2	+61.3%	**49.6**	61.0	102.7	+51.7%	100.7	+50.7%	109.3	+54.6%
R+W	152.9	169.4	**117.2**	-30.5%	144.2	-6.0%	141.3	-8.2%	80.2	86.6	**62.7**	-28.0%	72.2	-11.1%	77.6	-3.4%
I+D+L	**65.1**	66.0	84.8	+23.2%	102.9	+36.7%	104.5	+37.7%	**38.1**	40.8	48.2	+20.9%	62.3	+38.8%	61.2	+37.7%
I+D+R	**81.0**	86.7	83.4	+2.9%	86.7	+6.6%	113.4	+28.6%	**38.2**	40.8	47.3	+19.3%	50.1	+23.8%	58.0	+34.2%
I+D+W	**80.5**	85.6	85.3	+5.6%	118.6	+32.1%	137.8	+41.6%	**39.2**	41.5	48.1	+18.5%	63.7	+38.5%	75.3	+47.9%
I+L+R	**67.2**	70.2	88.4	+24.0%	101.4	+33.7%	135.0	+50.2%	**37.4**	40.3	57.2	+34.7%	62.7	+40.4%	67.7	+44.8%
I+L+W	**65.9**	68.0	93.0	+29.2%	134.9	+51.2%	156.2	+57.8%	**37.9**	40.8	59.7	+36.4%	82.0	+53.7%	84.8	+55.2%
I+R+W	**86.7**	99.0	88.5	+2.0%	116.5	+25.6%	128.9	+32.7%	**38.2**	40.8	57.1	+33.1%	63.1	+39.5%	75.0	+49.1%
D+L+R	**72.9**	84.9	96.0	+24.1%	95.3	+23.5%	111.3	+34.5%	**45.2**	56.1	47.3	+4.4%	58.6	+22.9%	58.4	+22.6%
D+L+W	**76.5**	88.3	102.0	+25.0%	130.7	+41.5%	154.6	+50.5%	**47.9**	59.5	48.1	+0.3%	78.1	+38.6%	84.1	+43.0%
D+R+W	104.6	129.3	**97.0**	-7.8%	113.2	+7.6%	132.0	+20.8%	59.6	76.1	**47.1**	-26.6%	59.4	-0.4%	71.8	+17.0%
L+R+W	**76.3**	89.0	107.4	+29.0%	129.6	+41.1%	141.6	+46.1%	**46.7**	58.0	63.1	+26.0%	78.1	+40.2%	77.0	+39.3%
I+D+L+R	**64.4**	65.5	83.5	+22.9%	90.7	+29.0%	109.4	+41.2%	**37.6**	40.4	47.6	+21.0%	55.8	+32.6%	61.6	+38.9%
I+D+L+W	**65.4**	66.0	86.0	+23.9%	116.7	+43.9%	134.6	+51.4%	**38.2**	40.9	48.2	+20.7%	70.0	+45.4%	75.4	+49.3%
I+D+R+W	**79.2**	84.1	84.0	+5.8%	101.9	+22.3%	128.6	+38.5%	**38.3**	40.9	47.6	+19.5%	56.3	+32.0%	71.3	+46.3%
I+L+R+W	**65.1**	67.4	88.6	+26.5%	114.0	+42.9%	174.6	+62.7%	**37.4**	40.4	57.1	+34.4%	69.5	+46.1%	78.1	+52.1%
D+L+R+W	**73.9**	85.2	97.6	+24.3%	110.8	+33.3%	147.6	+50.0%	**45.8**	56.9	47.4	+3.3%	66.7	+31.3%	73.1	+37.3%
I+D+L+R+W	**64.7**	65.3	83.7	+22.7%	103.0	+37.2%	128.9	+49.8%	**37.7**	40.6	47.6	+20.8%	62.4	+39.6%	72.5	+48.0%

Here, $\mathbf{v}_0 \in \mathbb{R}_+^{d_c}$ is the element-wise consensus variance, and $u_0 \in \mathbb{R}_+$ is the scalar consensus uncertainty. We also define the mean reliability and mean posterior scale as

$$\bar{\rho} = \frac{1}{|\mathcal{A}|} \sum_{m \in \mathcal{A}} \rho_m, \quad \bar{a} = \frac{1}{|\mathcal{A}|} \sum_{m \in \mathcal{A}} a_m. \tag{13}$$

They are used as conditions for latent refinement.

This consensus is invariant to modality order and works with any available modality subset. A reliable and confident modality has high precision and contributes more to the human latent. An uncertain or unreliable modality has low precision and is suppressed. Appendix derives Eq. equation 11 from Gaussian posterior aggregation, and Appendix shows that it minimizes a reliability-weighted posterior disagreement objective.

3.6 Posterior Refinement in the Human State Space

The consensus latent $\mathbf{z}_0$ is a probability-guided estimate of the human state. It may still contain residual uncertainty due to sparse observations, wireless interference, occlusion, or missing modalities. HUPER therefore treats refinement as posterior correction in the human state space. The purpose is not to generate a new latent from noise, but to move an already meaningful human estimate toward a cleaner and more reliable one.

Human Latent Flow. Our default refinement module adapts Flow Matching (Lipman et al., 2022) to the human latent space, which we call Human Latent Flow. Given the uncertain source latent $\mathbf{z}_{\mathrm{s}} = \mathbf{z}_0$, we construct a cleaner target latent $\mathbf{z}_{\mathrm{c}}$ from a stronger posterior view, such as full-modality consensus or a less-corrupted augmented posterior. The target is detached from the refinement loss. For a transport time $\tau \sim \mathcal{U}(0, 1)$, we define

$$\mathbf{z}_\tau = (1 - \tau)\mathbf{z}_{\mathrm{s}} + \tau \mathbf{z}_{\mathrm{c}}, \quad \mathbf{v}^\star = \mathbf{z}_{\mathrm{c}} - \mathbf{z}_{\mathrm{s}}, \tag{14}$$

where $\mathbf{z}_\tau \in \mathbb{R}^{d_c}$ is an interpolated human latent and $\mathbf{v}^\star \in \mathbb{R}^{d_c}$ is the target transport velocity.

The posterior refiner predicts an uncertainty-conditioned vector field:

$$\hat{\mathbf{v}}_\theta = R_\theta(\mathbf{z}_\tau, \psi(\tau), u_0, \bar{\rho}, \bar{a}), \tag{15}$$

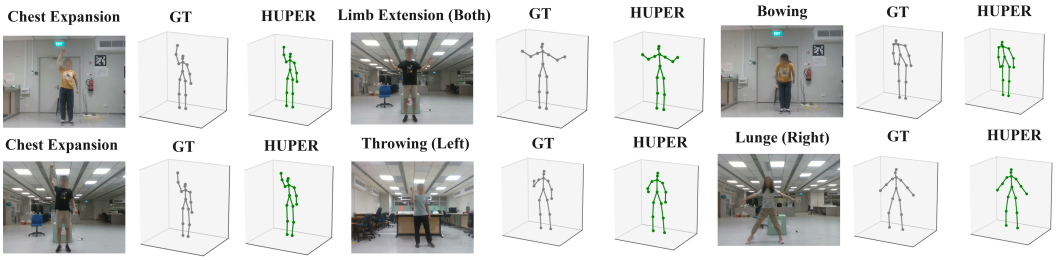


Figure 3: Qualitative comparison between GT and HUPER predictions on multiple actions.

where R_θ is the refiner, $\psi(\tau)$ is a time embedding, u_0 is the scalar consensus uncertainty, $\bar{\rho}$ is the mean reliability, and $\bar{a}$ is the mean posterior scale. The Flow Matching objective is

$$\mathcal{L}_{\text{flow}} = \|\hat{\mathbf{v}}_\theta - \mathbf{v}^\star\|_2^2 . \tag{16}$$

This objective directly trains R_θ to estimate how an uncertain human posterior should move toward a cleaner human posterior under the current uncertainty and reliability conditions.

At inference, HUPER is initialized with $\mathbf{z}^0 = \mathbf{z}_0$ and applies a small number of Euler updates:

$$\mathbf{z}^{k+1} = \mathbf{z}^k + \Delta\tau\, g(u_0) R_\theta(\mathbf{z}^k, \psi(\tau_k), u_0, \bar{\rho}, \bar{a}), \tag{17}$$

where $k \in \{0, \ldots, K_{\text{flow}} - 1\}$, $\Delta\tau = 1/K_{\text{flow}}$, and $g(u_0) = \text{clip}(u_0, 0, 1)$ is an uncertainty gate. The refined latent is

$$\tilde{\mathbf{z}} = \mathbf{z}^{K_{\text{flow}}}, \tag{18}$$

where $\tilde{\mathbf{z}} \in \mathbb{R}^{d_c}$ is the refined human latent. The update is conservative when the consensus is confident and stronger when the posterior is uncertain.

Uncertainty-scaled residual correction. We also keep a residual perturbation variant to connect with the DDPM-style (Ho et al., 2020) noise prediction module used in HUPER-DDPM. This variant corrupts $\mathbf{z}_0$ with uncertainty-scaled Gaussian noise and trains a residual predictor to estimate the injected perturbation. At inference, the predicted residual is subtracted from $\mathbf{z}_0$ as a single posterior sharpening step. Unlike Human Latent Flow, this variant corrects the posterior indirectly through noise prediction rather than explicitly learning the semantic transport direction. The detailed objective is given in Appendix.

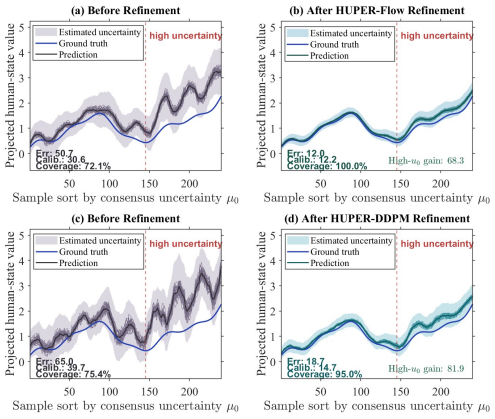


Figure 4: Uncertainty-aware refinement yields tighter uncertainty bands and more accurate predictions.

3.7 Task Adaptation and Structure Regularization

When pose annotations are available, we regularize the refined latent with body structure. A skeleton proxy head predicts 3D joints from $\tilde{\mathbf{z}}$:

$$\hat{\mathbf{P}}_{\text{pre}} = H_{\text{skel}}(\tilde{\mathbf{z}}) \in \mathbb{R}^{3\times J}, \tag{19}$$

where H_{skel} is the skeleton head and $\hat{\mathbf{P}}_{\text{pre}}$ is the proxy pose. We use a structure loss

$$\mathcal{L}_{\text{struct}} = \mathcal{L}_{\text{skel}} + \lambda_{\text{bone}} \mathcal{L}_{\text{bone}}, \tag{20}$$

where $\mathcal{L}_{\text{skel}}$ is a smooth L1 pose loss and $\mathcal{L}_{\text{bone}}$ is a bone length consistency loss. The full pretraining objective is

$$\mathcal{L}_{\text{pre}} = \mathcal{L}_{\text{HUPER}} + \beta_{\text{ref}}\mathcal{L}_{\text{ref}} + \beta_{\text{struct}}\mathcal{L}_{\text{struct}}, \tag{21}$$

where $\mathcal{L}_{\text{ref}}$ is $\mathcal{L}_{\text{flow}}$ for the vector-field refiner and $\mathcal{L}_{\text{diff}}$ for the residual perturbation refiner. The weights β_{ref} and β_{struct} are non-negative. If pose annotations are unavailable, the term is omitted. For downstream human activity recognition, $\tilde{\mathbf{z}}$ is passed to a classifier and optimized with cross-entropy loss. For human pose estimation, $\tilde{\mathbf{z}}$ is passed to a pose regressor and optimized with pose regression losses. Since Eq. equation 11 is defined for arbitrary $\mathcal{A}$, the same model can be evaluated under single-modality, partial-modality, and full-modality settings without architecture changes.

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate HUPER on MM-Fi (Yang et al., 2023b) and XRF55 (Wang et al., 2024) for HPE and HAR. MM-Fi provides RGB (I), depth (D), LiDAR (L), mmWave radar (R), and WiFi CSI (W), with 17-joint annotations for HPE and 27 action classes for HAR. XRF55 contains mmWave, WiFi, and RFID (RF) measurements across 55 action classes. Further details on dataset configurations, modality settings, and evaluation protocols are provided in the Appendix.

Table 2: HAR accuracy (%) on MM-Fi and XRF55. Imp._{X} and Imp._{B} denote the improvements over X-Fi and the baseline, respectively.

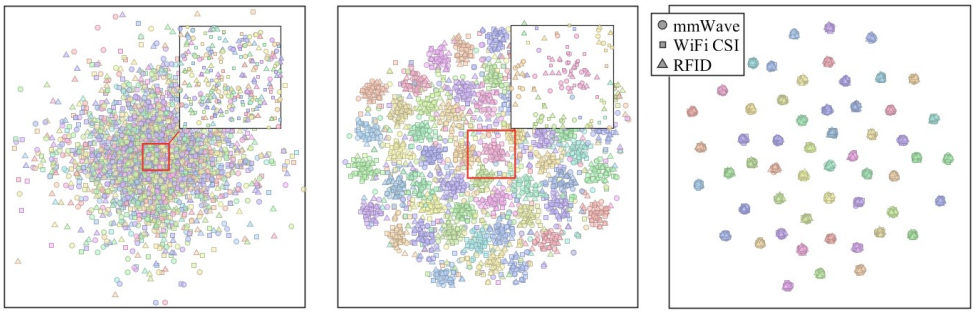
Modality	Flow	DDPM	X-Fi	$\text{Imp.}_{\text{X}}\uparrow$	Baseline	$\text{Imp.}_{\text{B}}\uparrow$
MM-Fi						
I+D	71.0	**71.8**	45.3	+25.7	45.2	+25.8
I+L	75.3	**79.1**	35.2	+40.1	23.8	+51.5
I+R	**88.0**	**88.0**	73.4	+14.6	66.9	+21.1
D+L	**53.2**	49.5	51.6	+1.6	49.7	+3.5
D+R	**81.4**	57.2	79.8	+1.6	74.1	+7.3
L+R	75.3	74.7	**88.7**	-13.4	85.6	-10.3
I+D+L	71.0	**75.6**	48.7	+22.3	43.0	+28.0
I+D+R	85.5	**85.7**	70.7	+14.8	69.1	+16.4
I+L+R	88.0	**88.4**	77.8	+10.2	68.2	+19.8
D+L+R	70.9	70.8	**80.5**	-9.6	72.9	-2.0
I+D+L+R	85.5	**86.0**	72.2	+13.3	72.6	+12.9
XRF55						
R+W	**96.6**	96.5	88.2	+8.4	86.8	+9.8
R+RF	**86.9**	82.6	86.5	+0.4	71.4	+15.5
W+RF	**90.4**	87.3	58.1	+32.3	55.6	+34.8
R+W+RF	**97.1**	96.7	89.8	+7.3	70.6	+26.5

Implementation details. HUPER-Flow and HUPER-DDPM share modality encoders, task heads, and a 256-dimensional posterior. Pretraining/downstream training uses AdamW for up to 50 epochs with cosine decay, learning rates of $3\times10^{-4}/10^{-3}$, weight decay of $5 \times 10^{-4}/10^{-4}$, and dropout of $0.1/0.2$. Random non-empty modality subsets are sampled during downstream training. Results are averaged over five independent seeds without best-run selection. All experiments use FP32 on an NVIDIA RTX 4090; the settings are provided in the Appendix.

4.2 Quantitative Results

Human pose estimation. HUPER-Flow achieves the best overall MM-Fi HPE performance in most modality settings (Table 1). With all five modalities, it obtains 64.7 mm MPJPE and 37.7 mm PA-MPJPE, reducing MPJPE by 22.7% over X-Fi. It also outperforms X-Fi in 9 of 10 dual-modality settings, with relative MPJPE reductions of 27.2% on I+L and 50.0% on L+W. These gains demonstrate the effectiveness of reliability-calibrated consensus in exploiting complementary geometric cues. The exception is R+W, where both modalities provide limited geometric information.

Human activity recognition. Consistent HAR improvements are observed on MM-Fi and XRF55 (Table 2). On MM-Fi, HUPER reaches 88.4% accuracy on I+L+R and improves I+L from 35.2% to 79.1%. On XRF55, HUPER-Flow achieves 96.6% on R+W and 97.1% with all modalities, outperforming X-Fi by 8.4 and 7.3 percentage points. The improvement from 58.1% to 90.4% on W+RF further demonstrates its effectiveness for heterogeneous RF sensing. Overall, the results show that human-centered posterior fusion remains effective across both visual–geometric and RF-only modality combinations.



(a) Baseline token feature (b) Consensus posterior z_0 (c) Refined human latent $\tilde{z}$

Figure 5: t-SNE visualization of latent spaces. HUPER forms compact, well-separated human-state clusters through posterior consensus z_0 and refinement $\tilde{z}$.

Table 3: Component ablation on XRF55 HAR. Single, Dual, and All denote average accuracy (%) over the corresponding modality subsets.

Method	Single	Dual	Full	All
X-Fi	60.7	77.6	89.8	73.5
HUPER w/o subset training	66.4	84.7	96.1	80.6
HUPER w/o env. disentanglement	68.9	86.2	96.0	82.1
HUPER w/o reliability	69.8	87.3	96.2	83.0
HUPER w/o posterior variance	70.6	88.1	96.3	83.8
HUPER with mean pooling	71.3	88.7	96.4	84.4
HUPER w/o refinement	72.1	89.6	96.5	85.1
HUPER-DDPM	73.8	91.4	96.7	86.8
HUPER-Flow	**76.2**	**93.1**	**97.1**	**88.9**

Table 4: Posterior refinement ablation on MM-Fi HPE and XRF55 HAR. We compare no refinement, DDPM residual correction, Flow without uncertainty conditioning, and the complete HUPER-Flow. Full and All Avg. denote full-modality performance and the cardinality-balanced average across modality-subset groups.

Refinement	MM-Fi Full ↓	MM-Fi All Avg. ↓	XRF55 Full ↑	XRF55 All Avg. ↑
No refinement	73.0	93.4	96.5	85.1
DDPM residual correction	70.8	88.7	96.7	86.8
Flow w/o uncertainty	68.6	84.2	97.0	87.7
HUPER-Flow	**64.7**	**79.6**	**97.1**	**88.9**

Qualitative results. Across diverse actions, HUPER preserves global body structure and major limb orientations, with errors mainly confined to fine-grained joints (Fig. 3). Uncertainty-aware refinement reduces prediction errors from 50.7 to 12.0 for HUPER-Flow and from 65.0 to 18.7 for HUPER-DDPM (Fig. 4). Meanwhile, posterior consensus and refinement transform entangled features into compact, well-separated human-state clusters (Fig. 5), demonstrating effective cross-modal alignment and refinement.

4.3 Ablation Studies

Each HUPER component contributes to performance across modality subsets (Table 3). The complete model achieves 88.9% all-subset accuracy on XRF55, compared with 73.5% for X-Fi. Replacing precision consensus with mean pooling or removing refinement reduces it to 84.4% and 85.1%, respectively. Uncertainty conditioning further reduces MM-Fi full/all-subset MPJPE from 68.6/84.2 to 64.7/79.6 mm and improves XRF55 all-subset accuracy from 87.7% to 88.9% (Table 4), validating uncertainty-aware posterior refinement.

5 Conclusion

We presented HUPER, a human-centered framework that aligns heterogeneous sensors through the underlying human state rather than modality tokens. HUPER separates human content from environment factors, forms a reliability-calibrated posterior consensus, and corrects residual uncertainty through human-state transport without generation from noise. Results on MM-Fi and XRF55 demonstrate robustness to heterogeneous and missing modalities, establishing human-centered posterior learning as a principled alternative to token alignment.

AI USE STATEMENT

We used generative AI tools to provide feedback on research methodology and experiments, assist with translation, and interpret results. These tools were not used to generate synthetic data; develop theoretical models or conceptual frameworks; formulate mathematical claims; provide essential components of mathematical proofs; write proofs; propose or refine hypotheses; design methodology or experiments; implement methods; clean or reformat data; or conduct qualitative or thematic analysis. The authors independently evaluated all methodological and experimental feedback, manually reviewed and revised all translations, and verified AI-assisted interpretations against the experimental results. We take full responsibility for all content, claims, and artifacts in this work.

REFERENCES

Sizhe An and Umit Y Ogras. Mars: mmwave-based assistive rehabilitation system for smart healthcare. *ACM Transactions on Embedded Computing Systems (TECS)*, 20(5s):1–22, 2021.

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*, pp. 889–894, 2022.

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*, 35: 27414–27426, 2022.

XB Bruce, Yan Liu, and Keith CC Chan. Multimodal fusion via teacher-student network for indoor action recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 3199–3207, 2021.

Zhongang Cai, Wanqi Yin, Ailing Zeng, Chen Wei, Qingping Sun, Wang Yanjun, Hui En Pang, Haiyi Mei, Mingyuan Zhang, Lei Zhang, et al. Smpler-x: Scaling up expressive human pose and shape estimation. *Advances in Neural Information Processing Systems*, 36:11454–11468, 2023.

Anjun Chen, Xiangyu Wang, Kun Shi, Shaohao Zhu, Bin Fang, Yingfeng Chen, Jiming Chen, Yuchi Huo, and Qi Ye. Immfusion: Robust mmwave-rgb fusion for 3d human body reconstruction in all weather conditions. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 2752–2758. IEEE, 2023.

Xinyan Chen and Jianfei Yang. X-fi: A modality-invariant foundation model for multimodal human sensing. In *International Conference on Learning Representations*, volume 2025, pp. 97378–97397, 2025.

Toan D. Gian, Tien Dac Lai, Thien Van Luong, Kok-Seng Wong, and Van-Dinh Nguyen. Hpe-li: Wifi-enabled lightweight dual selective kernel convolution for human pose estimation. In *European Conference on Computer Vision*, pp. 93–111. Springer, 2024.

Shenghong Dai, Shiqi Jiang, Yifan Yang, Ting Cao, Mo Li, Suman Banerjee, and Lili Qiu. Advancing multi-modal sensing through expandable modality alignment. *arXiv preprint arXiv:2407.17777*, 2024.

Avigyan Das, Pritam Sil, Pawan Kumar Singh, Vikrant Bhateja, and Ram Sarkar. Mmhar-ensemnet: A multi-modal human activity recognition model. *IEEE Sensors Journal*, 21(10):11569–11576, 2021. doi: 10.1109/JSEN.2020.3034614.

Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Omar Elharrouss, Noor Almaadeed, and Somaya Al-Maadeed. A review of video surveillance systems. *Journal of Visual Communication and Image Representation*, 77:103116, 2021.

Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15180–15190, 2023.

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

Jiaming Han, Kaixiong Gong, Yiyuan Zhang, Jiaqi Wang, Kaipeng Zhang, Dahua Lin, Yu Qiao, Peng Gao, and Xiangyu Yue. Onellm: One framework to align all modalities with language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26584–26595, 2024.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Jialian Li, Jingyi Zhang, Zhiyong Wang, Siqi Shen, Chenglu Wen, Yuexin Ma, Lan Xu, Jingyi Yu, and Cheng Wang. Lidarcap: Long-range marker-less 3d human motion capture with lidar point clouds. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 20502–20512, 2022.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, 2019.

Mengmeng Ma, Jian Ren, Long Zhao, Davide Testuggine, and Xi Peng. Are multimodal transformers robust to missing modality? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18177–18186, 2022.

Yunze Man, Liang-Yan Gui, and Yu-Xiong Wang. Bev-guided multi-modality fusion for driving perception. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21960–21969, 2023.

Timo von Marcard, Gerard Pons-Moll, and Bodo Rosenhahn. Human pose estimation from video and imus. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(8):1533–1547, 2016. doi: 10.1109/TPAMI.2016.2522398.

Jeffrey A Nanzer. A review of microwave wireless techniques for human presence detection and classification. *IEEE Transactions on Microwave Theory and Techniques*, 65(5):1780–1794, 2017.

Duc Thanh Nguyen, Wanqing Li, and Philip O Ogunbona. Human detection from images and videos: A survey. *Pattern Recognition*, 51:148–175, 2016.

Isura Nirmal, Abdelwahed Khamis, Mahbub Hassan, Wen Hu, and Xiaoqing Zhu. Deep learning for radio-based human sensing: Recent advances and future directions. *IEEE Communications Surveys & Tutorials*, 23(2):995–1019, 2021.

Georgios Pavlakos, Luyang Zhu, Xiaowei Zhou, and Kostas Daniilidis. Learning to estimate 3d human pose and shape from a single color image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 459–468, 2018.

Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 652–660, 2017.

Yiming Ren, Xiao Han, Chengfeng Zhao, Jingya Wang, Lan Xu, Jingyi Yu, and Yuexin Ma. Livehps: Lidar-based scene-level human pose and shape estimation in free environment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1281–1291, 2024.

Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1010–1019, 2016.

Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5693–5703, 2019.

Alexander Toshev and Christian Szegedy. Deeppose: Human pose estimation via deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1653–1660, 2014.

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.

Michalis Vrigkas, Christophoros Nikou, and Ioannis A Kakadiaris. A review of human activity recognition methods. *Frontiers in Robotics and AI*, 2:28, 2015.

Fei Wang, Yizhe Lv, Mengdie Zhu, Han Ding, and Jinsong Han. Xrf55: A radio frequency dataset for human indoor action analysis. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 8(1):1–34, 2024.

Qi Wang, Liang Zhan, Paul Thompson, and Jiayu Zhou. Multimodal learning with incomplete modalities by knowledge distillation. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 1828–1838, 2020.

Yuheng Wang, Haipeng Liu, Kening Cui, Anfu Zhou, Wensheng Li, and Huadong Ma. m-activity: Accurate and real-time human activity recognition via millimeter wave radar. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 8298–8302. IEEE, 2021.

Zihui Xue and Radu Marculescu. Dynamic multimodal fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2575–2584, 2023.

Zihui Xue, Zhengqi Gao, Sucheng Ren, and Hang Zhao. The modality focusing hypothesis: Towards understanding crossmodal knowledge distillation. *arXiv preprint arXiv:2206.06487*, 2022.

Jianfei Yang, Xinyan Chen, Han Zou, Chris Xiaoxuan Lu, Dazhuo Wang, Sumei Sun, and Lihua Xie. Sensefi: A library and benchmark on deep-learning-empowered wifi human sensing. *Patterns*, 4 (3), 2023a.

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. *Advances in Neural Information Processing Systems*, 36:18756–18768, 2023b.

Yiyuan Zhang, Kaixiong Gong, Kaipeng Zhang, Hongsheng Li, Yu Qiao, Wanli Ouyang, and Xiangyu Yue. Meta-transformer: A unified framework for multimodal learning. *arXiv preprint arXiv:2307.10802*, 2023.

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*, pp. 16259–16268, 2021.

Mingmin Zhao, Tianhong Li, Mohammad Abu Alsheikh, Yonglong Tian, Hang Zhao, Antonio Torralba, and Dina Katabi. Through-wall human pose estimation using radio signals. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7356–7365, 2018.

Jingxiao Zheng, Xinwei Shi, Alexander Gorban, Junhua Mao, Yang Song, Charles R Qi, Ting Liu, Visesh Chari, Andre Cornman, Yin Zhou, et al. Multi-modal 3d human pose estimation with 2d weak supervision in autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4478–4487, 2022.

Yue Zhou, Aichun Zhu, Caojie Xu, Fangqiang Hu, and Yifeng Li. Perunet: Deep signal channel attention in unet for wifi-based human pose estimation. *IEEE Sensors Journal*, 22(20):19750–19760, 2022.

Yunjiao Zhou, He Huang, Shenghai Yuan, Han Zou, Lihua Xie, and Jianfei Yang. Metafi++: Wifi-enabled transformer-based human pose estimation for metaverse avatar simulation. *IEEE Internet of Things Journal*, 10(16):14128–14136, 2023.

A Derivation and Theoretical Justification

The main paper presents HUPER from a modular perspective. This appendix gives the derivations and additional objectives that support the design. We show that reliability-calibrated consensus arises from Gaussian posterior aggregation, and that the refinement objectives can be understood as uncertainty adaptive posterior correction in the human state space.

A.1 Notation

The notation follows the main paper. The set $\mathcal{M}$ contains all candidate modalities. The subset $\mathcal{A} \subseteq \mathcal{M}$ contains modalities available for one sample. For modality m, $\mathbf{x}_m$ is the raw observation, $\mathbf{f}_m$ is the encoded feature, $\mathbf{c}$ is the latent human content variable, and $\mathbf{s}_m$ is the environment latent. The posterior estimated by modality m is a diagonal Gaussian $q_m(\mathbf{c} \mid \mathbf{x}_m) = \mathcal{N}(\boldsymbol{\mu}_m, \mathrm{diag}(\boldsymbol{\sigma}_m^2))$. The scalar ρ_m denotes reliability, and $\boldsymbol{\pi}_m$ denotes reliability-calibrated precision. The index $j \in \{1, \ldots, d_c\}$ denotes one latent dimension. All divisions between vectors are element wise unless stated otherwise.

A.2 Derivation of Reliability-Calibrated Consensus

Assume that each available modality provides a conditionally independent Gaussian estimate of the same human content variable. For one dimension j, modality m predicts

$$q_m(c_j \mid \mathbf{x}_m) = \mathcal{N}(\mu_{m,j}, \sigma_{m,j}^2), \tag{22}$$

where c_j is the jth dimension of $\mathbf{c}$, $\mu_{m,j}$ is the mean, and $\sigma_{m,j}^2$ is the variance. If all modality estimates are equally reliable, the product posterior is

$$q(c_j \mid \{\mathbf{x}_m\}_{m\in\mathcal{A}}) \propto \prod_{m\in\mathcal{A}} q_m(c_j \mid \mathbf{x}_m). \tag{23}$$

The product of Gaussian experts remains Gaussian. Its variance and mean are

$$v_j^\star = \left(\sum_{m\in\mathcal{A}} \frac{1}{\sigma_{m,j}^2} \right)^{-1}, \quad z_j^\star = v_j^\star \sum_{m\in\mathcal{A}} \frac{\mu_{m,j}}{\sigma_{m,j}^2}. \tag{24}$$

This shows that posterior fusion should be precision weighted. In practice, a modality may be over confident due to scene bias or sensor noise. HUPER therefore rescales precision by reliability:

$$\pi_{m,j} = \frac{\rho_m}{\sigma_{m,j}^2 + \varepsilon}. \tag{25}$$

Substituting Eq. equation 25 into the Gaussian product solution gives

$$z_{0,j} = \frac{\sum_{m\in\mathcal{A}} \pi_{m,j}\mu_{m,j}}{\sum_{m\in\mathcal{A}} \pi_{m,j} + \varepsilon}, \quad v_{0,j} = \frac{1}{\sum_{m\in\mathcal{A}} \pi_{m,j} + \varepsilon}. \tag{26}$$

Vectorizing Eq. equation 26 yields

$$\mathbf{z}_0 = \frac{\sum_{m\in\mathcal{A}} \boldsymbol{\pi}_m \odot \boldsymbol{\mu}_m}{\sum_{m\in\mathcal{A}} \boldsymbol{\pi}_m + \varepsilon}, \quad \mathbf{v}_0 = \frac{1}{\sum_{m\in\mathcal{A}} \boldsymbol{\pi}_m + \varepsilon}. \tag{27}$$

The scalar consensus uncertainty is

$$u_0 = \frac{1}{d_c} \sum_{j=1}^{d_c} \sqrt{v_{0,j}}. \tag{28}$$

A.3 Optimality of Precision Weighted Consensus

The consensus can also be derived as the minimizer of a reliability weighted disagreement objective. For each dimension j, define

$$\mathcal{J}_j(z) = \frac{1}{2} \sum_{m\in\mathcal{A}} \pi_{m,j}(z - \mu_{m,j})^2, \tag{29}$$

where $z \in \mathbb{R}$ is a candidate consensus value. Its derivative is

$$\frac{\partial \mathcal{J}_j(z)}{\partial z} = \sum_{m \in \mathcal{A}} \pi_{m,j}(z - \mu_{m,j}). \tag{30}$$

Setting the derivative to zero yields

$$z = \frac{\sum_{m \in \mathcal{A}} \pi_{m,j} \mu_{m,j}}{\sum_{m \in \mathcal{A}} \pi_{m,j}}. \tag{31}$$

Since the second derivative is $\sum_{m \in \mathcal{A}} \pi_{m,j} > 0$ when at least one modality has positive precision, the solution is the unique global minimizer.

A.4 Reliability as a Robust Gate

Reliability ρ_m acts as a robust gate on posterior precision. If modality m is unreliable, then $\rho_m \to 0$ and $\boldsymbol{\pi}_m \to \mathbf{0}$. Its posterior mean is then nearly removed from the consensus. If modality m is reliable and confident, then ρ_m is large and $\boldsymbol{\sigma}_m$ is small, yielding high precision.

For a fixed dimension j, the sensitivity of the consensus to modality mean $\mu_{m,j}$ is

$$\frac{\partial z_{0,j}}{\partial \mu_{m,j}} = \frac{\pi_{m,j}}{\sum_{n \in \mathcal{A}} \pi_{n,j} + \varepsilon}. \tag{32}$$

Thus, the influence of a modality is exactly proportional to its normalized reliability-calibrated precision.

A.5 Posterior Refinement Objectives

Let $\mathbf{z}_0$ denote the consensus latent and let u_0 denote its scalar uncertainty. HUPER refines $\mathbf{z}_0$ because posterior aggregation may still contain residual uncertainty caused by missing modalities, sparse sensing, or environment-induced noise. This section gives the detailed objectives of the two refinement realizations used in the paper. The default realization is a Flow Matching formulation in the human latent space, which we call Human Latent Flow. The alternative realization follows a DDPM-style noise prediction objective and serves as a residual perturbation variant.

Human Latent Flow. The default refinement module learns an uncertainty-conditioned vector field that transports an uncertain human posterior toward a cleaner human posterior. The source latent is $\mathbf{z}_{\mathrm{s}} = \mathbf{z}_0$. The cleaner target latent $\mathbf{z}_{\mathrm{c}}$ is computed from a stronger posterior view, such as a full-modality consensus or a less-corrupted augmented view, and is detached from the refinement loss. For $\tau \sim \mathcal{U}(0, 1)$, the interpolation path and target velocity are

$$\mathbf{z}_\tau = (1 - \tau)\mathbf{z}_{\mathrm{s}} + \tau \mathbf{z}_{\mathrm{c}}, \quad \mathbf{v}^\star = \mathbf{z}_{\mathrm{c}} - \mathbf{z}_{\mathrm{s}}. \tag{33}$$

The refiner predicts

$$\hat{\mathbf{v}}_\theta = R_\theta(\mathbf{z}_\tau, \psi(\tau), u_0, \bar{\rho}, \bar{a}), \tag{34}$$

and is trained with the Flow Matching loss

$$\mathcal{L}_{\mathrm{flow}} = \mathbb{E}_\tau \left[\| R_\theta(\mathbf{z}_\tau, \psi(\tau), u_0, \bar{\rho}, \bar{a}) - \mathbf{v}^\star \|_2^2 \right]. \tag{35}$$

Under squared error regression, the optimal refiner satisfies

$$R_\theta^\star(\mathbf{z}_\tau, \psi(\tau), u_0, \bar{\rho}, \bar{a}) = \mathbb{E}\left[\mathbf{z}_{\mathrm{c}} - \mathbf{z}_{\mathrm{s}} \mid \mathbf{z}_\tau, \tau, u_0, \bar{\rho}, \bar{a} \right]. \tag{36}$$

Thus, Human Latent Flow estimates the conditional correction direction from an uncertain human posterior to a cleaner human posterior. This directly matches the goal of posterior refinement: moving an already meaningful human-state estimate toward a more reliable one.

At inference, we solve the learned flow with a small number of Euler steps:

$$\mathbf{z}^{k+1} = \mathbf{z}^k + \Delta\tau \, g(u_0) R_\theta(\mathbf{z}^k, \psi(\tau_k), u_0, \bar{\rho}, \bar{a}), \tag{37}$$

where $g(u_0) = \mathrm{clip}(u_0, 0, 1)$ controls the correction strength. The final output is $\tilde{\mathbf{z}}_{\mathrm{flow}} = \mathbf{z}^{K_{\mathrm{flow}}}$.

DDPM-style residual perturbation. The residual perturbation variant approaches posterior correction from the noise side. It corresponds to the DDPM-style noise prediction module used in HUPER-DDPM, but is applied only in the human latent space and does not perform full reverse sampling from pure noise. Let $t \in \{1, \dots, T\}$ be a perturbation step and let $\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i$. We corrupt the consensus latent by

$$\mathbf{z}_t = \sqrt{\bar{\alpha}_t}\mathbf{z}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\xi}, \tag{38}$$

where $\mathbf{z}_t \in \mathbb{R}^{d_c}$ is the perturbed latent and

$$\begin{aligned} \boldsymbol{\xi} &= \gamma(u_0)\boldsymbol{\epsilon}, \\ \boldsymbol{\epsilon} &\sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{d_c}), \\ \gamma(u_0) &= \mathrm{clip}(1+u_0, 1, 3). \end{aligned} \tag{39}$$

Here, $\boldsymbol{\xi} \in \mathbb{R}^{d_c}$ is an uncertainty-scaled perturbation and $\boldsymbol{\epsilon}$ is standard Gaussian noise. A residual predictor estimates

$$\hat{\boldsymbol{\xi}} = D_\theta(\mathbf{z}_t, \phi(t), u_0, \bar{\rho}, \bar{a}), \tag{40}$$

where D_θ is the DDPM-style residual predictor, $\phi(t)$ is a sinusoidal embedding of the perturbation step, and $\hat{\boldsymbol{\xi}} \in \mathbb{R}^{d_c}$ is the predicted perturbation. The training loss is

$$\mathcal{L}_{\text{diff}} = \mathbb{E}_{t,\boldsymbol{\epsilon}}\left[\|D_\theta(\mathbf{z}_t, \phi(t), u_0, \bar{\rho}, \bar{a}) - \boldsymbol{\xi}\|_2^2\right]. \tag{41}$$

For squared error regression, the optimal predictor is

$$D_\theta^\star(\mathbf{z}_t, \phi(t), u_0, \bar{\rho}, \bar{a}) = \mathbb{E}\left[\boldsymbol{\xi} \mid \mathbf{z}_t, t, u_0, \bar{\rho}, \bar{a}\right]. \tag{42}$$

At inference, this variant applies a single residual correction,

$$\tilde{\mathbf{z}}_{\text{diff}} = \mathbf{z}_0 - u_0\hat{\boldsymbol{\xi}}_0, \quad \hat{\boldsymbol{\xi}}_0 = D_\theta(\mathbf{z}_0, \phi(T), u_0, \bar{\rho}, \bar{a}). \tag{43}$$

This predictor estimates conditional residual corruption rather than a semantic transport direction. Therefore, it provides a strong DDPM-based refinement variant, while Human Latent Flow offers a more direct objective for posterior correction.

Discussion. Both objectives arise naturally once multimodal fusion is viewed as probabilistic human posterior learning. The vector-field objective learns where an uncertain posterior should move, while the residual objective learns what uncertainty-scaled corruption should be removed. Since HUPER starts from a meaningful consensus latent rather than pure noise, the vector-field objective provides a more direct supervision signal for posterior correction and is used as the default variant in the main experiments.

A.6 Complete Calibration Losses

Let B be the batch size. For sample i and modality m, $\boldsymbol{\mu}_m^{(i)}$ denotes the posterior mean. The normalized mean is

$$\bar{\boldsymbol{\mu}}_m^{(i)} = \frac{\boldsymbol{\mu}_m^{(i)}}{\|\boldsymbol{\mu}_m^{(i)}\|_2 + \varepsilon}. \tag{44}$$

Let $\mathcal{P} = \{(m,n) \mid m, n \in \mathcal{A}, m < n\}$ be the set of unordered modality pairs.

For a pair (m,n), the contrastive loss from m to n is

$$\mathcal{L}_{m\to n}^{\text{con}} = -\frac{1}{B}\sum_{i=1}^{B} \log \frac{\exp\left(\bar{\boldsymbol{\mu}}_m^{(i)\top}\bar{\boldsymbol{\mu}}_n^{(i)}/\tau_{\text{con}}\right)}{\sum_{b=1}^{B}\exp\left(\bar{\boldsymbol{\mu}}_m^{(i)\top}\bar{\boldsymbol{\mu}}_n^{(b)}/\tau_{\text{con}}\right)}, \tag{45}$$

where $\tau_{\text{con}} > 0$ is the contrastive temperature. The symmetric content alignment loss is

$$\mathcal{L}_{\text{con}} = \frac{1}{|\mathcal{P}|}\sum_{(m,n)\in\mathcal{P}} \frac{\mathcal{L}_{m\to n}^{\text{con}} + \mathcal{L}_{n\to m}^{\text{con}}}{2}. \tag{46}$$

The posterior distribution loss is

$$\mathcal{L}_{\text{dist}} = \frac{1}{|\mathcal{P}|} \sum_{(m,n)\in\mathcal{P}} \left(\|\boldsymbol{\mu}_m - \boldsymbol{\mu}_n\|_2^2 + \|\boldsymbol{\sigma}_m - \boldsymbol{\sigma}_n\|_2^2 \right). \tag{47}$$

The uncertainty weight for pair (m, n) is

$$w_{mn} = \text{clip}\left(\exp\left(-\frac{a_m + a_n}{2} \right), w_{\min}, 1 \right), \tag{48}$$

where $w_{\min} > 0$ is a lower bound. The uncertainty weighted alignment loss is

$$\mathcal{L}_{\text{ualign}} = \frac{1}{|\mathcal{P}|} \sum_{(m,n)\in\mathcal{P}} w_{mn} \|\bar{\boldsymbol{\mu}}_m - \bar{\boldsymbol{\mu}}_n\|_2^2. \tag{49}$$

The reconstruction loss is

$$\mathcal{L}_{\text{rec}} = \frac{1}{|\mathcal{A}|} \sum_{m\in\mathcal{A}} \|\hat{\mathbf{f}}_m - \mathbf{f}_m\|_2^2. \tag{50}$$

If $d_c = d_s$, the content and environment latents are encouraged to be complementary by

$$\mathcal{L}_{\text{orth}} = \frac{1}{|\mathcal{A}|} \sum_{m\in\mathcal{A}} \left(\frac{\boldsymbol{\mu}_m^\top \mathbf{s}_m}{(\|\boldsymbol{\mu}_m\|_2 + \varepsilon)(\|\mathbf{s}_m\|_2 + \varepsilon)} \right)^2. \tag{51}$$

For modality m, the detached consensus from the remaining modalities is

$$\mathbf{q}_{-m} = \frac{1}{|\mathcal{A}| - 1} \sum_{n\in\mathcal{A}, n\neq m} \bar{\boldsymbol{\mu}}_n. \tag{52}$$

The target reliability is

$$\rho_m^\star = \exp\left(-\frac{\left\| \bar{\boldsymbol{\mu}}_m - \frac{\mathbf{q}_{-m}}{\|\mathbf{q}_{-m}\|_2 + \varepsilon} \right\|_2^2}{\tau_\rho} \right), \tag{53}$$

where $\tau_\rho > 0$ controls target sharpness. The reliability loss is

$$\mathcal{L}_{\text{rel}} = \frac{1}{|\mathcal{A}|} \sum_{m\in\mathcal{A}} (\rho_m - \rho_m^\star)^2. \tag{54}$$

The variance regularization is

$$\mathcal{L}_{\text{var}} = \frac{1}{|\mathcal{A}|} \sum_{m\in\mathcal{A}} \frac{1}{2d_c} \sum_{j=1}^{d_c} \left(\mu_{m,j}^2 + \sigma_{m,j}^2 - \log \sigma_{m,j}^2 - 1 \right). \tag{55}$$

The complete calibration loss is

$$\begin{aligned} \mathcal{L}_{\text{HUPER}} =& \lambda_{\text{con}}\mathcal{L}_{\text{con}} + \lambda_{\text{dist}}\mathcal{L}_{\text{dist}} + \lambda_{\text{ualign}}\mathcal{L}_{\text{ualign}} \\ &+ \lambda_{\text{rec}}\mathcal{L}_{\text{rec}} + \lambda_{\text{orth}}\mathcal{L}_{\text{orth}} + \lambda_{\text{rel}}\mathcal{L}_{\text{rel}} + \lambda_{\text{var}}\mathcal{L}_{\text{var}}. \end{aligned} \tag{56}$$

Single-modality handling. The pairwise posterior objectives require at least two available modalities. When $|\mathcal{A}| = 1$, the modality-pair set $\mathcal{P}$ is empty and the leave-one-modality-out reliability target is unavailable. We therefore define

$$\mathcal{L}_{\text{con}} = \mathcal{L}_{\text{dist}} = \mathcal{L}_{\text{ualign}} = \mathcal{L}_{\text{rel}} = 0. \tag{57}$$

Single-modality samples remain optimized by the task, reconstruction, variance-regularization, orthogonality, and applicable refinement objectives. This conditional definition avoids empty-set normalization and division by zero while enabling training and inference with arbitrary non-empty modality subsets.

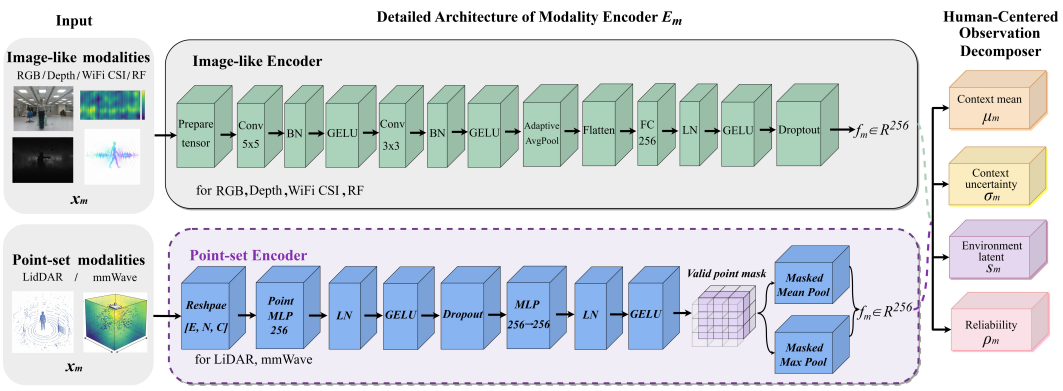


Figure 6: Detailed architecture of the modality encoder E_m. HUPER uses two lightweight encoder templates according to input structure. Image-like modalities, including RGB, depth, WiFi CSI, and dense RF representations, are processed by a convolutional encoder with normalization, activation, adaptive pooling, and projection. Point-set modalities, including LiDAR and mmWave point clouds, are processed by point-wise MLP layers followed by valid-point masking, masked mean pooling, and masked max pooling. Both branches output a fixed-dimensional modality feature $\mathbf{f}_m \in \mathbb{R}^{256}$, which is then passed to the human-centered observation decomposer to predict posterior mean $\boldsymbol{\mu}_m$, posterior scale $\boldsymbol{\sigma}_m$, environment latent $\mathbf{s}_m$, and reliability score ρ_m.

A.7 Structure Losses

Given ground truth pose $\mathbf{P} \in \mathbb{R}^{3\times J}$, the skeleton proxy loss is

$$\mathcal{L}_{\text{skel}} = \text{SmoothL1}(\hat{\mathbf{P}}_{\text{pre}}, \mathbf{P}). \tag{58}$$

Let $\mathcal{E}$ be the set of skeletal edges. For edge $(p, q) \in \mathcal{E}$, define

$$\begin{aligned} \hat{\ell}_{p,q} &= \left\| \hat{\mathbf{P}}_{\text{pre},:,p} - \hat{\mathbf{P}}_{\text{pre},:,q} \right\|_2, \\ \ell_{p,q} &= \left\| \mathbf{P}_{:,p} - \mathbf{P}_{:,q} \right\|_2. \end{aligned} \tag{59}$$

The bone loss is

$$\mathcal{L}_{\text{bone}} = \frac{1}{|\mathcal{E}|} \sum_{(p,q)\in\mathcal{E}} \text{SmoothL1}(\hat{\ell}_{p,q}, \ell_{p,q}). \tag{60}$$

B Network Configuration and Implementation Details

This section provides additional details on datasets, training objectives, baseline formulations, evaluation metrics, and implementation settings used in our experiments.

B.1 Modality Encoder Architecture

The modality encoder E_m is the first stage of HUPER and provides the interface between heterogeneous raw observations and the human-centered posterior decomposer. Although different sensors have distinct input structures, the role of E_m is unified: it maps each raw modality input $\mathbf{x}_m$ to a fixed-dimensional feature vector $\mathbf{f}_m \in \mathbb{R}^{d_f}$. This allows the subsequent decomposer to predict $(\boldsymbol{\mu}_m, \boldsymbol{\sigma}_m, \mathbf{s}_m, \rho_m)$ from a shared feature format, while still preserving modality-specific low-level inductive biases.

Image-like encoder. For image-like modalities, including RGB, depth, WiFi CSI, and dense RF representations when applicable, the encoder treats the input as a spatial or time-frequency tensor. Given an input $\mathbf{x}_m$, the image-like encoder applies two convolutional blocks followed by adaptive

pooling and projection:

$$\begin{aligned}\mathbf{U}_m^{(1)} &= \mathrm{ConvBlock}_{5,s2}^{32}(\mathbf{x}_m),\\ \mathbf{U}_m^{(2)} &= \mathrm{ConvBlock}_{3,s1}^{64}(\mathbf{U}_m^{(1)}),\\ \mathbf{r}_m &= \mathrm{Flatten}\left(\mathrm{AvgPool}_{4\times4}(\mathbf{U}_m^{(2)})\right),\\ \mathbf{f}_m &= \mathrm{Dropout}\left(\mathrm{GELU}\left(\mathrm{LN}\left(\mathrm{FC}(\mathbf{r}_m)\right)\right)\right).\end{aligned} \tag{61}$$

Here, $\mathrm{ConvBlock}_{5,s2}^{32}$ denotes $\mathrm{GELU}(\mathrm{BN}(\mathrm{Conv}_{5\times5,s=2}^{32}(\cdot)))$, where the convolution uses kernel size 5×5, stride 2, and 32 output channels. $\mathrm{ConvBlock}_{3,s1}^{64}$ is defined analogously with a 3×3 convolution, stride 1, and 64 output channels. FC, LN, and Dropout project the pooled feature to the final dimension $d_f = 256$.

Point-set encoder. For point-set modalities, including LiDAR and mmWave point clouds, the input is reshaped as $\mathbf{x}_m \in \mathbb{R}^{B\times N\times C}$, where B is the batch size, N is the number of points, and C is the point feature dimension. The encoder applies point-wise MLP layers to each valid point:

$$\begin{aligned}\mathbf{H}_m^{(1)} &= \mathrm{Dropout}\left(\mathrm{GELU}\left(\mathrm{LN}\left(\mathrm{MLP}_{C\to d_f}(\mathbf{x}_m)\right)\right)\right),\\ \mathbf{H}_m^{(2)} &= \mathrm{GELU}\left(\mathrm{LN}\left(\mathrm{MLP}_{d_f\to d_f}(\mathbf{H}_m^{(1)})\right)\right).\end{aligned} \tag{62}$$

A valid-point mask $\mathbf{M}_m \in \{0,1\}^{B\times N}$ is used to remove padded or invalid points during pooling. We compute masked mean and max pooled features:

$$\begin{aligned}\mathbf{p}_m^{\mathrm{mean}} &= \frac{\sum_{n=1}^{N} M_{m,n}\mathbf{H}_{m,n}^{(2)}}{\sum_{n=1}^{N} M_{m,n}+\varepsilon},\\ \mathbf{p}_m^{\mathrm{max}} &= \max_{n:M_{m,n}=1} \mathbf{H}_{m,n}^{(2)},\end{aligned} \tag{63}$$

where $\varepsilon > 0$ is a numerical stabilizer. The final point-set feature is:

$$\mathbf{f}_m = \frac{1}{2}\left(\mathbf{p}_m^{\mathrm{mean}} + \mathbf{p}_m^{\mathrm{max}}\right), \quad \mathbf{f}_m \in \mathbb{R}^{d_f}. \tag{64}$$

The masked mean pooling captures global point-set statistics, while masked max pooling preserves salient local responses. Their average provides a stable representation for sparse and variable-size point observations.

Connection to the human-centered decomposer. Both encoder templates produce a fixed-dimensional feature $\mathbf{f}_m \in \mathbb{R}^{256}$. This feature is then passed to the human-centered observation decomposer:

$$\mathbf{f}_m \longrightarrow (\boldsymbol{\mu}_m, \boldsymbol{\sigma}_m, \mathbf{s}_m, \rho_m). \tag{65}$$

The posterior mean $\boldsymbol{\mu}_m$ and scale $\boldsymbol{\sigma}_m$ define the modality posterior $q_m(\mathbf{c} \mid \mathbf{x}_m)$, the latent $\mathbf{s}_m$ captures environment and modality-specific factors, and ρ_m estimates the sample-level reliability of modality m. Therefore, E_m does not directly perform multimodal fusion. Instead, it provides modality-specific evidence that is converted into a comparable probabilistic human state representation by the decomposer.

B.2 Datasets and Protocols

MM-Fi for human pose estimation. MM-Fi is a large-scale multimodal human sensing dataset with synchronized observations from multiple sensing modalities. For the HPE task, we use five modalities: RGB images (I), depth images (D), LiDAR point clouds (L), mmWave radar (R), and WiFi CSI (W). Each synchronized frame is annotated with 3D whole-body keypoints, where the human body is represented by $J = 17$ joints. We evaluate HUPER under all non-empty modality subsets, including single-modality, dual-modality, triple-modality, four-modality, and full-modality settings. This protocol tests whether the learned human posterior space can generalize to arbitrary modality combinations.

MM-Fi for human activity recognition. For HAR on MM-Fi, we use four modalities: RGB images (I), depth images (D), LiDAR point clouds (L), and mmWave radar (R). The task contains 27 action categories, including daily activities and rehabilitation exercises. Following prior multimodal human sensing settings, WiFi CSI is not used for MM-Fi HAR because each WiFi sample covers a short temporal window that may be insufficient to distinguish action-level patterns. We evaluate all non-empty subsets of the four modalities.

XRF55 for human activity recognition. XRF55 is a radio-frequency human activity recognition dataset with 55 action classes collected from multiple subjects. It contains three sensing modalities: mmWave radar (R), WiFi CSI (W), and RFID phase series (RF). Compared with MM-Fi, XRF55 focuses on wireless and radio-frequency sensing, making it suitable for evaluating HUPER under strongly heterogeneous and noisy sensing conditions. We evaluate HUPER under all non-empty subsets of the three modalities and follow the original split protocol of the dataset.

B.3 Training Objectives

Human activity recognition. For HAR, the refined human latent $\tilde{\mathbf{z}}$ is passed to a classifier C_{HAR}, which predicts the action distribution:

$$\hat{\mathbf{y}} = C_{\mathrm{HAR}}(\tilde{\mathbf{z}}), \tag{66}$$

where $\hat{\mathbf{y}} \in \mathbb{R}^K$ is the predicted class probability vector and K is the number of action classes. We optimize the cross-entropy loss:

$$\mathcal{L}_{\mathrm{HAR}} = -\sum_{k=1}^{K} y_k \log \hat{y}_k, \tag{67}$$

where $\mathbf{y}$ is the one-hot ground-truth label.

Human pose estimation. For HPE, the refined human latent $\tilde{\mathbf{z}}$ is passed to a pose regressor R_{HPE}:

$$\hat{\mathbf{P}} = R_{\mathrm{HPE}}(\tilde{\mathbf{z}}), \tag{68}$$

where $\hat{\mathbf{P}} \in \mathbb{R}^{3\times J}$ is the predicted 3D pose and $J = 17$ is the number of body joints. Given the ground-truth pose $\mathbf{P} \in \mathbb{R}^{3\times J}$, we optimize a pose regression loss:

$$\mathcal{L}_{\mathrm{HPE}} = \frac{1}{J}\sum_{j=1}^{J} \left\| \hat{\mathbf{P}}_{:,j} - \mathbf{P}_{:,j} \right\|_2^2. \tag{69}$$

When pose annotations are available during pretraining, we also use the skeleton proxy loss and bone-length consistency loss described in the Structure Losses section above.

HUPER pretraining objective. In addition to downstream task losses, HUPER uses the human-oriented posterior learning objective described in the main paper. The full pretraining loss is:

$$\mathcal{L}_{\mathrm{pre}} = \mathcal{L}_{\mathrm{HUPER}} + \beta_{\mathrm{ref}}\mathcal{L}_{\mathrm{ref}} + \beta_{\mathrm{struct}}\mathcal{L}_{\mathrm{struct}}, \tag{70}$$

where $\mathcal{L}_{\mathrm{HUPER}}$ contains posterior alignment, posterior distribution consistency, reconstruction, and reliability calibration terms. The refinement loss $\mathcal{L}_{\mathrm{ref}}$ is $\mathcal{L}_{\mathrm{flow}}$ for the vector-field refiner and $\mathcal{L}_{\mathrm{diff}}$ for the residual perturbation refiner. The structure loss $\mathcal{L}_{\mathrm{struct}}$ is used when pose annotations are available.

B.4 Evaluation Metrics

HAR metric. For human activity recognition, we report classification accuracy:

$$\mathrm{Acc} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}\left[\arg\max_k \hat{y}_k^{(i)} = y^{(i)}\right], \tag{71}$$

where N is the number of evaluation samples, $\hat{\mathbf{y}}^{(i)}$ is the predicted class distribution, and $y^{(i)}$ is the ground-truth class label. We also report the validation loss and the average reliability score predicted by HUPER for each modality subset.

Table 5: Missing-modality robustness on MM-Fi HPE, grouped by the number of available modalities. Partial Avg. averages all incomplete subsets containing at least two modalities.

Modalities	Flow ↓	DDPM ↓	X-Fi ↓	Baseline1 ↓	Baseline2 ↓
2 modalities	90.6	108.8	105.0	124.1	139.2
3 modalities	77.7	86.7	92.6	113.0	131.5
4 modalities	69.6	73.6	87.9	106.8	139.0
5 modalities	64.7	65.3	83.7	103.0	128.9
Partial Avg.	81.2	92.9	96.6	116.2	136.1

HPE metrics. For human pose estimation, we report MPJPE and PA-MPJPE. MPJPE measures the average Euclidean distance between predicted and ground-truth joints:

$$\mathrm{MPJPE} = \frac{1}{NJ}\sum_{i=1}^{N}\sum_{j=1}^{J}\left\|\hat{\mathbf{P}}_{:,j}^{(i)} - \mathbf{P}_{:,j}^{(i)}\right\|_2 . \tag{72}$$

PA-MPJPE computes MPJPE after rigid Procrustes alignment between the predicted pose and the ground-truth pose:

$$\text{PA-MPJPE} = \frac{1}{NJ}\sum_{i=1}^{N}\sum_{j=1}^{J}\left\|\mathcal{T}^{(i)}\left(\hat{\mathbf{P}}_{:,j}^{(i)}\right) - \mathbf{P}_{:,j}^{(i)}\right\|_2 , \tag{73}$$

where $\mathcal{T}^{(i)}$ is the optimal similarity transformation estimated by Procrustes alignment for sample i. Lower MPJPE and PA-MPJPE indicate better pose estimation performance.

B.5 Baseline Formulation

We compare HUPER with the reported baseline and X-Fi results under the same multimodal human sensing setting.

Feature-level fusion baseline. The feature-level fusion baseline first extracts modality-specific features and then concatenates or averages the available modality features. The fused representation is mapped to the task space by an MLP. This baseline directly combines modality features after extraction and does not explicitly model human-centered posterior uncertainty.

Decision-level fusion baseline. The decision-level fusion baseline trains or evaluates each modality independently and then averages the final modality-level predictions. For HAR, the averaged prediction is the action probability distribution. For HPE, the averaged output is the predicted 3D pose. This strategy does not learn cross-modal posterior consensus and treats the available modalities in a relatively uniform manner.

Comparison with X-Fi. X-Fi is a modality-invariant multimodal human sensing framework that supports arbitrary modality combinations through modality-invariant fusion. Unlike X-Fi, HUPER focuses on the alignment object itself. Instead of directly aligning heterogeneous tokens or features, HUPER maps each modality into a probabilistic estimate of the underlying human state and performs reliability-calibrated posterior consensus in the human state space.

B.6 Implementation Details

Datasets and protocols. We evaluate HUPER on XRF55-HAR, MMFi-HAR, and MMFi-HPE using the original dataset protocols. XRF55-HAR follows the official-all split and uses mmWave radar, WiFi CSI, and RFID for 55-class activity recognition. MMFi-HAR and MMFi-HPE follow the random split of Protocol 3. MMFi-HAR uses RGB, depth, mmWave radar, and LiDAR for 27-class activity recognition, while MMFi-HPE additionally includes WiFi CSI and predicts 17 three-dimensional body joints. The MM-Fi split seed is fixed to 0, whereas the XRF55 split seed is fixed to 3407. These data partitions remain unchanged across all runs.

Table 6: Missing-modality robustness on HAR. Accuracy is reported in percentage.

Dataset	Modalities	Flow ↑	DDPM ↑	X-Fi ↑	Baseline ↑
MM-Fi	2 modalities	74.0	70.1	62.3	57.6
MM-Fi	3 modalities	78.9	80.1	69.4	63.3
MM-Fi	4 modalities	85.5	86.0	72.2	72.6
MM-Fi	Partial Avg.	76.0	74.1	65.2	59.9
XRF55	2 modalities	91.3	88.8	77.6	71.3
XRF55	3 modalities	97.1	96.7	89.8	70.6

Table 7: Modality encoder ablation on MM-Fi HPE and XRF55 HAR.

Encoder	MM-Fi HPE		XRF55 HAR	
	Full ↓	All Avg. ↓	Full ↑	Dual Avg. ↑
Shared MLP	67.8	94.5	94.1	86.0
Shared Transformer	67.4	88.3	95.6	89.0
Mod.-specific w/o pooling	65.8	83.7	96.5	91.2
Mod.-aware (ours)	**64.7**	**79.6**	**97.1**	**93.1**

Training configuration. HUPER-Flow and HUPER-DDPM share the same modality encoders and task heads, with a posterior dimension of 256. Both pretraining and downstream adaptation run for at most 50 epochs using AdamW and cosine annealing to a minimum learning rate of 10^{-6}. Pretraining uses a learning rate of 3×10^{-4}, weight decay of 5×10^{-4}, and dropout of 0.1; downstream training uses 10^{-3}, 10^{-4}, and 0.2, respectively. XRF55 uses a batch size of 32 throughout, whereas MM-Fi uses 64 for pretraining and 32 for downstream training. We evaluate HUPER using training seeds $\{3407, 3408, 3409, 3410, 3411\}$ while keeping the dataset partitions fixed. Non-empty modality subsets, including single-modality inputs, are randomly sampled during downstream training and exhaustively evaluated at test time. Each reported HUPER result is the arithmetic mean over the five runs. Within each run, checkpoint selection uses validation performance only, and the test set is never used for model selection.

Human posterior consensus. For each modality m, the reliability-calibrated precision is:

$$\boldsymbol{\pi}_m = \frac{\rho_m}{\boldsymbol{\sigma}_m^2 + \varepsilon}. \tag{74}$$

The consensus latent is computed as:

$$\mathbf{z}_0 = \frac{\sum_{m \in \mathcal{A}} \boldsymbol{\pi}_m \odot \boldsymbol{\mu}_m}{\sum_{m \in \mathcal{A}} \boldsymbol{\pi}_m + \varepsilon}, \tag{75}$$

where $\mathcal{A}$ denotes the available modality subset. The consensus variance $\mathbf{v}_0$ and scalar uncertainty u_0 are computed following the formulation defined above.

Posterior refinement. The posterior refiner R_θ corrects the consensus latent in the human state space. In the default vector-field variant, R_θ predicts an uncertainty-conditioned correction field from an uncertain posterior toward a cleaner posterior:

$$\hat{\mathbf{v}}_\theta = R_\theta(\mathbf{z}_\tau, \psi(\tau), u_0, \bar{\rho}, \bar{a}). \tag{76}$$

$$\mathbf{z}^{k+1} = \mathbf{z}^k + \Delta\tau \, g(u_0) R_\theta(\mathbf{z}^k, \psi(\tau_k), u_0, \bar{\rho}, \bar{a}). \tag{77}$$

$$\hat{\boldsymbol{\xi}} = D_\theta(\mathbf{z}_t, \phi(t), u_0, \bar{\rho}, \bar{a}). \tag{78}$$

At inference, HUPER applies the Euler update in Eq. equation 77 and obtains the refined latent $\tilde{\mathbf{z}}$. In the residual perturbation variant, the refiner is implemented by the predictor D_θ that estimates uncertainty-scaled perturbation as defined in Eq. equation 78. The refined latent $\tilde{\mathbf{z}}$ is then passed to task-specific heads for HAR or HPE.

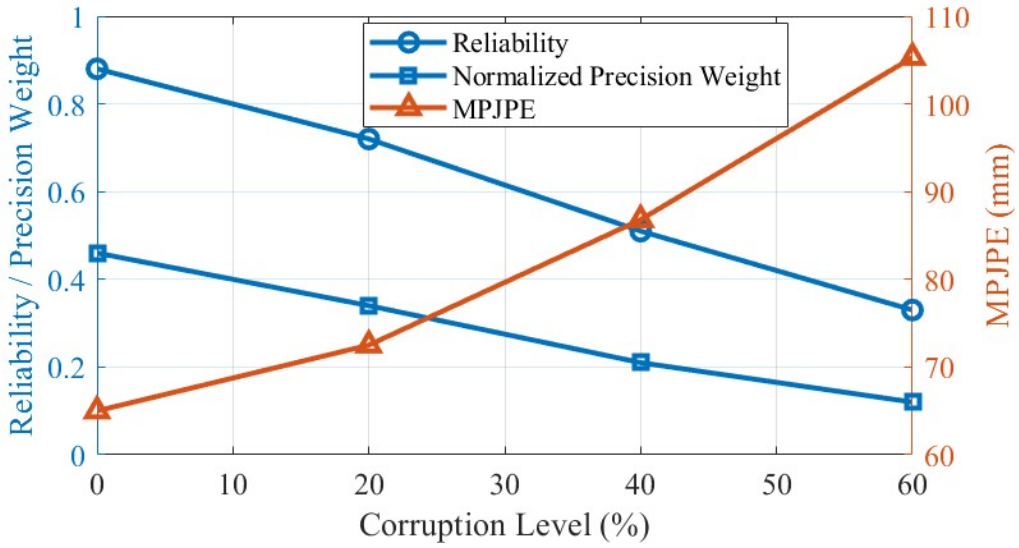


Figure 7: Reliability response under synthetic corruption, averaged over five independent runs. Increasing corruption reduces the predicted reliability and normalized precision weight of the corrupted modality while increasing MPJPE.

Table 8: Posterior decomposition ablation on XRF55 HAR. Accuracy is reported in percentage.

Method	Single Avg.	Dual Avg.	Full	All Avg.
HUPER w/o environment latent $\mathbf{s}_m$	68.9	86.2	96.0	82.1
HUPER w/o reconstruction loss $\mathcal{L}_{rec}$	72.4	89.0	96.6	85.0
HUPER w/o orthogonality loss $\mathcal{L}_{orth}$	74.1	90.5	96.9	86.3
Full HUPER	**76.2**	**93.1**	**97.1**	**88.9**

Subset evaluation. For each modality subset, only the selected modalities are used to estimate posteriors and compute the consensus latent. No architectural modification is required. This protocol evaluates whether HUPER can handle arbitrary missing-modality conditions using the same human-centered posterior aggregation rule.

C Additional Results

The main comparison tables report modality combinations containing at least two modalities. Partial Avg. averages all incomplete multimodal subsets containing between two and $|\mathcal{M}| - 1$ modalities. In contrast, All Avg. is the cardinality-balanced macro-average over all evaluated subset groups, including single- and full-modality settings.

Missing-modality robustness. To assess robustness to incomplete sensing, we group the results by the number of available modalities, as summarized in Tables 5 and 6. For MM-Fi HPE, HUPER-Flow consistently achieves the lowest average MPJPE across all modality-count groups. With only two available modalities, HUPER-Flow obtains 90.6 mm MPJPE, improving over HUPER-DDPM, X-Fi, Baseline1, and Baseline2 by 18.2, 14.4, 33.5, and 48.6 mm, respectively. The advantage remains clear as more modalities are available, and the partial-modality average is reduced from 96.6 mm for X-Fi to 81.2 mm for HUPER-Flow. These results indicate that reliability-calibrated posterior consensus improves robustness when sensing evidence is incomplete.

For HAR, HUPER also performs strongly under missing modalities. On MM-Fi, HUPER-Flow improves the partial-modality average accuracy from 65.2% for X-Fi to 76.0%, while HUPER-DDPM reaches 74.1%. On XRF55, HUPER-Flow obtains 91.3% accuracy under dual-modality settings and 97.1% under the full-modality setting, outperforming X-Fi by 13.7 and 7.3 percentage points, respectively. The larger gains on XRF55 suggest that posterior reliability is particularly useful for heterogeneous RF sensing, where modalities have different noise patterns and sensing ranges.

Effect of modality-aware encoders. To examine whether heterogeneous sensors require distinct inductive biases, we compare shared and modality-aware encoder designs in Table 7. The shared MLP performs worst, yielding 67.8 mm full-modality MPJPE and 94.5 mm all-subset average MPJPE on MM-Fi HPE. A shared Transformer improves the partial-modality result to 88.3 mm, but remains inferior to modality-specific encoding. Introducing modality-specific branches without mask pooling further reduces the errors to 65.8 and 83.7 mm, respectively. The complete modality-aware encoder achieves the best results: 64.7 mm full-modality MPJPE, 79.6 mm all-subset average MPJPE, 97.1% XRF55 full-modality accuracy, and 93.1% dual-modality accuracy. Compared with the shared MLP, it reduces all-subset average MPJPE by 14.9 mm and improves dual-modality accuracy by 7.1 percentage points. Its advantage over the variant without mask pooling is more pronounced under partial observations, reducing MPJPE by 4.1 mm. These results show that preserving input-specific structure and explicitly handling invalid point observations are important for learning robust posterior estimates from heterogeneous sensors.

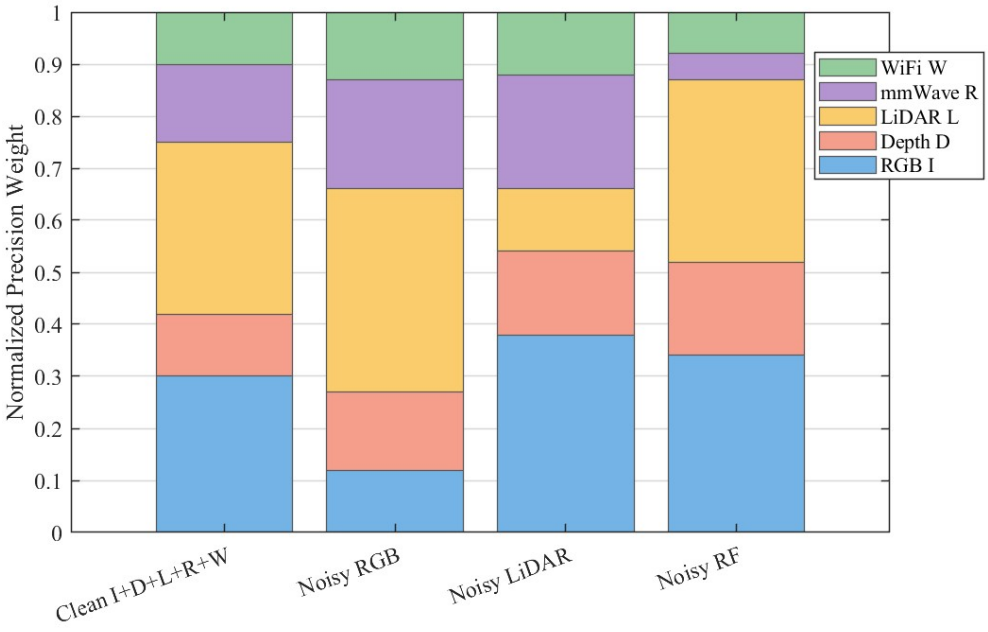


Figure 8: Reliability-calibrated precision weights under clean and corrupted inputs, averaged over five independent runs. Corrupting a modality suppresses its normalized contribution and shifts posterior consensus toward cleaner modalities.

Effect of posterior decomposition. The necessity of separating human content from modality- and environment-dependent factors is investigated through the decomposition ablations in Table 8. Removing the environment latent $\mathbf{s}_m$ causes the largest degradation: single-, dual-, and all-subset average accuracies decrease from 76.2%, 93.1%, and 88.9% to 68.9%, 86.2%, and 82.1%, respectively. This indicates that an explicit nuisance branch is particularly important under incomplete modality settings. Removing $\mathcal{L}_{\text{rec}}$ lowers the all-subset average to 85.0%, suggesting that feature reconstruction helps retain complementary information across the human and environment latents.

Removing $\mathcal{L}_{\text{orth}}$ produces a smaller but consistent decrease to 86.3%, showing that the orthogonality constraint further limits information leakage between the two representations. The complete decomposition achieves the best performance in every setting, reaching 97.1% with all modalities and 88.9% on average. Overall, these results support the joint use of an environment latent, reconstruction supervision, and orthogonality regularization for learning a cleaner human-centered posterior.

Reliability under corruption. We examine whether the learned reliability reflects sensing quality by progressively corrupting individual modalities. Each point in Figures 7 and 8 reports the mean over five independent runs. Increasing corruption consistently reduces the predicted reliability and normalized precision weight while increasing MPJPE. The corresponding precision redistribution shows that corrupting RGB, LiDAR, or RF suppresses its contribution and shifts the consensus toward cleaner modalities. These results demonstrate that reliability-calibrated precision responds systematically to sensing degradation and supports robustness to noisy observations.

Flow refinement analysis. We study the sensitivity of Human Latent Flow to its numerical solver by varying the number of Euler refinement steps, as shown in Fig. 9. Increasing K_{flow} from 1 to 4 improves MM-Fi HPE and XRF55 HAR, while using 8 steps gives little additional gain and slightly degrades HAR. We therefore use $K_{\text{flow}} = 4$ as a balance between refinement quality and computational cost.

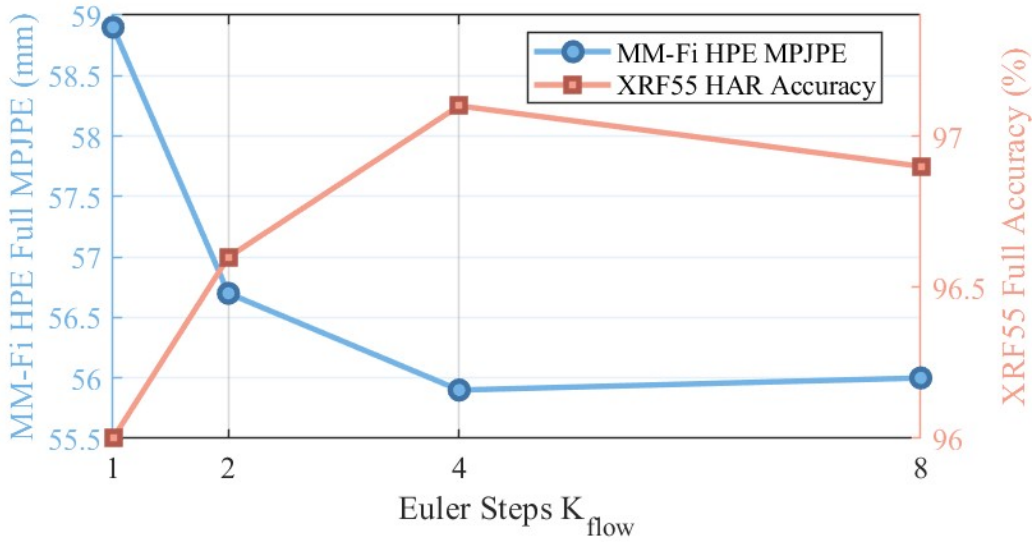


Figure 9: Sensitivity of Human Latent Flow to the number of Euler refinement steps. Increasing the number of steps from 1 to 4 improves both MM-Fi HPE and XRF55 HAR, while further increasing to 8 steps provides limited additional benefit. We therefore use $K_{\text{flow}} = 4$ as a practical trade-off between accuracy and refinement cost.

Flow refinement target. To identify effective transport supervision, we compare different cleaner-target constructions in Table 9. The self-target provides no informative cross-view correction and performs worst, obtaining 73.0 mm full-modality MPJPE, 93.4 mm all-subset average MPJPE, and 85.1% XRF55 all-subset accuracy. Replacing it with an augmented posterior or full-modality consensus progressively improves performance. Applying stop-gradient to the full-modality consensus yields the best results of 64.7 mm, 79.6 mm, and 88.9%, respectively. Relative to the self-target, the complete design reduces full-modality and all-subset average MPJPE by 8.3 and 13.8 mm and improves XRF55 accuracy by 3.8 percentage points. Stop-gradient prevents the cleaner target from co-adapting toward the uncertain source under the refinement loss, thereby providing a stable transport destination.

Table 9: Ablation of the cleaner target used for refinement.

Flow Target	MM-Fi HPE		XRF55 HAR
	Full ↓	All Avg. ↓	All Avg. ↑
Self target	73.0	93.4	85.1
Augmented posterior	69.7	84.8	87.2
Full-modality consensus	67.2	81.4	88.1
Consensus + stop-gradient	**64.7**	**79.6**	**88.9**

Additional qualitative comparison. The benefit of multimodal posterior consensus over individual wireless sensors is qualitatively demonstrated in Fig. 10. Single RF modalities often capture coarse body motion but exhibit larger errors in limb orientation and joint placement, especially for asymmetric actions such as lunges, twists, and picking up objects. In contrast, HUPER produces poses that are closer to the ground truth and more anatomically consistent across both upper- and lower-body joints. These qualitative results show that posterior consensus can combine complementary RF evidence and reduce the instability of individual wireless modalities.

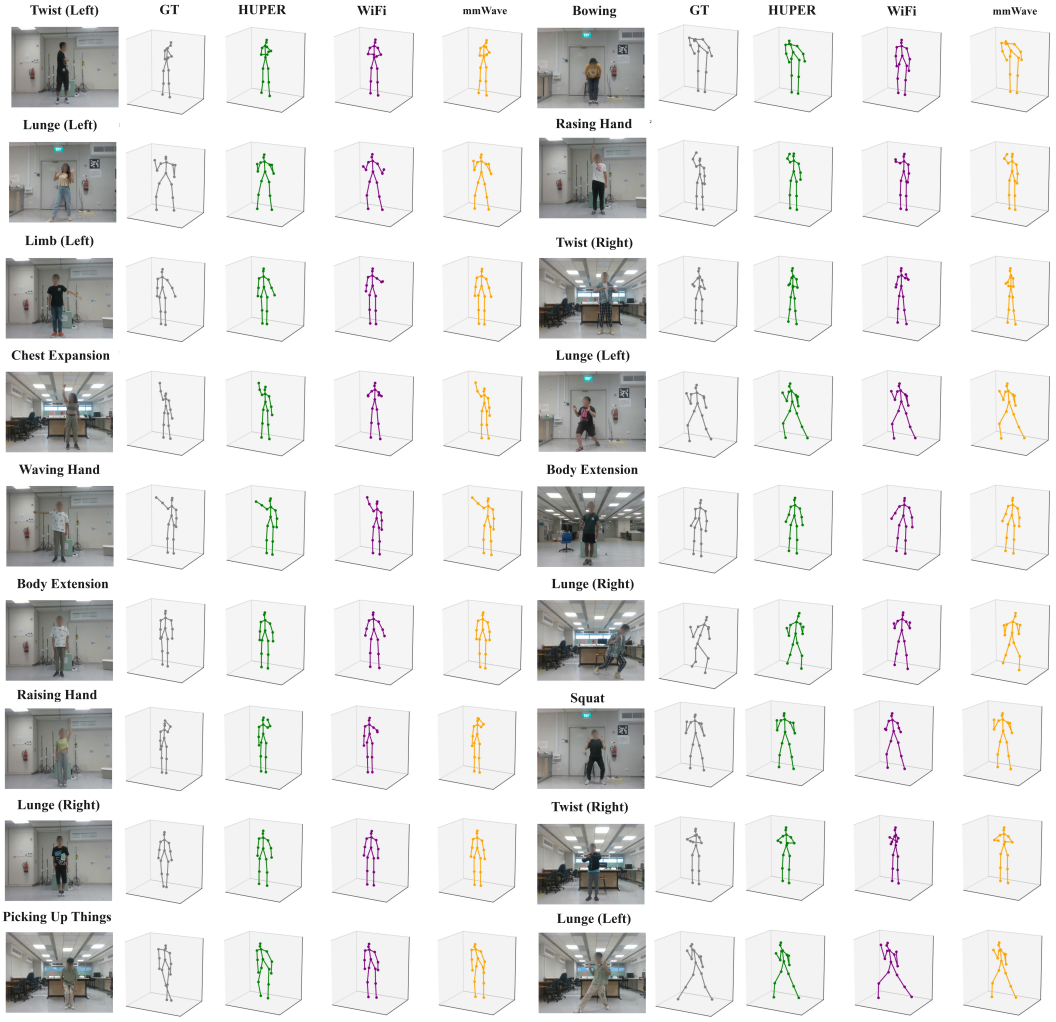


Figure 10: Qualitative comparison between GT, HUPER, WiFi-only and mmWave-only predictions on multiple actions.