WiCompass: Oracle-driven Data Scaling for mmWave Human Pose Estimation

Bo Liang[P], Chen Gong[P], Haobo Wang[P], Qirui Liu[P], Rungui Zhou[P], Fengzhi Shao[U], Yubo Wang[U], Wei Gao[PI], Kaichen Zhou[M], Guolong Cui[U], Chenren Xu[PK✉*]

[P]School of Computer Science, Peking University [U]University of Electronic Science and Technology of China
[PI]University of Pittsburgh [M]Electrical Engineering and Computer Science, Massachusetts Institute of Technology
[K]Key Laboratory of High Confidence Software Technologies, Ministry of Education (PKU)

Abstract

Millimeter-wave Human Pose Estimation (mmWave HPE) promises privacy but suffers from poor generalization under distribution shifts. We demonstrate that brute-force data scaling is ineffective for out-of-distribution (OOD) robustness; efficiency and coverage are the true bottlenecks. To address this, we introduce WiCompass, a coverage-aware data-collection framework. WiCompass leverages large-scale motion-capture corpora to build a universal pose space "oracle" that quantifies dataset redundancy and identifies underrepresented motions. Guided by this oracle, WiCompass employs a closed-loop policy to prioritize collecting informative missing samples. Experiments show that WiCompass consistently improves OOD accuracy at matched budgets and exhibits superior scaling behavior compared to conventional collection strategies. By shifting focus from brute-force scaling to coverage-aware data acquisition, this work offers a practical path toward robust mmWave sensing.

CCS Concepts

• **Human-centered computing → Ubiquitous and mobile computing systems and tools**.

Keywords

Millimeter-wave Radar Sensing, Human Pose Estimation, Coverage-aware Data Acquisition

ACM Reference Format:

Bo Liang, Chen Gong, Haobo Wang, Qirui Liu, Rungui Zhou, Fengzhi Shao, Yubo Wang, Kaichen Zhou, Wei Gao, Guolong Cui, Chenren Xu. 2026. WiCompass: Oracle-driven Data Scaling for mmWave Human Pose Estimation. In *The 32nd Annual International Conference on Mobile Computing and Networking (MobiCom '26), October 26–30, 2026, Austin, TX, USA.* ACM, New York, NY, USA, 17 pages. https://doi.org/10.1145/3795866.3796684

*✉: chenren@pku.edu.cn

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.

MobiCom '26, Austin, TX, USA

© 2026 Copyright held by the owner/author(s).

ACM ISBN 979-8-4007-2505-0/26/10
https://doi.org/10.1145/3795866.3796684

1 Introduction

Human Pose Estimation (HPE)—the task of localizing anatomical keypoints to recover body posture [1]—is a cornerstone technology for applications in human–computer interaction [2], automated healthcare [3], and immersive media [4]. The dominant approach relies on optical sensors such as RGB cameras which, despite high accuracy, suffer fundamental limitations: they compromise privacy, perform poorly under inconsistent lighting, and are susceptible to occlusion. In response, millimeter-wave (mmWave) radar has emerged as a compelling alternative. By nature, mmWave sensing is privacy-preserving, robust to lighting changes, and capable of penetrating many non-metallic materials [5–8].

Despite this promise, progress in mmWave HPE rests on a fragile data foundation. While state-of-the-art (SOTA) models achieve impressive in-distribution (ID) joint localization errors comparable to optical HPE [9–12], their performance degrades sharply under even modest data distribution shifts. This brittleness is stark in our leave-one-out test (§2.3): simply withholding a single action such as "Waving hand (left)" from the training set causes the human joint localization error on that unseen action to spike from roughly 50 mm to well over 120 mm. The disproportionate failure, even when semantically related motions ("Waving hand (right)" and "Raise hand (left)") remain in the training set, points to a critical problem: current models learn to interpolate expertly within a known distribution but fail to generalize across the true, diverse spectrum of human motion. This gap between performance in the lab and reliability in the wild is a critical barrier to deployment.

The intuitive response – simply scaling up data via additional real-world collection or synthetic generation – is a false panacea. Real-world acquisition is notoriously expensive in time, personnel, and setup costs [13–16]. Achieving

coverage comparable to vision HPE datasets [17–19] typically requires recruiting tens to hundreds of participants and capturing across multiple environments, with calendar time measured in months and total human-hours in the thousands. While data synthesis with simulators [20–22] can be faster, it often suffers a persistent "sim-to-real" gap, failing to capture the nuances of real-world physics and misaligning with in-the-wild data distributions [23]. More critically, our analysis shows that naively adding more data is not only inefficient but frequently ineffective. A progressive subset study (§2.3) on existing large-scale datasets reveals this redundancy: randomly discarding up to 70% of training samples of existing datasets results in less than a 2% drop in performance. These findings shift the focus from *data quantity* to *data efficiency*, and leads to the central question of this paper: *What is the optimal data acquisition strategy to maximize the generalization of mmWave HPE models?*

We address this challenge with WiCompass, an oracle-driven data collection framework for mmWave HPE, where the "oracle" is a high-coverage geometric prior over human poses. We distill motion priors from large-scale motion-capture (MoCap) corpora into a shared latent pose space. By projecting pose labels (skeletons) from both mmWave samples and reference MoCap into this space, we quantify coverage and redundancy via k-Nearest Neighbors (k-NN), identifying underrepresented poses and oversampled regions. We then convert these diagnostics into a closed-loop acquisition strategy that prioritizes informative targets—via real-world capture or simulation—under a fixed collection budget. Importantly, WiCompass does not attempt to align raw mmWave measurements with optical signals; it operates on modality-agnostic pose labels, so coverage analysis is not contingent on cross-modal feature alignment. *WiCompass is named to evoke a "compass" for data collection: it provides a principled guideline that steers mmWave HPE toward the most informative motions, rather than brute-force data scaling.* Our contributions are threefold:

- **Empirical diagnosis of the coverage bottleneck.** We identify insufficient data coverage—not model capacity—as the primary bottleneck in mmWave HPE (§2). Our analysis reveals severe out-of-distribution (OOD) fragility and extensive redundancy in current datasets.
- **An interpretable data coverage framework.** We propose a principled, backbone-independent framework to quantify coverage and redundancy in a latent pose space (§4 and §5). It provides visualizable diagnostics that pinpoint concrete gaps in existing data.
- **An actionable and coverage-efficient acquisition strategy.** Guided by these diagnostics, we design WiCompass, a closed-loop acquisition approach that targets underrepresented poses (§6). It delivers superior generalization with orders of magnitude fewer samples than baselines.

This study does not raise ethical concerns. All the code, datasets, and experiment scripts are publicly available at https://github.com/Galaxywalk/WiCompass

2 Pilot Study and Motivation

This section probes the main obstacles to progress in mmWave HPE. We first define the task, then present a pilot study showing that the dominant constraints are *data generalization* and *data efficiency*, rather than *model architectural capacity*.

Table 1: Open Source mmWave HPE Datasets.

Datasets	Data Format	Label Style	Size	Extra Annotations
MRI [24]	Point Cloud (PD)	COCO	160k	Subj, Scene, Action
mmBody [15]	PD	SMPL-X	67k	Subj, Scene
MMFi [14]	PD	Human3.6M	321k	Subj, Scene, Action
MilliPoint [25]	PD	Customized	213k	Subj, Action
HuPR [26]	Heatmap (HP)	LSPe	36k	No
RT-Pose [16]	HP & PD	Human3.6M	72k	(Multi-)Subj, Scene, Action

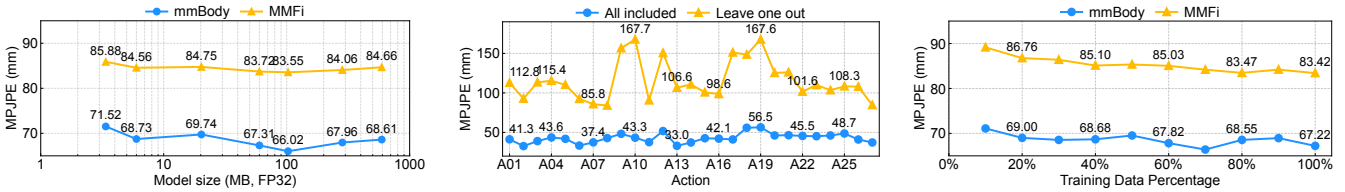
2.1 Task Definition

The objective of this work is to perform 3D HPE from mmWave radar data. We define the task as follows:

Input and Output. The deep learning model takes a mmWave point cloud of a human subject as *input* and regresses the 3D coordinates of the body joints as *output*. The input point cloud is generated from raw FMCW signals via a standard processing pipeline, which includes FFTs to create a radar heatmap and a Constant False Alarm Rate algorithm (CFAR) [27] for point extraction. For the output, we adopt the SMPL-X skeleton format [28] of 22 main joints, chosen for its compatibility with motion capture data over other common formats like COCO [29] and Human3.6M [17].

Datasets. An overview of publicly available mmWave HPE datasets is provided in Tab. 1. From this pool, we selected the mmBody [15] and MMFi [14] datasets for our experiments. Other datasets were intentionally excluded due to issues that would compromise experimental validity, such as poor data quality (e.g., misalignment between sensor readings and ground-truth labels) or the use of custom data formats incompatible with our processing pipeline.

Metrics. Our primary evaluation metric is the Mean Per-Joint Position Error (MPJPE), reflecting a model's absolute spatial accuracy. We relegate another common metric, Procrustes Aligned MPJPE (PA-MPJPE) to a secondary role. PA-MPJPE applies a rigid transformation (rotation, translation, scaling) to align poses before error calculation – a justifiable step for vision-based methods contending with 2D-to-3D scale ambiguities. This rationale is inapplicable to mmWave sensors, which provide direct and unambiguous 3D metric



(a) Model Performance vs. Model Size. **(b) Generalization Tests.** **(c) Dataset Efficiency.**

Figure 1: An overview of the pilot studies. (a) illustrates the relationship between model performance and size. (b) shows the results of leave-one-out generalization tests. (c) demonstrates the data efficiency of existing datasets.

Table 2: Comparison of MPJPE Errors.

	Original paper	X-Fi [12]	mmDiff [11]	**Our Model**
mmBody [15]	82.00	-	68.08	**67.22**
MMFi [14]	136.1	127.4	**82.73**	83.42

data. For this modality, any spatial error is a direct model failure, not a sensor artifact. PA-MPJPE's alignment masks these critical failures, yielding an overly optimistic score unrepresentative of real-world utility. We, therefore, contend that MPJPE is the truer measure of a deployable system's performance, and use PA-MPJPE exclusively as a diagnostic tool during model training.

2.2 Model is Not the Bottleneck

Motivated by scaling-law thinking in deep learning, we ask whether progress in mmWave HPE comes mainly from larger or more sophisticated models, or whether the true bottleneck lies elsewhere [30–32]. To test this, we use a scalable Point-Transformer model [33] with a transformer decoder and learnable joint queries. We vary model capacity by adjusting depth and width and train all variants on the mmBody and MMFi datasets (details in Appendix A) under an identical protocol. The trend in Fig. 1a is clear: error decreases at first but quickly plateaus around a model size of ~10 MB, after which adding more parameters gives only small gains or even worse performance. Compute is not a limiting factor, so this saturation cannot be explained by training budget. We then compare our baseline—the 102.64 M-parameter variant in Fig. 1a, also used in later experiments—against SOTA methods. As shown in Tab. 2, this straightforward architecture matches or even outperforms more complex designs on both MMFi and mmBody (*e.g.,* X-Fi [12] uses multi-modal pretraining, while mmDiff [11] adds physical and pose priors). Together, these results show that at the current data scale, model capacity—and by extension, compute—is not the limiting factor. This agrees with established scaling-law results highlighting the joint roles of model, data, and compute [31, 32, 34]. We therefore turn our focus to data coverage and redundancy, which we identify as the true bottlenecks for generalization in mmWave HPE.

2.3 The Dual Data Bottlenecks

Our analysis uncovers two data-level issues in current mmWave datasets: (i) poor out-of-distribution (OOD) generalization on unseen actions, and (ii) low data efficiency due to heavy redundancy.

Generalization Bottleneck. Although ID metrics for mmWave HPE look strong (see Tab. 2, even comparable to vision-based methods), models fail badly on unseen variations and tend to overfit the training distribution. To evaluate this systematically, we conduct a *leave-one-out* (LOO) action split on the MMFi dataset [14] [1]. In each round, the model trains on 26 actions and is tested on the held-out one, repeating this process for all 27 actions with the same training recipe as §2.2. For comparison, we also train an *all-included* (ID) model using all training splits. The results in Fig. 1b reveal several issues:

- **Severe overfitting from data sparsity.** While the ID MPJPE is 50.64 ± 7.67 mm,[2] the error on these held-out actions jumps to 122.10 ± 26.76 mm (~ 2.4×). This shows that the dataset lacks sufficient diversity to support generalization across unseen actions.
- **Insufficient variation prevents learning symmetries.** Failures on symmetric action pairs (e.g., "Lunge (toward left-front/right-front)", A9/A10; "Waving hand (left/right)", A17/A18) indicate that mirrored counterparts are treated as unrelated. Although both appear in the dataset, the model fails to learn the underlying symmetry, but merely employs a brittle memorization.
- **Outlier poses highlight imbalance.** Disproportionately high errors on complex actions such as "Picking up things" (A19) expose structural imbalance: the dataset is dominated by common upright postures, leaving rare but important motions underrepresented. Apparent wins (e.g.,

[1] MMFi provides annotations for 40 subjects, 27 actions, and 4 environments. We exclude mmBody because it lacks action annotations.
[2] The original dataset split is both cross-scene and cross-subject in Tab. 2, so the testing error decreases from 82.42 mm to 50.64 mm in this fully ID case.

"Bowing," A27) occur because key torso configurations are incidentally covered by another action ("Picking up things," A19), rather than from a robust representation of motion space. Such reliance on chance overlaps reflects insufficient diversity, producing models whose success is fragile and inconsistent.

Efficiency Bottleneck. A natural response to weak generalization is to scale up data, yet current collection practices are inefficient. We run a *progressive-subset* test—train the same model on random subsets of the training split of sizes of 100%, 90%, . . . , 10%, then evaluate on the original test set. As shown in Fig. 1c, performance stays almost unchanged until the subset shrinks below ∼30%: discarding about 70% of the training samples only increases error by <2%. This is not robustness but redundancy. Human-intuitive collection (e.g., pre-planned action lists, repeated takes of the same motion) oversamples already-dense regions of pose space while leaving important regions sparse. A large share of effort thus goes to near-duplicates that add little new information.

2.4 Summary and Motivation

Our pilot study rules out a model-centric explanation and identifies data as the primary obstacle to robust mmWave HPE. Two challenges emerge clearly:

- **Generalization Bottleneck.** Models do not learn invariant motion representations and can treat even mirrored actions as OOD, a gap that vision systems often close with massive, diverse datasets that implicitly teach spatial equivariances. It shows that mmWave datasets have limited coverage to various human motion data.
- **Efficiency Bottleneck.** Naively "scaling up" data under current practices yields highly redundant collections, consuming budget on near-duplicates while failing to cover informative, underrepresented motions.

This creates a paradox: we need substantially more diverse data, but existing pipelines acquire it inefficiently. To resolve this, we introduce WiCompass, an "oracle-driven" framework that aligns to a comprehensive pose latent space and uses quantitative coverage analysis to target collection toward maximally informative gaps—improving generalization without brute-force scaling.

3 Overview

The central challenge in mmWave HPE lies in the prohibitive cost of building datasets that are both large and truly diverse. Existing mmWave corpora remain narrow in scope, whereas the motion-capture community has curated broad repositories such as AMASS [35], which span a much richer spectrum of human motion. Our key idea is to treat this motion prior as an "oracle" that guides which mmWave samples should be collected next: instead of indiscriminately expanding data volume, we selectively target the samples that contribute most to generalization.

Fig. 2 depicts our framework, WiCompass, which operationalizes this principle through a compact, cross-modal pipeline with three stages. *(1) Shared latent space.* We begin by training a Vector Quantized Variational Autoencoder (VQ-VAE) model on AMASS to construct a modality-agnostic vocabulary of pose primitives. Both MoCap and mmWave poses are then embedded into this shared latent space (§4). *(2) Coverage analysis.* Within this space, we conduct a directional k-NN analysis to measure how much of the MoCap manifold is represented by the mmWave dataset and to reveal underrepresented regions. This analysis produces a multi-scale *coverage map*, and we further introduce a scale-free redundancy metric to quantify oversampling (§5). *(3) Guided acquisition.* Based on these diagnostics, WiCompass selects a batch of targets under a fixed budget using capped probability-proportional-to-size sampling: extreme outliers are first filtered out, after which candidate poses are sampled in proportion to their coverage density. The mmWave data for these targeted poses are then obtained via a unified interface supporting both real-world and simulation, and the resulting samples are merged back into the dataset (§6).

Together, these stages form a closed feedback loop: WiCompass identifies coverage gaps, translates them into concrete acquisition targets, and iteratively enriches dataset diversity while improving generalization efficiency.

4 Learning a Universal Vocabulary of Human Motion

Our pilot study highlights a critical bottleneck: mmWave HPE models suffer from poor generalization and redundancy. To overcome this, we advocate a shift in data scaling—from brute-force collection to a targeted, informed strategy. The key insight is that while existing mmWave datasets are sparse, the manifold of human motion is already comprehensively captured by large-scale MoCap datasets such as AMASS [35]. We can transfer this rich motion knowledge from the MoCap domain into the wireless domain.

This requires a common and high accuracy representation of human motion. Directly using raw 3D joint coordinates (66-D for 22 joints) is ill-suited: high dimensionality makes distances unreliable indicators of perceptual or semantic similarity between poses, and the space itself is unstructured and redundant. Instead, we learn a compact latent representation that encodes poses into a semantically meaningful form. In this section, we describe how we construct such a representation by training a "universal pose tokenizer" on large-scale MoCap data.

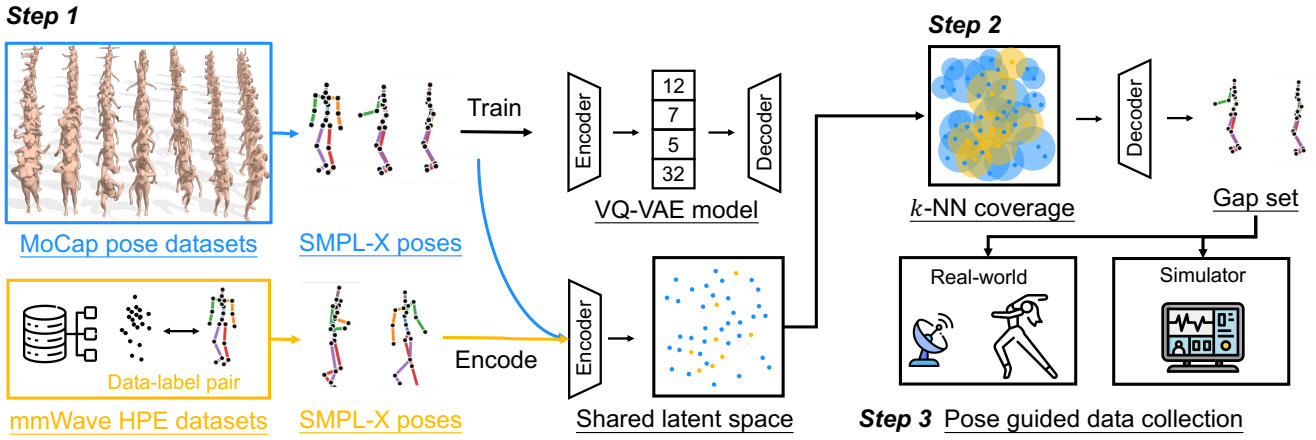


Figure 2: WiCompass Overview. Human poses of SMPL-X format from MoCap datasets and mmWave datasets are encoded into a shared latent space via VQ-VAE. Directional k-NN coverage identifies uncovered regions, which then guides pose-conditioned data collection in real or simulated settings.

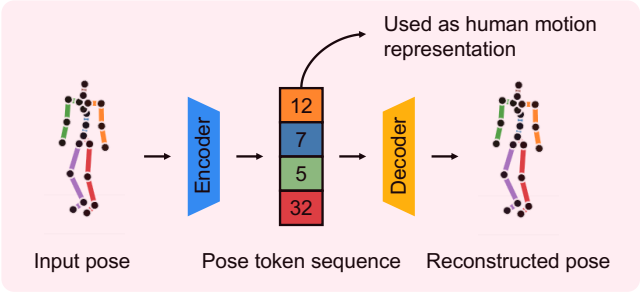


Figure 3: VQ-VAE Model Architecture.

4.1 A Discrete and Compositional Pose Representation

We employ a standard VQ-VAE architecture [36], trained exclusively on MoCap data, to learn a universal motion vocabulary (Fig. 3). A VQ-VAE is an autoencoder that compresses input poses into a latent space and reconstructs them from this compressed representation. Its distinctiveness lies in the latent space: rather than being continuous, it consists of a finite *codebook* of discrete *tokens*. Each pose must be reconstructed as a composition of these tokens, forcing the model to discover recurring motion patterns and store them as "words" in the codebook. The result is a meaningful, learned pose vocabulary. Representing any pose as a *token sequence* brings three key advantages:

- **Modality Agnosticism via Quantization.** Quantization maps any 3D pose—whether from optical MoCap or mmWave estimates—into a shared set of tokens. This yields a canonical vocabulary that is inherently modality-independent and enables direct comparison.
- **Tractable Distributional Analysis.** The discrete space makes coverage analysis straightforward: instead of comparing continuous distributions, we simply count tokens' distance. This renders the analysis computationally tractable, robust, and interpretable. Recent studies also show that discrete latent spaces enhance learning capacity and capture long-range motion dependencies [37–39].
- **Decoupling from Downstream Backbones.** Our latent space is learned through a generic reconstruction task, making it task-agnostic. In contrast, using features from a task-specific backbone (e.g., action classifiers) would introduce biases tied to that model's architecture and objective.

4.2 Model Architecture

The model architecture is shown in Fig. 3, which consists of an encoder E, a discrete codebook $C = \{\mathbf{c}_k\}_{k=1}^{K} \in \mathbb{R}^{K \times D}$, and a decoder D. The encoder maps an input 3D pose $\mathbf{x} \in \mathbb{R}^{J \times 3}$ to a sequence of continuous latent vectors $\mathbf{z}_e = E(\mathbf{x}) \in \mathbb{R}^{T \times D}$. Here, J denotes the number of joints (in our case, $J = 22$), T is the number of tokens (sequence length), and D is the dimension of each latent embedding. The codebook contains K distinct embedding vectors. The vector quantization mechanism, Q, then maps each continuous vector $z_{e,t}$ to its closest codebook embedding $\mathbf{c}_k$ via a nearest-neighbor lookup:

$$Q(z_{e,t}) = \mathbf{c}_k \quad \text{where} \quad k = \arg\min_{j \in \{1,\ldots,K\}} ||z_{e,t} - \mathbf{c}_j||_2^2 \tag{1}$$

The decoder D then attempts to reconstruct the original pose $\hat{\mathbf{x}} = D(Q(\mathbf{z}_e))$ from the sequence of quantized vectors.

4.3 Loss and Training

We train the model end-to-end on the AMASS dataset using the standard VQ-VAE objective, combining reconstruction loss with codebook commitment terms. The composite loss function balances reconstruction fidelity with the stability of the discrete latent space:

$$\mathcal{L} = \mathcal{L}_{\text{recon}} + ||\text{sg}(\mathbf{z}_e) - \mathbf{z}_q||_2^2 + \beta||\mathbf{z}_e - \text{sg}(\mathbf{z}_q)||_2^2 \quad (2)$$

where $z_q = Q(z_e)$ is the *quantized latent vector*, formed by replacing each continuous encoder output with its nearest codebook entry. The first term, $\mathcal{L}_{recon} = \|x - \hat{x}\|_2^2$, ensures reconstruction fidelity. The second term is VQ loss, which pulls codebook embeddings toward the encoder outputs, and the third term is the commitment loss (weighted by β), which encourages the encoder to commit to the discrete codebook vectors. To further improve stability, we update the codebook embeddings using an Exponential Moving Average (EMA) of the encoder outputs, which maintains a running average of the codebook embeddings and provides a more stable update rule compared to direct gradient-based updates [40].

Upon completion of training, the frozen encoder and codebook serve as our universal pose tokenizer. For any pose—whether from the reference optical MoCap dataset or a mmWave dataset—we can generate a canonical feature vector ϕ by looking up its token sequence in the codebook and concatenating the embeddings. This process projects poses from disparate sources into a single, unified latent space, enabling the direct distributional comparison detailed in §5. The decoder, however, is retained; it plays a crucial role in §6, where we use it to visualize the identified coverage gaps, allowing us to generate and inspect concrete examples of the missing poses that need to be collected.

5 Dataset Coverage Analysis

Having established a universal motion vocabulary in the latent space (§4), we now introduce a framework to quantify how well a given dataset occupies this space. Our goal is to move beyond anecdotal statements such as "this dataset lacks sitting motions" or "there is a lot of redundancy" and instead define metrics that: (i) measure how much of the motion manifold a dataset actually covers, and (ii) reveal how efficiently it uses its samples.

We build a k-Nearest Neighbors (k-NN)–based framework that provides a unified view of dataset quality from two complementary perspectives: (i) a *cross-dataset* analysis that identifies what is missing by comparing a mmWave dataset against a larger MoCap corpus, and (ii) an *intra-dataset* analysis that characterizes redundancy and internal coverage. In this subsection, we first define local neighborhoods in the latent manifold; later, we instantiate these neighborhoods into concrete coverage metrics.

5.1 Local Neighborhoods via k-NN

Our approach is inspired by manifold-based analyses in generative modeling [41–43]. In contrast to methods that estimate density directly (e.g., kernel density estimation) or rely on global clustering (e.g., k-means [41]), we use a local, nonparametric view that is better suited to high-dimensional latent spaces. We treat each latent pose $\phi \in \Phi$ as a point on the pose manifold and approximate its local support by a k-NN ball under cosine distance. Specifically, we define the cosine similarity style distance by:

$$d(\phi, \psi) = 1 - \frac{\phi^\top \psi}{\|\phi\|_2 \, \|\psi\|_2}, \quad (3)$$

and let $\text{NN}_k(\phi, \Phi)$ denote the k-th nearest neighbor of ϕ in Φ under d. The local radius around ϕ is

$$r_\phi^{(k)} = d\big(\phi, \text{NN}_k(\phi, \Phi)\big), \quad (4)$$

which we interpret as the size of the region over which a downstream model can reliably interpolate from ϕ. Intuitively, deep models obey a local smoothness assumption: they interpolate well within regions that are sufficiently dense in the training data, but can fail abruptly once test samples lie outside this region. The radius $r_\phi^{(k)}$ operationalizes this idea as a data-driven "interpolation zone" around each pose: a smaller k corresponds to a conservative generalizer that only handles poses very close to training examples, whereas a larger k corresponds to a stronger generalizer that can tolerate greater deviations. Rather than fixing a single value of k, we evaluate coverage across a range of k values. This multi-scale view exposes both fine-grained gaps (small k) and broader structural deficiencies (large k). A 2-D illustration of the k-NN neighborhood construction is shown in Fig. 4: when $k = 3$, for each sample, we find its two closest neighbors (excluding itself) from the same dataset (i.e., same color) as its support.

5.2 Cross-Dataset Analysis: Identifying Coverage Gaps

Given this notion of a local interpolation region, we define a directional *Coverage* metric that answers the question: *what fraction of poses in a reference set Φ_A are close to at least one pose in a target set Φ_B, relative to Φ_A's own notion of locality?*

Formally, a point $a_i \in \Phi_A$ is said to be covered by Φ_B if its nearest neighbor in Φ_B lies within its local radius $r_{a_i}^{(k)}$:

$$\text{covered}(a_i; \Phi_B) = \mathbb{I}\Big(d\big(a_i, \text{NN}_1(a_i, \Phi_B)\big) \le r_{a_i}^{(k)}\Big), \quad (5)$$

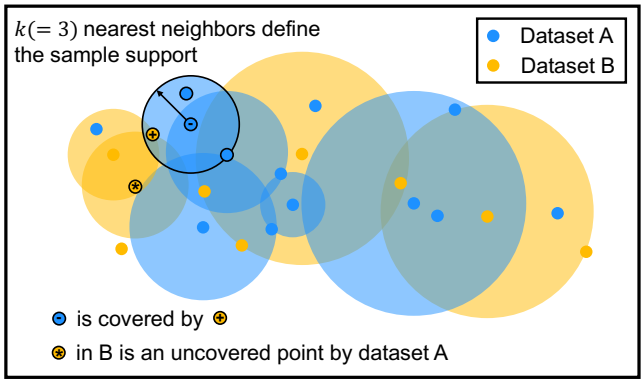


Figure 4: k-NN Coverage in the Latent Space.

where $\mathbb{I}(\cdot)$ is the indicator function and $\mathrm{NN}_1(a_i, \Phi_B)$ is the nearest neighbor of a_i in Φ_B under d. The directional coverage score is then

$$\mathrm{C}(\Phi_A \to \Phi_B) = \frac{1}{|\Phi_A|} \sum_{a_i \in \Phi_A} \mathbb{I}\Big(d\big(a_i, \mathrm{NN}_1(a_i, \Phi_B)\big) \le r_{a_i}^{(k)}\Big). \tag{6}$$

As illustrated in Fig. 4, a blue point from dataset A is covered if there exists a yellow point from B within its k-NN support region, whereas isolated yellow points outside A's support remain uncovered. A central byproduct of this analysis is the *gap set*, Φ_{G}: the subset of poses in the reference dataset that are not covered by the target dataset. When Φ_A is the large MoCap corpus and Φ_B is a mmWave dataset, Φ_{G} precisely pinpoints motions that are underrepresented in mmWave:

$$\Phi_{\mathrm{G}} = \left\{ a_i \in \Phi_{\mathrm{MoCap}} \,\middle|\, d\big(a_i, \mathrm{NN}_1(a_i, \Phi_{\mathrm{mmWave}})\big) > r_{a_i}^{(k)} \right\}. \tag{7}$$

In essence, a pose a_i from MoCap is flagged as a gap if we cannot find any pose in the mmWave dataset that is close enough to it, where close enough is adaptively determined by the local density of the motion manifold around a_i itself. This yields a precise and actionable candidate pool Φ_{G} for targeted data collection in WiCompass.

5.3 Intra-Dataset Analysis: Evaluating Redundancy

Next, we employ the k-NN coverage framework to analyze the internal structure of a single dataset. The core insight comes from the relationship between local data density, which is inversely proportional to the k-NN radius $r_\phi^{(k)}$, and information content:

- Samples in dense regions (small radii) are highly likely to be redundant. Adding a new sample that is close to many existing ones provides little new information and leads to the inefficiency we observed in our pilot study.
- Samples in sparse regions (large radii) are highly likely to be informative. They represent novel edge cases and are critical for improving a model's generalization.

While this principle is clear, using the raw k-NN radius $r_\phi^{(k)}$ directly as a comparable metric for redundancy is problematic, as its value is dependent on the dataset size n and the choice of k. To create a more principled and scale-free measure, we introduce the *Normalized Redundancy Index (NRI)*. The NRI converts the local radius into an estimate of the local "volume" required to contain k neighbors, thereby providing a scale-free density indicator [44, 45]. For each sample ϕ in a dataset of size n, we define its NRI as:

$$\mathrm{NRI}_\phi(k) \;=\; \frac{(n-1)\,[r_\phi^{(k)}]^{\,d_{\mathrm{eff}}}}{k}, \tag{8}$$

where d_{eff} is the effective intrinsic dimension of the latent space, estimated from the slope of an regression between $\log r^{(k)}$ and $\log k$ over multiple k, which follows from $r^{(k)} \propto (k/n)^{1/d}$ (Refer to Appendix B for the details). It allows us to estimate the local volume of the k-NN hypersphere correctly. A smaller NRI value signifies a higher local density and thus higher data redundancy. By analyzing the distribution of NRI values across a dataset, we can rigorously quantify its overall efficiency and identify over-sampled regions.

In summary, our framework provides two key outputs: (1) A concrete gap set (Φ_{G}) of missing poses, identified through cross-dataset analysis. (2) A principled NRI distribution to evaluate the internal efficiency of any dataset. The next section will focus on using this gap set to guide the high efficient data collection strategy.

6 Coverage-Driven Data Collection

The dataset analysis in §5 produces a *gap set* Φ_G: a large pool of candidate poses from the MoCap manifold that are insufficiently covered by the mmWave dataset. However, Φ_G is both high-volume and internally redundant. Blindly collecting all candidates would be inefficient and counterproductive. To address this, we design a coverage-driven data collection framework that is both *scalable* and *principled*.

As illustrated in Fig. 5, our framework supports two usage modes—*bootstrapping* a new dataset from scratch or *augmenting* an existing one—and two acquisition pathways: *real-world capture* or *simulation based synthesis*. This generality ensures that the method is deployable across diverse experimental contexts.

6.1 Target Selection via Capped-PPS Sampling

Given a collection budget of B samples, the key task is to select a target subset $\mathcal{T} \subset \Phi_G$ that maximizes information gain while avoiding redundancy and outliers. We adopt a

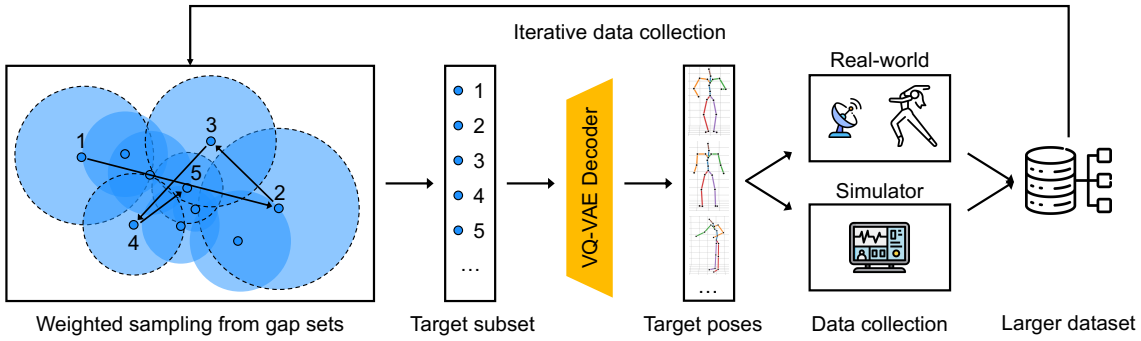


Figure 5: Coverage-driven Data Collection Workflow.

Capped Probability-Proportional-to-Size (Capped-PPS) sampling strategy, which leverages local density information while applying outlier control. Specifically, for each sample ϕ_i in Φ_G, we compute its k-NN radius $r_i^{(k)}$ as a proxy for local sparsity: samples with larger radii reside in underrepresented regions of the pose manifold. These radii are then converted into sampling weights:

$$w_i = \min\left(r_i^{(k)}, \mathrm{Q}_{TH}(\{r^{(k)}\})\right) \tag{9}$$

Here, TH is a percentile threshold and $\mathrm{Q}_{TH}(\{r^{(k)}\})$ is the corresponding quantile value, which that acts as a validity filter. In latent space analysis, extreme distance outliers often correspond to *physically implausible poses* or artifacts (e.g., severe self-intersections or floating limbs) rather than valid rare motions. By capping the weights at the TH-th percentile, we prevent a few extreme outliers from dominating the sampling budget and focus it on diverse yet plausible motions, rather than on implausible noise. [3]. Sampling is performed using the Efraimidis–Spirakis algorithm [46] for weighted sampling without replacement, where each candidate is assigned a priority score:

$$p_i = -\log u_i / w_i, \quad u_i \sim \mathcal{U}(0,1) \tag{10}$$

This method prioritizes sparse regions with higher sampling probability for diversity and avoids budget waste on implausible or outlier poses for robustness. It supports both data augmentation and bootstrapping. In augmentation mode, the candidate space is the known Φ_G; in bootstrapping mode, the candidate space is the full pose space.

6.2 Data Acquisition Pipeline

Once the B target embeddings $\mathcal{T}$ are selected, we instantiate them as samples using a flexible acquisition pipeline:

- **Pose Reconstruction.** For each latent sample $\phi \in \mathcal{T}$, the trained VQ-VAE decoder reconstructs a clean, canonical 3D skeleton $\mathbf{x}$. This ensures all downstream data generation is grounded in a unified latent representation.
- **Pose Realization and Data Capture.** The reconstructed skeleton is then realized for data capture in one of two modes. For real-world capture, the skeleton is rendered visually and mimicked by a human participant for radar capture. For synthetic capture, the skeleton is converted to an SMPL-X mesh and imported into a mmWave sensor simulator, which computes realistic radar returns based on pose, position, and radar parameters.
- **Stopping Criterion.** Unlike conventional pipelines with ambiguous termination points, WiCompass defines a clear stopping condition via the gap set Φ_G. As Φ_G represents the finite set of uncovered motions, collection concludes when Φ_G falls below a density threshold. In practice, we operate under a fixed budget B, strictly prioritizing the highest-weighted targets in Φ_G to maximize coverage.

After each acquisition round, the newly collected data is added to Φ_{mmWave}, and the coverage analysis in §5 is recomputed. This enables an *iterative closed-loop acquisition process*: each round targets new coverage gaps, steadily improving dataset diversity, efficiency, and generalization power.

7 Evaluation

Our evaluation is designed to provide a systematic and evidence-based validation of the proposed framework. To this end, we first report the end-to-end performance of our data collection method. Then we conduct a series of microbenchmarks and measurement studies to examine the validity of our latent space representation and metric reasonableness.

7.1 Experimental Setup

To ensure rigor and reproducibility, we design our experiments around a combination of large-scale motion capture priors, representative mmWave datasets, a standardized training environment, simulation pipeline, benchmark dataset, and real-world experiment setups.

[3] In our implementation, we set TH= 0.95 to filter out the last 5% long-tail outliers and k= 4 based on the effective intrinsic dimension of the pose latent space calculated in §B.

Vision MoCap Datasets. We adopt the AMASS dataset [35] as our comprehensive motion capture reference. AMASS aggregates dozens of optical HPE datasets into a unified SMPL-X parameterization and contains over 8.9M motion sequences. Its scale and diversity make it a suitable surrogate for the broader human pose manifold, serving as the reference prior for our coverage analysis.

MmWave HPE Datasets. Our evaluation focuses on two widely used mmWave HPE datasets: mmBody [15] and MMFi [14]. We choose them because they are the largest and most representative publicly available datasets with reliable annotations, while other resources either suffer from alignment and quality issues or adopt bespoke formats incompatible with our pipeline. mmBody provides SMPL-X parameterized labels of 22 keypoints human skeletons, whereas MMFi follows the Human3.6M skeleton (17 keypoints) with additional subject and action annotations. Since our VQ-VAE operates on SMPL-X inputs, we convert MMFi labels via a MoCap-to-SMPL-X pipeline [28, 47] to ensure consistency.

Model and Training Environment. Our universal pose tokenizer is a VQ-VAE with an MLP-Mixer backbone [48], trained exclusively on AMASS. The human pose estimator is a Point-Transformer [33] model identical to our pilot study (§2). All k-NN operations for coverage analysis are accelerated with FAISS-GPU [49]. Experiments are run on a server equipped with a Genuine Intel(R) CPU, 32 GB RAM, and six NVIDIA GeForce RTX 3090 GPUs (24 GB each). We use a batch size of 16 and train for 25 epochs.

Simulation Pipeline. We evaluate the data scaling efficiency under simulation environments to avoid domain gap and ensure fairness among different methods. We adapt RF Genesis [21], a ray-tracing-based simulator for mmWave HPE. The simulated radar is configured with a 77 GHz carrier frequency, 3.5 GHz bandwidth, a 3Tx–4Rx antenna array, and 256 ADC samples. It operates at 10 frames per second, with each frame comprising 128 chirps. The original simulator takes textual pose descriptions, generates SMPL-X meshes via a diffusion model [50], and outputs radar range Doppler spectra and point clouds. We modify its pipeline to accept single-frame skeleton inputs directly, bypassing text-to-pose synthesis. We manually validate the generated point clouds for geometric plausibility and cross-verify them against reference SMPL-X poses. To evaluate out-of-distribution (OOD) generalization, we construct a challenging benchmark of 40k poses uniformly sampled from the latent pose space of Φ_{AMASS}, which represents all possible and reasonable human poses. This benchmark is strictly held out from all training and validation, ensuring no data leakage.

Real-World Data Collection Pipeline. We also build a real-world acquisition pipeline. Our mmWave front-end is a 3Tx–4Rx FMCW radar [51] operating at 79 GHz with 6 GHz bandwidth and 15 frames per second, placed about 3 meters in front of the subject. It is paired with an Intel RealSense D455 RGB-D camera and a four-camera vision motion capture system [52, 53] that provides 3D ground-truth poses. The cameras are rigidly mounted around the sensing area, and we perform joint spatial and temporal calibration to align radar, RGB-D, and MoCap coordinate frames, ensuring accurate human-pose–mmWave point-cloud correspondence. The RGB-D stream is used only for calibration; supervision comes solely from the vision MoCap labels.

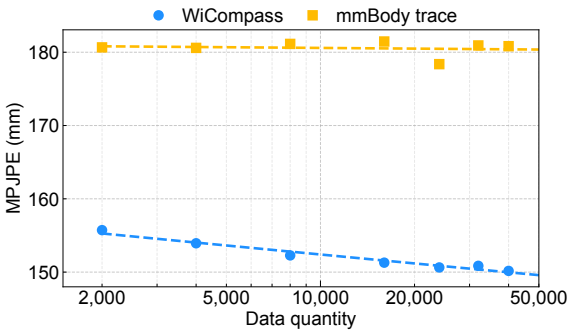


Figure 6: Scaling Laws Comparison on Synthetic Data.

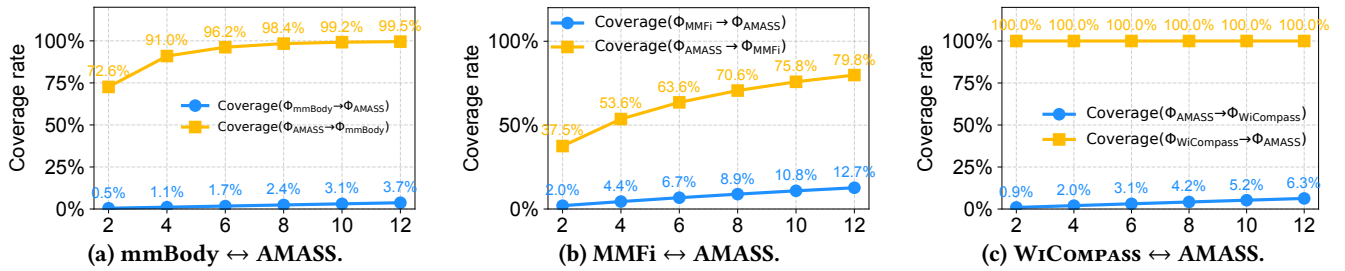
7.2 Data Scaling Efficiency

We ask whether our coverage-aware acquisition policy yields better OOD scaling than the conventional sequential-action protocol under the same data budget. To answer this, we evaluate the data scaling efficiency of WiCompass with simulation data. We compare two acquisition strategies, each generating up to 40k samples (matching the size of the mmBody training set) and evaluated on our OOD generalization benchmark: (i) our proposed WiCompass, following the sampling weights in Eqn. 9, and (ii) a baseline designed to mirror existing mmWave collection practices.

Since we cannot replicate the hardware and experimental conditions of existing open sourced datasets, all comparisons are conducted in a controlled simulation environment. Specifically, we reproduce the mmBody dataset's sampling trajectory by feeding its ground-truth skeleton labels into the simulation pipeline, producing a "simulated mmBody" dataset. This baseline reflects the prevailing methodology in existing datasets: predefined actions are captured sequentially without regard to motion diversity or distributional coverage. Each strategy yields a separate synthetic dataset, on which we train identical models using the same configuration; we then measure OOD MPJPE on the held-out generalization benchmark.

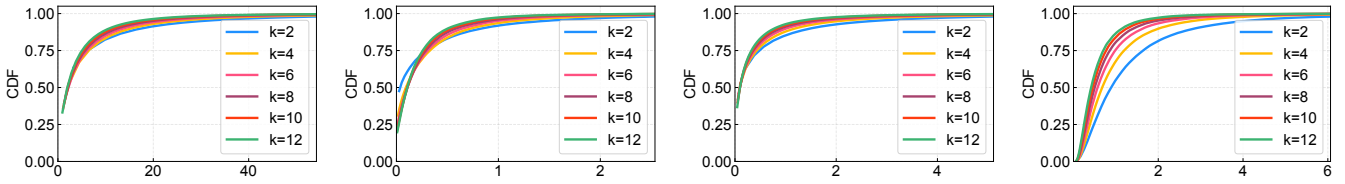
Results. To quantify how performance evolves with scale, we adopt a power-law model for data scaling [31]:

$$E(D) = AD^{\alpha_D}, \tag{11}$$



(a) mmBody ↔ AMASS. (b) MMFi ↔ AMASS. (c) WiCompass ↔ AMASS.

Figure 7: Cross-Dataset Coverage vs. k.



(a) AMASS. (b) mmBody. (c) MMFi. (d) WiCompass.

Figure 8: NRI Distribution for Intra-Dataset Coverage Analysis.

where D is the dataset size, $E(D)$ is the testing error, A is a fitted constant, and α_D is the data-scaling exponent that characterizes how rapidly performance changes with additional data samples. We fit this model via nonlinear least-squares regression on log–log transformed data, and the resulting curves in Fig. 6 reveal a clear difference between coverage-aware and conventional collection. For WiCompass, the best-fit curve is $E_{\text{WiCompass}}(D) \approx 169.57\, D^{-0.0116}$, with a strong log–log correlation ($r \approx -0.98$), indicating a stable power-law regime. In contrast, the mmBody-trace baseline is fit by $E_{\text{mmBody-trace}}(D) \approx 181.86\, D^{-0.0008}$, with a much weaker correlation ($r \approx -0.15$). As we increase the simulated dataset size, WiCompass yields a consistent monotonic decrease in OOD MPJPE, with an order-of-magnitude larger (more negative) scaling exponent than the baseline. This indicates that additional samples collected under the coverage-aware policy remain informative even at tens of thousands of frames. Using these fits, WiCompass reduces OOD MPJPE by roughly 25–30 mm compared to the baseline across 2k–40k samples, and the gap widens as D grows.

7.3 Dataset Coverage Measurement

We next relate the scaling behaviors to coverage metrics in the latent pose space. We use the k-NN coverage metric (§5) to evaluate the pose coverage of existing mmWave datasets (mmBody, MMFi) and our collected dataset WiCompass against the AMASS reference, and to quantify their internal redundancy.

Cross-Dataset Coverage. Coverage is defined as the fraction of MoCap poses that are supported—i.e., have at least one mmWave sample within their local neighborhood of radius $r^{(k)}$. Fig. 7 shows directional coverage as a function of k. Key observations include:

- Coverage increases smoothly with k for all datasets, indicating that our metric behaves consistently.
- mmBody covers a narrow subregion of AMASS. In Fig. 7a, coverage of AMASS by mmBody (blue) reaches only 3.7% at $k = 12$, while the reverse direction (yellow) exceeds 99%. This suggests that mmBody occupies a narrow, but valid, subset of the MoCap manifold.
- MMFi exhibits limited and noisy coverage. In Fig. 7b, coverage of MMFi by AMASS (yellow) reaches 79.8% at $k = 12$, below mmBody. We attribute this to label noise introduced during Human3.6M-based annotation—specifically, missing joints (e.g., neck, clavicles, ankles) cause alignment errors despite the use of optimization-based format conversion. Nevertheless, the reverse coverage (blue) still reveals distinctive motion regions that MMFi fails to capture, highlighting opportunities for future data acquisition.
- WiCompass achieves superior coverage efficiency. In Fig. 7c, coverage of WiCompass by AMASS (yellow) is 100% since WiCompass is generated from the AMASS pose space. More importantly, coverage of AMASS by WiCompass (blue) consistently exceeds that of mmBody and comes within a factor of two of MMFi, while using roughly one-eighth as many samples. This indicates that our sampling method achieves broader and more representative coverage per unit of data budget.

Intra-Dataset Coverage. To assess internal redundancy, we compute the cumulative distribution function (CDF) of NRI, as shown in Fig. 8. For each dataset, we set the x-axis limits from $\min_k Q_0(\text{NRI}_k)$ to $\max_k Q_{98}(\text{NRI}_k)$ across all plotted k

values to suppress extreme outliers while keeping all CDF curves visible. Key observations include:

- AMASS shows broad coverage with informative samples. Fig. 8a reveals a wide NRI spread, reflecting diverse and sparsely sampled poses. It serves as a reference baseline for well-structured motion coverage.
- MmBody (Fig. 8b) suffers from high redundancy. Its CDF rises steeply with uniformly low NRI values, confirming that mmBody contains many highly redundant samples. MMFi (Fig. 8c) exhibits moderate redundancy.
- WiCompass encourages sparse, informative sampling. Fig. 8d exhibits the broadest NRI distribution among all mmWave datasets, validating that our coverage-driven acquisition strategy effectively targets underrepresented regions of the motion manifold.

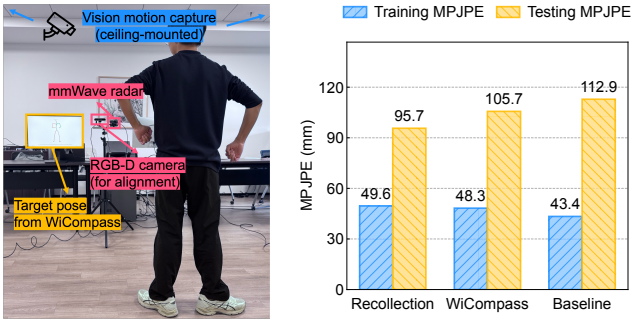
The consistency of NRI curves across different neighborhood sizes k further supports its utility as a *robust and scale-insensitive metric* for evaluating intra-dataset redundancy.

Metrics as Generalization Proxies. We further validate that these metrics serve as robust, predictive proxies for generalization—rather than post-hoc descriptors—through two steps: independent diagnosis and predictive confirmation. First, the metrics correctly diagnose the failures observed in our pilot study on existing datasets (§2.3). For instance, the poor OOD generalization aligns with our finding that mmBody covers only 3.7% of the pose manifold (Fig. 7a). Likewise, the finding that 70% of data is discardable aligns with our NRI analysis, which exposes massive structural redundancy (Fig. 8b). This confirms the metrics are valid independent of WiCompass. Second, the metrics accurately predict WiCompass's efficiency gains. By targeting the identified coverage gaps, WiCompass achieves both broader metric distribution and significantly lower OOD error (§7.2). This direct correlation between metric improvement and error reduction confirms that k-NN coverage and NRI are predictive proxies that effectively guide data quality.

7.4 Real-world Validation

We conduct a closed-loop real-world feasibility validation, demonstrating that WiCompass can instantiate concrete acquisition targets and drive practical mmWave HPE data collection that improves generalization under a fixed budget.

Protocol and Acquisition Strategies. We choose bodyweight exercises as a challenging case study because its large-amplitude, whole-body motions are substantially more diverse than the upright postures dominating existing mmWave datasets. Two volunteers performed exercise sequences to form a held-out benchmark of 8k frames, which is strictly reserved for testing. Under an identical budget of 8k training frames, we compare three acquisition strategies: (i) Recollection. The same motion family (bodyweight exercises) is captured again, providing the closest possible distributional match to the benchmark; (ii) WiCompass. We project the target motion family into the universal latent pose space and use our acquisition policy to select informative target poses; the volunteer then mimics these on-screen visualizations for radar capture; (iii) baseline. Following common practice in prior datasets (e.g., MMFi), we collect data by sequentially recording 10 common actions (i.e., standing, walking, jogging, turning, squatting down and up, waving left/right, raising left/right hand, and jumping).



(a) Setups. **(b) Generalization Errors.**

Figure 9: Real-world Validation.

Results. Fig. 9 reports MPJPE on the training split and on the held-out benchmark. All methods achieve comparable training errors (43.4–49.6 mm), but their generalization differs substantially. The recollection set attains the lowest test error (95.7 mm), as expected: by construction, its acquisition list exactly matches the poses in the benchmark, so the training and test distributions are identical. It therefore serves as an oracle upper bound rather than a realistic data-collection strategy. In this context, WiCompass achieves a test MPJPE of 105.7 mm, moving significantly closer to the recollection oracle, and clearly outperforming the baseline (112.9 mm). Notably, the baseline attains the lowest training error (43.4 mm) yet the worst test error, underscoring that conventional action-list collection can be sample-inefficient for generalization. Although absolute errors remain relatively high due to the limited real-world scale (8k frames), this feasibility study demonstrates an end-to-end, practical closed-loop pipeline and provides direct evidence that WiCompass prioritizes more informative real mmWave samples under a fixed collection budget.

7.5 VQ-VAE Microbenchmarks

The latent space extracted from the VQ-VAE bottleneck layer is the foundation of our framework. We evaluate whether the learned pose vocabulary is compact, representative, and semantically interpretable.

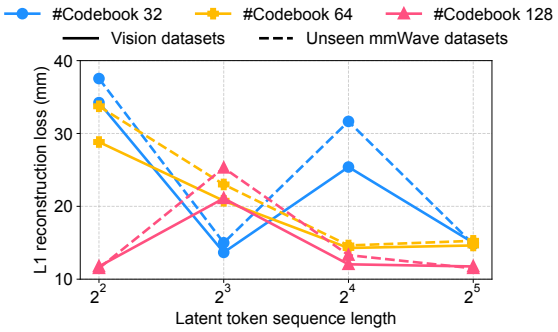


Figure 10: Reconstruction Fidelity of VQ-VAE Model.

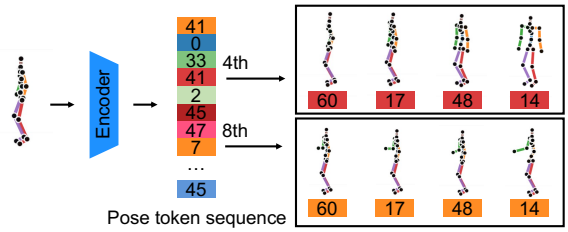


Figure 11: Pose Token Influence Visualization.

Reconstruction Fidelity of VQ-VAE Model. The goal of the VQ-VAE model is to encode human poses into a discrete latent space that is both semantically meaningful and computationally efficient. Two hyperparameters are critical: the codebook size K and the latent sequence length T. Smaller K and T maximize compression efficiency and enforce semantic compactness, but overly small values risk under-representing the pose manifold and causing high reconstruction error. We compute the L1 reconstruction loss on two datasets: the AMASS test set and the unseen mmBody dataset. Fig. 10 plots the average reconstruction error across different latent configurations, varying the codebook size and sequence length. We make the following remarks:

- Generally, the reconstruction fidelity decreases with larger K or T. At $K = 64$, $T = 16$, the reconstruction error stabilizes at approximately 10 mm on both AMASS and mmBody. This error is on par with the groundtruth error of MoCap systems [54, 55]. After this point, increasing K or T slightly improves fidelity but yields diminishing returns. Therefore, we select $K = 64$, $T = 16$ as the default configuration in our all experiments.
- These results confirm that the latent representation generalizes across modalities, supporting downstream use in coverage analysis.

Qualitative Visualization. Fig. 11 illustrates token-level influence by modifying one latent token at a time in mmBody poses. The results show that tokens correspond to distinct motion primitives—for example, the fourth token controls global orientation, while the eighth affects the right elbow. This demonstrates the semantic interpretability of the codebook, consistent with prior findings in pose-space studies [37–39]. Additional visualization cases, including poses from existing mmWave HPE datasets as well as those uncovered relative to AMASS, are provided in Appendix C.

8 Related Work

MmWave HPE. Recent mmWave HPE researches follow three major directions. *(1) Signal representation.* Models either operate directly on raw radar cubes with 3D CNNs or Transformers, project RF signals into 2D range–angle or range–Doppler views for heatmap prediction, or reconstruct sparse point clouds for point-based backbones with temporal aggregation [15, 56, 57]. *(2) Structure-aware inference.* Body priors improve robustness under sparsity and occlusion, including graph–attention hybrids for part refinement and diffusion-style denoisers that regularize kinematics [11, 26]. *(3) Multi-view and cross-modality fusion.* Multi-view radar mitigates directionality and self-occlusion via learned fusion, while vision-derived labels bootstrap radar-only models without manual annotation [26, 58]. Despite these advances, all approaches depend on existing datasets so struggle with poor generalization to unseen actions, subjects, and environments. Our work tackles this limitation from a distributional perspective, orthogonal to architectural design.

MmWave HPE Data Scaling. Scaling datasets and simulators has been another major driver of progress. Public corpora such as mRI, mmBody, MMFi, MilliPoint, HuPR, and RT-Pose enable supervised or cross-modality training, but their coverage remains uneven and limited across subjects, actions, and environments [14–16, 24–26], as we measured in the §2.3 and §7.3. Synthetic pipelines—ranging from video-to-radar translation [59–61] to hybrid physics–generative methods [21, 62]—expand cheaply but face sim-to-real gaps (multipath, materials, hardware) and often impose distribution shift problem [23]. Our approach is complementary: rather than indiscriminately enlarging datasets, we provide a distribution-aware criterion that directs both real and synthetic collection toward underrepresented regions.

Active Learning and Core-Set. Our objective of optimizing data collection parallels active learning (AL) and core-set selection. Classical AL methods optimize sample efficiency by querying the most informative points from a large unlabeled pool using uncertainty or diversity metrics (e.g., entropy, margins, query-by-committee) [63–66], while core-set methods emphasize geometric diversity through k-center or submodular objectives [67–72]. Crucially, however, these approaches are fundamentally *selective*: they assume a large, pre-collected pool of candidate sensor data and operate by reweighting or filtering this pool. In contrast, WiCompass adopts an *oracle-guided, plan-based acquisition* paradigm.

Rather than scoring an existing pool, it constructs an acquisition plan directly in the MoCap oracle pose space and then executes this plan in the radar domain, synthesizing novel, informative targets that do not yet exist as mmWave measurements. This shifts the workflow from "filtering a pool" to "synthesizing a plan," bypassing the need for massive and blind data pre-collection.

Pose Latent Space. Latent variable models, particularly VQ-VAEs, are established tools in computer vision for motion generation and compression. We emphasize that our contribution is not the architecture itself, but the *novel repurposing* of its latent space as a diagnostic framework for wireless sensing. While prior work restricts latent spaces primarily to instance-level tasks (e.g., reconstruction or generation) [28, 37–39, 73], we utilize the discrete latent space as a *distributional probe*. By projecting disparate sources—optical MoCap priors and mmWave pose labels—into this shared density manifold, we translate abstract concepts of "dataset coverage" and "redundancy" into computable metrics (k-NN radius and NRI). This enables rigorous quantification of mmWave dataset quality in a manner inaccessible to raw signal analysis alone.

9 Discussion

Latent Space Completeness. The utility of our latent pose space hinges on the diversity of its underlying motion data. We leverage AMASS [35], a de facto human motion prior aggregating 40 hours of diverse captures—including daily actions (CMU, KIT), object interactions (GRAB), and extreme postures (MOYO). However, completeness remains an ideal; culturally specific gestures and rare outdoor behaviors are still underrepresented, as evidenced by the coverage never achieve 100% in Fig. 7. Furthermore, VQ-VAE discretization, while enabling tractable analysis, inevitably merges subtle variations into shared codewords, reducing sensitivity to fine-grained differences. Future work should explore hierarchical quantization to capture more nuanced motion structures.

Sim-to-Real Gap. Our simulator serves as an efficient testbed for evaluating the *relative* sample efficiency of acquisition strategies. We acknowledge that it does not fully capture real-world complexities such as multipath propagation and hardware noise; thus, absolute performance metrics in simulation may not directly transfer to deployment. However, the selection signal in WiCompass relies on modality-agnostic pose labels in the shared latent space rather than raw radar features. This design ensures that our target selection policy remains invariant to the sim-to-real gap, even if the downstream radar training is domain-dependent. Furthermore, ensuring identical simulation protocols across all baselines guarantees that observed gains arise strictly from the acquisition strategy rather than data artifacts.

Evaluation Scale and Extrapolated Conclusions. Our experiments at the 40k-sample scale (simulation) and 8k-sample scale (real-world) demonstrate that WiCompass offers a superior acquisition strategy. However, verifying whether the observed log–log scaling behavior persists at the massive scales seen in vision and language models remains an open question. MmWave sensing involves fundamentally different data sparsity patterns, and we are currently constrained by the prohibitive cost of large-scale real-world collection. While our fitted curves provide meaningful extrapolations within the tested regime, we caution against generalizing these trends to arbitrarily large scales without further validation. Future work will focus on large-scale, multi-round real-world studies to explore these limits.

Generalization Across Multiple Dimensions. This work primarily addresses the *pose diversity* (geometric coverage) to ensure the model observes the comprehensive manifold of human kinematics. We recognize that real-world robustness also demands generalization across physical environments (e.g., multipath and clutter). We view this "sensing gap" as distinct from the kinematic gap addressed by WiCompass; they are orthogonal problems that must be solved in tandem. A key strength of our framework is its extensibility: the proposed coverage analysis are domain-agnostic. Future work can extend this framework to environmental latent spaces, enabling the joint targeting of kinematic and environmental diversity for holistic wireless sensing.

10 Conclusion

This paper shows that the key bottleneck in mmWave HPE lies not in model/algorithm but in data. By quantifying coverage and redundancy in a modality-agnostic latent space, we transform vague concerns about diversity into actionable diagnostics and a principled acquisition strategy. Our coverage-driven loop consistently improves generalization comparing to the baseline trajectories. More broadly, the lesson for wireless sensing is clear: prioritize data quality over scale, treat datasets as engineered artifacts with measurable coverage, and guide costly RF data collection with inexpensive cross-modal priors. By shifting attention to improving data quality, WiCompass provides a reproducible path toward robust, data-driven wireless sensing.

Acknowledgments

This work is supported by National Key R&D Program of China (Grant No. 2023YFF0725004), National Natural Science Foundation of China (Grant No. U25B2040, 62502012, and 62272010), and the Postdoctoral Fellowship Program and China Postdoctoral Science Foundation (Grant No. GZC20251070). Chenren Xu is the corresponding author.

References

[1] 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.

[2] Gines Hidalgo, Yaadhav Raaj, Haroon Idrees, Donglai Xiang, Hanbyul Joo, Tomas Simon, and Yaser Sheikh. Single-network whole-body pose estimation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6982–6991, 2019.

[3] Sword Health. A better way to treat pain. https://swordhealth.com/, 2025. Accessed: 2025-08-09.

[4] Google. Pose landmarker | mediapipe. https://ai.google.dev/edge/mediapipe/solutions/vision/pose_landmarker, 2025. Accessed: 2025-08-09.

[5] 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.

[6] Cong Shi, Li Lu, Jian Liu, Yan Wang, Yingying Chen, and Jiadi Yu. mpose: Environment-and subject-agnostic 3d skeleton posture reconstruction leveraging a single mmwave device. *Smart Health*, 23:100228, 2022.

[7] Hongfei Xue, Qiming Cao, Chenglin Miao, Yan Ju, Haochen Hu, Aidong Zhang, and Lu Su. Towards generalized mmwave-based human pose estimation through signal augmentation. In *Proceedings of the 29th Annual International Conference on Mobile Computing and Networking*, pages 1–15, 2023.

[8] Shuting Hu, Siyang Cao, Nima Toosizadeh, Jennifer Barton, Melvin G Hector, and Mindy J Fain. mmpose-fk: A forward kinematics approach to dynamic skeletal pose estimation using mmwave radars. *IEEE sensors journal*, 24(5):6469–6481, 2024.

[9] 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 *Proceedings of the 19th Annual International Conference on Mobile Systems, Applications, and Services*, pages 269–282, 2021.

[10] Hao Kong, Xiangyu Xu, Jiadi Yu, Qilin Chen, Chenguang Ma, Yingying Chen, Yi-Chao Chen, and Linghe Kong. m3track: mmwave-based multi-user 3d posture tracking. In *Proceedings of the 20th Annual International Conference on Mobile Systems, Applications and Services*, pages 491–503, 2022.

[11] 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*, pages 1–18. Springer, 2024.

[12] Xinyan Chen and Jianfei Yang. X-fi: A modality-invariant foundation model for multimodal human sensing. *ArXiv*, 2024.

[13] 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.

[14] 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, 2023.

[15] 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. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 3501–3510, 2022.

[16] 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.

[17] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE transactions on pattern analysis and machine intelligence*, 2013.

[18] Hanbyul Joo, Hao Liu, Lei Tan, Lin Gui, Bart Nabbe, Iain Matthews, Takeo Kanade, Shohei Nobuhara, and Yaser Sheikh. Panoptic studio: A massively multiview system for social motion capture. In *The IEEE International Conference on Computer Vision (ICCV)*, 2015.

[19] Zhongang Cai, Daxuan Ren, Ailing Zeng, Zhengyu Lin, Tao Yu, Wenjia Wang, Xiangyu Fan, Yang Gao, Yifan Yu, Liang Pan, et al. Humman: Multi-modal 4d human dataset for versatile sensing and modeling. In *European Conference on Computer Vision*, pages 557–577. Springer, 2022.

[20] Cong Yu, Zhi Wu, Dongheng Zhang, Zhi Lu, Yang Hu, and Yan Chen. Rfgan: Rf-based human synthesis. *IEEE Transactions on Multimedia*, 25:2926–2938, 2022.

[21] Xingyu Chen and Xinyu Zhang. Rf genesis: Zero-shot generalization of mmwave sensing through simulation-based data synthesis and generative diffusion models. In *Proceedings of the 21st ACM Conference on Embedded Networked Sensor Systems*, pages 28–42, 2023.

[22] Guoxuan Chi, Zheng Yang, Chenshu Wu, Jingao Xu, Yuchong Gao, Yunhao Liu, and Tony Xiao Han. Rf-diffusion: Radio signal generation via time-frequency diffusion. In *Proceedings of the 30th Annual International Conference on Mobile Computing and Networking*, pages 77–92, 2024.

[23] Chen Gong, Bo Liang, Wei Gao, and Chenren Xu. Data can speak for itself: Quality-guided utilization of wireless synthetic data. *arXiv preprint arXiv:2506.23174*, 2025.

[24] Sizhe An, Yin Li, and Umit Ogras. mRI: Multi-modal 3d human pose estimation dataset using mmwave, RGB-d, and inertial sensors. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022.

[25] Han Cui, Shu Zhong, Jiacheng Wu, Zichao Shen, Naim Dahnoun, and Yiren Zhao. Milipoint: A point cloud dataset for mmwave radar. *Advances in Neural Information Processing Systems*, 36:62713–62726, 2023.

[26] 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.

[27] Duo Zhang, Xusheng Zhang, Shengjie Li, Yaxiong Xie, Yang Li, Xuanzhi Wang, and Daqing Zhang. Lt-fall: The design and implementation of a life-threatening fall detection and alarming system. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 7(1):1–24, 2023.

[28] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10975–10985, 2019.

[29] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[30] Joel Hestness, Sharan Narang, Newsha Ardalani, Gregory Diamos, Heewoo Jun, Hassan Kianinejad, Md Mostofa Ali Patwary, Yang Yang, and Yanqi Zhou. Deep learning scaling is predictable, empirically. *arXiv preprint arXiv:1712.00409*, 2017.

[31] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint*

arXiv:2001.08361, 2020.

[32] Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. Scaling vision transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12104–12113, 2022.

[33] 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.

[34] Margaret Li, Sneha Kudugunta, and Luke Zettlemoyer. (mis)fitting: A survey of scaling laws. *arXiv preprint arXiv:2502.18969*, 2025.

[35] Naureen Mahmood, Nima Ghorbani, Nikolaus F Troje, Gerard Pons-Moll, and Michael J Black. Amass: Archive of motion capture as surface shapes. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5442–5451, 2019.

[36] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.

[37] Chuan Guo, Xinxin Zuo, Sen Wang, and Li Cheng. Tm2t: Stochastic and tokenized modeling for the reciprocal generation of 3d human motions and texts. In *European Conference on Computer Vision*, pages 580–597. Springer, 2022.

[38] Jianrong Zhang, Yangsong Zhang, Xiaodong Cun, Yong Zhang, Hongwei Zhao, Hongtao Lu, Xi Shen, and Ying Shan. Generating human motion from textual descriptions with discrete representations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14730–14740, 2023.

[39] Muhammad Usama Saleem, Ekkasit Pinyoanuntapong, Pu Wang, Hongfei Xue, Srijan Das, and Chen Chen. Genhmr: Generative human mesh recovery. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 6749–6757, 2025.

[40] Ali Razavi, Aaron Van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with vq-vae-2. *Advances in neural information processing systems*, 32, 2019.

[41] Mehdi SM Sajjadi, Olivier Bachem, Mario Lucic, Olivier Bousquet, and Sylvain Gelly. Assessing generative models via precision and recall. *Advances in neural information processing systems*, 31, 2018.

[42] Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved precision and recall metric for assessing generative models. *Advances in neural information processing systems*, 32, 2019.

[43] Ahmed Alaa, Boris Van Breugel, Evgeny S Saveliev, and Mihaela Van Der Schaar. How faithful is your synthetic data? sample-level metrics for evaluating and auditing generative models. In *International conference on machine learning*, pages 290–306. PMLR, 2022.

[44] Don O Loftsgaarden and Charles P Quesenberry. A nonparametric estimate of a multivariate density function. *The Annals of Mathematical Statistics*, 36(3):1049–1051, 1965.

[45] Gérard Biau and Luc Devroye. *Lectures on the nearest neighbor method*, volume 246. Springer, 2015.

[46] Pavlos S Efraimidis and Paul G Spirakis. Weighted random sampling with a reservoir. *Information processing letters*, 97(5):181–185, 2006.

[47] XinTao Lv et al. Inter-x: Towards versatile human-human interaction analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2024.

[48] Zigang Geng, Chunyu Wang, Yixuan Wei, Ze Liu, Houqiang Li, and Han Hu. Human pose as compositional tokens. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 660–671, 2023.

[49] Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with GPUs. In *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, pages 1–9. IEEE, 2017.

[50] Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H Bermano. Human motion diffusion model. *arXiv preprint arXiv:2209.14916*, 2022.

[51] DuoPucetan. Duopucetan radar. https://web.archive.org/web/20251221040934/https://duopucetan.com/pd.jsp?recommendFromPid=0&id=18&fromMid=522http://duopucetan.com/pd.jsp?recommendFromPid=0&id=18&fromMid=522, 2025. Accessed: 2025-12-20.

[52] Qing Shuai, Chen Geng, Qi Fang, Sida Peng, Wenhao Shen, Xiaowei Zhou, and Hujun Bao. Novel view synthesis of human interactions from sparse multi-view videos. In *SIGGRAPH Conference Proceedings*, 2022.

[53] Easymocap - make human motion capture easier. Github, 2021.

[54] Lars Mündermann, Stefano Corazza, and Thomas P Andriacchi. The evolution of methods for the capture of human movement leading to markerless motion capture for biomechanical applications. *Journal of neuroengineering and rehabilitation*, 3(1):6, 2006.

[55] Jochen Tautges, Arno Zinke, Björn Krüger, Jan Baumann, Andreas Weber, Thomas Helten, Meinard Müller, Hans-Peter Seidel, and Bernd Eberhardt. Motion reconstruction using sparse accelerometer data. *ACM Transactions on Graphics (ToG)*, 30(3):1–12, 2011.

[56] Lin Chen and Guoli Wang. Cpformer: End-to-end multi-person human pose estimation from raw radar cubes with transformers. *IEEE Sensors Journal*, 2025.

[57] Niraj Prakash Kini, Ruey-Horng Shiue, Ryan Chandra, Wen-Hsiao Peng, Ching-Wen Ma, and Jenq-Neng Hwang. Transhupr: Cross-view fusion transformer for human pose estimation using mmwave radar. In *Proc. British Machine Vision Conference (BMVC 2024)*, 2024.

[58] Jaeho Choi, Soheil Hor, Shubo Yang, and Amin Arbabian. Mvdoppler-pose: Multi-modal multi-view mmwave sensing for long-distance self-occluded human walking pose estimation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 27750–27759, 2025.

[59] Karan Ahuja, Yue Jiang, Mayank Goel, and Chris Harrison. Vid2doppler: Synthesizing doppler radar data from videos for training privacy-preserving activity recognition. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–10, 2021.

[60] Kaikai Deng, Dong Zhao, Qiaoyue Han, Zihan Zhang, Shuyue Wang, Anfu Zhou, and Huadong Ma. Midas: Generating mmwave radar data from videos for training pervasive and privacy-preserving human sensing tasks. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 7(1):1–26, 2023.

[61] Xiaotong Zhang, Zhenjiang Li, and Jin Zhang. Synthesized millimeter-waves for human motion sensing. In *Proceedings of the 20th ACM Conference on Embedded Networked Sensor Systems*, pages 377–390, 2022.

[62] Jiamu Li, Dongheng Zhang, Zhi Wu, Cong Yu, Yadong Li, Qi Chen, Yang Hu, Qibin Sun, and Yan Chen. Sbrf: A fine-grained radar signal generator for human sensing. *IEEE Transactions on Mobile Computing*, 2024.

[63] David D Lewis. A sequential algorithm for training text classifiers: Corrigendum and additional data. In *Acm Sigir Forum*, volume 29, pages 13–19. ACM New York, NY, USA, 1995.

[64] H Sebastian Seung, Manfred Opper, and Haim Sompolinsky. Query by committee. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 287–294, 1992.

[65] Neil Houlsby, Ferenc Huszár, Zoubin Ghahramani, and Máté Lengyel. Bayesian active learning for classification and preference learning. *arXiv preprint arXiv:1112.5745*, 2011.

[66] Yarin Gal, Riashat Islam, and Zoubin Ghahramani. Deep bayesian active learning with image data. In *International conference on machine learning*, pages 1183–1192. PMLR, 2017.

[67] Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. *arXiv preprint arXiv:1708.00489*, 2017.

[68] Kai Wei, Rishabh Iyer, and Jeff Bilmes. Submodularity in data subset selection and active learning. In *International conference on machine learning*, pages 1954–1963. PMLR, 2015.
[69] Jordan T Ash, Chicheng Zhang, Akshay Krishnamurthy, John Langford, and Alekh Agarwal. Deep batch active learning by diverse, uncertain gradient lower bounds. *arXiv preprint arXiv:1906.03671*, 2019.
[70] Baharan Mirzasoleiman, Jeff Bilmes, and Jure Leskovec. Coresets for data-efficient training of machine learning models. In *International Conference on Machine Learning*, pages 6950–6960. PMLR, 2020.
[71] Krishnateja Killamsetty, Sivasubramanian Durga, Ganesh Ramakrishnan, Abir De, and Rishabh Iyer. Grad-match: Gradient matching based data subset selection for efficient deep model training. In *International Conference on Machine Learning*, pages 5464–5474. PMLR, 2021.
[72] Krishnateja Killamsetty, Durga Sivasubramanian, Ganesh Ramakrishnan, and Rishabh Iyer. Glister: Generalization based data subset selection for efficient and robust learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 8110–8118, 2021.
[73] Kaichen Zhou, Yuhan Wang, Grace Chen, Gaspard Beaudouin, Fangneng Zhan, Paul Pu Liang, and Mengyu Wang. Page-4d: Disentangled pose and geometry estimation for 4d perception. In *The Fourteenth International Conference on Learning Representations*, 2025.
[74] Elizaveta Levina and Peter Bickel. Maximum likelihood estimation of intrinsic dimension. *Advances in neural information processing systems*, 17, 2004.
[75] Elena Facco, Maria d'Errico, Alex Rodriguez, and Alessandro Laio. Estimating the intrinsic dimension of datasets by a minimal neighborhood information. *Scientific reports*, 7(1):12140, 2017.

A Model Architecture and Scaling

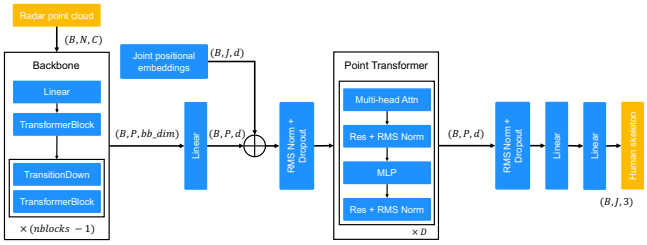


Figure 12: Our mmWave HPE Model.

Model Structure. Our model, shown in Fig. 12, consists of three stages: (i) a lightweight backbone encodes radar points into per-point features, (ii) a fixed set of learnable *joint tokens* are concatenated with point tokens and processed by a Transformer encoder stack, and (iii) two small MLP heads decode the joint tokens into compact features and final 3D coordinates.

Scaling Strategy. We vary model size using a *depth-first compound scaling* rule: the primary axis is the number of Transformer layers (D); the per-head dimension is fixed at $\text{dim}_{\text{head}} = 64$; the number of heads is modestly increased so that the model width is $d = \text{heads} \times 64$; and the feed-forward width follows $\text{MLP} = 4d$. In parallel, we apply a mild scaling to the backbone by increasing both the number of blocks and its internal channel size (bb_dim) to avoid an early bottleneck. Across all scales, tokenization (joint tokens J plus point tokens) and training recipe are kept fixed, with RMSNorm and dropout ($p = 0.1$) for stability.

Scaling Grid. Our study covers depths $D \in \{2, 3, 5, 8, 12, 16\}$, heads $\in \{3, 4, 6, 8, 10, 12\}$ (widths $d \in \{384, 512, 640, 768\}$), and backbone blocks $\in \{2, 3, 4, 5, 6, 8\}$ with bb_dim increasing from 32 to 192. For each configuration, we also log parameter counts and FLOPs to relate model size to performance to plot Fig. 1a.

B NRI Algorithm

Motivation. The raw k-NN radius $r_\phi^{(k)}$ is a convenient, non-parametric proxy for local density, but it is not directly comparable across datasets because its magnitude depends on both the dataset size n and the neighborhood size k. We therefore employ a *scale-free* redundancy measure that preserves the geometric intuition of $r_\phi^{(k)}$ while removing these extrinsic dependencies.

Definition. For a dataset $\Phi = \{\phi_i\}_{i=1}^n$ embedded in the shared latent space (same encoder and metric as in the main text), the *Normalized Redundancy Index* (NRI) of sample ϕ_i at scale k is

$$\text{NRI}_i(k) = \frac{(n-1)\,[r_i^{(k)}]^{\,d_{\text{eff}}}}{k},$$

where $r_i^{(k)}$ is the distance from ϕ_i to its k-th nearest neighbor in $\Phi \setminus \{\phi_i\}$, and d_{eff} is the effective intrinsic dimension of the latent space. Operationally, d_{eff} is estimated from the slope of a robust regression between $\log r^{(k)}$ and $\log k$ across multiple k values (median-based trimming), following the small-ball scaling $r^{(k)} \propto (k/n)^{1/d}$ [74, 75].

Why NRI is Scale Free. Classical k-NN density arguments state that, in a small neighborhood around ϕ,

$$\mathbb{E}[\#\text{NN within } r] \approx (n-1)\, f(\phi)\, C_d\, r^d,$$

where $f(\phi)$ is the local density and C_d is the unit-ball constant under the chosen metric. Solving for the radius at a fixed neighbor count k gives $r_\phi^{(k)} \approx \left(k/[(n-1)\, f(\phi)\, C_d]\right)^{1/d}$. Substituting this into the definition above yields

$$\text{NRI}_\phi(k) \;\approx\; \frac{1}{f(\phi)\, C_d},$$

which is *independent of n and (to first order) k*. In other words, NRI is a monotone transform of the standard k-NN density estimator [44, 45]: smaller NRI indicates higher density and, consequently, higher redundancy.

Implementation Details. We compute $r_i^{(k)}$ for a set of scales k (or, equivalently, for fixed coverage rates $\alpha = k/(n-1)$) in the same latent space [74, 75] and metric as the main text. To avoid numerical issues under duplicates, we floor radii by a small ε before taking logarithms. We estimate

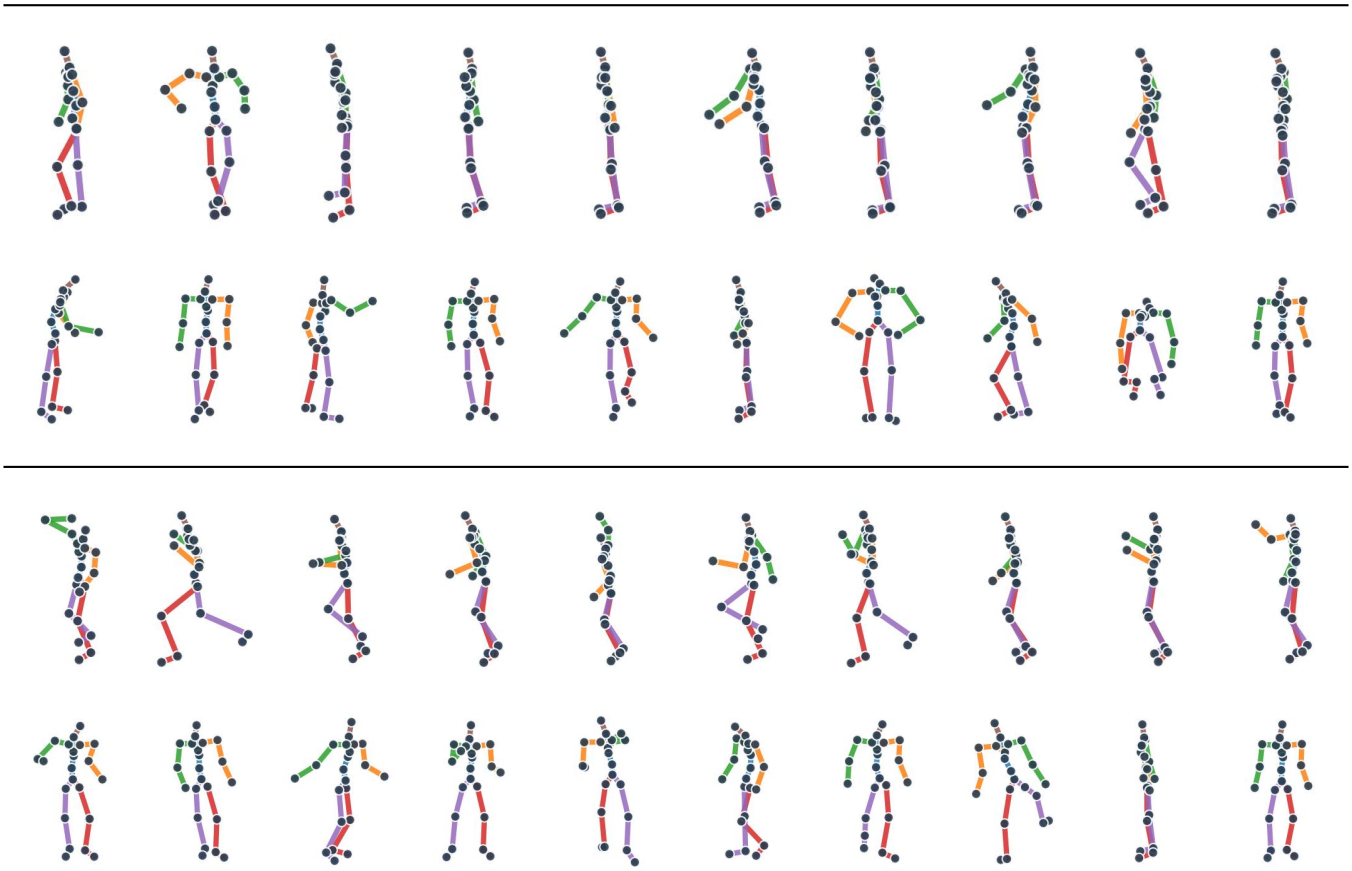


Figure 13: Pose visualizations (randomly selected). From top to bottom: (1) mmBody poses, (2) poses in AMASS but uncovered by mmBody ($k = 10$), (3) MMFi poses, and (4) poses in AMASS but uncovered by MMFi ($k = 12$).

d_{eff} via a robust linear fit of $\log r^{(k)}$ versus $\log k$ (with mild trimming of extreme k if needed), and then evaluate $\text{NRI}_i(k)$ per sample.

C Uncovered Cases of mmWave HPE Datasets

To qualitatively illustrate the coverage gaps identified by our cross-dataset analysis, we visualize several poses from the *gap set* (Φ_G). These examples, shown in Fig. 13, represent poses that are present in the comprehensive AMASS dataset but are underrepresented in existing mmWave datasets. This visualization provides a concrete illustration of the abstract concept of coverage gaps and intuitively demonstrates the kind of informative data our guided collection strategy aims to acquire.