# EquiFusion: Kinematics-Agnostic Human Motion Prediction via Equivariant Latent Diffusion

Cecilia Curreli[1,2] Florian Hofherr[1,2] Dominik Muhle[1,2]
Abhishek Saroha[1,2] Riccardo Marin[1,2] Daniel Cremers[1,2]

[1] Technical University of Munich, Munich, Germany
[2] Munich Center for Machine Learning, Munich, Germany

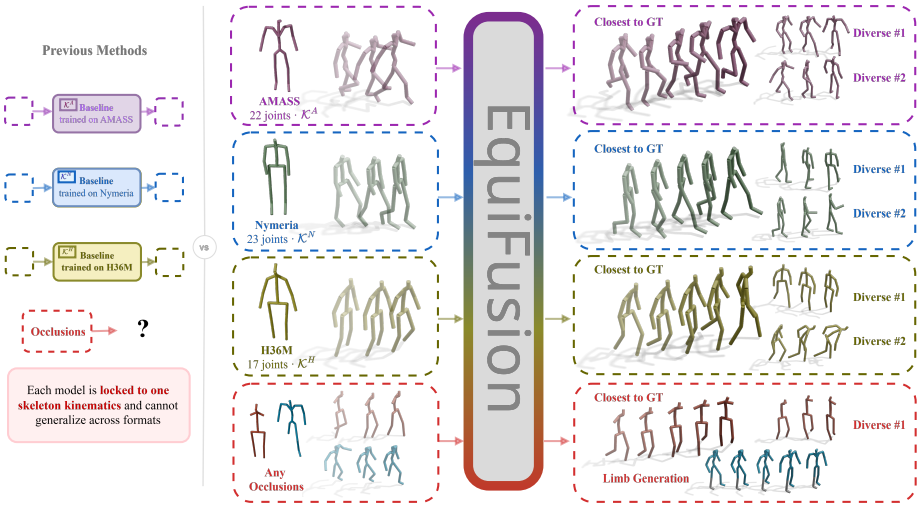


**Fig. 1: EquiFusion**. We introduce the first model for stochastic human motion prediction that generalizes to unseen skeleton parameterization, i.e. *kinematics*. While previous methods require a trained instance for each dataset or better kinematics, with a single model we unlock training on multiple datasets and inference on motion parametrized with different kinematics. EquiFusion is the first SHMP model to handle **zero-shot** novel kinematics and occluded limbs without explicit training, while achieving state-of-the-art results on established benchmarks.

**Abstract.** Existing Stochastic 3D Human Motion Prediction models are fundamentally constrained by hard-coding the skeleton kinematics, severely limiting generalization, preventing cross-dataset training, and requiring complex data retargeting. We introduce *EquiFusion*, the first kinematics-agnostic model to solve this bottleneck, implementing a latent diffusion model with a permutation equivariant architecture. EquiFusion treats the kinematics' connectivity as an explicit input parameter, ensuring its internal computations are inherently agnostic to joint ordering and graph structure. This novel design enables truly cross-dataset generalization to unseen kinematics and unlocks novel zero-shot directions, such as motion prediction from partial or occluded observations

and targeted limb generation. EquiFusion achieves state-of-the-art results on major benchmarks, being up to 75% more compact than previous kinematics-specific methods, while achieving faster training and inference. EquiFusion thus establishes a new, flexible standard for robust human motion prediction.

## 1 Introduction

Predicting future human motion from past observations is a core component of human intelligence and a prerequisite for mature spatial AI. Due to the inherent ambiguity of human intent, the field has shifted from deterministic toward Stochastic Human Motion Prediction (SHMP), which models a distribution of diverse, physically plausible futures from an observation, with applications including human-robot collaboration, autonomous navigation, and augmented telepresence. While recent advances in generative models [11, 34, 76] have further popularized SHMP probabilistic formulations [7, 15, 19, 85], a fundamental bottleneck remains: kinematics rigidity. Diverse datasets (e.g., Human3.6M [38], AMASS [63], and more recently Nymeria [62]) often employ different motion-capturing technologies, resulting in diverse skeleton parametrizations *i.e. kinematics*. Motion retargeting [6, 35, 47], *i.e.*, converting a motion to a different kinematics parametrization, is often not possible without introducing errors and drifts [55, 93, 94]. The downstream impact on SHMP methods is tremendous. Previous works [7, 15, 19, 21, 64, 85, 109] have hard-coded kinematics as a network structural prior, resulting in a plethora of skeleton-specific networks. Training a network for every different kinematics is impractical, inefficient, and does not generalize to new skeletal configurations.

We break these dataset-specific boundaries by addressing settings involving heterogeneous kinematic structures. We propose two *zero-shot kinematics* inference tasks for SHMP: predicting motion for (i) full-body kinematics unseen at training time, and (ii) kinematics with occlusions or missing parts (*i.e.*, partial observations). Both these tasks address real-world use cases and important milestones toward mature spatial AI and foundation human motion models. To date, this has been unaddressed in the SHMP literature, while deterministic HMP solutions (*e.g.*, hand-crafted training to address specific occlusions or skeletal conversions [17, 102]) are impractical and require manual engineering. Can we provide such flexibility for SHMP?

In this paper, we answer this question by proposing the first kinematics-agnostic method for SHMP. To handle arbitrary kinematics chains, a model's learned weights must be independent of the cardinality of the joint set. Most existing works violate this principle by employing joint-dependent weights or temporal-only transformers that treat joints as feature channels. We identify permutation equivariance as the key mathematical property to achieve this independence. We thus present EquiFusion , the first kinematics-agnostic approach to SHMP, implemented as an equivariant latent diffusion model that explicitly incorporates the kinematics connectivity as a model input. By design, we achieve

robustness to partial and unseen kinematics without requiring explicit exposure to such corruptions during training, and we demonstrate this through extensive experiments. Our design allows us to break the single-dataset barrier, enabling a single model to be trained simultaneously on heterogeneous datasets such as AMASS and Nymeria for the first time, unlocking the future potential of "foundation models" in human motion. By leveraging the inductive bias of equivariance with formal guarantee [10,61], EquiFusion achieves state-of-the-art performance on standard benchmarks even when trained on single-dataset distributions while being *75% smaller* than the closest competitor [19], and significantly faster in training and inference. These improvements are particularly remarkable in zero-shot scenarios, where our model demonstrates a performance boost of at least 25% across all metrics and up to 70% in terms of fidelity. Our contributions are:

- We formalize and analyze for the first time the challenge of zero-shot kinematics SHMP, introducing two novel tasks of high relevance for the real-world setting, while opening new applications and use cases: training on datasets having multiple kinematics, zero-shot handling of occlusion, and zero-shot limb generation without ad-hoc training.
- We identify joint-ordering permutation equivariance as a key property to enable kinematics-agnostic SHMP and propose a principled architecture that inherently satisfies this requirement.
- We present EquiFusion, which not only achieves zero-shot kinematics by design but outperforms existing SHMP methods in extensive experiments. Our method achieves state-of-the-art results on standard benchmarks (AMASS, Human3.6M) with 75% fewer parameters than the closest competitor. Additionally, our model's complexity remains invariant to the number of joints, kinematics, or datasets, offering superior scalability.

See https://ceveloper.github.io/publications/equifusion/ for models and code.

## 2 Related Works

### 2.1 Human Kinematics Representations

In SHMP, human motions are represented as the temporal evolution of body joints; we refer to them as skeleton kinematics $\mathcal{K}$, which are generally inherited from the Motion Capture (MoCap) system used to register the human movement. Hence, depending on the MoCap system, the kinematic system can differ in the number and positions of the joints, resulting in a plethora of different formats. Examples are: the AMASS [63] dataset uses the SMPL [60] format, and is parametrized by 22 joints; H36M [38] kinematics comprises 17 joints and does not include the foot joints; Nymeria [62] has been captured by META's ARIA devices [24] and parametrized with XSens [1] by 23 joints, exhibiting different spine and hip modeling from the aforementioned formats (see Fig. 1 for visualizations). With the increasing availability and affordability of wearable sensors [24,33,73,97], new kinematic configurations are expected to emerge [27,105], making it crucial to address multiple kinematics, since differences in kinematic structure hinder data combination during training [24].

### 2.2 Stochastic Human Motion Prediction

Human Motion Prediction (HMP) aims to predict future motion from past observations. While deterministic HMP [18, 49, 51] forecasts a single future, Stochastic HMP (SHMP) predicts multiple plausible ones. SHMP is an ill-posed problem with multiple solutions, which has been modeled by different generative approaches involving GANs [8, 44, 58], VAEs [12, 32, 64, 96, 104], diverse sampling [21, 103, 109], and lately denoising diffusion models [7, 15, 19, 78, 85, 100]. All such models are trained and evaluated independently for each kinematics $\mathcal{K}^i$, which means that each requires ad hoc training, computational waste, hyperparameter engineering, and lacks generalization at inference time to novel kinematics $\mathcal{K}^j$ with $j \neq i$, or occluded data (*i.e.*, partial kinematics). Digesting motions in different kinematics formats has been tackled in other computer vision tasks, such as human pose estimation from video [82], 2D-to-3D keypoints lifting [20], unconditional animal motion generation [28], motion classification [47], or character animation via text [37,57]. Occluded motions and training on multiple kinematics have already been investigated by deterministic HMP, but to the best of our knowledge, only explicitly, by directly training for occlusions [17,102] or for multiple skeletons without supporting zero-shot inference on new kinematics [81]. Yet, no method has been proposed for SHMP, which is the core of our work and a mandatory stage towards foundation models for SHMP. We provide an extended discussion on existing methods in the Appendices A.2 to A.4. We will build our framework on top of a newly designed permutation equivariant diffusion model. Diffusion methods equivariant to SE(3) or permutation groups have provided promising results for the task of drug molecule generation [45, 70, 83], but they haven't found broad application in computer vision, especially not in HMP or SHMP to date. A reasonable alternative to enable flexibility of existing approaches would be solving for skeleton retargeting, and we discuss such techniques in the next section.

### 2.3 Kinematics Conversion and Motion Retargeting

As current SHMP approaches do not support inference on kinematics $\mathcal{K}^j$ different from the training kinematics $\mathcal{K}^i$, the only way to allow inference on $\mathcal{K}^j$ is to first convert the input motion from kinematics $\mathcal{K}^j$ to the same motion with $\mathcal{K}^i$. This conversion challenge is commonly addressed as *skeleton retargeting*. The large literature on this topic ranges from robotics [22, 26, 108] to computer graphics [6,36,47,53] and specializes in isomorphic [3,16,22,25,31,39,46,55,68,69, 86, 93, 94], homeomorphic [2, 36, 110], and non-homeomorphic graphs [14, 35, 40, 42, 47, 52, 54, 67, 98]. For the SHMP use case, only *non-homeomorphic* techniques can be considered, since kinematics can have different end-effectors and a varying number of joints. As not all approaches are applicable to SHMP's data structure, in this paper, we consider the retargeting approach from Holden *et al.* [35], which is already established in SHMP by HumanMAC [15]. More recent approaches for non-homeomorphic retargeting on robotics and character animation are not applicable to our case. These require joint rotations of end-effectors and reference poses [47], skinning [54], meshes [79,98], or both [40] – even without the skeleton

itself [54, 98] – which are not always available in SHMP. Or, they present a too high reconversion error (ca. 100mm [67] or 90mm [14], in both cases, code is not available). We discuss retargeting approaches and each graph category in detail in Appendix A.1, and here we maintain a high-level contextualization, taking as an example the AMASS and H36M skeletons depicted in Fig. 1. These kinematics are non-homeomorphic to each other, so the conversion is not bijective and causes: **(a) Error accumulation**, as converting H36M (17 joints) to AMASS (22 joints) requires adding non-existent feet (with errors of 22.18 mm or 2.27 mm depending on the conversion direction), **(b) Distribution shift**: additionally, this kinematics conversion shifts the input to a distribution that differs from the one seen by the SHMP models at train time, resulting in additional noise for the models. As SHMP approaches achieve precision up to 70 mm, the additional retargeting error is not negligible.

### 2.4 Human Pose Representations

Another relevant aspect of SHMP is the motion parametrization. In SHMP, motion sequences are conventionally represented as trajectories of 3D joint coordinates in Euclidean space [7, 21, 64, 96, 100, 109]. While such representation is flexible and a natural output of upstream pipelines [13, 59, 84, 101], *e.g.* human tracking in videos [71, 74, 77, 112], it is under-constrained and allows non-realistic predictions. This often results in limb stretching or jitter, requiring subsequent SHMP approaches to directly measure [19] and address [7, 19, 21] this issue. Instead of making the model learn from data what we already know *i.e.* human bones have fixed length, we propose to include this prior directly in the parametrization. Intuitively, consistent limb lengths are guaranteed by design in rotation-based representation, where the joint positions are expressed as angles from a canonical pose, and the bone lengths are fixed along a sequence. Among the many possible representations for rotations [29], some common choices are quaternions [80] or axis-angles [48, 60, 87, 88] as in SMPL [60]. However, rotations also tie the pose representation to a specific skeleton kinematic chain and rest pose, which can cause instabilities [7]. In this work, we adopt a spatial representation, where every joint is encoded by the relative direction from its parent, *i.e.* its *bone direction*. In this way, predictions are guaranteed to have perfectly consistent bones by design, and networks do not have to learn this additional prior. This representation is general and can be applied to both any SHMP kinematics and any SHMP method.

## 3 Methodology

### 3.1 Problem Definition

In stochastic human motion prediction (SHMP), a motion sequence is represented as a trajectory $\mathbf{M} \in \mathbb{R}^{T \times J \times 3}$, describing the temporal evolution of $J$ body joints over $T$ frames. Given an observed motion $\mathbf{X} = \mathbf{M}_{0:T_P} \in \mathbb{R}^{T_P \times J \times 3}$ of length $T_P$, the objective is to generate multiple plausible future trajectories

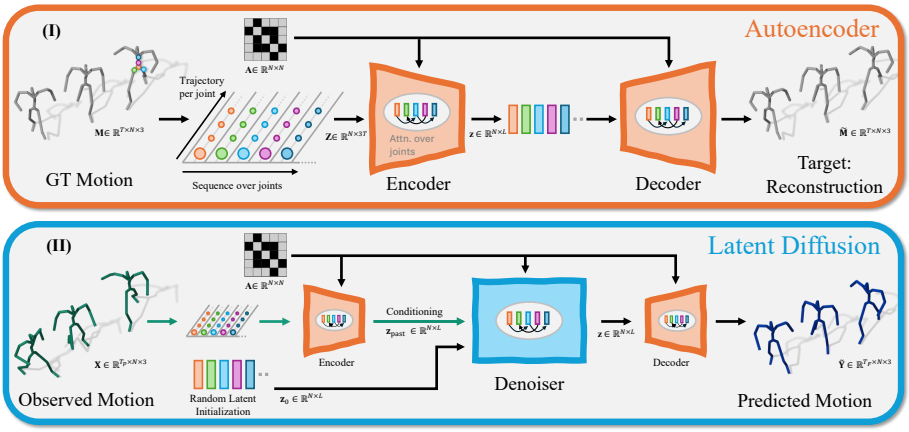


**Fig. 2: EquiFusion is the first skeleton-agnostic model for SHMP, implemented as a novel equivariant latent diffusion model that adds the adjacency matrix A of the motion kinematics as input**. **(I)** A transformer autoencoder learns a latent space $\boldsymbol{z}$ equivariant to joint order permutation. **(II)** A denoiser predicts future motion in this latent domain conditioned on the observed past $\mathbf{X}$.

$\tilde{\mathbf{Y}} \in \mathbb{R}^{T_F \times J \times 3}$ over a prediction horizon of $T_F$ frames. Each motion sequence is described with respect to a kinematic structure $\mathcal{K}$, as already discussed in Sec. 2.1. Formally, we define a kinematics $\mathcal{K} := (V_{\boldsymbol{\tau}}, E)$ as an undirected graph with semantically labeled vertices $V_{\boldsymbol{\tau}}$ and edges $E$. We refer to nodes as joints and edges as limbs or bones, and we define the cardinality $|\mathcal{K}|$ equal to the number of joints. Each vertex $v_{\boldsymbol{\tau},i}$ represents a joint with a connected semantic convention label that distinguishes among joint types and kinematics conventions. The kinematic configuration $\mathcal{K}$ is fixed within each dataset across all motion sequences, so we usually refer to a kinematics with the name of the dataset: *e.g.* $\mathcal{K}^A$ for AMASS [63], $\mathcal{K}^H$ for H36M [38], and $\mathcal{K}^N$ for Nymeria [62], where different individuals in the same dataset have the same kinematics $\mathcal{K}$ but different bone lengths. Typically, models are evaluated on the test set of the dataset they were trained on, *in-domain*, and within the same kinematics. The popular mesh SMPL model, which underlies the AMASS kinematics, has led to smaller data collections [66, 90] that share the kinematics $\mathcal{K}^A$. These datasets are occasionally used for evaluation on out-of-distribution motions, i.e., zero-shot motion for SHMP. For the first time in SHMP, we investigate **zero-shot kinematics**, the case where inference kinematics differ from those seen at training. While both evaluations are theoretically independent, the nature of the datasets leads to zero-shot kinematics, which often implies zero-shot motion. Zero-shot kinematics also covers *partial* kinematics, representing limbs that are occluded or missing due to physical impairments. While partiality is highly relevant for real-world applications, data collection is not straightforward, and there are no specific SHMP datasets to date. Partial motions are thus usually investigated by masking limbs of motions parametrized with existing full-body kinematics.

### 3.2 Kinematics-Agnostic SHMP

While kinematics-agnostic models have been investigated in other domains, such as character animation from text [28,57], motion classification [47], and robotics [5], the problem of addressing multiple skeleton kinematics has not been addressed or formalized in SHMP so far. Existing SHMP approaches assume a single, fixed skeleton kinematics inherited from the training dataset [7,15,19,21,30,36,85,109]. Instead, we are interested in a model that natively supports 1) training on multiple kinematics, and 2) inference on novel kinematics, partial or full-body, not seen at training time, i.e., in a zero-shot kinematics setting. We call such a model *kinematics-agnostic*.

Formally, let $\mathbb{K}_{\text{train}} = \{\mathcal{K}^1, \mathcal{K}^2, \dots, \mathcal{K}^S\}$ be the set of kinematic settings seen a training time by a method $f$, $S$ the number of train skeletons, and $\mathbb{K}_{\text{test}}$ the set of kinematics used for evaluation. When we perform evaluation on a motion under the kinematic setting $\mathcal{K}^{'} \in \mathbb{K}_{\text{test}}$ with $\mathbb{K}_{\text{train}} \cap \mathbb{K}_{\text{test}} = \emptyset$, we regard this as a zero-shot scenario. Previous SHMP approaches have been limited to $N = 1$ and $\mathbb{K}_{\text{train}} = \mathbb{K}_{\text{test}}$ due to their kinematics-specific design decisions, requiring a trained instance for each kinematics. While zero-shot kinematics may be achieved at least in some cases through overly convoluted engineering, we advocate for a method that achieves zero-shot kinematics by design.

Intuitively, the core requirement is that the model must accommodate kinematic chains with an arbitrary number of joints. We formalize this as a constraint on the learnable parameters $\Theta$.

**Lemma 1** *To handle arbitrarily sized kinematic chains $\mathcal{K}$, the learned parameters $\Theta$ of a model cannot be dependent on the number of joints* i.e. *the cardinality of the input $\mathcal{K}$:*

$$\frac{d|\Theta|}{dJ} = 0 \qquad \text{i.e.} \qquad |\Theta| \in O(1) \tag{1}$$

Indeed, in the trivial case $\mathbf{W} \in \mathbb{R}^{J \times J}$, the weights do not generalize to any new $|\mathcal{K}^{'}| > J$. And generally, allocating independent parameters $\mathbf{W}_j$ for each joint $j \in 1 \dots J$ lets the parametrization $\Theta$ grow with $J$ and change whenever the kinematic chain changes, violating the Lemma. Yet previous works in both deterministic [51,65,111] and stochastic HMP [19,80,85] intentionally learn joint-dependent weights to extract strong dataset- [111] and kinematics-specific [19,80] priors, and thus cannot cannot satisfy Lemma 1 (we prove this by counterexample for each architecture class in Appendix B.2).

We recognize that this requirement is naturally fulfilled by models $f$ that are permutation equivariant $f(\mathbf{P}\mathbf{X}) = \mathbf{P}f(\mathbf{X})$ with respect to arbitrary joint reordering $\mathbf{P} \in \mathbb{R}^{J \times J}$.

**Theorem 2.** *Let $o(\mathbf{X}) = \mathbf{WXG}$ be a general network operation for feature extraction on an input $\mathbf{X} \in \mathbb{R}^{J \times F}$. If $o(\mathbf{X})$ is permutation equivariant under joint reordering, i.e. $\mathbf{P}\,o(\mathbf{X}) = o(\mathbf{P}\mathbf{X})$ for any permutation $\mathbf{P} \in \mathbb{R}^{J \times J}$, then the number of learned parameters $|\Theta|$ is constant in $J$, and Lemma 1 is satisfied.*

In other words, weights must be shared across all joints, and the model must behave consistently under any reordering of these instances. Permutation equivariance is therefore sufficient, though not necessary, for kinematics-agnosticism. We report the full proof in Appendix C.1. Since equivariance is preserved under composition, a network $f$ built entirely from equivariant operations $o$ is end-to-end permutation equivariant. With this motivation, we decide to implement our kinematics-agnostic solution as a permutation equivariant model. While we already mentioned that previous work do not fulfill the Lemma, we also prove numerically and mathematically in Appendix C.4 that their architectural designs are not equivariant. We present an intuitive high level explanation on why this is the case in Appendix B.1 and in the next section.

### 3.3 EquiFusion

*Overview.* Based on the previously presented findings on permutation equivariance fulfilling the premise of a kinematics-agnostic model, we implement EquiFusion as an equivariant latent diffusion model (Fig. 2). We design a novel end-to-end equivariant framework consisting of (i) an autoencoder mapping motion sequences $\mathbf{M}$ to and from the latent space $\boldsymbol{z} \in \mathbb{R}^{J \times L}$, and (ii) a denoiser that predicts future motions in the latent domain $\boldsymbol{z_\theta}$ conditioned on the embedding $\boldsymbol{z}_{past} = \mathrm{e}(\mathbf{X})$ of the input past $\mathbf{X}$.

*Equivariant Latent Diffusion* The generative process of diffusion models [34] is known to be computationally expensive in input space [23, 72]. To gain in efficiency we operate in a lower-dimensional latent space [7, 19] and opt for latent diffusion models (LDM) [76]. While previous approaches learn a mapping $f_{\mathcal{K}}(\mathbf{X}) = \tilde{\mathbf{Y}}$ for a fixed skeletal kinematics $\mathcal{K}$, we support operations on different kinematics out-of-the-box, by making the connectivity $\mathbf{A}$ of the kinematics $\mathcal{K}$ an explicit input to the model. We thus design a novel framework architecture that is end-to-end formally permutation equivariant w.r.t. its inputs:

$$f(\mathbf{X}, \mathbf{A}) = \tilde{\mathbf{Y}}\,, \quad \text{with} \quad f(\mathbf{P}\mathbf{X}, \mathbf{P}\mathbf{A}\mathbf{P}^\top) = \mathbf{P} f(\mathbf{X}, \mathbf{A}). \tag{2}$$

Equivariance in LDMs requires an equivariant denoiser and a latent space that preserves input permutations [56,89,95].Although this space can be learned non-deterministically (e.g., via VAEs [76]), we adopt a deterministic approach to improve training stability [106]. At inference, new latent variables are sampled from a univariate Gaussian distribution. Since this sampling is i.i.d., sample-wise equivariance is guaranteed only in a deterministic setting where the noise is fixed and permuted accordingly. Indeed, in generative models, permutation equivariance holds at a distribution level. We provide a detailed discussion in Appendix C.3. Differently from previous LDM in SHMP [7, 19], we implement the autoencoder as a transformer rather than a recurrent network, gaining in inference speed. We follow the training paradigm of LDM [76] adapted to SHMP [7, 19]. Further details and equations in Appendix D.

*EquiFusion's Architecture.* In looking for a permutation equivariant (PEQ) architecture, the reader may already be thinking that Graph Convolution Networks (GCN) [43] naturally fulfill the requirement [41]. However, while many SHMP models are based on GCN [7,19,51,80,85], no one of them enjoys PEQ. The reason is that independently extracted features are not aggregated according to the connectivity of the graph [43], but according to learned weights $\mathbf{W} \in \mathbb{R}^{J \times J}$ that explicitly depend on the number and position of joints [19,51,75]. Such practice in deterministic [111] and SHMP [19, 80, 85] showed advantages in leveraging dataset-specific priors (*e.g.* action [111] or joint-specific [19,80]). However, such implicit bias is detrimental in our case. We advocate for a kinematics-agnostic formulation aligning with [43]. Specifically, for an initial input $\mathbf{Z} \in \mathbb{R}^{J \times 3T}$ , we implement a convolution $c$ as:

$$c(\mathbf{Z}, \mathbf{A}) = \mathbf{W}\mathbf{Z}\mathbf{G_1} + \mathbf{Z}\mathbf{G_0} + \boldsymbol{b}, \quad \text{with} \quad \mathbf{W} = \mathbf{D}^{-1}\mathbf{A} \tag{3}$$

where the weights $\mathbf{G_0}, \mathbf{G_1} \in \mathbb{R}^{F_\text{i} \times F_\text{o}}$ learn to extract features for a joint itself or its neighbours respectively, $\boldsymbol{b} \in \mathbb{R}^{F_\text{o}}$ is learned, $\mathbf{D}$ the normalizing degree matrix [43] of $\mathbf{A}$, $F_\text{i}$ and $F_\text{o}$ are the input and output feature dimensions. This operation fulfills equivariance not only with respect to a vector $\mathbf{X}$, but also with respect to the input matrix $\mathbf{A}$ [89], which excludes otherwise compelling layers [107]. We want to consider the skeleton graph on a global scale in addition to the local scale of Eq. (3), and do so via Graph Attention (GAT) [92] or multi-head self-attention [91] with joints as token dimension.

$$\text{Att}(\mathbf{Z}, \mathbf{A})_h = \text{softmax}\left(c_Q(\mathbf{Z}, \mathbf{A}) c_K(\mathbf{Z}, \mathbf{A})^\top / \sqrt{F_o}\right) c_V(\mathbf{Z}, \mathbf{A}) \tag{4}$$

Particularly, we employ the operation in Eq. (3) to compute $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$ for attention on each of the $h$ heads. Building on these operations, we design a novel end-to-end equivariant architecture. Noticeably, by definition, attention layers [91] without positional encodings are PEQ along their token dimension. But in transformer-based SHMP approaches [7, 15, 21, 64, 85, 96, 109], or approaches that treat joints as features [15,50,78], tokens are reserved for the time dimension instead of the joints (against Lemma 1). We provide mathematical proof that our architecture is end-to-end equivariant in Appendix C.2, together with experimental results on equivariance for the baselines. To the best of our knowledge, an end-to-end PEQ architecture over joint orderings has not been demonstrated before for SHMP.

*Additional Advantages.* While equivariance may be learned via extensive augmentation (as for molecule generation [4, 99]), we chose to fulfill it mathematically by design, reducing model and learning complexity [10, 61]: on just a single dataset [38], our kinematics-agnostic approach achieves state-of-the-art results with 75% fewer parameters than the latest baseline SkelDiff [19], requiring around half of the training and inference time (see Fig. 3 and discussion in Sec. 4.2). Furthermore, our equivariance property allows us to train natively on multiple datasets $\mathbb{K}_\text{train} = \{\mathcal{K}^A, \mathcal{K}^N\}$ leveraging for the first time the two largest SHMP datasets at once, AMASS and Nymeria.

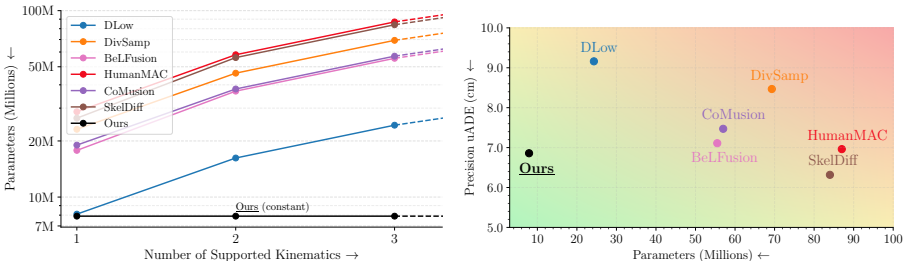


**Fig. 3: State-of-the-art with 75% fewer parameters**: (left) baselines scale with the number of supported datasets i.e. kinematics, we do not; (right) we achieve SOTA precision, averaged over three datasets [38, 62, 63], with a single model instance.

### 3.4 Directions as Motion Parametrization

Motion representation is a critical choice in SHMP, especially for our aims of ensuring compatibility across disparate datasets and motion formats. Despite its importance, this remains underinvestigated in SHMP literature, where 3D joint absolute coordinates (Sec. 2.4) remain the de facto standard. While flexible, this representation suffers from limb stretching and physically inconsistencies [19,21]. Conversely, joints' rotation angles relative to the kinematic chain parent(as in SMPL [60]) preserve the body structure by design, but require a canonical pose, which is not always available. We propose a robust middle ground by representing the joints as the relative direction vector w.r.t. the parent joint, *i.e.* $\mathbf{L}_t \in \mathbb{R}^{B\times 3}$, where each limb $i$ is defined by the vector between a joint $i$ and its parent.

$$\mathbf{L}_t^i = \mathbf{M}_t^i - \mathbf{M}_t^{\text{parent}(i)}, \quad i = 1, \dots, B\,. \tag{5}$$

We report the details in Appendix D.3. This parametrization has several advantages: 1) it is easily derived from both positions and angles pose representations, facilitating cross-dataset training; 2) it eliminates limb stretching by rescaling relative distances between joints based on past observations at inference time; 3) it remains numerically stable during training [9] without singularities typical of rotation spaces [29,80]. We use the average limb length of the training data as a heuristic for generating missing limbs. Empirical results in Tab. 5 demonstrate that this representation improves realism and diversity metrics also for other SHMP methods, without requiring any change to existing SHMP architectures.

## 4 Experiments

### 4.1 Experimental Settings

*Datasets.* We follow the evaluation settings of [7, 15, 19, 85], and include an additional dataset [62]: in addition to the protocols involving the highly diverse AMASS (A) dataset [63], and the widely employed, but involving only 7 subjects, H36M (H) [38], we also include Nymeria (N) [62], after adapting it to SHMP.

**Table 1: Evaluation of zero-shot kinematics on H36M ($\mathcal{K}^H$) [38].** We support inference on novel kinematics out-of-the-box, while previous approaches require additional kinematics conversion [15,35]. Baselines are trained on AMASS [63] (A, $\mathcal{K}^A$). We are the first to natively support multiple kinematics and thus present a model trained additionally on Nymeria [62] (N, $\mathcal{K}^N$). The best results are highlighted in **bold**, second-best are underlined. Conventional SHMP metrics rank identically, see Tab. 19.

| | | Precision ↓ | | | MM GT ↓ | | Div ↑ | Realism ↓ | | Body Real ↓ | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Units | $\mathcal{K}$ | cm | cm | deg° | cm | cm | m | – | – | % | % |
| Method | new | $u$ADE | $u$FDE | MAE | $u$MMA | $u$MMF | $u$APD | CMD | FID | str | jit |
| ZeroVel | ✓ | 11.77 | 17.88 | 6.753 | 13.74 | 18.56 | 0.000 | 22.822 | - | 0.00 | 0.00 |
| TPK [96]+ [35] | ✗ | 13.81 | 16.13 | 22.276 | 14.60 | 16.22 | 1.469 | 10.051 | 3.773 | 19.55 | 0.46 |
| DLow [109]+ [35] | ✗ | 12.71 | 14.80 | 21.887 | 13.60 | 14.97 | 2.060 | 9.204 | 2.875 | 20.26 | 0.53 |
| GSPS [64]+ [35] | ✗ | 9.29 | 11.91 | 8.107 | 10.85 | 12.35 | 2.069 | 7.409 | 1.735 | 11.51 | 0.38 |
| DivSamp [21]+ [35] | ✗ | 9.27 | 12.61 | 8.374 | 11.42 | 13.27 | **4.210** | 47.783 | 5.629 | 18.47 | 1.01 |
| BeLFusion [7]+ [35] | ✗ | 9.24 | 11.62 | 8.200 | 10.93 | 12.16 | 1.305 | 8.031 | 1.195 | 9.81 | 0.34 |
| CoMusion [85]+ [35] | ✗ | 10.07 | 12.15 | 21.066 | 12.49 | 12.96 | 2.070 | 8.587 | 1.426 | 15.98 | 0.51 |
| SkelDiff [19]+ [35] | ✗ | 10.81 | 14.99 | 14.947 | 12.75 | 15.42 | 0.992 | 7.616 | 5.252 | 11.25 | 0.28 |
| EquiFusion(A) | ✓ | 7.86 | 10.47 | 5.861 | 10.66 | 11.61 | 1.973 | **7.061** | 0.691 | **0.00** | **0.00** |
| EquiFusion(A+N) | ✓ | **7.71** | **10.21** | **5.683** | **10.58** | **11.36** | 1.797 | 7.349 | **0.504** | **0.00** | **0.00** |

Details of this process are reported in Appendix G.2, but overall the final size is comparable to AMASS and the quality of the motions is less dynamic. While the notation $\mathcal{K}$ denotes exclusively kinematics, the letters A,H,N denote training data, automatically implying the corresponding kinematics. See Sec. 3.1 and Sec. 2.1 for details on kinematics. We test models trained on A for the zero-shot motion scenario of MoYoga ($\mathcal{K}^A$) [90].

*Metrics.* Conventionally employed metrics [7, 19, 109] can be categorized into precision, diversity, realism, and body realism (see Appendix F for extensive definitions). Current metrics ADE, FDE, APD cannot be used to compare methods across datasets because they are dependent on the number of joints and cannot be measured in meters i.e. they include a cofactor $\sim \sqrt{J}$. Hence we introduce corresponding revised, unified versions $u$ADE, $u$FDE, $u$APD measured in centimeters ($u$ADE, $u$FDE with their multimodal counterpart) or meters ($u$APD). Note that $u$ADE is mathematically equivalent to the mean per-joint projection error (MPJPE), widely employed in other vision tasks. Since they differ by just a cofactor, unified metrics rank identical to conventional metrics, which are provided for every table in the appendix.

*Baselines.* SOTA baselines [7, 15, 19, 21, 64, 85, 96, 109] do not support novel kinematics out-of-the-box, while we do so by design. To enable evaluation in a zero-shot kinematics setting for baselines, as described in Sec. 2.3, we perform kinematics conversions following Holden *et al.* established by [15]. We include the ZeroVelocity baseline (ZeroVel), a competitive algorithmic baseline that repeats the last frame of the past observation for all future frames [109].

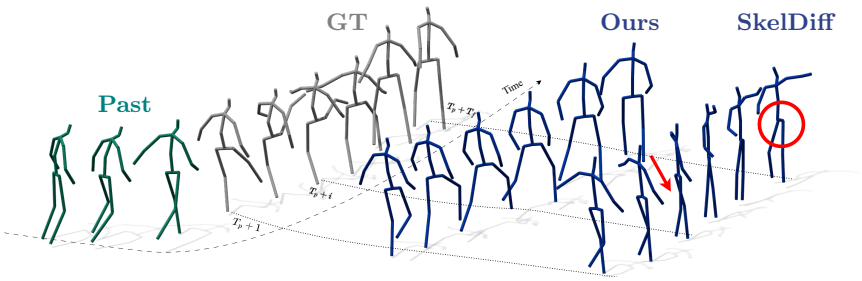


**Fig. 4: Qualitatives for zero-shot kinematics on H36M($\mathcal{K}^H$) [38].** We report the prediction closest to the GT for our method and SkelDiff [19], which does not support novel skeletons natively and is thus combined with a retargeting procedure [35]. While this challenging dynamic kick is not reproduced by either works, our prediction is realistic and coherent with the observation. SkelDiff is unable to generate a semantically close motion and cannot recover from the degradation introduced by retargeting.

### 4.2 Comparison

*Zero-Shot full-body Kinematics* . In the main body of this paper, we concentrate on the realistic case that the SHMP model is trained on the kinematic $\mathcal{K}^A$ corresponding to the dataset with the largest data amount available, AMASS [63], and inference is performed on a skeleton kinematics $\mathcal{K}^H$ with fewer joints and a smaller dataset, H36M [38]. This is the most advantageous setting for a SHMP model, as it is trained with a stronger prior (we discuss the more challenging, reverse scenario in Appendix H.2). We report quantitative results in Tab. 1. Current SOTA approaches do not support this zero-shot kinematics scenario on $\mathcal{K}^H$ out-of-the-box, and the input motion must first be converted via retargeting [35] to the kinematics supported by the model (see discussion in Sec. 2.3). Isolating the retargeting error from the baseline error completely is not possible, but via triangle inequality we estimate an upper bound. We see that baselines can recover from input retargeting: SkelDiff's total error 10.81 cm is around half of the upper bound 22 cm (further analysis in Appendix H.3). Our model instead, Tab. 1, can operate natively on any kinematics and achieves the best results, with an improvement up to 27% for precision metrics, and 72% for realism. As already discussed by previous works [7, 19], evaluating the multidimensional HMP problem is complex, and the numerous metrics are often complementary: a high diversity scores (APD) can be originated by unrealistic and ill-posed motions, as shown by the poor realism and body realism results of the VAE-based method DivSamp [21]. When considering the latest diffusion method, SkelDiff [19], our generated motions are almost twice as diverse. Additionally, our motion representation as bone direction delivers perfect body realism by definition, and we discuss it further in Sec. 4.2. A qualitative result can be seen in Fig. 4.

*Multi-dataset Training.* Our key ability to train on multiple kinematics translates into the ability to train on any dataset simultaneously. We leverage for the first time the two largest SHMP datasets, AMASS (A) and Nymeria (N), and report results for our method trained on the combination of both (A+N). As expected,

**Table 2:** Out-of-distribution MoYoga [90] for models trained on AMASS [63].

| | | MoYoga ($\mathcal{K}^A$) | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | Precision ↓ | | | Div↑ | Real↓ | B. Real↓ | |
| Method | $\mathcal{K}$ train | ADE | FDE | MAE | APD | CMD | str | jit |
| ZeroVel | - | 0.709 | 1.187 | 7.954 | 0.000 | 20.333 | 0.00 | 0.00 |
| SkelDiff [19] | A | 0.567 | 0.892 | 8.048 | **13.304** | 15.710 | 7.25 | 0.29 |
| EquiFusion | A | 0.501 | **0.780** | 6.982 | 13.153 | 11.961 | **0.00** | **0.00** |
| EquiFusion | A+N | **0.492** | 0.786 | **6.596** | 12.467 | **7.095** | 0.00 | 0.00 |

**Table 3:** Occlusion of a random limb (leg or arm) at inference on AMASS test set.

| | | AMASS ($\mathcal{K}^A$) | | | | | |
|---|---|---|---|---|---|---|---|
| | | Precision ↓ | | | Div ↑ | B. Real ↓ | |
| Method | $\mathcal{K}$ train | ADE | FDE | MAE | APD | str | jit |
| SkelDiff [19] | A | - | - | - | - | - | - |
| SkelDiff [19]+rp | A | 0.574 | 0.727 | **6.996** | 8.890 | 8.15 | 0.27 |
| SkelDiff [19]+sl | A | 0.567 | 0.683 | 7.162 | 9.274 | 5.64 | 0.23 |
| EquiFusion | A | 0.553 | 0.618 | 10.190 | **10.152** | 0.00 | 0.00 |
| EquiFusion | A+N | **0.499** | **0.553** | 8.777 | 9.099 | **0.00** | **0.00** |

training on an additional kinematic type (i.e. $\mathcal{K}^N$) and more data unlocks better performance. Our gains in this cross-dataset setting do not come from increased data diversity alone: we see that only 2% of our 29%ADE improvement and 27% of our 90% FID come from the multi-dataset training. The only difference between EquiFusion(A), trained only on AMASS, and EquiFusion(A+N) is the training data, while the architecture and hyperparameters remain the same. We find the fact remarkable, highlighting the need for methods that natively reason over any dataset and kinematics and opening future discussion on the impact of different motion distributions on training.

*Scalability to Any Kinematics.* Fig. 3 highlights the impact on scaling and efficiency of our method. The kinematics-agnostic property grants us two advantages: 1) the number of training parameters does not increase with the number of joints $J$ of the training kinematics $\mathcal{K}$ (*e.g.* SkelDiff requires more parameters for AMASS compared to H36M as AMASS has more joints); 2) we scale constantly with the number of supported datasets or kinematics because a single instance of our method can handle all kinematics, full-body or partial, that currently exists or will be released in the future. Instead, as shown on the left plot of Fig. 3, baselines require a new instance for each kinematics. Therefore, when comparing instances on a single dataset (AMASS), we are 75% more compact (70% on H36M) than the latest baseline SkelDiff [19], while when considering three datasets (right plot of Fig. 3), we are 90% more compact. Our inference and training time are halved, despite training on multiple datasets (Appendix H.1).

*Out-of-distribution Motions on MoYoga ($\mathcal{K}^A$).* In Tab. 2 we report quantitative evaluation on the MoYoga dataset [90], containing out-of-distribution motions recorded via MoCap. Since this dataset shares the same kinematics as the training dataset AMASS ($\mathcal{K}^A$), baselines allow testing without retargeting. Our architecture outperforms the most recent and competitive baseline, SkelDiff [19], showcasing the stronger generalization capability of our model and again the advantages of using a richer prior. This is best highlighted in the qualitative results provided on our webpage.

*Zero-Shot Partial Kinematics.* Performing SHMP with partial input skeletons in a zero-shot setting, without training for it specifically, constitutes an exciting opportunity opened by our method. We also find it particularly relevant for applications, as it allows, depending on the upstream acquisition (MoCap or video), to represent uncertainty or occlusions. EquiFusion natively supports

**Table 4:** Ablations for methods trained on multiple datasets, AMASS($\mathcal{K}^A$) and Nymeria($\mathcal{K}^N$), tested on AMASS ($\mathcal{K}^A$) as usual for single-kinematics and on H36M ($\mathcal{K}^H$) for zero-shot kinematics.

| | AMASS ($\mathcal{K}^A$) | | | H36M ($\mathcal{K}^H$) | | |
|---|---|---|---|---|---|---|
| Method | ADE↓ | MAE ↓ | CMD↓ | ADE↓ | MAE ↓ | CMD↓ |
| Modified [19] | **0.500** | 7.007 | **19.233** | 0.670 | 12.065 | 31.461 |
| Ours w/o Eq | 0.539 | 6.974 | 22.294 | 0.467 | 7.245 | 9.559 |
| Ours | 0.512 | 6.525 | 19.699 | **0.380** | 5.685 | **8.086** |
| Ours on $\mathbf{P}+\mathbf{P}\epsilon$ | 0.512 | 6.525 | 19.699 | **0.380** | 5.685 | **8.086** |
| Ours on $\mathbf{P}$ | 0.513 | 6.529 | 19.686 | **0.380** | **5.680** | 8.088 |
| Ours+$S$ | 0.508 | **6.456** | 19.492 | 0.381 | 5.708 | 8.526 |

**Table 5:** Models trained on AMASS [63] with two different motion parametrization: the proposed bone directions **L** or the conventional 3D positions **M**.

| | | AMASS ($\mathcal{K}^A$) | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | Precision ↓ | | | Div ↑ | Real ↓ | B. Real ↓ | |
| | Mot. | ADE | FDE | MAE | APD | CMD | str | jit |
| SkelDiff [19] | **M** | **0.480** | **0.545** | **6.124** | 9.456 | 11.417 | 3.15 | 0.20 |
| SkelDiff [19] | **L** | 0.496 | 0.546 | 6.193 | **9.960** | **9.143** | **0.00** | **0.00** |
| Ours | **M** | 0.501 | 0.561 | 6.551 | 8.348 | 13.963 | 3.58 | 0.27 |
| Ours | **L** | **0.498** | **0.559** | **6.173** | **8.413** | **12.530** | **0.00** | **0.00** |

flexible kinematics, and so also this scenario. For quantitative evaluation, we randomly remove full limbs (i.e., legs or arms consisting of three joints) from every input sequence in the AMASS test set and pass them to the networks (Tab. 3). To let the most recent and competitive baselines operate in this case, we adopt two heuristics. In the first, we complete the missing information with the one from the rest pose (+rp), simulating a "mask" effect. In the second step, we replicate the motion observed from the symmetric counterpart (+sl), resulting in more realistic motions. Despite the symmetric completion enhancing the diversity of predictions and improving body realism, we observe that precision is still lacking. Our model instead predicts a reasonable future regardless of the missing part. Also in this case, relying on a wider motion prior (A+N) helps overcoming the missing information. Qualitative results are presented on our website. Apart from its applicative relevance to face occlusions, we also foresee our flexibility handy for tackling marginalized categories, such as individuals with diverse body types and those with missing limbs.

*Conventional Single-kinematics SHMP.* Beyond cross-kinematics, our approach achieves state-of-the-art competitive performance on the in-domain, same kinematics evaluation of conventional SHMP benchmarks, *i.e.* on the designed test set of each dataset, for the kinematics belonging to that dataset. Besides AMASS [63] in Tab. 17 and H36M [38] in Tab. 16, we additionally train and evaluate the latest SHMP diffusion models on Nymeria in Tab. 18. The precision $u$ADE averaged over all three datasets is reported in Fig. 3. Remarkably, while different methods require manual tuning of hyperparameters for each dataset [7,15,19,85], our method employs only a single configuration.

*Ablations.* We first validate our approach in Tab. 4. We show that a model whose weights are learned in dependence of joints, despite being trained with data augmentation on multiple kinematics on the same amount of data (A+N), does achieve competitive performance on the training kinematics (A), but fails at zero-shot kinematics on H36M. This shows in current settings, data alone does not guarantee kinematics-agnostic models. See Appendix G.1 for how we adapt the most competitive baseline [19] to multidataset training. Breaking our equivariance property by inserting positional encoding (*Ours w/o Eq*) leads to failure at zero-shot kinematics: equivariance guarantees kinematics-agnostic ca-

pabilities. We verify that our model remains equivariant under permutations of joints, adjacency, and noise (*Ours on* $\mathbf{P}+\mathbf{P}\epsilon$). We empirically validate end-to-end equivariance under stochastic sampling (i.e. we permute joints and adjacency on the same noise realization and compare outputs) as (*Ours on* $\mathbf{P}$). Furthermore, providing explicit joint semantics (e.g., limb side or type) is non-trivial (permutation equivariance must hold) and yields no gains (*Ours+S*); the motion's temporal information alone suffices to distinguish limbs, as human degrees of freedom are invariant. In Tab. 5 we validate the choice of using bone directions as motion representation, showing its contribution to both our method and the baseline SkelDiff. In both cases, such a representation has a limited impact on precision and diversity, while leading to improved realism by 10% and achieving perfectly consistent limb length by design. Further metrics and more detailed ablations are reported in the appendix.

## 5 Conclusion

*Limitations and Future Work* We introduced a novel motion parameterization, bone directions, whose formulation guarantees perfect adherence to bone length constraints. However, this representation handles only missing joints which are leaves of the kinematic chain; it does not support, e.g., a missing elbow when the hand joint is present. While such occlusions can be represented naturally in the input adjacency matrix, extending bone directions to handle them remains open. Beyond this, several broader directions warrant exploration. Training on multiple datasets with differing motion distributions raises open questions about how data quality, particularly the prevalence of dynamic movements, affects performance at inference time in both in-domain and cross-dataset scenarios. Finally, future work may integrate skeletons beyond human kinematics, investigating bipedal and quadrupedal animal motion.

*Conclusions* We presented EquiFusion, a novel and fundamentally more generalizable approach to Stochastic 3D Human Motion Prediction (SHMP). By formulating the skeleton kinematics as an explicit input and designing an end-to-end permutation-equivariant architecture for a latent diffusion model, we successfully sever SHMP models' reliance on graph structures hard-coded at training time. This paradigm enables a single model to generalize zero-shot to unseen datasets and eliminates the need for expensive and inaccurate data retargeting. Beyond resolving the generalization bottleneck, EquiFusion achieves state-of-the-art performance with remarkably enhanced efficiency and supports novel tasks such as partial motion prediction and targeted limb generation. This represents a fundamental shift from kinematics-specific models toward a truly kinematics-agnostic, generalizable SHMP framework.

*Acknowledgments* This work was supported by the European Research Council (ERC) Advanced Grant SIMULACRON. Thanks to Maolin Gao and Felix Wimbauer for proofreading, Thomas Dagès for the detailed and constructive suggestions, Stefania Zunino and the CVG team for their unwavering support.

## References


1. Movella XSens MVN Link motion capture, https://www.movella.com/products/motion-capture/xsens-mvn-link 3
2. Aberman, K., Li, P., Lischinski, D., Sorkine-Hornung, O., Cohen-Or, D., Chen, B.: Skeleton-aware networks for deep motion retargeting. ACM Transactions on Graphics (TOG) **39**(4), 62–1 (2020) 4
3. Aberman, K., Wu, R., Lischinski, D., Chen, B., Cohen-Or, D.: Learning character-agnostic motion for motion retargeting in 2d. arXiv preprint arXiv:1905.01680 (2019) 4
4. Abramson, J., Adler, J., Dunger, J., Evans, R., Green, T., Pritzel, A., Ronneberger, O., Willmore, L., Ballard, A.J., Bambrick, J., et al.: Accurate structure prediction of biomolecular interactions with alphafold 3. Nature **630**(8016), 493–500 (2024) 9
5. AlAttar, A., Chappell, D., Kormushev, P.: Kinematic-model-free predictive control for robotic manipulator target reaching with obstacle avoidance. Frontiers in Robotics and AI **9**, 809114 (2022) 7
6. Alibeigi, M., Ahmadabadi, M.N., Araabi, B.N.: A fast, robust, and incremental model for learning high-level concepts from human motions by imitation. IEEE Transactions on Robotics **33**(1), 153–168 (2016) 2, 4
7. Barquero, G., Escalera, S., Palmero, C.: Belfusion: Latent diffusion for behavior-driven human motion prediction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2317–2327 (2023) 2, 4, 5, 7, 8, 9, 10, 11, 12, 14
8. Barsoum, E., Kender, J., Liu, Z.: Hp-gan: Probabilistic 3d human motion prediction via gan. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops. pp. 1418–1427 (2018) 4
9. Bie, X., Guo, W., Leglaive, S., Girin, L., Moreno-Noguer, F., Alameda-Pineda, X.: Hit-dvae: Human motion generation via hierarchical transformer dynamical vae. arXiv preprint arXiv:2204.01565 (2022) 10
10. Bietti, A., Venturi, L., Bruna, J.: On the sample complexity of learning under geometric stability. Advances in neural information processing systems **34**, 18673–18684 (2021) 3, 9
11. Brooks, T., Peebles, B., Holmes, C., DePue, W., Guo, Y., Jing, L., Schnurr, D., Taylor, J., Luhman, T., Luhman, E., et al.: Video generation models as world simulators. OpenAI Blog **1**(8), 1 (2024) 2
12. Cai, Y., Wang, Y., Zhu, Y., Cham, T.J., Cai, J., Yuan, J., Liu, J., Zheng, C., Yan, S., Ding, H., et al.: A unified 3d human motion synthesis model via conditional variational auto-encoder. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 11645–11655 (2021) 4
13. Cao, Z., Simon, T., Wei, S.E., Sheikh, Y.: Realtime multi-person 2d pose estimation using part affinity fields. In: CVPR (2017) 5
14. Cao, Z., Liu, B., Li, S., Zhang, W., Chen, H.: G-dream: Graph-conditioned diffusion retargeting across multiple embodiments. arXiv preprint arXiv:2505.20857 (2025) 4, 5
15. Chen, L.H., Zhang, J., Li, Y., Pang, Y., Xia, X., Liu, T.: Humanmac: Masked motion completion for human motion prediction. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9544–9555 (2023) 2, 4, 7, 9, 10, 11, 14

16. Choi, K.J., Ko, H.S.: Online motion retargetting. The Journal of Visualization and Computer Animation **11**(5), 223–235 (2000) 4
17. Cui, Q., Sun, H.: Towards accurate 3d human motion prediction from incomplete observations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4801–4810 (2021) 2, 4
18. Cui, Q., Sun, H., Yang, F.: Learning dynamic relationships for 3d human motion prediction. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6519–6527 (2020) 4
19. Curreli, C., Muhle, D., Saroha, A., Ye, Z., Marin, R., Cremers, D.: Nonisotropic gaussian diffusion for realistic 3d human motion prediction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1871–1882 (2025) 2, 3, 4, 5, 7, 8, 9, 10, 11, 12, 13, 14
20. Dabhi, M., Jeni, L.A., Lucey, S.: 3d-lfm: Lifting foundation model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10466–10475 (2024) 4
21. Dang, L., Nie, Y., Long, C., Zhang, Q., Li, G.: Diverse human motion prediction via gumbel-softmax sampling from an auxiliary space. In: Proceedings of the 30th ACM International Conference on Multimedia. pp. 5162–5171 (2022) 2, 4, 5, 7, 9, 10, 11, 12
22. Delhaisse, B., Esteban, D., Rozo, L., Caldwell, D.: Transfer learning of shared latent spaces between robots with similar kinematic structure. In: 2017 International Joint Conference on Neural Networks (IJCNN). pp. 4142–4149. IEEE (2017) 4
23. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. Advances in neural information processing systems **34**, 8780–8794 (2021) 8
24. Engel, J., Somasundaram, K., Goesele, M., Sun, A., Gamino, A., Turner, A., Talattof, A., Yuan, A., Souti, B., Meredith, B., et al.: Project aria: A new tool for egocentric multi-modal ai research. arXiv preprint arXiv:2308.13561 (2023) 3
25. Feng, A., Huang, Y., Xu, Y., Shapiro, A.: Automating the transfer of a generic set of behaviors onto a virtual character. In: Motion in Games: 5th International Conference, MIG 2012, Rennes, France, November 15-17, 2012. Proceedings 5. pp. 134–145. Springer (2012) 4
26. Figuera, X., Park, S., Ahn, H.: Redefining data pairing for motion retargeting leveraging a human body prior. In: 2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 4800–4807. IEEE (2024) 4
27. Fritsche, O., Camacho, S., Hossain, M.S.B., Halpenny, T., Archniegas, C., Dranetz, J., Hadley, D., Guo, Z., Choi, H.: Ultra-mocap: A multimodal imu and semg dataset for upper body joint kinematics analysis. Authorea Preprints (2025) 3
28. Gat, I., Raab, S., Tevet, G., Reshef, Y., Bermano, A.H., Cohen-Or, D.: Anytop: Character animation diffusion with any topology. In: Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers. pp. 1–10 (2025) 4, 7
29. Geist, A.R., Frey, J., Zhobro, M., Levina, A., Martius, G.: Learning with 3d rotations, a hitchhiker's guide to so (3). arXiv preprint arXiv:2404.11735 (2024) 5, 10
30. Gil, O., Sanfeliu, A.: Human motion trajectory prediction using the social force model for real-time and low computational cost applications. In: Iberian Robotics conference. pp. 235–247. Springer (2023) 7

31. Gleicher, M.: Retargetting motion to new characters. In: Proceedings of the 25th annual conference on Computer graphics and interactive techniques. pp. 33–42 (1998) 4
32. Gu, C., Yu, J., Zhang, C.: Learning disentangled representations for controllable human motion prediction. Pattern Recognition **146**, 109998 (2024) 4
33. Guzov, V., Chibane, J., Marin, R., He, Y., Saracoglu, Y., Sattler, T., Pons-Moll, G.: Interaction replica: Tracking human–object interaction and scene changes from human motion. In: 2024 International Conference on 3D Vision (3DV). pp. 1006–1016. IEEE (2024) 3
34. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020) 2, 8
35. Holden, D., Saito, J., Komura, T.: A deep learning framework for character motion synthesis and editing. ACM Transactions on Graphics (TOG) **35**(4), 1–11 (2016) 2, 4, 11, 12
36. Hu, L., Zhang, Z., Zhong, C., Jiang, B., Xia, S.: Pose-aware attention network for flexible motion retargeting by body part. IEEE Transactions on Visualization and Computer Graphics (2023) 4, 7
37. Huang, Z., Feng, H., Sun, Y.T., Guo, Y.C., Cao, Y.P., Sheng, L.: Animax: Animating the inanimate in 3d with joint video-pose diffusion models. In: Proceedings of the SIGGRAPH Asia 2025 Conference Papers. pp. 1–13 (2025) 4
38. Ionescu, C., Papava, D., Olaru, V., Sminchisescu, C.: Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments (2014), http://vision.imar.ro/human3.6m. 2, 3, 6, 9, 10, 11, 12, 14
39. Jang, H., Kwon, B., Yu, M., Kim, S.U., Kim, J.: A variational u-net for motion retargeting. In: SIGGRAPH Asia 2018 Posters. pp. 1–2 (2018) 4
40. Jang, I., Choi, S., Hong, S., Kim, C., Noh, J.: Geometry-aware retargeting for two-skinned characters interaction. ACM Transactions on Graphics (TOG) **43**(6), 1–17 (2024) 4
41. Keriven, N., Peyré, G.: Universal invariant and equivariant graph neural networks. Advances in neural information processing systems **32** (2019) 9
42. Kim, W., Li, T., Ha, S.: Moreflow: Motion retargeting learning through unsupervised flow matching. arXiv preprint arXiv:2509.25600 (2025) 4
43. Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907 (2016) 9
44. Kundu, J.N., Gor, M., Babu, R.V.: Bihmp-gan: Bidirectional 3d human motion prediction gan. In: Proceedings of the AAAI conference on artificial intelligence. vol. 33, pp. 8553–8560 (2019) 4
45. Laabid, N., Rissanen, S., Heinonen, M., Solin, A., Garg, V.: Equivariant denoisers cannot copy graphs: Align your graph diffusion models. In: International Conference on Learning Representations. vol. 2025, pp. 100256–100294 (2025) 4
46. Lee, J., Shin, S.Y.: A hierarchical approach to interactive motion editing for human-like figures. In: Proceedings of the 26th annual conference on Computer graphics and interactive techniques. pp. 39–48 (1999) 4
47. Lee, S., Kang, T., Park, J., Lee, J., Won, J.: Same: Skeleton-agnostic motion embedding for character animation. In: SIGGRAPH Asia 2023 Conference Papers. pp. 1–11 (2023) 2, 4, 7
48. Li, C., Chibane, J., He, Y., Pearl, N., Geiger, A., Pons-Moll, G.: Unimotion: Unifying 3d human motion synthesis and understanding. In: 2025 International Conference on 3D Vision (3DV). pp. 240–249. IEEE (2025) 5

49. Li, M., Chen, S., Chen, X., Zhang, Y., Wang, Y., Tian, Q.: Actional-structural graph convolutional networks for skeleton-based action recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3595–3603 (2019) 4
50. Li, M., Chen, S., Liu, Z., Zhang, Z., Xie, L., Tian, Q., Zhang, Y.: Skeleton graph scattering networks for 3d skeleton-based human motion prediction. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 854–864 (2021) 9
51. Li, M., Chen, S., Zhao, Y., Zhang, Y., Wang, Y., Tian, Q.: Dynamic multiscale graph neural networks for 3d skeleton based human motion prediction. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 214–223 (2020) 4, 7, 9
52. Li, P., Starke, S., Ye, Y., Sorkine-Hornung, O.: Walkthedog: Cross-morphology motion alignment via phase manifolds. In: ACM SIGGRAPH 2024 Conference Papers. pp. 1–10 (2024) 4
53. Li, T., Won, J., Clegg, A., Kim, J., Rai, A., Ha, S.: Ace: Adversarial correspondence embedding for cross morphology motion retargeting from human to nonhuman characters. In: SIGGRAPH Asia 2023 Conference Papers. pp. 1–11 (2023) 4
54. Liao, Z., Yang, J., Saito, J., Pons-Moll, G., Zhou, Y.: Skeleton-free pose transfer for stylized 3d characters. In: European Conference on Computer Vision. pp. 640–656. Springer (2022) 4, 5
55. Lim, J., Chang, H.J., Choi, J.Y.: Pmnet: Learning of disentangled pose and movement for unsupervised motion retargeting. In: 30th British Machine Vision Conference (BMVC 2019). British Machine Vision Association, BMVA (2019) 2, 4
56. Lin, P., Chen, P., Jiao, R., Mo, Q., Cen, J., Huang, W., Liu, Y., Huang, D., Lu, Y.: Equivariant diffusion for crystal structure prediction. arXiv preprint arXiv:2512.07289 (2025) 8
57. Liu, Q., Lv, K., Dong, K., Xue, J., Niu, Z., Wang, J.: Text-to-any-skeleton motion generation without retargeting. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 12926–12936 (2025) 4, 7
58. Liu, Z., Lyu, K., Wu, S., Chen, H., Hao, Y., Ji, S.: Aggregated multi-gans for controlled 3d human motion prediction. In: Proceedings of the AAAI conference on artificial intelligence. vol. 35, pp. 2225–2232 (2021) 4
59. Lohit, S., Anirudh, R., Turaga, P.: Recovering trajectories of unmarked joints in 3d human actions using latent space optimization. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 2342–2351 (2021) 5
60. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: SMPL: A skinned multi-person linear model. ACM Trans. Graphics (Proc. SIGGRAPH Asia) **34**(6), 248:1–248:16 (Oct 2015) 3, 5, 10
61. Lyle, C., van der Wilk, M., Kwiatkowska, M., Gal, Y., Bloem-Reddy, B.: On the benefits of invariance in neural networks. arXiv preprint arXiv:2005.00178 (2020) 3, 9
62. Ma, L., Ye, Y., Hong, F., Guzov, V., Jiang, Y., Postyeni, R., Pesqueira, L., Gamino, A., Baiyya, V., Kim, H.J., et al.: Nymeria: A massive collection of multimodal egocentric daily motion in the wild. In: European Conference on Computer Vision. pp. 445–465. Springer (2024) 2, 3, 6, 10, 11
63. Mahmood, N., Ghorbani, N., Troje, N.F., Pons-Moll, G., Black, M.J.: Amass: Archive of motion capture as surface shapes. In: Proceedings of the IEEE/CVF

international conference on computer vision. pp. 5442–5451 (2019) 2, 3, 6, 10, 11, 12, 13, 14
64. Mao, W., Liu, M., Salzmann, M.: Generating smooth pose sequences for diverse human motion prediction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13309–13318 (2021) 2, 4, 5, 9, 11
65. Mao, W., Liu, M., Salzmann, M., Li, H.: Learning trajectory dependencies for human motion prediction. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9489–9497 (2019) 7
66. von Marcard, T., Henschel, R., Black, M., Rosenhahn, B., Pons-Moll, G.: Recovering accurate 3d human pose in the wild using imus and a moving camera. In: European Conference on Computer Vision (ECCV) (sep 2018) 6
67. Mourot, L., Hoyet, L., Clerc, F.L., Hellier, P.: Humot: human motion representation using topology-agnostic transformers for character animation retargeting. arXiv preprint arXiv:2305.18897 (2023) 4, 5
68. Musoni, P., Marin, R., Melzi, S., Castellani, U.: A functional skeleton transfer. Proceedings of the ACM on Computer Graphics and Interactive Techniques **4**(3), 1–15 (2021) 4
69. Musoni, P., Marin, R., Melzi, S., Castellani, U.: Reposing and retargeting unrigged characters with intrinsic-extrinsic transfer. In: STAG. pp. 21–30 (2021) 4
70. Nan, X., Liu, X., You, X., Du, Y., Ji, C., Song, J.: Pgdiff: A physics-guided equivariant diffusion model for structure-based drug design. In: 2025 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). pp. 687–692 (2025) 4
71. Park, J.S., Manocha, D.: Hmpo: Human motion prediction in occluded environments for safe motion planning. arXiv preprint arXiv:2006.00424 (2020) 5
72. Patterson, D., Gonzalez, J., Le, Q., Liang, C., Munguia, L.M., Rothchild, D., So, D., Texier, M., Dean, J.: Carbon emissions and large neural network training. arXiv preprint arXiv:2104.10350 (2021) 8
73. Petrov, I.A., Guzov, V., Marin, R., Aksan, E., Chen, X., Cremers, D., Beeler, T., Pons-Moll, G.: Echo: Ego-centric modeling of human-object interactions. In: Proceedings of the European conference on computer vision (ECCV) (2026) 3
74. Phu, K.A., Hoang, V.D., et al.: Predicting occluded skeletal joints via tracking-based feature extraction. Neurocomputing p. 131004 (2025) 5
75. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125 **1**(2), 3 (2022) 9
76. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models (2022) 2, 8
77. Roudsarabi, N., Behrad, A.R.: Solving occlusion problem in 3d human motion reconstruction. In: 2008 International Symposium on Telecommunications. pp. 701–706. IEEE (2008) 5
78. Saadatnejad, S., Rasekh, A., Mofayezi, M., Medghalchi, Y., Rajabzadeh, S., Mordan, T., Alahi, A.: A generic diffusion-based approach for 3d human pose prediction in the wild (2023) 4, 9
79. Saito, J., Li, J., de Ruyter, M., Guerrero, M., Lim, E., Hassani, E., Ribera, R.B., Moon, H., Dadela, M., Di Lucca, M., et al.: Soma: Unifying parametric human body models. arXiv preprint (2026) 4
80. Salzmann, T., Pavone, M., Ryll, M.: Motron: Multimodal probabilistic human motion forecasting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6457–6466 (2022) 5, 7, 9, 10

81. Sárándi, I., Hermans, A., Leibe, B.: Learning 3d human pose estimation from dozens of datasets using a geometry-aware autoencoder to bridge between skeleton formats. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 2956–2966 (2023) 4
82. Sárándi, I., Pons-Moll, G.: Neural localizer fields for continuous 3d human pose and shape estimation. Advances in Neural Information Processing Systems **37**, 140032–140065 (2025) 4
83. Schneuing, A., Harris, C., Du, Y., Didi, K., Jamasb, A., Igashov, I., Du, W., Gomes, C., Blundell, T.L., Lio, P., et al.: Structure-based drug design with equivariant diffusion models. Nature Computational Science **4**(12), 899–909 (2024) 4
84. Simon, T., Joo, H., Matthews, I., Sheikh, Y.: Hand keypoint detection in single images using multiview bootstrapping. In: CVPR (2017) 5
85. Sun, J., Chowdhary, G.: Comusion: Towards consistent stochastic human motion prediction via motion diffusion. European Conference on Computer Vision (2024) 2, 4, 7, 9, 10, 11, 14
86. Tak, S., Ko, H.S.: A physically-based motion retargeting filter. ACM Transactions on Graphics (ToG) **24**(1), 98–117 (2005) 4
87. Tang, J., Yang, H., Chen, T., Hu, J.F.: Stochastic human motion prediction with memory of action transition and action characteristic. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 1883–1893 (2025) 5
88. Tevet, G., Gordon, B., Hertz, A., Bermano, A.H., Cohen-Or, D.: Motionclip: Exposing human motion generation to clip space. In: European Conference on Computer Vision. pp. 358–374. Springer (2022) 5
89. Thiede, E.H., Hy, T.S., Kondor, R.: The general theory of permutation equivarant neural networks and higher order graph variational encoders. arXiv preprint arXiv:2004.03990 (2020) 8, 9
90. Tripathi, S., Müller, L., Huang, C.H.P., Omid, T., Black, M.J., Tzionas, D.: 3D human pose estimation via intuitive physics. In: Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4713–4725 (2023) 6, 11, 13
91. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017) 9
92. Veličković, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., Bengio, Y.: Graph attention networks. arXiv preprint arXiv:1710.10903 (2017) 9
93. Villegas, R., Ceylan, D., Hertzmann, A., Yang, J., Saito, J.: Contact-aware retargeting of skinned motion. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9720–9729 (2021) 2, 4
94. Villegas, R., Yang, J., Ceylan, D., Lee, H.: Neural kinematic networks for unsupervised motion retargetting. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 8639–8648 (2018) 2, 4
95. Wad, T., Sun, Q., Pranata, S., Jayashree, K., Zhang, H.: Equivariance and invariance inductive bias for learning from insufficient data. In: European Conference on Computer Vision. pp. 241–258. Springer (2022) 8
96. Walker, J., Marino, K., Gupta, A., Hebert, M.: The pose knows: Video forecasting by generating pose futures. In: Proceedings of the IEEE international conference on computer vision. pp. 3332–3341 (2017) 4, 5, 9, 11
97. Wang, H., Basu, A., Durandau, G., Sartori, M.: A wearable real-time kinetic measurement sensor setup for human locomotion. Wearable technologies **4**, e11 (2023) 3

98. Wang, J., Li, X., Liu, S., De Mello, S., Gallo, O., Wang, X., Kautz, J.: Zero-shot pose transfer for unrigged stylized 3d characters. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8704–8714 (2023) 4, 5
99. Wang, Y., Hu, K., Gupta, S., Ye, Z., Wang, Y., Jegelka, S.: Understanding the role of equivariance in self-supervised learning. Advances in Neural Information Processing Systems **37**, 127483–127510 (2024) 9
100. Wei, D., Sun, H., Li, B., Lu, J., Li, W., Sun, X., Hu, S.: Human joint kinematics diffusion-refinement for stochastic motion prediction. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 6110–6118 (2023) 4, 5
101. Wei, S.E., Ramakrishna, V., Kanade, T., Sheikh, Y.: Convolutional pose machines. In: CVPR (2016) 5
102. Xu, C., Tan, R.T., Tan, Y., Chen, S., Wang, X., Wang, Y.: Auxiliary tasks benefit 3d skeleton-based human motion prediction. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9509–9520 (2023) 2, 4
103. Xu, S., Wang, Y.X., Gui, L.Y.: Diverse human motion prediction guided by multi-level spatial-temporal anchors. In: European Conference on Computer Vision. pp. 251–269. Springer (2022) 4
104. Yan, X., Rastogi, A., Villegas, R., Sunkavalli, K., Shechtman, E., Hadap, S., Yumer, E., Lee, H.: Mt-vae: Learning motion transformations to generate multimodal human dynamics. In: Proceedings of the European conference on computer vision (ECCV). pp. 265–281 (2018) 4
105. Yang, Z., Leite, C.S., Xiao, Y.: Strengthsense: A dataset of imu signals capturing everyday strength-demanding activities. arXiv preprint arXiv:2511.02027 (2025) 3
106. Yao, J., Yang, B., Wang, X.: Reconstruction vs. generation: Taming optimization dilemma in latent diffusion models. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 15703–15712 (2025) 8
107. Ying, C., Cai, T., Luo, S., Zheng, S., Ke, G., He, D., Shen, Y., Liu, T.Y.: Do transformers really perform badly for graph representation? Advances in neural information processing systems **34**, 28877–28888 (2021) 9
108. Yoon, T., Kang, D., Kim, S., Cheng, J., Ahn, M., Coros, S., Choi, S.: Spatio-temporal motion retargeting for quadruped robots. IEEE Transactions on Robotics (2025) 4
109. Yuan, Y., Kitani, K.: Dlow: Diversifying latent flows for diverse human motion prediction. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16. pp. 346–364. Springer (2020) 2, 4, 5, 7, 9, 11
110. Zhang, H., Chen, Z., Xu, H., Hao, L., Wu, X., Xu, S., Zhang, Z., Wang, Y., Xiong, R.: Semantics-aware motion retargeting with vision-language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2155–2164 (2024) 4
111. Zhong, C., Hu, L., Zhang, Z., Ye, Y., Xia, S.: Spatio-temporal gating-adjacency gcn for human motion prediction. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6447–6456 (2022) 7, 9
112. Zhou, L., Meng, X., Liu, Z., Wu, M., Gao, Z., Wang, P.: Human pose-based estimation, tracking and action recognition with deep learning: A survey. arXiv preprint arXiv:2310.13039 (2023) 5