

Geometric Neural Distance Fields for Learning Human Motion Priors

Zhengdi Yu¹, Simone Foti¹, Linguang Zhang², Amy Zhao²,

Cem Keskin², Stefanos Zafeiriou¹, Tolga Birdal¹

¹Imperial College London

²Meta Reality Labs

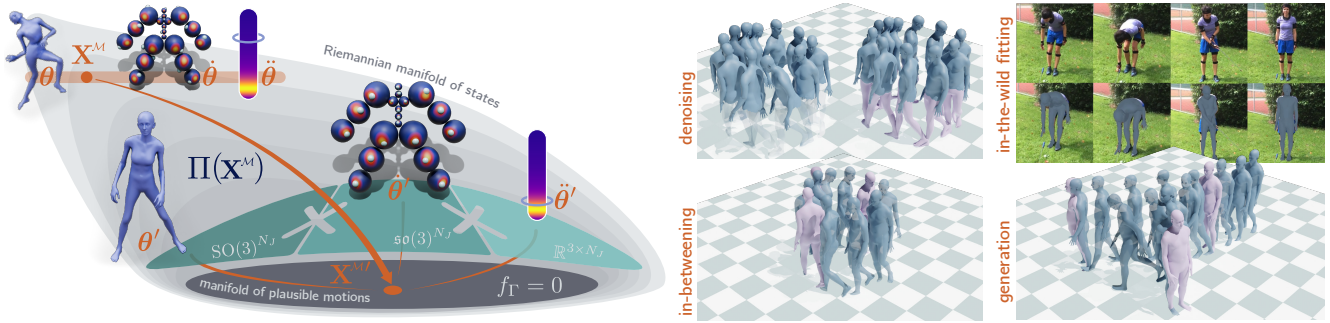


Figure 1. **NRMF** is a general-purpose, expressive and robust unconditional motion prior. It models the space of plausible **poses** (θ), **transitions** ($\dot{\theta}$) and **accelerations** ($\ddot{\theta}$) on the zero-level set of a **geometric neural distance field**. This implicitly captures the data distribution. Poses are depicted alongside their transitions and accelerations, which are visualized as blue dots onto the per-joint distributions of learned transitions and as blue rings around the magnitude distribution of all accelerations. We develop projection (Π) and integration algorithms to deploy NRMF into several applications as shown on the right.

Abstract

We introduce *Neural Riemannian Motion Fields (NRMF)*, a novel 3D generative human motion prior that enables robust, temporally consistent, and physically plausible 3D motion recovery. Unlike existing VAE or diffusion-based methods, our higher-order motion prior explicitly models the human motion in the zero level set of a collection of neural distance fields (NDFs) corresponding to pose, transition (velocity), and acceleration dynamics. Our framework is rigorous in the sense that our NDFs are constructed on the product space of joint rotations, their angular velocities, and angular accelerations, respecting the geometry of the underlying articulations. We further introduce: (i) a novel adaptive-step hybrid algorithm for projecting onto the set of plausible motions, and (ii) a novel geometric integrator to “roll out” realistic motion trajectories during test-time-optimization and generation. Our experiments show significant and consistent gains: trained on the AMASS dataset, NRMF remarkably generalizes across multiple input modalities and to diverse tasks ranging from denoising to motion in-betweening and fitting to partial 2D / 3D observations.

1. Introduction

Motion is intrinsic to human experience. As humans, we are in perpetual motion, engaging with our surroundings in rich

and diverse ways. Accurately recovering 3D human motion from sparse or noisy data is therefore not only a fundamental challenge in computer vision, but also a critical enabler for a myriad of applications, from immersive virtual reality and animation to clinical diagnostics, and sports analysis. Despite remarkable progress in learning-based techniques and model-based reconstructions, current methods still struggle to deliver reliable, temporally consistent, and physically plausible 3D motion recovery while being able to generate diverse and *lifelike* motions. The challenges stem from the intrinsic complexity of human motion, compounded by occlusions, rapid movements, and inherent ambiguities in 2D observations.

To address these challenges, we focus on the problem of building *general-purpose, expressive and robust* models that capture the full dynamics of *realistic* human motion. A conventional way to handle sequences of articulations is to model static poses, frame-by-frame as done in NRDF [18], PoseNDF [43] or DPoser [27]. Yet, processing frames independently jeopardizes generalization and temporal consistency. As a remedy, modern models of motion have shifted toward dynamics-based methods, many of which rely either on VAEs [33, 39] suffering from *posterior-collapse* and increased accumulation of errors over time (*drift*) or leverage the recent diffusion modeling [38, 42], which while promising, tend to capture short pose sequences in clean settings or require additional conditioning such as action or textual descriptions [1, 32]. Diffusion-based models further suf-

fer from increased inference times and their conditioning to observations (*inversion*) is far from being trivial. While RoHM [51] tries to bridge this gap, as we will demonstrate, it suffers from low estimation accuracy, due to the limited handling of higher-order motion dynamics, such as acceleration, causing oversmoothed transitions and drift. MoManifold [11] factors in the accelerations. Yet, beyond being limited to short, fixed time windows, its *decoupled manifold* treats each joint’s acceleration in isolation, neglecting the essential inter-dependencies of the body joints.

In this work, we introduce **Neural Riemannian Motion Fields (NRMF)**, a novel human motion prior that models the manifold of plausible motions as a collection of three neural distance fields (NDFs), each corresponding to 0th-order (pose), 1st-order (pose transitions), and crucially, 2nd-order (acceleration) dynamics, captured as the zero level sets of dedicated conditioned, implicit neural networks (Fig. 2b). We rigorously respect the underlying geometry of articulated motion by operating in the product space of joint rotations, angular velocities, and angular accelerations. To map an arbitrary articulated motion onto a plausible one (Fig. 2a), we introduce a three-stage adaptive-step hybrid gradient descent. We further devise a novel *geometric* and *robust integrator* to *rollout* physically realistic motion sequences, while correcting errors in all dynamic components. Together, these enable a plethora of applications as exemplified in Fig. 1.

After training on the large AMASS [28] motion capture dataset, we extensively evaluate NRMF on numerous tasks including motion estimation, denoising, in-betweening, and fitting to various input modalities, over various datasets including AMASS, i3DB [29], 3DPW [44], EgoBody [50] and PROX [16]. Our quantitative results show **significant and consistent improvements** across several standard metrics for assessing human motion. Notably, a qualitative investigation into the learned pose, transition and acceleration distributions reveals the expressivity of our method. To summarize, our contributions are:

1. We propose to model poses, transitions and accelerations as the zero level sets of corresponding product NDFs.
2. We introduce new algorithms for projecting onto the zero level sets as well as for rolling out physically plausible motion sequences.
3. We ensure that all our algorithms respect the geometry of articulated rotations.
4. From sampling motions to fitting 2D/3D observations, we demonstrate that NRMF surpasses all prominent state-of-the-art methods across several tasks and metrics.

We additionally present a condensed primer on the geometry of articulated motions in our suppl. material as a self-contained background. We will make our implementation publicly available upon publication.

2. Related Work

Unconditional human pose and motion priors. Human bodies have been modeled by unconditional priors of various different kinds including Gaussian Processes [46], VAEs [30], NDFs [9, 18, 43], diffusion models [10, 27]. Yet, per-frame inference is not sufficient to capture the full complexity of motion. To this end, various works such as MotionVAE [24], HuMoR [33] and PhaseMP [39] leveraged VAEs in an autoregressive fashion. With the proliferation of diffusion models and transformers, modern approaches such as [19, 22, 38, 42, 47] leverage transformer-based diffusion frameworks to generate diverse motion. Unlike these works, we take a fundamentally different approach and offer a strong generative prior that is deployable for recovering human motion from observations.

Monocular human pose and motion recovery. Approaches for estimating human pose and shape from monocular observations can be categorized into regression-based and optimization-based. Regression-based methods predict pose directly from images or videos using deep networks [2, 6, 14, 21, 41, 45, 49] which are typically based on the parametric models of SMPL [26], SMPL-H [35] or SMPL-X [30], but often lack temporal consistency, and fail under occlusion. Optimization-based approaches [33, 39, 51] fit parametric models to observations while enforcing temporal priors, achieving better robustness but at the cost of high computation, and sensitivity to initialization. Recent methods also integrate learned motion priors via diffusion [37, 42], improving plausibility and infilling. RoHM [51] reconstructs motion via diffusion denoising of root and local pose spaces with image and physics guidance. However, it still suffers from oversmoothing and cannot fully exploit higher-order dynamics. Our method NRMF explicitly models pose, velocity, and acceleration via neural distance fields on Riemannian motion manifolds, enabling physically plausible recovery across diverse inputs.

3. Neural Riemannian Motion Fields

We start by explaining **Riemannian Motion Fields** to model realistic articulated motions in Sec. 3.1 and introduce our novel projection algorithm to map onto this space while adhering to the manifold of joint rotations, transitions and accelerations. We then propose NRMF and a novel method for sampling articulated motions, to generate desired training data, in Sec. 3.2. Before delving into method specifics, we briefly review the $SO(3)$ -kinematics and provide a detailed, self-contained exposition in our suppl. material.

Background. We are particularly interested in the non-commutative Lie group of finite rigid body rotations:

Definition 1 ($SO(3)$). *The set of rotation matrices admit a group structure under matrix multiplication:*

$$SO(3) = \{\mathbf{R} \in \mathbb{R}^{3 \times 3} : \mathbf{R}^\top \mathbf{R} = \mathbf{I}, \det(\mathbf{R}) = 1\}. \quad (1)$$

We now briefly characterize the properties of $\text{SO}(3)$.

Proposition 1. *The tangent space of $\text{SO}(3)$ at $\mathbf{R}$ is defined throughout the paper as the left-invariant vector fields:*

$$\mathcal{T}_{\mathbf{R}}\text{SO}(3) = \{\mathbf{R}\boldsymbol{\Omega} : \boldsymbol{\Omega} \in \mathfrak{so}(3)\}, \quad (2)$$

where $\mathfrak{so}(3)$ is given by the set of skew-symmetric tensors:

$$\mathfrak{so}(3) = \{\boldsymbol{\Omega} \in \mathbb{R}^{3 \times 3} : \boldsymbol{\Omega}^\top = -\boldsymbol{\Omega}\}. \quad (3)$$

Proof. Differentiating the constraints that specify $\text{SO}(3)$ leads us to the definition of skew-symmetric matrices:

$$\frac{d(\mathbf{R}^\top \mathbf{R})}{dt} = \dot{\mathbf{R}}^\top \mathbf{R} + \mathbf{R}^\top \dot{\mathbf{R}} = 0 \implies \boldsymbol{\Omega} = -\boldsymbol{\Omega}^\top, \quad (4)$$

where $\boldsymbol{\Omega} = \mathbf{R}^\top \dot{\mathbf{R}}$ is a skew symmetric matrix and the vector $\dot{\mathbf{R}} = \mathbf{R}\boldsymbol{\Omega}$ is tangent to $\mathbf{R}$. $\square$

We also need the notions of angular velocity and acceleration on $\text{SO}(3)$ to accurately model articulated motions.

Definition 2 (Angular velocity). *The angular velocity of the geodesic motion on $\text{SO}(3)$ is represented as a time (t) dependent curve $\boldsymbol{\omega}_t := \omega(t) := \text{skew}(\boldsymbol{\Omega}(t))$ extracted from:*

$$\boldsymbol{\Omega}_t = \mathbf{R}_t^{-1} \dot{\mathbf{R}}_t, \quad (5)$$

where $\text{skew}(\cdot)$ transforms the matrix into a skew-vector and the **angular velocity matrix** $\boldsymbol{\Omega}_t := \boldsymbol{\Omega}(t)$ is skew-symmetric.

Proposition 2 (Angular acceleration). *The time-derivative of angular velocity defines the angular acceleration:*

$$\frac{d[\boldsymbol{\omega}_t]_x}{dt} := [\dot{\boldsymbol{\omega}}_t]_x := \ddot{\mathbf{R}}_t = \mathbf{R}_t (\boldsymbol{\Omega}_t^2 + \dot{\boldsymbol{\Omega}}_t), \quad (6)$$

where $\boldsymbol{\Omega}_t^2 + \dot{\boldsymbol{\Omega}}_t$ is also skew-symmetric and $\dot{\mathbf{R}}_t := d\mathbf{R}_t/dt$.

In practice, we use a central differencing scheme to estimate the angular velocity and acceleration. As we show in our suppl. material, this is a good approximation of the true acceleration computed via log-maps and *parallel transport*.

Our work differs significantly from modeling motion in pure $\text{SO}(3)$ [3, 15] also because of the specific *product manifold* structure which we discuss next.

3.1. Modeling of Plausible Articulated Motions

We consider an input motion sequence $\{\mathbf{X}_t\}_{t=0}^T$ composed of individual states $\mathbf{X}_t$:

Definition 3 (State). *We represent the state of a moving articulated body as a time dependent function $x(t) : t \rightarrow \mathbf{X}_t$ producing the matrix $\mathbf{X}_t = x(t)$ composed of a root translation $\mathbf{t}_r \in \mathbb{R}^3$, body joint angles $\boldsymbol{\theta} \in \text{SO}(3)^{N_J}$, their*

velocities $\dot{\boldsymbol{\theta}} \in \text{TSO}(3)^{N_J}$ and angular accelerations $\ddot{\boldsymbol{\theta}} \in \mathbb{R}^{3 \times N_J}$.

$$\mathbf{X} = [\mathbf{t}_r \ \boldsymbol{\theta} \ \dot{\boldsymbol{\theta}} \ \ddot{\boldsymbol{\theta}}] \in (\mathbb{R}^3 \times \mathcal{M}), \quad (7)$$

where N_J refers to the total number of joints excluding the hand joints ($N_J = 22$ for SMPL model) and $\mathcal{M}$ denotes the product space of the state vector including the root orientation: $\mathcal{M} = (\text{SO}(3)^{N_J} \times \mathfrak{so}(3)^{N_J} \times \mathbb{R}^{3 \times N_J})$.

The first part of the state involving root translation is known as the **global motion**, whereas the rest parameterize the **local motion**, the motion of the skeleton. Part of the state, $(\mathbf{t}_r, \boldsymbol{\theta})$, parameterizes the articulated object model (e.g., SMPL for humans [26, 35] or MANO for hands [34]) which is a differentiable function $M(\mathbf{t}_r, \boldsymbol{\theta}, \boldsymbol{\beta})$ mapping joints $\mathbf{J} \in \mathbb{R}^{3 \times J}$ and shape parameters $\boldsymbol{\beta} \in \mathbb{R}^{N_\beta}$ to the vertices $\mathbf{V} \in \mathbb{R}^{3 \times N_V}$ of an articulated object, where $N_V = 6890$ for SMPL human bodies.

Instead of parameterizing the entire motion sequence as done in [17], we propose to model the current state (pose, transition and acceleration) while respecting the geometry:

Proposition 3 (Riemannian Motion Fields (RMFs)). *Given the state, we model the manifold of instantaneous **realistic and plausible motions** of an articulated body as the zero level set of an implicit function $f_\Gamma : \mathcal{M} \rightarrow \mathbb{R}_+^3$:*

$$\mathcal{S} = \{\mathbf{X}^\mathcal{M} \in \mathcal{M} \mid f_\Gamma(\mathbf{X}^\mathcal{M}) = 0\}, \quad (8)$$

such that the range of f_Γ (parameterized by Γ) represents the unsigned geodesic distance to the closest plausible pose on the manifold. To account for the interdependence of pose, velocity and acceleration we disentangle f_Γ into conditional components as:

$$f_\Gamma = \begin{bmatrix} f_\Phi^R(\boldsymbol{\theta}) \\ f_\Psi^\omega(\boldsymbol{\theta} \mid \boldsymbol{\theta}) \\ f_\Xi^\omega(\ddot{\boldsymbol{\theta}} \mid \boldsymbol{\theta}, \dot{\boldsymbol{\theta}}) \end{bmatrix} \begin{matrix} \text{(pose field)} \\ \text{(transition field)} \\ \text{(acceleration field)} \end{matrix} \quad \text{(RMF)}$$

where $f_\Phi^R : \text{SO}(3)^{N_J} \rightarrow \mathbb{R}_+$, $f_\Psi^\omega : \mathfrak{so}(3)^{N_J} \rightarrow \mathbb{R}_+$ and $f_\Xi^\omega : \mathbb{R}^{3 \times N_J} \rightarrow \mathbb{R}_+$ account for the plausibility in pose, velocity and acceleration, respectively. $\Gamma = [\Phi, \Psi, \Xi]$.

We are now ready to state our iterations for projecting onto this zero level set specified by f_Γ :

Proposition 4 (RMF-Grad). *A geometry-aware single step projection of a given state onto our motion manifold involves the updates:*

$$\begin{aligned} \boldsymbol{\theta}_{t+1} &:= \Pi^R(\boldsymbol{\theta}_t) = \text{Exp}_{\boldsymbol{\theta}_t} \left(-\alpha_\theta f_\Phi^R(\boldsymbol{\theta}_t) \frac{\text{grad} f_\Phi^R(\boldsymbol{\theta}_t)}{\|\text{grad} f_\Phi^R(\boldsymbol{\theta}_t)\|} \right) \\ \dot{\boldsymbol{\theta}}_{t+1} &:= \Pi_t^\omega(\dot{\boldsymbol{\theta}}_t) = \dot{\boldsymbol{\theta}}_t - \alpha_\theta f_\Psi^\omega(\dot{\boldsymbol{\theta}}_t) \frac{\nabla f_\Psi^\omega(\dot{\boldsymbol{\theta}}_t \mid \boldsymbol{\theta}_t)}{\|\nabla f_\Psi^\omega(\dot{\boldsymbol{\theta}}_t \mid \boldsymbol{\theta}_t)\|} \\ \ddot{\boldsymbol{\theta}}_{t+1} &:= \Pi_t^\omega(\ddot{\boldsymbol{\theta}}_t) = \ddot{\boldsymbol{\theta}}_t - \alpha_\omega f_\Xi^\omega(\ddot{\boldsymbol{\theta}}_t) \frac{\nabla f_\Xi^\omega(\ddot{\boldsymbol{\theta}}_t \mid \boldsymbol{\theta}_t, \dot{\boldsymbol{\theta}}_t)}{\|\nabla f_\Xi^\omega(\ddot{\boldsymbol{\theta}}_t \mid \boldsymbol{\theta}_t, \dot{\boldsymbol{\theta}}_t)\|}, \end{aligned} \quad (9)$$

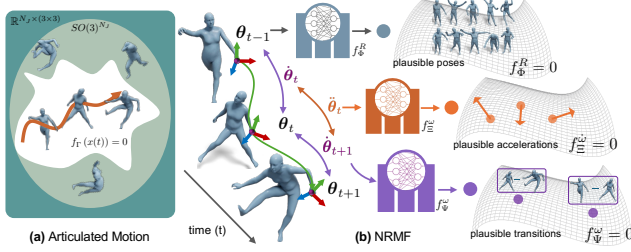


Figure 2. Neural Riemannian Motion Fields (NRMF) models the motion in the zero-level-set of three disjoint distance fields.

where α_* corresponds to individual learning rates and grad and Exp denote the **Riemannian gradient** and **Exponential maps** as we provide in our suppl. material.

Remark 1. Note that under ideal conditions where transition (velocities) are obtained computationally, we expect:

$$\Pi^\omega(\dot{\theta}) \approx (\Pi^R(\theta)^\top \Pi^R(\dot{\theta})), \quad (10)$$

and leverage the right hand side, accounting for the use of noisy input velocities. Likewise, $\Pi^\omega(\ddot{\theta}) \approx d(\Pi^\omega(\dot{\theta}))/dt$.

A geometry-aware robust integrator. We now provide a novel *deterministic* integration algorithm for **rolling out** motion sequences starting from a plausible initial state. Equivalently, it will allow us to *rebuild* a consistent state sequence from the projected measurements, a key step in motion generation and test-time-optimization as we will explain later.

Proposition 5 (RMF-Integrator). For a moving articulated body $x(t) : t \rightarrow \mathbf{X}_t$, the “ideal” continuous dynamics are governed by the differential equations in Eqs. (5) and (6). We then apply a geometric, projected **Euler integration** to those factoring in our priors. Given a sequence of plausible accelerations, this leads to the following discrete updates:

$$\dot{\theta}_{t+1} = \Pi^\omega(\dot{\theta}_t + \lambda_t \ddot{\theta}_t) \quad (11)$$

$$\theta_{t+1} = \Pi^R(\text{Exp}_{\theta_t}(\alpha_t [\dot{\theta}_t]_x)). \quad (12)$$

In practice, we first update the velocities before integrating for the rotations.

The pseudocode for our integrator is provided in our suppl. material.

Remark 2 (Relation to known operators and integrators). $\Pi^R(\theta_t)$ corresponds to NRDF [18] while the other two projectors resemble Euclidean PoseNDF [43] updates but for velocities and accelerations. Beyond ensuring plausibility of motion, our RMF-Integrator can be thought of as a **denoising** of the trajectory and its derivatives or a **drift correction** that helps mitigate the accumulation of errors.

3.2. Learning Riemannian Motion Fields (NRMF)

We now describe how we construct $\mathcal{S}$, *i.e.*, learn all components of f_Γ indicating distances to the zero level set.

Proposition 6 (NRMF). As shown in Fig. 2, we model f_Γ hierarchically by three individual neural fields corresponding to different orders of derivatives in the equation of motion, trained to predict the distance to the closest example from dataset $\mathcal{D} = [\mathcal{D}_\theta, \dot{\mathcal{D}}_\theta, \ddot{\mathcal{D}}_\theta]$ as follows¹:

$$\Gamma^* = \begin{cases} \Phi^* = \arg \min_{\Phi} \sum_{i=1}^N \|f_\Phi^R(\theta_i) - \min_{\theta' \in \mathcal{D}_\theta} d_{\text{SO}(3)}^{N_J}(\theta_i, \theta')\| \\ \Psi^* = \arg \min_{\Psi} \sum_{i=1}^N \|f_\Psi^\omega(\dot{\theta}_i) - \min_{\dot{\theta}' \in \dot{\mathcal{D}}_\theta} d_{\text{so}(3)}^{N_J}(\dot{\theta}_i, \dot{\theta}')\| \\ \Xi^* = \arg \min_{\Xi} \sum_{i=1}^N \|f_\Xi^\omega(\ddot{\theta}_i) - \min_{\ddot{\theta}' \in \ddot{\mathcal{D}}_\theta} d_{\mathbb{R}^3}^{N_J}(\ddot{\theta}_i, \ddot{\theta}')\| \end{cases}.$$

We call f_Γ^* , learned in this way, a neural Riemannian motion field.

Each individual function is modeled by a combination of a hierarchical network and an MLP decoder, similar to NRDF [18] and Pose-NDF [43]. We obtain $\text{grad} f_\Gamma^*$ via backprop, additionally followed by egrad2rgrad for the poses themselves. Given a dataset $\mathcal{D}_\theta = \{\theta_i\}_{1 \leq i \leq N}$ of articulated poses, we obtain the dataset of transitions ($\dot{\mathcal{D}}_\theta$) and accelerations ($\ddot{\mathcal{D}}_\theta$) by computing the angular velocities and accelerations as explained earlier.

Remark 3. The first part of this equation corresponds to the recent NRDF [18], which builds upon PoseNDF [43] and correctly treats the Riemannian geometry of articulated 3D poses using quaternions. The acceleration field in the last part vaguely resembles MoManifold [11], while properly using the joints orientations instead of 3D positions. The **transition prior** in the middle row bears similarities with various works which model transitions [33, 39], inspired mostly by [33]. A majority of these works use some form of autoencoding, which is fundamentally different from our approach. To the best of our knowledge, we are the firsts to properly model all components of the kinematics, while respecting the geometry of parameters.

4. NRMF as Versatile Generative Priors

In this section, we introduce multiple downstream applications. Utilizing our NRMF as a generative prior, we can enable: 1) test-time optimization, 2) motion generation and 3) motion in-betweening, as shown in Fig. 1.

Test-time optimization. We introduce a novel runtime optimization objective to recover a plausible (likely under the

¹Note, for brevity, we dropped the conditionals.

NRMF) motion sequence $\{\mathbf{X}_t\}_{t=0}^T$ conditioned on a given sequence of 2D or 3D data observations, *e.g.* joint detections, point clouds, described as $\mathbf{O}_{0:T}$. Our multi-stage optimization starts from a proper initialization using the pose prior, by minimizing a HuMoR-like objective in stage I:

$$E_I(\mathbf{t}_r, \boldsymbol{\theta}, \boldsymbol{\beta}) = \mathcal{L}_{\text{data}} + \lambda_{\boldsymbol{\beta}} \mathcal{L}_{\boldsymbol{\beta}} + \lambda_{\boldsymbol{\theta}} \mathcal{L}_{\boldsymbol{\theta}} + \lambda_{\text{reg}} \mathcal{L}_{\text{reg}} \quad (13)$$

where $\mathcal{L}_{\boldsymbol{\beta}} = \|\boldsymbol{\beta}\|^2$ is the shape prior². $\mathcal{L}_{\boldsymbol{\theta}} = f_{\Phi}^R(\boldsymbol{\theta}_i)$ is our pose prior term predicting the distance to the nearest plausible pose³. We further employ additional regularization terms $\mathcal{L}_{\text{reg}}$ to constrain the bone lengths and smoothness of the skeleton as described in [33]. Finally, $\mathcal{L}_{\text{data}}$ denotes the task-dependent data terms and will be different when given different types of inputs: 2D, 3D joints or 3D point clouds. We provide complete details of these in our supplementary material.

While optimizing E_I yields refined estimates $\mathbf{X}$, it cannot handle motion ambiguities since the pose prior only captures 0th-order distribution. After initialization, we factor in our transition and acceleration priors as $\mathcal{L}_{\dot{\boldsymbol{\theta}}}$ and $\mathcal{L}_{\ddot{\boldsymbol{\theta}}}$:

$$\mathcal{L}_{\dot{\boldsymbol{\theta}}} := f_{\Psi}^{\omega}(\dot{\boldsymbol{\theta}}_i) \quad \mathcal{L}_{\ddot{\boldsymbol{\theta}}} := f_{\Xi}^{\omega}(\ddot{\boldsymbol{\theta}}_i) \quad (14)$$

By further combining foot-floor contact physics constraint terms into $\mathcal{L}_{\text{reg}}$ in stage II, the full objective function reads:

$$E'_{II}(\mathbf{t}_r, \boldsymbol{\theta}, \boldsymbol{\beta}) = \mathcal{L}_{\text{data}} + \lambda_{\boldsymbol{\beta}} \mathcal{L}_{\boldsymbol{\beta}} + \lambda_{\boldsymbol{\theta}} \mathcal{L}_{\boldsymbol{\theta}} + \lambda_{\text{reg}} \mathcal{L}_{\text{reg}} + \lambda_{\dot{\boldsymbol{\theta}}} \mathcal{L}_{\dot{\boldsymbol{\theta}}} + \lambda_{\ddot{\boldsymbol{\theta}}} \mathcal{L}_{\ddot{\boldsymbol{\theta}}} \quad (15)$$

Moreover, we apply the rollout within each iteration to further increase precision. We provide more details in the supplementary material.

Motion generation & in-betweening. Our motion prior can be utilized to generate plausible and diverse motions. To this end, inspired by [11, 33], we start from a random initialization of all motion components. In particular, given an initial motion state $\mathbf{X}_0$, we randomly sample a noisy transition $\dot{\boldsymbol{\theta}}_0$ and generate the next $\boldsymbol{\theta}_1$ using the transition. By repeating this procedure in a rollout manner, we obtain a noisy partial motion sequence with $\{\boldsymbol{\theta}_t, \dot{\boldsymbol{\theta}}_t\}_{t=0}^T$. The noisy acceleration $\ddot{\boldsymbol{\theta}}$ is then computed based on it to compose the complete initial noisy motion sequences $\{\tilde{\mathbf{X}}_t\}_{t=0}^T$. Next, we run our RMF-Integrator taken the sequence as input to denoise and obtain the $\{\mathbf{X}_t\}_{t=0}^T$, which is subsequently treated as the initial noisy observation for test-time optimization (Sec. 4) similar to **MoManifold** to obtain the final clean motion from the randomly composed disordered starting point. For in-betweening, given a sequence of incomplete observations, we preprocess the unobserved poses with [17, 40] and then apply the optimization process with E_I and E'_I treating

²We will later optimize for the shape parameters ($\boldsymbol{\beta}$) per sequence.

³Note that for VPoser [31] $\mathcal{L}_{\boldsymbol{\theta}} = \|\mathbf{z}\|_2^2$ where $\mathbf{z}$ is the latent code.

Table 1. Motion and shape estimation from noisy 3D observations (**motion denoising**).

Method	Pos. Err. (mm)			Motion Prop.			
	All	Legs	Vtx	Contact	Freq	Dist (mm)	Acc Err
MVAE [24]	26.1	32.3	44.6	-	1.72%	1.4	6.74
HuMoR [33]	22.7	26.1	35.5	0.97	1.18%	0.8	4.67
PhaseMP [39]	26.6	28.7	39.1	0.95	1.33%	1.2	6.47
RoHM [51]	20.9	24.1	32.4	0.97	1.12%	0.6	2.61
MDM [42]	25.4	26.7	40.1	0.95	1.32%	0.5	6.24
VPoser-t [30]	36.5	44.2	52.4	-	1.53%	0.7	7.21
PoseNDF-t [43]	24.6	27.5	32.7	-	1.42%	0.9	6.97
DPoser-t [27]	25.1	27.9	32.4	-	1.35%	0.8	7.13
NRDF-t [18]	22.8	22.4	28.1	-	1.31%	0.7	6.89
DPoser + T-PoseNDF	19.1	21.1	25.9	0.97	1.31%	0.7	3.29
DPoser + T-NRDF	17.3	17.9	23.3	0.95	1.22%	0.5	3.13
VPoser + T-PoseNDF	21.4	23.5	31.6	0.97	1.20%	0.6	3.47
VPoser + T-NRDF	19.0	20.8	27.5	0.96	1.29%	0.6	3.18
PoseNDF + T-PoseNDF	18.6	19.2	26.5	0.98	1.28%	0.6	3.24
PoseNDF + T-NRDF	17.0	18.2	25.2	0.96	1.26%	0.7	3.08
NRDF + T-PoseNDF	18.8	20.5	25.4	0.96	1.26%	0.7	3.19
NRDF + T-NRDF	16.7	17.5	22.6	0.98	1.22%	0.5	2.97
Motion-NDF	18.1	19.8	23.5	0.98	1.22%	0.5	2.64
NRMF (ours)	16.4	17.1	19.9	0.98	1.17%	0.5	2.25

the observed frames as noisy data observation inspired by [48]. Unlike [42], our method is not restricted to clean input motions, it can be applied to partial masked body infilling, and it also deduces natural transitions between masked between-frame with $f_{\Xi}^{\omega}(\dot{\boldsymbol{\theta}}_i)$, while respecting the 2nd-order dynamics via the acceleration prior $f_{\Xi}^{\omega}(\ddot{\boldsymbol{\theta}}_i)$.

5. Experiments

We now assess the performance of our method across various downstream tasks: (i) motion estimation from 3D observations, (ii) motion refinement for human mesh recovery, (iii) motion estimation from RGB(-D) in-the-wild (2D) observations and (iv) motion generation and in-betweening. Additionally, we conduct ablation studies to demonstrate the contribution of each prior. We recommend viewing the [supplementary videos](#) for qualitative results. Further dataset and experimental details are in the supplementary material.

5.1. Setup

Datasets. In this work, we train our motion prior with the AMASS [28]. For evaluation, we leverage the 3DPW [44], i3DB [44], EgoBody [50] and PROX [16] for a fair comparison with previous methods.

Baselines. As pose prior baselines, we employ VPoser [30], DPoser [27], Pose-NDF [43], and NRDF [18]. We also append *-t* (*e.g.* *DPoser-t*) to indicate that only the corresponding 0-order pose prior is used in the initialization phase, without the final optimization stage which involves higher-order priors (*i.e.* transition & acceleration prior). We also adopted the Euclidean PoseNDF [43] and geometric NRDF [18] to model transitions (T-) and accelerations (A-). We compare our method with existing human motion priors [24, 33, 39, 42, 51]. We further augment MDM [42] with infilling capabilities to tackle motion denoising. To achieve

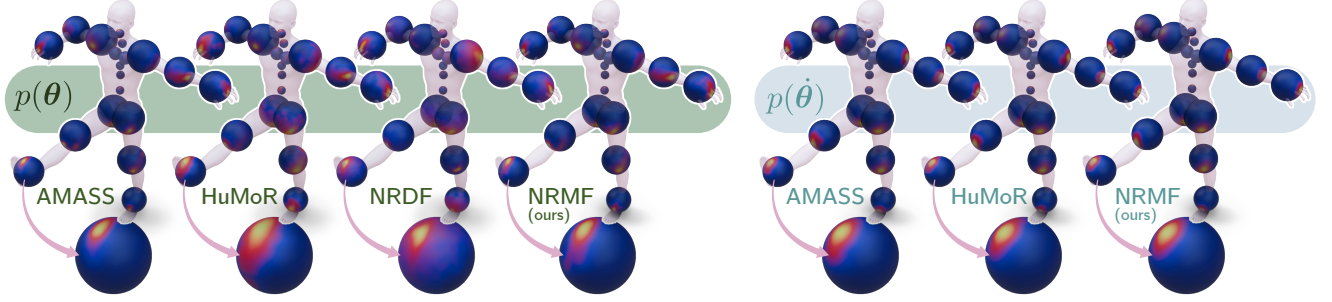


Figure 3. **Comparison of pose and transition distributions.** Each distribution is estimated from the points obtained intersecting rotated skeletal bones with a sphere centered at the joint. The pose distributions $p(\theta)$ are obtained applying θ rotations (left), while the transition distributions $p(\hat{\theta})$ applying $\hat{\theta}$ (right). Each sphere effectively displays the range of motion of each joint.

this, we train MDM with the same data split on AMASS. In ablations, we use the state-of-the-art human pose estimation (HPE) model SMPLer-X [6] as the initialization baseline and apply our motion priors to quantify the impact of each prior in optimization. For a fair comparison, all the methods we compared are trained and tested with the same data split as ours and [18, 30, 43]. For PROX dataset, we follow RoHM [51] to use off-the-shelf regressors [13, 23, 36] to obtain per-frame initialization for a fair comparison. We provide more details in our supplementary material.

Evaluation metrics. To comprehensively evaluate our method, we categorize our metrics across multiple tasks:

- **Motion estimation and refinement:** We measure global 3D joint positional error to evaluate the accuracy of estimated joint positions and meshes. We report the Mean Per Joint Position Error (MPJPE) and Procrustes-aligned Mean Per Joint Position Error (PA-MPJPE) in mm , and further distinguishing the MPJPE in the global coordinate system (GMPJPE) between visible (Vis), occluded (Occ), Leg (toes, ankles, and knees), all body parts (All), and vertices (Vtx) in mm .
- **Motion quality and plausibility:** We assess floor contact plausibility by reporting the frequency (Freq) and mean distance (Dist) of floor penetrations in cm , along with the overall contact rate (Contact). Mean per-joint acceleration $\|Acc\|$ and Acceleration Error (Acc Err) are measured in mm/s^2 to evaluate motion smoothness. To assess motion diversity and fidelity, we report Average Pairwise Distance (APD) and Fréchet Distance (FID), respectively. Specifically, we evaluate FID for pose (FID_p) and for motion (FID_m) following [17].

5.2. Evaluation & Results

As introduced in Sec. 3.2, with accurate learned neural (Riemannian) distance fields, we are able to map any noisy motion observations (e.g. noisy 3D mocap data) onto a plausible one through optimization with our motion prior while maintaining the alignment with the original observation. In this section, we leverage NRMF as a set of motion prior

Table 2. Motion and shape estimation from partial 3D observations.

Method	Positional Error (mm)					Motion Prop.			
	Vis	Occ	All	Legs	Vtx	Contact	Freq	Dist (mm)	Acc Err
MVAE [24]	23.9	191.5	105.6	188.6	114.5	-	3.15%	3.0	6.91
HuMoR (VPoser) [33]	14.6	120.3	67.9	118.1	76.8	0.89	3.31%	2.6	4.78
PhaseMP [39]	38.7	177.5	110.1	175.5	118.4	0.89	3.28%	3.1	6.29
RoHM [51]	12.9	88.4	51.3	86.7	58.4	0.91	3.08%	1.6	2.43
MDM [42]	27.8	173.6	101.2	170.6	112.7	0.88	3.18%	2.6	6.23
VPoser-t [30]	6.7	207.6	108.8	205.4	118.3	-	16.77%	22.8	7.18
PoseNDF-t [43]	5.9	204.1	104.6	200.9	113.7	-	15.49%	23.1	7.01
DPoser-t [27]	4.9	211.2	109.4	209.3	117.5	-	15.65%	22.5	6.95
NRDF-t [18]	4.2	194.5	100.2	192.7	109.1	-	15.23%	20.9	6.73
DPoser + T-PoseNDF	11.2	111.5	61.8	109.2	72.4	0.86	3.67%	3.4	3.36
DPoser + T-NRDF	11.0	102.8	56.7	100.8	66.3	0.87	3.12%	2.9	3.19
VPoser + T-PoseNDF	13.4	108.7	60.5	106.9	71.7	0.88	3.29%	3.1	3.41
VPoser + T-NRDF	12.0	102.3	57.8	100.2	67.9	0.87	3.19%	2.2	3.13
PoseNDF + T-PoseNDF	11.5	100.2	56.3	97.6	65.8	0.88	3.17%	2.8	3.35
PoseNDF + T-NRDF	11.6	97.3	54.9	95.4	64.2	0.88	3.23%	2.6	3.11
NRDF + T-PoseNDF	11.1	108.1	59.1	105.8	69.5	0.87	3.16%	2.5	3.27
NRDF + T-NRDF	10.5	87.1	48.5	84.8	58.6	0.89	3.17%	2.1	3.01
Motion-NDF	11.6	99.5	55.2	97.3	63.9	0.89	3.17%	2.6	2.72
NRMF (ours)	9.7	83.8	46.3	81.6	56.7	0.89	3.05%	1.9	2.31

terms in an optimization-based pipeline for various downstream tasks from different observation.

Motion estimation from 3D observations. For a fair comparison, we follow [33] to extract (i) noisy and (ii) partial 3D keypoints from the AMASS dataset and sample 3s (90 frames). We conduct both the motion estimation accuracy and plausibility evaluation of (i) motion denoising and (ii) fitting to partial 3D observation.

To perform motion denoising, we align the recovered human body well with the observations and preserve the plausibility of the motion. To test this, we sample **noisy keypoints** from the AMASS dataset with a Gaussian noise of $4cm$. As shown in Tab. 1, we compare NRMF with state-of-the-art pose [18, 27, 30, 43] and motion priors [24, 33, 39, 42, 51]. It can be seen that our method consistently outperforms the SotA. Previous methods typically only focus on 0-order or 1st-order cues either neglecting the high-order information [18, 27, 30, 43] thus producing unrealistic, oversmooth, and implausible motion transitions, or causing error accumulation over time (drift) [33]. In contrast, our method incorporates priors for all essential components of dynamics to fill in the plausibility gap while maintaining the fidelity of each frame, which is supported by the comparison of the output distribution in Fig. 3.

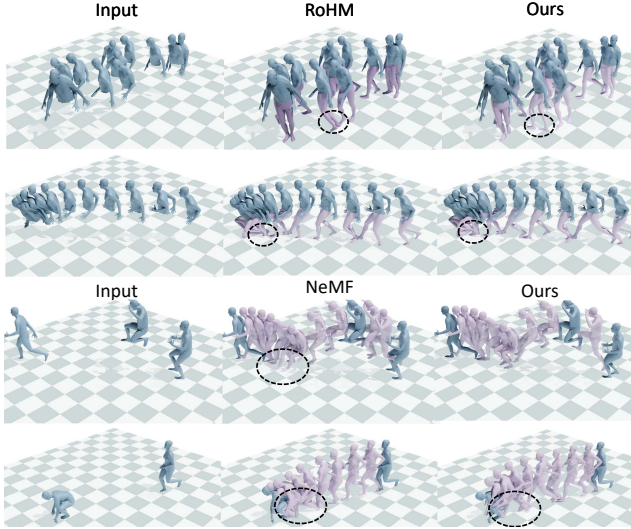


Figure 4. Qualitative comparison with state-of-the-art methods [17, 51] on motion estimation from **noisy partial** observation (top two) and motion in-between (bottom two) where the **purple body parts** represent the recovered motion.

Moreover, as shown in Tab. 1, methods that properly treat the geometry of articulations tend to perform better, while PoseNDF based (Euclidean) projections fail under larger noise. MDM [42] and RoHM [51] present comparable results for contact but yield low estimation accuracy.

To test motion estimation under occlusions (*i.e.* **partial observation**) we sample the AMASS dataset with a standard mask height of $0.9m$ following [33] to yield partial 3D observations. Tab. 2 reveals that while applying only a pose prior term (*i.e.* - t) gives lower errors for visible body parts, the model suffers from over-smoothing and higher reconstruction and acceleration errors when deducing invisible body parts over time, due to the inherent negligence of transitions. In contrast, our NRMF consistently outperforms both baselines and prior state-of-the-art methods. This demonstrates the effectiveness of our approach in modeling motion dynamics: by jointly incorporating pose, transition, and acceleration priors, NRMF achieves improved pose accuracy while producing more temporally coherent and physically plausible motions, and successfully mitigates error accumulation. Qualitative results in Fig. 4 and Supp. Mat. further support these observations.

Motion refinement for human mesh recovery. Our motion prior can also be used with regression-based human pose estimation methods as an efficient optimization-based refinement module for mesh recovery from videos. Tab. 3 compares our method with the SotA baselines of [18, 27, 30, 51] using SMPLer-X [6] and [7] as initialization. Notably, while optimization improves pixel alignment, it does not guarantee plausible motion due to the lack of constraints and biased optimization without appropriate pri-

Table 3. Fitting to 2D observations on 3DPW. We compare the refinement results after optimizing SMPLer-X [6] with different prior terms. MPJPE and MPVPE are in millimeters.

Method	MPJPE (mm)	MPVPE (mm)	Acc Err (m/s^2)	Trans Err ($\times 10^{-3}$)
SMPLer-X [6]	82.65	94.23	23.71	31.63
+ No prior	84.67	96.82	26.75	34.54
+ VPoser [30]	79.98	91.53	25.82	30.58
+ DPoser [27]	75.45	87.04	27.54	28.69
+ PoseNDF [43]	73.49	84.61	25.12	29.77
+ NRDF [18]	71.88	83.23	24.31	26.38
+ T-NRDF	66.98	76.94	9.89	7.98
+ A-NRDF	70.88	81.92	6.73	11.87
+ RoHM [51]	69.78	79.72	9.13	12.37
+ NRMF (full)	66.13	75.61	6.52	5.67

ors. Specifically, only applying a pose prior (0-order) usually leads to large errors in transition and acceleration due to the absence of higher-order motion constraints. Notably, the ablations highlight the significance of each prior term separately. By integrating acceleration and transition, we improve the Acc Err and Trans Err significantly. Applying our **NRMF(full)** model for refinement will boost the per-frame mesh reconstruction accuracy while preserving plausible and realistic motion transitions.

Motion estimation from RGB(-D) in-the-wild observations. As discussed in Sec. 3.2, our method is readily applicable to real-world cases, enabling robust motion recovery from various modalities including in-the-wild RGB(-D) videos, *i.e.* partial and noisy 2D & 3D observations. We conduct our evaluation on i3DB (RGB), EgoBody (RGB) and PROX (RGB & RGB-D) datasets. Specifically, we leverage the off-the-shelf 2D pose estimation [7], human mask segmentation [8], along with a plane detection [25] model as observations during optimization following [33]. As shown in Tabs. 5 to 7, conventional pose priors perform reasonably well on visible joints but struggle in occluded regions. Relying solely on 0-order pose information leads them to overfit visible joints, producing temporally inconsistent and physically implausible reconstructions. In contrast, NRMF incorporates richer dynamics through transition and acceleration priors, yielding substantially lower position and acceleration errors on both visible and occluded joints compared to existing motion [33, 51] and pose [18, 27, 43] priors. Qualitative comparisons in Fig. 5 highlight NRMF’s superior motion recovery, while Fig. 3 shows that the pose and transition distributions learned by NRMF better match the ground-truth distribution.

Motion Generation and In-betweening. For motion generation, we follow the protocol of [33], sampling 50 motion sequences (2sec each) for each initial state. As shown in Tab. 4, VPoser achieves the lowest FID and APD scores, but produces less diverse and expressive motions. In contrast, DPoser yields high diversity (APD) at the cost of low plausibility (FID). NRDF strikes a balance between fidelity and diversity. Among the motion priors, our method achieves the best overall FID for motion while maintaining competi-

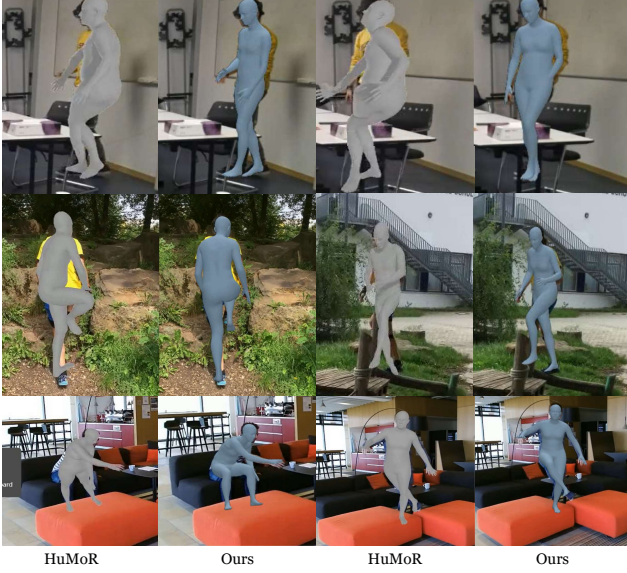


Figure 5. Qualitative comparison with state-of-the-art methods for in-the-wild motion estimation.

Table 4. Quantitative analysis of pose and motion generation.

Method	APD $\uparrow$ (cm)	FID _p $\downarrow$	FID _m $\downarrow$	$\ \text{Acc}\ $ $\downarrow$
GMM [5]	16.28	0.435	-	-
GAN-S [12]	15.68	0.201	-	-
VPoser [30]	10.75	0.113	-	-
Pose-NDF [43]	18.54	1.124	-	-
DPoser [27]	36.25	2.344	-	-
NRDF [18]	24.87	0.706	-	-
MVAE [24]	85.26	1.475	10.571	8.15
HuMoR [33]	98.47	0.278	9.147	4.96
Motion-NDF	93.46	0.894	8.351	-
NRMF (ours)	96.37	0.718	5.317	3.18

Table 5. Motion and shape from RGB sequences on i3DB.

Method	Global Joint Error (cm)				Root-Aligned Joint Error (cm)			
	Vis	Occ	All	Legs	Vis	Occ	All	Legs
VIBE [20]	90.05	192.55	116.46	121.61	12.06	23.78	15.08	21.65
MVAE [24]	37.54	50.63	40.91	44.42	16.00	28.32	19.17	26.63
HuMoR [33]	26.37	34.79	28.22	31.41	12.02	21.73	14.51	20.84
VPoser-t	28.33	40.97	31.59	35.16	12.77	26.48	16.31	25.64
PoseNDF-t	29.17	42.36	32.64	36.28	13.19	27.12	17.03	26.14
NRDF-t	27.85	39.48	30.82	34.13	12.53	25.88	15.92	24.89
DPoser-t	28.12	40.23	31.32	34.74	12.63	26.05	16.17	25.08
NRMF (ours)	22.35	25.12	24.67	23.68	10.27	17.51	11.97	16.73

tive APD, indicating that NRMF generates realistic and diverse motions. Additionally, NRMF produces the lowest mean acceleration norm, reflecting stable motion with reduced jitter. For in-betweening, we follow a similar protocol to [17] and [11], focusing on sparse keyframe supervision and infilling refinement. Specifically, we provide a sparse set of keyframe observations corresponding either to the two fixed endpoints, to 10% and 20% of the total frames, or to a 30% scheme divided across the start, middle, and end of the sequence similar to the protocol in [42]. We compare our method against [17, 33] and SLERP. Our

Table 6. Motion estimation from RGB(-D) input on PROX.

Method	RGB-D				RGB			
	$\ \text{Acc}\ $	Dist	FID _m	Freq	$\ \text{Acc}\ $	Dist	FID _m	Freq
HuMoR [33]	1.88	35.67	11.65	6.13%	2.35	41.92	12.47	10.52%
PhaseMP [39]	-	-	-	-	1.80	46.96	-	-
RoHM [51]	1.79	6.67	12.16	3.21%	2.19	14.31	13.65	4.65%
VPoser-t	3.45	53.49	21.89	10.66%	3.24	54.96	19.64	13.95%
PoseNDF-t	2.84	48.77	16.45	11.35%	2.93	50.13	16.87	13.46%
NRDF-t	2.67	45.91	14.58	9.97%	2.77	46.78	14.32	12.98%
DPoser-t	3.32	49.45	17.32	10.89%	3.08	51.35	16.98	12.21%
NRMF (Ours)	1.48	5.78	8.95	2.78%	1.73	16.21	14.13	5.12%

Table 7. Motion estimation results on EgoBody dataset.

Method	Positional Error (cm)				Motion Prop.			
	Vis	Occ	Legs	All	Vtx	Freq	Dist	Acc
HuMoR [33]	75.3	129.5	124.6	108.85	105.94	7.31%	12.65	5.36
RoHM [51]	60.1	123.6	116.9	98.33	93.81	6.08%	2.15	2.97
VPoser-t [30]	65.5	139.4	129.7	110.30	116.14	21.77%	17.34	7.18
PoseNDF-t [43]	62.1	135.6	131.4	106.41	99.05	19.49%	16.98	7.01
DPoser-t [27]	68.1	141.3	138.6	112.14	118.47	22.65%	18.35	6.95
NRDF-t [18]	60.3	131.6	133.6	103.03	109.22	19.23%	15.26	6.73
NRMF (Ours)	57.5	112.6	109.8	92.10	90.06	4.78%	3.78	1.58

method explicitly leverages high-order motion dynamics, enabling robust interpolation even under extremely sparse constraints. Fig. 4 illustrates a qualitative comparison in the two-endpoint setting, where our method produces more natural and plausible interpolations. Additional qualitative results, as well as quantitative evaluations across different sparsity levels, are provided in the Supp. Mat..

6. Conclusion

In summary, our work presented Neural Riemannian Motion Fields (NRMF), a novel framework for modeling 3D human motion while rigorously respecting the underlying geometry of articulated motions. By decomposing motion into its constituent components—pose, transition, and acceleration—and representing these as the zero-level sets of dedicated neural distance fields, NRMF acts as a strong inductive bias bridging the gap between learning and physical modeling of human motion. Our proposed projection and geometric integration algorithms help ensuring temporal consistency and robustness against noise and enable versatile applications ranging from motion denoising and in-betweening to test-time optimization for mesh recovery. Extensive experiments on diverse benchmarks consistently confirm that NRMF outperforms existing methods, delivering enhanced accuracy and more lifelike trajectories.

Limitation & future work. Our iterative updates can incur significant runtime leading to minutes for processing a sequence of < 1 min. Moreover, our projected-integrators lack rigorous theoretical understanding, preventing us from certifying their optimality. As such, our work leaves ample room for future study. *Learning-to-optimize* can be an effective future direction, while principled sampling algorithms such as *Riemannian Langevin MCMC* [4] can be employed as integrators.

Acknowledgments

T. Birdal, S. Zafeiriou, S. Foti and Z. Yu acknowledge support from the Engineering and Physical Sciences Research Council [grant EP/X011364/1]. T. Birdal was supported by a UKRI Future Leaders Fellowship [grant number MR/Y018818/1] as well as a Royal Society Research Grant RG/R1/241402.

References

- [1] Samaneh Azadi, Akbar Shah, Thomas Hayes, Devi Parikh, and Sonal Gupta. Make-an-animation: Large-scale text-conditional 3d human motion generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15039–15048, 2023. 1
- [2] Fabien Baradel, Romain Brégier, Thibault Groueix, Philippe Weinzaepfel, Yannis Kalantidis, and Grégory Rogez. Posebert: A generic transformer module for temporal 3d human modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(11):12798–12815, 2022. 2
- [3] Lennart Bastian, Mohammad Rashed, Nassir Navab, and Tolga Birdal. Forecasting continuous non-conservative dynamical systems in so (3). In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025. 3
- [4] Tolga Birdal and Umut Simsekli. Probabilistic permutation synchronization using the riemannian structure of the birkhoff polytope. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11105–11116, 2019. 8
- [5] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J Black. Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part V 14*, pages 561–578. Springer, 2016. 8
- [6] Zhongang Cai, Wanqi Yin, Ailing Zeng, Chen Wei, Qingping Sun, Wang Yanjun, Hui En Pang, Haiyi Mei, Mingyuan Zhang, Lei Zhang, Chen Change Loy, Lei Yang, and Ziwei Liu. SMPLer-X: Scaling up expressive human pose and shape estimation. In *Advances in Neural Information Processing Systems*, 2023. 2, 6, 7
- [7] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019. 7
- [8] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017. 7
- [9] Julian Chibane, Gerard Pons-Moll, et al. Neural unsigned distance fields for implicit function learning. *Advances in Neural Information Processing Systems*, 33:21638–21652, 2020. 2
- [10] Hai Ci, Mingdong Wu, Wentao Zhu, Xiaoxuan Ma, Hao Dong, Fangwei Zhong, and Yizhou Wang. Gfpose: Learning 3d human pose prior with gradient fields. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4800–4810, 2023. 2
- [11] Ziqiang Dang, Tianxing Fan, Boming Zhao, Xujie Shen, Lei Wang, Guofeng Zhang, and Zhaopeng Cui. Momanifold: Learning to measure 3d human motion via decoupled joint acceleration manifolds. In *35th British Machine Vision Conference 2024, BMVC 2024, Glasgow, UK, November 25–28, 2024. BMVA*, 2024. 2, 4, 5, 8
- [12] Andrey Davydov, Anastasia Remizova, Victor Constantin, Sina Honari, Mathieu Salzmann, and Pascal Fua. Adversarial parametric pose prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10997–11005, 2022. 8
- [13] Yao Feng, Vasileios Choutas, Timo Bolkart, Dimitrios Tzionas, and Michael J Black. Collaborative regression of expressive bodies using moderation. In *2021 International Conference on 3D Vision (3DV)*, pages 792–804. IEEE, 2021. 6
- [14] Shubham Goel, Georgios Pavlakos, Jathushan Rajasegaran, Angjoo Kanazawa, and Jitendra Malik. Humans in 4D: Reconstructing and tracking humans with transformers. In *ICCV*, 2023. 2
- [15] Adrian Haarbach, Tolga Birdal, and Slobodan Ilic. Survey of higher order rigid body motion interpolation methods for keyframe animation and continuous-time trajectory estimation. In *2018 International Conference on 3D Vision (3DV)*, pages 381–389. IEEE, 2018. 3
- [16] Mohamed Hassan, Vasileios Choutas, Dimitrios Tzionas, and Michael J. Black. Resolving 3D human pose ambiguities with 3D scene constraints. In *International Conference on Computer Vision*, pages 2282–2292, 2019. 2, 5
- [17] Chengan He, Jun Saito, James Zachary, Holly Rushmeier, and Yi Zhou. Nemf: Neural motion fields for kinematic animation. *Advances in Neural Information Processing Systems*, 35:4244–4256, 2022. 3, 5, 6, 7, 8
- [18] Yannan He, Garvita Tiwari, Tolga Birdal, Jan Eric Lenssen, and Gerard Pons-Moll. Nrdf: Neural riemannian distance fields for learning articulated pose priors. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 1, 2, 4, 5, 6, 7, 8
- [19] Yiheng Huang, Hui Yang, Chuanchen Luo, Yuxi Wang, Shibiao Xu, Zhaoxiang Zhang, Man Zhang, and Junran Peng. Stablemofusion: Towards robust and efficient diffusion-based motion generation framework. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 224–232, 2024. 2
- [20] Muhammed Kocabas, Nikos Athanasiou, and Michael J. Black. Vibe: Video inference for human body pose and shape estimation. In *The IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 8
- [21] Muhammed Kocabas, Chun-Hao P Huang, Otmar Hilliges, and Michael J Black. Pare: Part attention regressor for 3d human body estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11127–11137, 2021. 2
- [22] Nilesh Kulkarni, Davis Rempe, Kyle Genova, Abhijit Kundu, Justin Johnson, David Fouhey, and Leonidas Guibas.

- Nifty: Neural object interaction fields for guided human motion synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 947–957, 2024. 2
- [23] Zhihao Li, Jianzhuang Liu, Zhensong Zhang, Songcen Xu, and Youliang Yan. Cliff: Carrying location information in full frames into human pose and shape estimation. In *European Conference on Computer Vision*, pages 590–606. Springer, 2022. 6
- [24] Hung Yu Ling, Fabio Zinno, George Cheng, and Michiel van de Panne. Character controllers using motion vaes. In *ACM Transactions on Graphics (Proceedings of ACM SIGGRAPH)*. ACM, 2020. 2, 5, 6, 8
- [25] Chen Liu, Kihwan Kim, Jinwei Gu, Yasutaka Furukawa, and Jan Kautz. Planercnn: 3d plane detection and reconstruction from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4450–4459, 2019. 7
- [26] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16, 2015. 2, 3
- [27] Junzhe Lu, Jing Lin, Hongkun Dou, Yulun Zhang, Yue Deng, and Haoqian Wang. Dposer: Diffusion model as robust 3d human pose prior. *arXiv preprint arXiv:2312.05541*, 2023. 1, 2, 5, 6, 7, 8
- [28] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. AMASS: Archive of motion capture as surface shapes. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5441–5450, 2019. 2, 5
- [29] Aron Monszpart, Paul Guerrero, Duygu Ceylan, Ersin Yumer, and Niloy J. Mitra. iMapper: Interaction-guided scene mapping from monocular videos. *ACM SIGGRAPH*, 2019. 2
- [30] 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. 2, 5, 6, 7, 8
- [31] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3D hands, face, and body from a single image. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 10975–10985, 2019. 5
- [32] Mathis Petrovich, Michael J Black, and Gül Varol. Tmr: Text-to-motion retrieval using contrastive 3d human motion synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9488–9497, 2023. 1
- [33] Davis Rempe, Tolga Birdal, Aaron Hertzmann, Jimei Yang, Srinath Sridhar, and Leonidas J. Guibas. Humor: 3d human motion model for robust pose estimation. In *International Conference on Computer Vision (ICCV)*, 2021. 1, 2, 4, 5, 6, 7, 8
- [34] Javier Romero, Dimitris Tzionas, and Michael J Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics*, 36(6), 2017. 3
- [35] Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 36(6), 2017. 2, 3
- [36] István Sárádi, Timm Linder, Kai Oliver Arras, and Bastian Leibe. Metrabs: metric-scale truncation-robust heatmaps for absolute 3d human pose estimation. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 3(1):16–30, 2020. 6
- [37] Yonatan Shafir, Guy Tevet, Roy Kapon, and Amit H Bermano. Human motion diffusion as a generative prior. *arXiv preprint arXiv:2303.01418*, 2023. 2
- [38] Yoni Shafir, Guy Tevet, Roy Kapon, and Amit Haim Bermano. Human motion diffusion as a generative prior. In *The Twelfth International Conference on Learning Representations*, 2024. 1, 2
- [39] Mingyi Shi, Sebastian Starke, Yuting Ye, Taku Komura, and Jungdam Won. Phasemp: Robust 3d pose estimation via phase-conditioned human motion prior. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14725–14737, 2023. 1, 2, 4, 5, 6, 8
- [40] Ken Shoemake. Animating rotation with quaternion curves. In *Proceedings of the 12th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH)*, pages 245–254. ACM, 1985. 5
- [41] Yu Sun, Qian Bao, Wu Liu, Tao Mei, and Michael J. Black. TRACE: 5D Temporal Regression of Avatars with Dynamic Cameras in 3D Environments. In *CVPR*, 2023. 2
- [42] Guy Tevet, Sigal Raab, Brian Gordon, Yoni Shafir, Daniel Cohen-or, and Amit Haim Bermano. Human motion diffusion model. In *The Eleventh International Conference on Learning Representations*, 2023. 1, 2, 5, 6, 7, 8
- [43] Garvita Tiwari, Dimitrije Antic, Jan Eric Lenssen, Nikolaos Sarafianos, Tony Tung, and Gerard Pons-Moll. Pose-ndf: Modeling human pose manifolds with neural distance fields. In *European Conference on Computer Vision (ECCV)*, 2022. 1, 2, 4, 5, 6, 7, 8
- [44] Timo von Marcard, Roberto Henschel, Michael Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *European Conference on Computer Vision (ECCV)*, 2018. 2, 5
- [45] Xiangyu Xu, Lijuan Liu, and Shuicheng Yan. Smpler: Taming transformers for monocular 3d human shape and pose estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(5):3275–3289, 2023. 2
- [46] Sijie Yan, Zhizhong Li, Yuanjun Xiong, Huahan Yan, and Dahua Lin. Convolutional sequence generation for skeleton-based action synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4394–4402, 2019. 2
- [47] Hua Yu, Weiming Liu, Jiapeng Bai, Xu Gui, Yaqing Hou, YewSoon Ong, and Qiang Zhang. Towards efficient and diverse generative model for unconditional human motion syn-

- thesis. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 2535–2544, 2024. [2](#)
- [48] Zhengdi Yu, Stefanos Zafeiriou, and Tolga Birdal. Dynhamr: Recovering 4d interacting hand motion from a dynamic camera. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. [5](#)
- [49] Hongwen Zhang, Yating Tian, Yuxiang Zhang, Mengcheng Li, Liang An, Zhenan Sun, and Yebin Liu. Pymaf-x: Towards well-aligned full-body model regression from monocular images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(10):12287–12303, 2023. [2](#)
- [50] Siwei Zhang, Qianli Ma, Yan Zhang, Zhiyin Qian, Taein Kwon, Marc Pollefeys, Federica Bogo, and Siyu Tang. Ego-body: Human body shape and motion of interacting people from head-mounted devices. In *European conference on computer vision*, pages 180–200. Springer, 2022. [2](#), [5](#)
- [51] Siwei Zhang, Bharat Lal Bhatnagar, Yuanlu Xu, Alexander Winkler, Petr Kadlecek, Siyu Tang, and Federica Bogo. Rohm: Robust human motion reconstruction via diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14606–14617, 2024. [2](#), [5](#), [6](#), [7](#), [8](#)