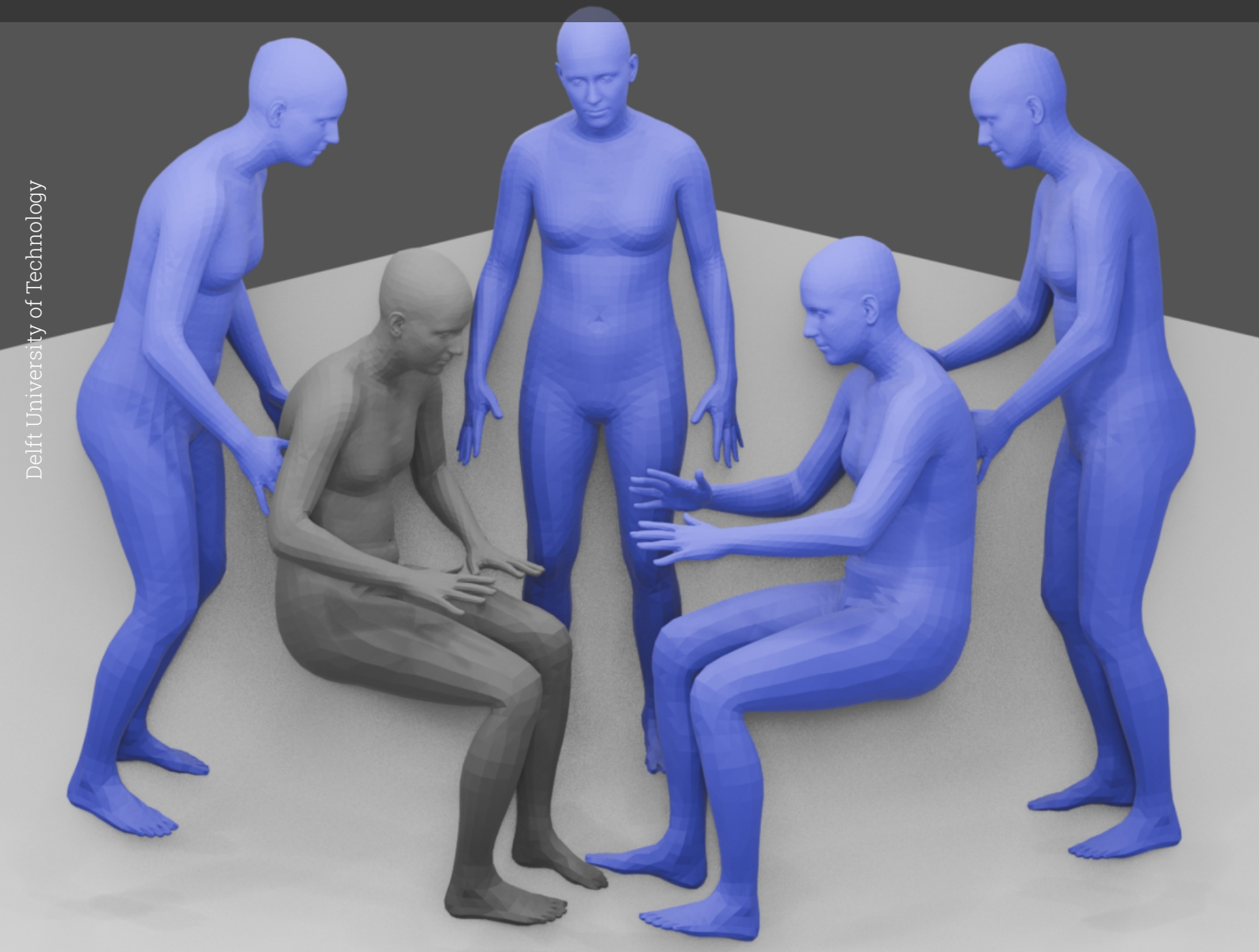
Human Interaction Pose Prior via Conditional Normalizing Flows

MSc Computer Science

Dmytro Maksymchuk



Delft University of Technology

TU Delft

Human Interaction Pose Prior via Conditional Normalizing Flows

by

Dmytro Maksymchuk

Student Number:	5500605
Thesis advisor:	Dr. Petr Kellnhofer
Graduation committee member:	Dr. L. Cavalcante Siebert
Research Group:	Computer Graphics & Visualization
Project Duration:	November 2025 - July 2026
Faculty:	Electrical Engineering, Mathematics and Computer Science

Style: TU Delft Report Style, modifications by Daan Zwaneveld

TU Delft

Abstract

Close physical interactions such as hugs, handshakes, and collaborative tasks are commonplace in everyday life, yet 3D reconstruction methods still struggle to capture them accurately. Most approaches treat each person independently, producing reconstructions where bodies unrealistically intersect, drift apart, or collapse to averaged poses, particularly under occlusion. Furthermore, without an explicit probability model over interactions, there is no principled way to assess whether a reconstruction is plausible in the first place.

We address this by learning an explicit probabilistic prior over close human-human interactions. Using a conditional normalizing flow, we model how one person's pose and position relate to the other's, yielding a tractable likelihood that can be evaluated for any pair of poses at low computational cost, unlike diffusion models, which require iterative sampling chains, or VAEs, whose likelihoods are lower bounds rather than exact densities. Our model serves three purposes: it generates diverse and realistic reaction poses given an observed person, scores the plausibility of arbitrary interaction pairs, and acts as a differentiable regularizer in image-to-mesh optimization to improve two-person 3D reconstruction from a single RGB image. Experiments demonstrate strong performance in generating diverse reaction poses and in discriminating plausible from implausible interaction pairs. Used as an optimization prior, our method improves two-person image-to-mesh reconstruction, with the largest gains in the relative placement of the two bodies.

Acknowledgments

I would first like to thank my supervisor, Dr. Petr Kellnhofer, for his expertise and continuous guidance throughout this research. He was always ready to discuss my progress and help steer the work in the right direction, and I am very grateful for his support.
I would also like to thank Dr. Luciano Cavalcante Siebert for serving on my thesis committee and for taking the time to review my work.
Finally, I want to thank my family, friends, and my girlfriend for their support throughout this thesis. Their encouragement kept me going, especially during the more difficult stretches of this process.

Dmytro Maksymchuk
Delft, July 2026

Contents

Abstract i

Acknowledgments ii

1 Introduction 1

2 Related Work 3
2.1 Human Pose Representations and Pose Priors . . . 3
2.2 Human-Human Interaction Modeling . . . 4
2.3 Conditional Generative Models and Normalizing Flows . . . 4

3 Method 6
3.1 Two-Person Pose Representation . . . 6
3.2 Model Definition . . . 7
3.3 Architecture . . . 7
3.4 Training Objective . . . 8
3.5 Implementation Details . . . 8

4 Experiments 10
4.1 Conditional Generation . . . 11
4.2 Plausibility Scoring . . . 13
4.3 Prior-Guided Image-to-Mesh Optimization . . . 18
4.3.1 Ablation: Prior Without a Strong Initialization . . . 20

5 Discussion 22
5.1 Generative Quality and Multimodality . . . 22
5.2 Plausibility Scoring . . . 23
5.3 Application to Image-to-Mesh Optimization . . . 24
5.4 Cross-Dataset Generalization . . . 24
5.5 Limitations . . . 26

6 Reflection 27
6.1 Expressiveness vs Tractability in Broader CS and ML Field . . . 27
6.2 Real-World Impact and Responsible Use . . . 28
6.2.1 Research Prototype and Path to Deployment . . . 28
6.2.2 Applications and Beneficiaries . . . 28
6.2.3 Risks and Stakeholder Considerations . . . 28

7 Conclusion 30

References 31

A Supplementary Material 34
A.1 Use of AI tools . . . 34
A.2 Additional Qualitative Results . . . 35
A.2.1 Learned Likelihood in Noisy and Mismatched Cases . . . 35
A.2.2 Learned Likelihood Under Spatial Transformation . . . 36

1

Introduction

Many everyday activities involve close physical interactions between people, such as greetings, sports, or collaborative work. Accurately capturing and modeling such human–human interactions (HHI) in 3D is important for a range of applications. In AR/VR, applications such as VR telepresence and social avatar animation require not just per-person pose estimates but also a coherent model of how two people are positioned and oriented relative to each other, since independently reconstructed bodies tend to overlap or drift, breaking the plausibility of the shared virtual space [13]. In motion capture for film, games, or sports analysis, closely interacting people cause severe inter-person occlusions and ambiguities, so high-fidelity capture requires jointly reconstructing all bodies and explicitly modeling mutual contacts and relative placement, as emphasized by close-interaction MoCap and reconstruction systems [9, 31].

Despite recent progress in 3D human pose estimation, most work focuses on single-person reconstruction [2, 14, 28]. Multi-person methods typically extend single-person pipelines to detect and fit multiple bodies in an image or video [35, 34, 7], but they usually treat people independently, focusing on per-person accuracy and coarse relative depth rather than the fine-grained structure of close interactions. As a result, reconstructed bodies can intersect unrealistically, drift apart, or converge to averaged poses under occlusion.

In contrast to these person-centric approaches, a separate line of work addresses these limitations by introducing learned priors specifically over two-person interactions. DualHuman [8] introduces a ”reaction prior” that maps the pose distribution of one person, encoded in a motion VAE latent space, to a predicted distribution over the interacting partner’s pose, and uses this mapping to regularize a two-person motion capture pipeline from monocular video. However, this prior models a single distribution per input and can therefore only produce a single best-fit reaction pose. In practice, this means the model cannot represent multiple plausible responses to the same reference pose. For example, a person being embraced could realistically step forward, lean back, or raise their arms, yet the method collapses to one averaged prediction. In contrast, BUDDI [26] learns a diffusion-based proxemics prior over the joint SMPL-X parameters of both people, conditioned on an initial estimate from the regression-based predictor BEV [35], capturing realistic spatial arrangements and pose combinations, and uses it as a data-driven prior in single-image reconstruction. Neither approach provides an explicit, tractable probability density over an interaction representation that can be used both for sampling and for plausibility scoring.

In several downstream applications, including pose hypothesis ranking in monocular reconstruction, anomaly detection in surveillance or sports analysis, and optimization-based image-to-mesh fitting, it is necessary not just to generate plausible interactions but to quantitatively score them. Existing priors based on VAEs, diffusion models, or GANs provide implicit or approximate likelihoods, or rely on surrogate scores such as reconstruction error, making none of these use cases straightforward to address. In addition, many methods tend to average over multimodal pose spaces. Given a pose of the reference person, they produce a single, often smoothed pose for the target person, even though multiple

different reaction configurations may be plausible. A model that both preserves this multimodality and produces explicit densities would be a better tool for HHI reasoning.

In this work, we propose a conditional normalizing flow (CNF) as an explicit interaction prior for close human–human interaction, jointly modeling one person's pose and relative placement with respect to their partner, conditioned on the other person's pose. We represent each person's body pose using VPoser [27], a variational autoencoder trained on SMPL-X body poses that encodes each pose into a compact 32-dimensional approximately Gaussian latent code, ensuring any decoded pose is anatomically plausible. We define a conditional normalizing flow over the latent code of the target person and the relative translation and rotation between bodies, conditioned on the latent code of the reference person. This results in a conditional probability density that can be easily evaluated for any interaction and used to draw diverse samples of the target person's reaction given a fixed reference person. By operating in a compact latent space that already encodes anatomically plausible single-person poses, the flow focuses on modeling interaction structure and multimodality rather than the validity of joint angles.

We show that this conditional flow can be used in two ways: as a generative model for sampling diverse, realistic reactions, and as a plausibility metric for optimization. Given a conditioning pose, we sample multiple candidate reactions and evaluate them quantitatively against held-out data. Second, we use the learned density as a prior in two scenarios: scoring the plausibility of real versus mismatched or noisy pose pairs, and measuring how well the model separates plausible from implausible interactions. We further demonstrate SMPLify-style [2] image-to-mesh optimization, where the explicit log-density term guides the refinement of an initial two-person reconstruction estimate towards more plausible interactions, similar to the prior-guided optimization used in BUDDI [26]. Together, these results show that an explicit, conditional density over a position-aware latent representation can serve as both a generative model and an interaction prior for reasoning about close human–human interactions.

In summary, this thesis makes the following contributions:

- **A conditional normalizing flow prior for close HHI.** We apply a conditional normalizing flow to the problem of two-person interaction modeling, jointly capturing one person's pose and relative placement given their partner, providing a tractable and exact likelihood.
- **Multimodal conditional generation.** We show that the flow captures the multimodal structure of reaction poses, producing qualitatively distinct samples for the same conditioning pose.
- **Interaction plausibility scoring.** We demonstrate that the learned density serves as a reliable plausibility score, successfully discriminating real interactions from mismatched and noise-corrupted pairs.
- **Prior-guided image-to-mesh optimization.** We apply the model in an optimization setting, using the learned likelihood as a differentiable regularizer in a SMPLify-style pipeline, and show that it improves two-person relative placement accuracy over both the BEV baseline and keypoint-only refinement.

2

Related Work

We cover three areas of related work. We first review human body representations and pose priors, which form the foundation that our method builds on. We then discuss existing approaches to human-human interaction modeling, stating our contribution relative to prior work. Finally, we introduce conditional normalizing flows, the generative framework at the core of our model.

2.1. Human Pose Representations and Pose Priors

A human body in 3D can be represented at varying levels of abstraction. One of the simplest representations is a skeleton: a set of joints connected by bones, where pose is described by joint angles or positions. While skeletons are lightweight and widely used in pose estimation, they do not capture body shape or surface geometry, limiting their usefulness for tasks that require modeling physical contact or body volume.

A common way to obtain richer representations is to parameterize a body model instead of only using skeleton joints. Parametric models such as SMPL [22] and SMPL-X [27] describe a human by pose and shape parameters that manipulate a skinned vertex-based mesh template. This enables realistic deformations of the mesh under joint movement and gives access to both skeleton joints and surface vertices. These models are fit to thousands of 3D body scans using PCA-based blendshapes and linear regression, and are widely used in recent work on monocular human capture. In this work we also adopt SMPL-X as the underlying body model so that the interaction prior operates in a space that already captures realistic full-body geometry rather than just joint locations.

Fitting such parametric models to images is an under-constrained problem: many pose and shape configurations can explain the same 2D observations, especially under occlusion. Pose priors address this by constraining the solution to plausible regions of body-pose space, either as a regularizer during optimization or as a learned distribution for sampling. Early pose priors for such body models typically relied on statistical assumptions over joint rotations. For example, SMPLify [2] optimizes SMPL pose and shape to fit 2D keypoints while regularizing with a Gaussian pose prior and a learned mixture-model prior over joint rotations obtained from mocap data, combined with a Gaussian shape prior and a differentiable interpenetration term. These priors help prevent anatomically impossible poses and self-intersections, but they capture unimodal statistics and do not explicitly model the complex multimodal structure of human pose space.

More recent work learns expressive pose priors in a latent space. VPoser [27] trains a variational autoencoder on SMPL/SMPL-X body poses and encodes each pose into a low-dimensional latent code that is approximately Gaussian, with the decoder mapping latent vectors back to valid poses. This latent prior has been shown to significantly improve SMPL-X fitting in SMPLify-X [27]. However, the Gaussian assumption tends to bias solutions toward mean poses [36], which can lead to over-smoothed or averaged predictions in ambiguous configurations. To better capture the structure of pose space, PoseNDF [36] models a manifold of plausible poses as the zero level set of a neural unsigned distance field, predicting both distance and gradient so that any pose can be projected onto the manifold via gradient

descent. However, this achieves a distance-to-manifold rather than a normalized likelihood and therefore does not provide a tractable probability density. Diffusion-based pose prior DPoser [23] learns a diffusion model over SMPL joint rotations from a large mocap dataset and uses it as a regularizer for pose generation, human mesh recovery, pose completion, and motion denoising within a SMPLify-style optimization. While this produces a strong and versatile prior, the diffusion process defines an implicit distribution that is easy to sample from and to use in iterative optimization but expensive to evaluate as a normalized density, making direct likelihood scoring of individual poses intractable.

In addition to single-frame pose priors, motion prior models such as HuMoR [29] learn a generative model of 3D human motion over SMPL parameters and use it as a temporal prior for robust pose and shape estimation from noisy observations. HuMoR improves robustness to occlusion and sensor noise by modeling motion in a latent space, but this provides only an implicit likelihood and focuses on single-person motion rather than multi-body configurations.

We build directly upon these single-person foundations: we adopt SMPL-X as the body representation and use VPoser's frozen latent space as the pose parameterization, placing our model within an already anatomically plausible pose space so it can focus fully on learning the structure of two-person interaction.

2.2. Human-Human Interaction Modeling

Most methods for 3D human pose and mesh estimation still treat people independently, even in multi-person scenes, which leads to unrealistic relative placement, missing contacts, and interpenetrations when individuals are in close proximity. Recent work therefore learns explicit interaction priors over pairs of people, typically defined in the space of SMPL/SMPL-X parameters or skeletons, and uses these priors either to generate plausible interactions or to regularize reconstruction from images or videos.

BUDDI [26] introduces a "generative proxemics" prior over pairs of SMPL-X bodies using a denoising diffusion model trained on fitted 3D meshes from interaction datasets. The diffusion process operates directly on the joint parameter vector of two SMPL-X bodies and can sample diverse configurations, or refine noisy multi-person regressor outputs (in their case BEV [35]) during single-image optimization via a score-distillation-style loss term. While BUDDI shows that such a learned proxemics prior can significantly improve pairwise reconstruction quality, the model defines an implicit score-based prior. It is easy to sample from or to use inside an iterative optimization, but it does not provide a tractable, closed-form likelihood for a given interaction instance.

DualHuman [8] focuses on closely interacting two-person motion capture from monocular video and introduces "reaction priors" that model the pose distribution of one person given the other over time. More specifically, it uses a motion VAE to obtain a latent representation of single-person motion clips and an invertible neural network to bi-directionally map between the latent distributions of the "action" and "reaction" characters, which are then injected into a transformer-based two-person pose estimator as both a training regularizer and a test-time optimization prior. This captures temporal dependencies and helps in cases of severe occlusions, but the reaction prior is in motion-latent space and is mainly used to predict a single best reaction trajectory rather than to define a normalized conditional density that can be evaluated for any static pose pairs.

Our model addresses the gap left by both approaches. We learn a conditional normalizing flow over the VPoser latent code of the target person and the relative spatial transform between bodies, conditioned on the reference person's latent code. Unlike BUDDI, this yields an exact, tractable log-likelihood that can be evaluated for any pose pair without running a diffusion chain, making it suitable as a differentiable prior in optimization. Unlike DualHuman, the normalizing flow is not constrained to a unimodal Gaussian output per conditioning input and can represent the full multimodal distribution of plausible reactions.

2.3. Conditional Generative Models and Normalizing Flows

Several families of generative models have been applied to human pose and interaction modeling. GANs [10] can produce sharp samples but are unstable to train and provide no likelihood, making them unsuitable as a probabilistic prior in optimization. VAEs [16] offer a structured latent space and

approximate posteriors, but their likelihoods are lower bounds rather than exact densities, and the Gaussian assumption encourages pose-averaging behavior, a known problem in pose estimation [36]. Diffusion models [12] learn expressive implicit distributions and have been applied to pose priors [23] and interaction priors [26], but evaluating the likelihood of a specific sample requires expensive numerical integration over the reverse diffusion chain, making them impractical as explicit regularizers in optimization loops.

Normalizing flows [30] address these limitations by learning a bijective mapping from data space to a simple base distribution via the change-of-variables formula, yielding exact, tractable likelihoods that are cheap to evaluate at test time. Architectures such as RealNVP [6] and Glow [17] scale this to high-dimensional data using affine coupling layers. Conditional normalizing flows extend this to model $p(y \mid c)$ by conditioning the invertible mapping on an input c, enabling exact conditional density evaluation for any pair (y, c). Recent work on pose anomaly detection [11] demonstrates that flows over pose representations can assign meaningful probabilities to human skeletal poses. These properties: exact likelihoods, efficient sampling, and differentiability, make conditional flows a good choice for our interaction prior, where we need to both sample diverse reactions and score the plausibility of arbitrary pose pairs.

3

Method

Our goal is to learn an explicit probability distribution over the pose of a target person and the relative configuration of a pair, given the pose of a reference person. We represent each person using a SMPL-X body model and the VPoser latent space, and we model a conditional density over the VPoser latent of the target person and the relative translation and orientation between the two people, given the latent of the reference person. This section first describes the representation and variables used for interaction modeling, then introduces the conditional normalizing flow that parameterizes the interaction prior.

3.1. Two-Person Pose Representation

We represent each person using the SMPL-X body model [27], which maps pose, shape, and global transformation parameters to a body mesh with joints and surface vertices. Each person in role $r \in \{\text{ref}, \text{tgt}\}$ (reference, target) is denoted as:

- $\theta_r \in \mathbb{R}^{21\times 3}$ the axis-angle body pose parameters,
- $\beta_r \in \mathbb{R}^{16}$ the shape parameters,
- $R_r \in \mathrm{SO}(3)$ the global orientation,
- $t_r \in \mathbb{R}^3$ the global translation of the root joint.

The SMPL-X model is a differentiable function

$$M(\theta_r, \beta_r, R_r, t_r) \to (V_r, J_r),$$

that outputs the mesh vertices V_r and joint locations J_r given these parameters.

Directly learning a prior in the high-dimensional pose space θ_r is challenging due to complex joint angle constraints and joint correlations. To address this, we leverage VPoser [27], a pre-trained variational autoencoder that maps body poses to a compact 32-dimensional latent space. More specifically, we use its frozen encoder $E_{\text{vp}}(\theta_r) = z_r \in \mathbb{R}^{32}$ and decoder $\theta_r = D_{\text{vp}}(z_r)$ without modification. The resulting latent space is approximately Gaussian and already enforces single-person anatomical plausibility. Our contribution is the interaction prior learned within this latent space.

We explicitly model the relative positioning between the conditioning and target people. We define the relative translation from conditioning to target as

$$t_{\text{rel}} = R_{\text{c}}^{\top}(t_{\text{t}} - t_{\text{c}}) \in \mathbb{R}^3,$$

Similarly, we define the relative orientation as

$$R_{\text{rel}} = R_{\text{c}}^{\top} R_{\text{t}} \in \mathrm{SO}(3),$$

which we parameterize with a 6D representation $r_{\text{rel}} \in \mathbb{R}^6$ described in [42].

In our main model, we focus on learning a prior over $(z_{\mathsf{t}}, t_{\mathsf{rel}}, r_{\mathsf{rel}})$ conditioned on z_{c}; the symmetric case where conditioning and target roles are swapped is handled at training time (Section 3.4).

The normalizing flow is used to learn the density:

$$p_\phi(z_{\mathsf{t}}, t_{\mathsf{rel}}, r_{\mathsf{rel}} \mid z_{\mathsf{c}}),$$

which supports tractable training via the change-of-variables formula and efficient sampling via its inverse. The concatenated *target* vector is $y = [z_{\mathsf{t}};\, t_{\mathsf{rel}};\, r_{\mathsf{rel}}] \in \mathbb{R}^{41}$ and the *conditioning* variable is $c = z_{\mathsf{c}}$.

3.2. Model Definition

We introduce a sequence of K invertible transformations

$$u = f_\phi(y; c) = f_K \circ f_{K-1} \circ \cdots \circ f_1(y; c),$$

where each $f_k(\cdot;\, h)$ is an invertible affine coupling transform whose parameters depend on a learned conditioning feature $h = g_\psi(c)$. Rather than a fixed standard Gaussian, the base distribution is a conditional diagonal Gaussian whose parameters are themselves predicted from h:

$$p_0(u \mid h) = \mathcal{N}\big(u;\, \mu_\psi(h),\, \mathrm{diag}(\sigma_\psi(h))^2\big)\,.$$

By the change-of-variables formula, the conditional density over y is

$$p_\phi(y \mid c) = p_0(f_\phi(y;\, h) \mid h) \left|\det \frac{\partial f_\phi(y;\, h)}{\partial y}\right| \tag{3.1}$$

where $h = g_\psi(c)$. In this setting, $p_\phi(y \mid c)$ models the distribution over $(z_{\mathsf{t}}, t_{\mathsf{rel}}, r_{\mathsf{rel}})$ given z_{c}. Figure 3.1a illustrates this forward pass.

Sampling N diverse reactions from the prior, as illustrated in Figure 3.1b and summarized in Algorithm 1, proceeds as follows. The reference pose θ_A is encoded into the VPoser latent $z_A = E_{\mathsf{vp}}(\theta_A)$ and passed through the conditioning encoder to obtain $h = g_\psi(z_A)$. Base samples are then drawn from $p_0(u \mid h)$ and mapped through the inverse flow to interaction vectors $y^{(i)} = f_\phi^{-1}(u^{(i)};\, h)$. Each $y^{(i)}$ is split and denormalized to recover the VPoser latent $z_B^{(i)}$, relative translation $t_{\mathsf{rel}}^{(i)}$, and 6D rotation $r_{\mathsf{rel}}^{(i)}$, which are then decoded to a body pose $\theta_B^{(i)}$ and rotation matrix $R_{\mathsf{rel}}^{(i)}$. A temperature $\tau > 0$ scales the noise at sampling time: $\tau < 1$ concentrates samples near the conditional mode, while $\tau > 1$ increases diversity at the cost of plausibility.

(a) Forward pass (training and scoring): y is mapped to $u = f_\phi(y;\, h)$ and log-likelihood is evaluated via the change-of-variables formula.

(b) Inverse pass (sampling): base sample $u \sim p_0(u \mid h)$ is mapped back through $f_\phi^{-1}(u;\, h)$ to yield z_t, t_{rel}, and r_{rel}.

Figure 3.1: Forward and inverse passes of the conditional normalizing flow.

3.3. Architecture

We implement f_ϕ as a stack of $K = 8$ flow blocks. Each block consists of three components applied in sequence:

1. **ActNorm**: a learned per-dimension affine normalization, initialized on the first training batch to produce zero-mean, unit-variance activations, which stabilizes early training.
2. **Affine coupling layer**: an invertible transformation in the RealNVP style, described in detail below.
3. **Fixed random permutation**: a parameter-free reordering of dimensions with zero log-determinant, which ensures all dimensions interact across the stack of coupling layers.

Following [38], we condition the flow by first projecting z_{c} through a small MLP encoder $h = g_\psi(z_{\mathsf{c}})$ and then passing the resulting feature h to both the coupling layers and the base distribution. This allows the model to learn a more expressive conditioning representation than the raw VPoser latent alone.

Algorithm 1 Sampling from the Interaction Prior

Require: Reference pose $\theta_A \in \mathbb{R}^{21\times 3}$, number of samples N, temperature $\tau > 0$
Ensure: $\{(\theta_B^{(i)},\, t_{\text{rel}}^{(i)},\, R_{\text{rel}}^{(i)})\}_{i=1}^{N}$

$z_A \leftarrow E_{\text{vp}}(\theta_A)$ ▷ Encode reference pose
$h \leftarrow g_\psi(z_A)$ ▷ Conditioning features
for $i = 1$ **to** N **do**
  $u^{(i)} \sim \mathcal{N}(\mu_\psi(h),\, \tau^2\, \mathrm{diag}(\sigma_\psi(h))^2)$ ▷ $\tau > 0$ user-chosen
  $y^{(i)} \leftarrow f_\phi^{-1}(u^{(i)};\, h)$ ▷ Inverse flow
  $(z_B^{(i)},\, t_{\text{rel}}^{(i)},\, r_{\text{rel}}^{(i)}) \leftarrow \mathrm{split_and_denorm}(y^{(i)})$ ▷ Split and recover metric scale
  $\theta_B^{(i)} \leftarrow D_{\text{vp}}(\hat{z}_B^{(i)})$; $R_{\text{rel}}^{(i)} \leftarrow \mathrm{Rot6dToMat}(r_{\text{rel}}^{(i)})$ ▷ Decode pose and rotation
end for
return $\{(\theta_B^{(i)},\, t_{\text{rel}}^{(i)},\, R_{\text{rel}}^{(i)})\}_{i=1}^{N}$

Each affine coupling layer splits the input y into two parts (y_1, y_2) and updates one part conditioned on the other, in the RealNVP style:

$$y_1' = y_1, \quad y_2' = y_2 \odot \exp(s_\phi(y_1, h)) + t_\phi(y_1, h),$$

where s_ϕ and t_ϕ are scale and translation networks implemented as three-layer MLPs, taking the non-updated part y_1 and the conditioning feature h as input. To prevent exploding scale values, s_ϕ is passed through a $\tanh$ before exponentiation, bounding the scale transformation. Each layer swaps the roles of y_1 and y_2 so that all dimensions are transformed across the stack.

The base distribution is a conditional diagonal Gaussian

$$p_0(u \mid h) = \mathcal{N}(u;\, \mu_\psi(h),\, \mathrm{diag}(\sigma_\psi(h))^2),$$

where μ_ψ and $\log \sigma_\psi$ are predicted by a linear projection from the conditioning feature h [38]. This lets the base distribution shift and scale with the reference pose, providing an additional source of conditioning beyond the coupling layers.

We structure y such that the pose latent z_{t} and the relative configuration $(t_{\text{rel}}, r_{\text{rel}})$ are both transformed across layers, allowing the flow to learn their joint correlations. All components are differentiable, which is required for using $\log p_\phi$ as a prior term in optimization.

3.4. Training Objective

Given a dataset of paired poses and relative transforms $\{(z_{\text{t}}^{(n)}, z_{\text{c}}^{(n)}, t_{\text{rel}}^{(n)}, r_{\text{rel}}^{(n)})\}_{n=1}^{N}$, we train the conditional flow by maximum likelihood. For each sample, we form $y^{(n)} = [z_{\text{t}}^{(n)};\, t_{\text{rel}}^{(n)};\, r_{\text{rel}}^{(n)}]$, compute the conditioning feature $h^{(n)} = g_\psi(z_{\text{c}}^{(n)})$ and the base encoding $u^{(n)} = f_\phi(y^{(n)};\, h^{(n)})$, then minimize the negative log-likelihood

$$\mathcal{L}_{\text{flow}}(\phi) = -\frac{1}{N}\sum_{n=1}^{N}\left[\log p_0\Big(u^{(n)} \mid h^{(n)}\Big) + \log\left|\det \frac{\partial f_\phi(y^{(n)};\, h^{(n)})}{\partial y^{(n)}}\right|\right]$$

Since $p_0(\cdot \mid h)$ is a diagonal Gaussian and the Jacobian of each coupling layer is triangular, both terms are tractable.

Because either person in an interaction can serve as conditioning or target, we train the flow in both directions. Hence, for each pair we also construct $y' = [z_{\text{c}}; t_{\text{rel}}'; r_{\text{rel}}']$ with the inverse relative transform from target to conditioning, condition on $c' = z_{\text{t}}$, and add its log-likelihood term to the loss. This encourages the model to treat conditioning and target symmetrically.

3.5. Implementation Details

Model hyperparameters. The flow consists of $K = 8$ blocks, each containing an ActNorm layer, an affine coupling layer, and a fixed random permutation. The coupling networks are three-layer MLPs

with hidden dimension $d_h = 128$ and ELU activations. The conditioning encoder g_ψ is a two-layer MLP: $\mathbb{R}^{32} \xrightarrow{\text{ELU}} \mathbb{R}^{64} \xrightarrow{\text{ELU}} \mathbb{R}^{32}$. The base distribution head is a single linear layer $\mathbb{R}^{32} \to \mathbb{R}^{82}$, predicting $(\mu_\psi(h), \log \sigma_\psi(h))$ for the 41-dimensional base distribution; $\log \sigma_\psi$ is clamped to $[-7, 5]$ for numerical stability. The VPoser v2 encoder and decoder are used frozen throughout; no VPoser parameters are updated during training. The full model has **1.2M** parameters in total (VPoser: **936K** frozen; normalizing flow: **304K** trainable).

Input normalization. The relative translation t_{rel} and relative orientation r_{rel} are standardized before being passed to the flow. Per-dimension mean and standard deviation are computed on the training split and applied to normalize the spatial components of $y = [z_t; \hat{t}_{\text{rel}}; \hat{r}_{\text{rel}}]$, where hats denote standardized quantities. The VPoser latents z_t and z_c are not re-normalized, as the VPoser latent space is already approximately unit Gaussian by construction. Denormalization is applied after sampling to recover metric-scale translations and valid rotation representations.

Training. The model is trained by minimizing the symmetric negative log-likelihood (Section 3.4). We use Adam [15] with a learning rate of 10^{-4}, a batch size of 512, and train for up to 300 epochs with early stopping (patience 10) on the validation NLL. The best checkpoint is selected by minimum validation loss. All experiments are run on a single NVIDIA GeForce RTX 3060 Ti.

4

Experiments

We evaluate the learned interaction prior in three experiments. First, we assess it as a generative model by measuring the diversity and realism of sampled reactions given an observed person (Section 4.1). Second, we evaluate it as a plausibility scorer by testing how well the learned density separates real interactions from mismatched and noisy pose pairs (Section 4.2). Third, we use it as a differentiable prior in an image-to-mesh optimization and measure whether it improves 3D reconstruction quality for interacting pairs (Section 4.3).

Dataset We train and evaluate the generative and plausibility experiments (Sections 4.1 and 4.2) on Inter-X [39], a large-scale dataset of close two-person interactions captured in a multi-view studio. Inter-X contains over 8,000 interaction sequences across 40 subjects, covering a wide range of interaction types including handshakes, hugs, and collaborative tasks, with ground-truth SMPL-X parameters per frame. We use the official train/test split for both training the CNF prior and the DH-like baseline, and report all generative and plausibility results on the Inter-X test set. Inter-X does not provide RGB images, so it cannot be used for the image-to-mesh experiment.

For the image-to-mesh experiment (Section 4.3), we instead use Hi4D [40], which provides synchronized RGB images absent from Inter-X. Hi4D is not used for the generative and plausibility experiments because its ~ 100 sequences are too few for the distributional metrics (KID) reported there: a sequence-level split leaves too little training data, and a window-level split does not guarantee sufficient temporal separation for KID. Full dataset details and the train/test split are described in Section 4.3.

Baselines. In the first two evaluations we compare against two baselines: a DualHuman-inspired INN-based reaction prior and BUDDI, a diffusion-based proxemics prior. Both are trained on the same Inter-X training split as our model.

DH-like baseline. An invertible neural network (INN) is a neural network whose layers are bijective by construction, so the mapping and its exact inverse are both available in closed form. DualHuman [8] uses an INN to bi-directionally map the distribution parameters of one person's motion latent to those of their interaction partner, and applies the learned mapping as a prior to regularize pose estimation.

We do not re-implement DualHuman directly: their motion VAE is trained on interaction-level data that encodes global translation and orientation alongside body pose, whereas VPoser is trained on single-person body poses only and carries no spatial information. Extending the DH design to also model $(t_{\mathrm{rel}}, r_{\mathrm{rel}})$ would require training a new VAE on interaction data that jointly encodes pose and relative placement; without one, spatial quantities can only be predicted by direct regression, producing point estimates with no distributional representation. The DH design works for pose because VPoser latents have a natural Gaussian structure, with the encoder outputting $(\mu, \log \sigma)$ and the INN mapping one such distribution to another, and no equivalent encoder exists for spatial quantities. We therefore implement a pose-only version: an INN f_ϕ that predicts the VPoser distribution parameters of the target person

from those of the reference person,

$$(\tilde{\mu}_{\mathrm{tgt}},\, \log \tilde{\sigma}_{\mathrm{tgt}}) = f_\phi(\mu_{\mathrm{ref}},\, \log \sigma_{\mathrm{ref}}),$$

trained bidirectionally on the same Inter-X data by minimizing

$$\mathcal{L} = \left\| (\tilde{\mu}_{\mathrm{tgt}},\, \log \tilde{\sigma}_{\mathrm{tgt}}) - (\mu_{\mathrm{tgt}},\, \log \sigma_{\mathrm{tgt}}) \right\|_2^2,$$

where $(\mu_{\mathrm{tgt}}, \log \sigma_{\mathrm{tgt}})$ are the VPoser-encoded parameters of the ground-truth target pose. This baseline isolates the key architectural question: does replacing a Gaussian-to-Gaussian INN mapping with a full normalizing flow result in a better interaction prior? We include a pose-only ablation of our CNF so that both models are evaluated on equal footing.

BUDDI. BUDDI [26] is a denoising diffusion model that learns the *joint* distribution over the full SMPL-X body parameters (global orientation, body pose, and translation) of two interacting people simultaneously, operating directly in body-parameter space rather than in a VPoser latent. The original BUDDI was trained on close-interaction data from FlickrCI3D Signatures, CHI3D, and Hi4D; we retrain it on the same Inter-X training split for a fair comparison. Unlike our model, which is explicitly a conditional distribution $p(\text{person B} \mid \text{person A})$, BUDDI models both people jointly. To draw a reaction sample for a given reference person, we use inpainting: at each denoising step t, person A's parameters are re-noised to the current noise level via the forward diffusion process $q(x_t \mid x_0)$, then substituted back into the joint state before the denoiser runs. This keeps the joint input $(x_{A,t},\, x_{B,t})$ on the marginal distribution the model was trained on at every step, while person B denoises freely toward a plausible reaction.

4.1. Conditional Generation

We evaluate the interaction prior as a generative model: given an observed reference person in an Inter-X interaction frame, how diverse and how realistic are the reactions a model can generate for the target person?

Setup. We sample 100 frames uniformly from the Inter-X test set, treating one person as the fixed reference and generating 20 reaction samples per model per frame. These samples are used to measure diversity (APD) and realism (KID) against the real Inter-X distribution. For our CNF prior, samples are drawn with temperature $\tau = 0.5$, which scales the base distribution by τ^2. This value was chosen empirically: lower temperatures reduce sample diversity, while temperatures above ≈ 0.7 begin to produce implausible looking poses. The DH-like baseline samples from its predicted Gaussian directly, with no temperature parameter.

Diversity metric. To quantify how varied the sampled poses are for a fixed reference, we compute the *average pairwise distance* (APD) between generated target poses. Given n reaction samples $\{\theta_{\mathsf{tgt}}^{(i)}\}_{i=1}^{n}$ represented as vectorized joint positions, we define

$$\mathrm{APD} = \frac{2}{n(n-1)} \sum_{1 \le i < j \le n} \left\| \theta_{\mathsf{tgt}}^{(i)} - \theta_{\mathsf{tgt}}^{(j)} \right\|_2 .$$

APD is computed over decoded body pose only, so all four models are evaluated on equal footing regardless of whether they predict spatial layout. Higher APD indicates that the model explores a wider region of pose space for the same reference person; a very low APD suggests mode collapse.

Realism metrics. Diversity alone does not guarantee good generative quality. To measure how close the generated interactions are to the real Inter-X distribution, we apply the Kernel Inception Distance (KID) [1] to two different interaction representations. KID is an unbiased maximum mean discrepancy (MMD) estimator with a cubic polynomial kernel $k(x, y) = (\gamma x^\top y + c)^3$; lower values indicate that generated samples are statistically closer to the real distribution.

Pose KID. We represent each interaction as a 64-dimensional pose vector

$$v_{\mathrm{pose}} = \left[z_A,\; z_B \right] \in \mathbb{R}^{64},$$

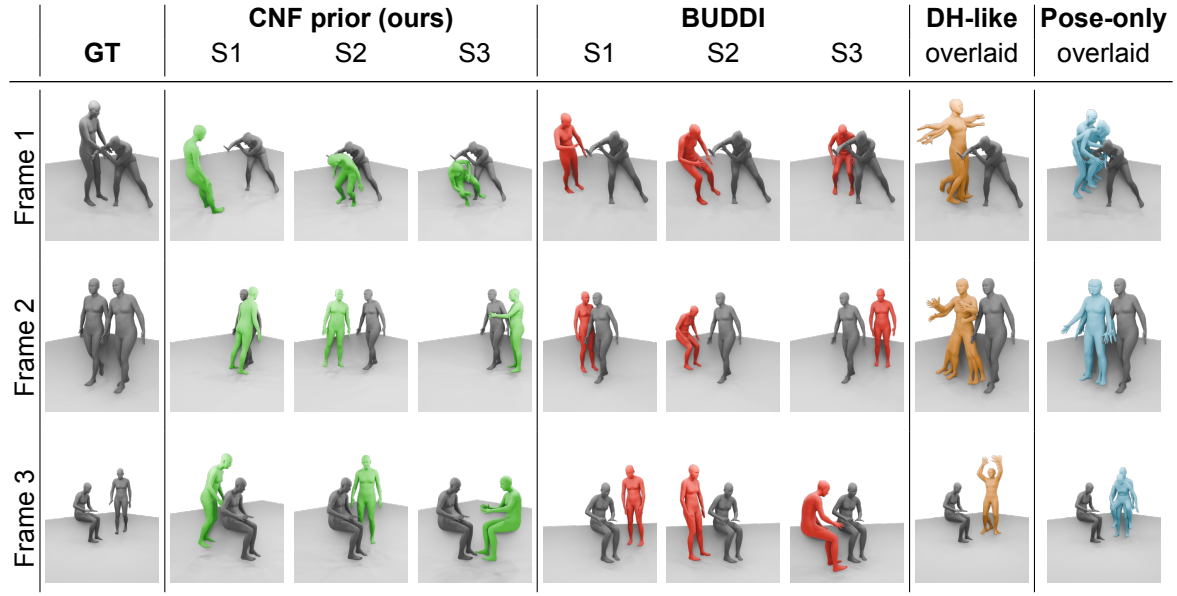


Figure 4.1: Qualitative comparison of sampled reactions for three test frames (rows) selected for variety of close interactions. In each image, the **grey** mesh is the conditioning signal (reference person), and the generated reactions are shown in color. Per frame and model, 10 reactions are sampled at τ=0.5 for CNF models; three are shown per method. **CNF prior (S1–S3):** three individual samples showing diverse modes of the conditional distribution. **BUDDI (S1–S3):** three individual samples. **DH-like baseline:** all samples overlaid. **CNF pose-only:** all samples overlaid.

Table 4.1: Generative evaluation on Inter-X. Higher APD indicates greater sample diversity; lower KID indicates closer alignment to the real interaction distribution. Interaction KID additionally captures spatial arrangement (relative translation and orientation). "—" indicates the model does not predict spatial layout.

		Pose KID ↓		Interaction KID ↓	
Model	APD ↑	mean	std	mean	std
DH-like baseline	0.165	0.262	0.026	—	—
CNF (ours) pose-only	0.224	0.072	0.012	—	—
BUDDI	0.167	**0.052**	0.064	**0.106**	0.078
CNF (ours)	**0.326**	0.211	0.042	0.269	0.141

formed by concatenating the VPoser latents of the reference and target persons. KID is then computed between real Inter-X pose vectors and those of the generated reactions. This variant is computed for all four models, since every model produces a body pose for the target person.

Interaction KID. To also capture the spatial arrangement between people, we extend the pose vector with the relative translation and orientation, forming a 73-dimensional interaction vector

$$v_{\text{int}} = \left[z_A,\ z_B,\ t_{\text{rel}},\ r_{\text{rel}}\right] \in \mathbb{R}^{73},$$

where $t_{\text{rel}} \in \mathbb{R}^3$ and $r_{\text{rel}} \in \mathbb{R}^6$ are computed as in Section 3.1. This variant is computed only for our CNF prior and BUDDI, since the DH-like baseline and CNF pose-only ablation do not predict spatial layout.

For both KID variants, we compare against a reference set of 10,000 real Inter-X frames sampled uniformly from the test split. Each repetition generates 100 new interaction samples per model; we run 20 such repetitions and report mean and standard deviation of KID across them. Quantitative results are summarised in Table 4.1.

Qualitative results. Figure 4.1 shows three test frames with reaction samples from each model. For CNF models, 10 reactions were sampled at τ=0.5; three are shown per method. DH-like baseline and CNF pose-only samples are shown overlaid.

4.2. Plausibility Scoring

Modeling an explicit conditional density over interactions allows it to be used as a plausibility score. Given a pair of poses and their relative configuration, the model should assign higher likelihood to realistic interactions and lower likelihood to implausible ones. To quantify this, we evaluate how well our CNF prior separates real Inter-X test frames from mismatched and noisy pose pairs, and compare against the DH-like baseline and BUDDI [26]. We treat plausibility scoring as a binary classification problem and report standard metrics: ROC AUC and pairwise accuracy.

The DH-like baseline produces $\log p(z_{tgt} \mid z_{ref})$ under a predicted diagonal Gaussian. Our CNF model produces $\log p_\phi(z_{tgt}, t_{rel}, r_{rel} \mid z_{ref})$, a joint pose-and-placement score. BUDDI produces a denoising reconstruction error as described below. These scores are not numerically comparable across models, but each serves as a scalar binary classifier.

Evaluation setup. We sample 1 000 pairs from the Inter-X test set and evaluate on three pair types:

- **Positive:** both people taken from the same Inter-X test frame.
- **Mismatched:** the reference person from one frame paired with a target from a randomly shuffled frame. This preserves individual pose realism but violates interaction-level constraints; because targets are drawn from entirely different sequences, these pairs are typically implausible in both joint pose and spatial configuration.
- **Noisy:** the target person's pose perturbed with independent Gaussian noise at $\sigma \in \{0.02, 0.05, 0.10, 0.15\}$, leaving global translation and orientation unchanged. All models have access to the same information here; performance differences reflect only the quality of the learned conditional pose distribution.

Plausibility models. Our (CNF) model encodes both people into VPoser latents z_{ref} and z_{tgt}, computes t_{rel} and r_{rel} from their global transforms, and packs the target variables $y = [z_{tgt};\, t_{rel};\, r_{rel}]$. Conditioning on z_{ref}, the flow produces the exact log-density via the change-of-variables formula (Eq. 3.1). We use this scalar $\log p_\phi$ as the plausibility score where higher values indicate more plausible interactions.

The DH-like baseline encodes z_{ref}, predicts $(\tilde{\mu}_{tgt}, \tilde{\sigma}_{tgt})$ via the Invertible Neural Network, encodes z_{tgt}, and evaluates $\log \mathcal{N}(z_{\mathrm{tgt}} \mid \tilde{\mu}_{\mathrm{tgt}}, \tilde{\sigma}_{\mathrm{tgt}})$ as its score.

BUDDI was originally designed as an optimizer: during image-to-mesh reconstruction it diffuses the current body estimate and uses a single denoise step to pull the estimate toward the learned interaction distribution, similar to a score-distillation-style loss [26]. It does not expose an explicit density and was not designed to score arbitrary pairs. To include it as a baseline in this discriminative task we adapt its denoiser into a plausibility surrogate, inspired by the same diffuse-denoise intuition: if a pair (A, B) is plausible, the denoiser should be able to recover B accurately from the noised joint pair.

Concretely, we corrupt both persons' parameters with DDPM noise at a fixed timestep $t{=}500$ and run a single denoiser forward pass to predict the clean $\hat{x}_0$ for the whole pair. Because the noise draw is random, the reconstruction error varies across samples, so we average over $K{=}5$ independent noise draws to reduce variance. We selected $t{=}500$ by sweeping over a range of timesteps on a held-out set and keeping the most discriminative one; no timestep produced a signal far above chance, and $t{=}500$ was the strongest among them. The plausibility score is the negative weighted reconstruction MSE of person B, averaged across all noise draws:

$$s_{\text{BUDDI}} = -\frac{1}{K}\sum_{k=1}^{K} \mathrm{MSE}_W\left(\hat{x}_{0,B}^{(k)},\, x_{0,B}\right),$$

where MSE_W is a weighted MSE over orientation, pose, and translation with weights $1{:}10{:}1$ (taken from the original BUDDI optimizer) and $\hat{x}_{0,B}^{(k)}$ is the single-step denoised prediction for person B on the k-th noise draw. Higher scores (closer to zero) indicate that the denoiser reconstructs person B accurately given the noised joint pair, suggesting the interaction lies near the learned data manifold. This adaptation is not a standard technique and using diffusion models as direct plausibility scorers remains an open research question.

Table 4.2: Plausibility discrimination (higher is better). AUC / Pairwise Accuracy.

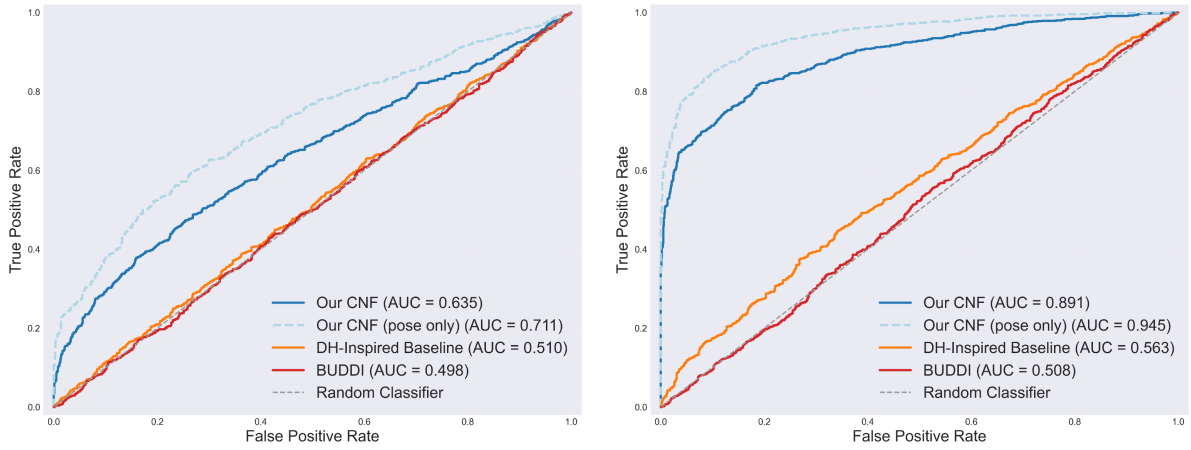
Model	Mismatched	Noise σ			
		0.02	0.05	0.10	0.15
DH-like baseline	0.623 / 0.633	0.510 / 0.608	0.563 / 0.722	0.677 / 0.854	0.796 / 0.917
CNF (ours) pose-only	0.678 / 0.667	**0.711** / **0.901**	**0.945** / **0.991**	**0.993** / **0.999**	**0.998** / **1.000**
CNF (ours)	**0.722** / **0.727**	0.635 / 0.894	0.891 / 0.986	0.984 / **0.999**	0.997 / **1.000**
BUDDI	0.530 / 0.530	0.498 / 0.481	0.508 / 0.513	0.523 / 0.567	0.536 / 0.555

Scoring and metrics. Each frame produces a scalar plausibility score for each model. We evaluate on two binary discrimination tasks:

1. **Positive vs. mismatched:** distinguish true Inter-X interactions from mismatched pairs, which tests whether the model captures the structure of close interactions beyond per-person pose realism.
2. **Positive vs. noisy:** the primary controlled comparison between the models. Only pose is perturbed and spatial configuration is unchanged. Performance differences directly show the quality of the learned conditional pose distribution.

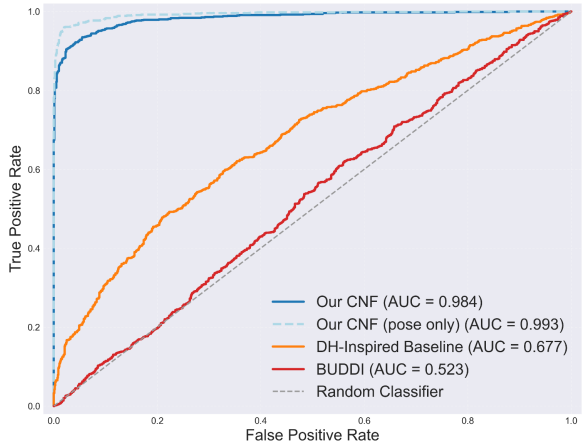
For each task and model we compute ROC AUC and average pairwise accuracy, reported in Table 4.2. ROC curves for the noise-discrimination task are shown in Figure 4.2; the mismatch-discrimination ROC is shown in Figure 4.2e.

Qualitative analysis. Figure 4.3 shows three example frames alongside their noisy and mismatched variants, with the symmetric log-probability assigned by our CNF prior displayed beneath each rendering. Scores generally decrease as noise increases, and mismatched pairs typically receive lower scores than the lightly noised ground-truth pairs. Additional examples are provided in Appendix A.2.1. Figure 4.4 illustrates the spatial sensitivity of the prior: as person B is swept in equal angular steps around person A, the assigned probability peaks near the ground-truth orientation and drops for configurations that are implausible for the interaction type. A further rotation example is shown in Appendix A.2.2.

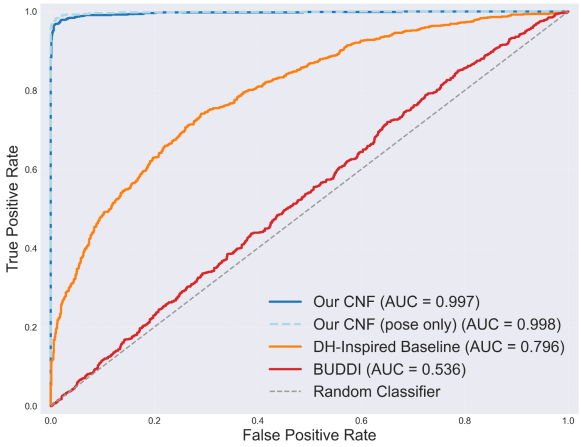


(a) Noise σ=0.02

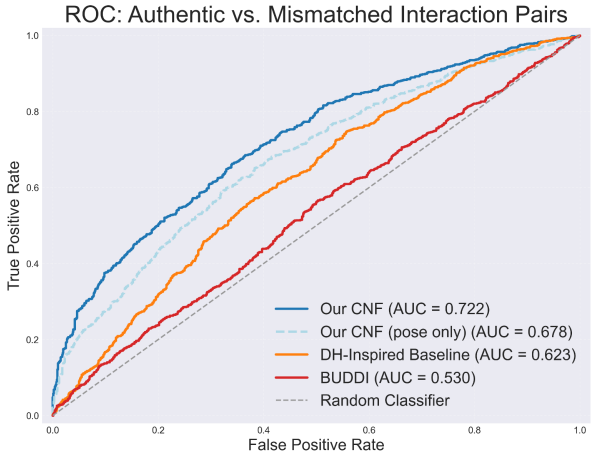
(b) Noise σ=0.05



(c) Noise σ=0.10



(d) Noise σ=0.15



(e) Mismatched pairs

Figure 4.2: ROC curves for positive vs. noisy interactions (top four) and positive vs. mismatched discrimination (bottom).

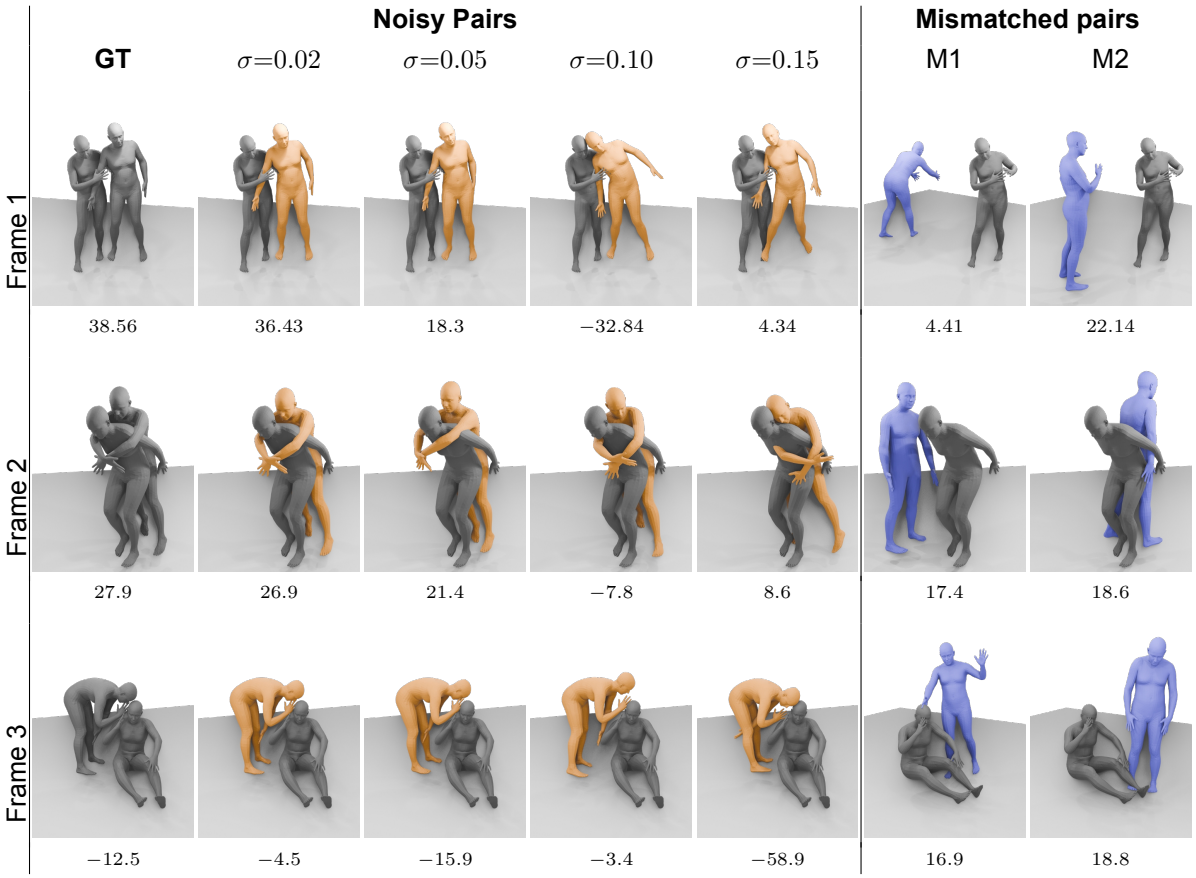
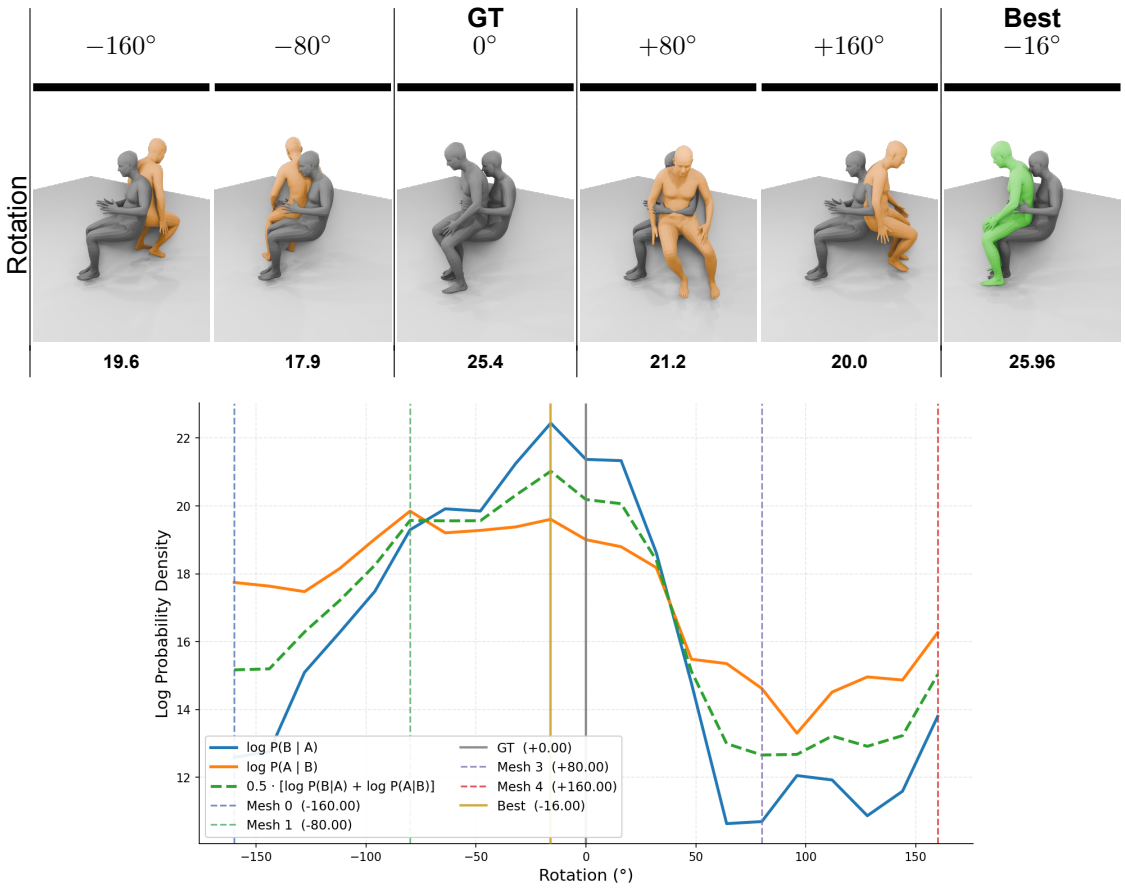


Figure 4.3: Qualitative evaluation of interaction pair plausibility scoring using our CNF prior on three test frames (rows). Ground truth frame is followed by same pair with increasing Gaussian pose noise added to person B ($\sigma \in \{0.02, 0.05, 0.10, 0.15\}$). In two right-most columns are mismatched pairs. Values below each rendering show the symmetric log-probability $\frac{1}{2}[\log p(B|A) + \log p(A|B)]$ assigned by the model.



Symmetric log-probability vs. rotation angle

Figure 4.4: Symmetric log-probability $\frac{1}{2}[\log p(B|A) + \log p(A|B)]$ as person B orbits around person A in equal angular steps. Columns show two orientations before the original (GT) and two after, with the highest-scoring orientation ($-16°$) separated on the right (Best). The bottom panel shows the log-probability density over the range of 20 equally sized steps of transformation.

4.3. Prior-Guided Image-to-Mesh Optimization

Beyond sampling, we use the learned interaction prior as a differentiable regularizer in an image-to-mesh optimization inspired by SMPLify [2]. Given a single RGB image of two interacting people, we refine an initial SMPL-X estimate by minimizing an energy that combines 2D keypoint reprojection with a prior term from the conditional flow. The two terms act jointly throughout optimization. The prior penalizes implausible relative configurations caused by occlusion or depth ambiguity, steering the solution towards a realistic interaction region, while the data term ensures it remains consistent with the observed 2D positions of the visible joints.

Data and initialization. Hi4D [40] is a dataset of closely interacting two-person motions captured at 30 fps with synchronized multi-view RGB cameras. It contains around 100 motion sequences across 20 subject pairs performing interactions such as hugging, dancing, and fighting, with ground-truth SMPL-X fits and synchronized RGB images. We train a separate instance of the CNF prior on the Hi4D training split.

Train/test split. Because 100 sequences are too few for a reliable sequence-level split, we partition at the window level. Each sequence is divided into consecutive non-overlapping windows of $W{=}30$ frames (1 second), discarding any incomplete tail. All windows are pooled, shuffled with a fixed random seed, and partitioned by ratio into train, validation, and test splits, so no individual frame appears in more than one split. Windows from the same sequence may appear in different partitions. For evaluation we use only frames at positions 10 and 20 within each test window, yielding 199 evaluation frames; selecting interior frames avoids window boundaries, which are temporally closest to frames that may appear in a neighbouring training window.

We evaluate on the Hi4D test set described above. Each of the 199 evaluation frames provides a reference RGB view, ground-truth SMPL-X parameters (GT), and an initial BEV prediction [35] which serves as our starting point and baseline. We compute 2D joint detections with OpenPose [3], which outputs per-person 2D keypoints with per-keypoint confidence scores $w_{r,k}$. To assign OpenPose detections to the two BEV persons, we compute a cost between the BEV-projected joints of each person and each OpenPose detection, using only keypoints above a confidence threshold, and select the assignment that minimizes the total cost. If no reliable detection is found for a person, their data term is dropped for that frame and optimization proceeds with the prior term alone.

Optimization variables and procedure. We optimize the interaction pair parameters

$$\Theta = \{\theta_{\mathsf{t}}, \theta_{\mathsf{c}}, R_{\mathsf{t}}, R_{\mathsf{c}}, t_{\mathsf{t}}, t_{\mathsf{c}}\},$$

initializing from BEV. At each step we: (1) encode $(\theta_{\mathsf{t}}, \theta_{\mathsf{c}})$ into VPoser latents $(z_{\mathsf{t}}, z_{\mathsf{c}})$; (2) compute the relative translation t_{rel} and orientation r_{rel}; (3) evaluate the interaction prior; and (4) project both body skeletons into the image to compare against detected keypoints. We optimize using Adam for 200 iterations.

Energy. The total energy is

$$E(\Theta) = \lambda_{\mathsf{data}}\, E_{\mathsf{data}} + \lambda_{\mathsf{prior}}\, E_{\mathsf{prior}} + \lambda_{\mathsf{vp}}\, E_{\mathsf{vp}},$$

with $\lambda_{\mathsf{data}} = 0.001$, $\lambda_{\mathsf{prior}} = 1.0$, $\lambda_{\mathsf{vp}} = 0.1$.

Data term. We penalize the weighted reprojection error between projected joints $\hat{J}_{r,k}(\Theta)$ and OpenPose detected 2D keypoints $j^{2D}_{r,k}$ with per-keypoint confidence $w_{r,k}$:

$$E_{\mathsf{data}}(\Theta) = \sum_{r\in\{\mathsf{t},\mathsf{c}\}} \sum_{k} w_{r,k} \left\| \hat{J}_{r,k}(\Theta) - j^{2D}_{r,k} \right\|^2.$$

Prior term. The prior term penalizes configurations with low probability under the learned flow, evaluated symmetrically in both role assignments:

$$E_{\mathsf{prior}}(\Theta) = -\log p_\phi(z_{\mathsf{t}}, t_{\mathsf{rel}}, r_{\mathsf{rel}} \mid z_{\mathsf{c}}) - \log p_\phi(z_{\mathsf{c}}, t'_{\mathsf{rel}}, r'_{\mathsf{rel}} \mid z_{\mathsf{t}}),$$

Table 4.3: Image-to-mesh refinement on Hi4D. We report mean ± std over the test set for MPJPE, PA-MPJPE$_{\text{per}}$, PA-MPJPE$_{\text{2p}}$, contact precision, contact recall, and contact F1. The upper block shows refinement from BEV initialization; the lower block shows the ablation starting from a template pose initialized from 2D keypoints only.

Method	MPJPE ↓	PA-MPJPE$_{\text{per}}$ ↓	PA-MPJPE$_{\text{2p}}$ ↓	Precision ↑	Recall ↑	F1 ↑
BEV	74.06	**50.41**	93.16	**0.220**	0.134	0.134
BEV + OpenPose	74.87	52.73	89.65	0.191	0.185	**0.146**
BEV + OpenPose + prior (ours)	**73.62**	51.83	**87.25**	0.188	**0.187**	0.140
Ablation: template initialization						
OpenPose + VPoser	162.80	104.90	179.58	0.119	0.074	**0.068**
OpenPose + prior (ours)	**159.44**	**92.58**	**165.17**	0.091	0.075	0.054

where t'_{rel} and r'_{rel} denote the relative translation and orientation recomputed with person B as the reference frame.

VPoser regularization. To keep the optimized poses on the learned pose manifold, we penalize the ℓ_2 norm of the VPoser latent codes for both people:

$$E_{\text{vp}}(\Theta) = \|z_{\text{t}}\|_2^2 + \|z_{\text{c}}\|_2^2.$$

Since the VPoser prior is approximately $\mathcal{N}(0, I)$, this encourages the latents to remain close to the origin and discourages anatomically implausible poses.

Comparisons. We compare three main configurations:

- **BEV:** raw BEV predictions [35], serving as the initialization and a baseline.
- **BEV + OpenPose:** optimization with $\lambda_{\text{prior}} = 0$ – not using our interaction prior.
- **BEV + OpenPose + prior (ours):** combining OpenPose keypoints with the interaction prior.

Results for all three configurations are reported in Table 4.3. Figure 4.5 shows frames where the prior has the most visible effect on spatial configuration.

Metrics. We evaluate reconstruction quality using four metrics that capture per-person accuracy, interaction-level accuracy, and the ability to reconstruct body contact correctly.

MPJPE. Mean Per-Joint Position Error measures the average Euclidean distance between predicted and ground-truth 3D joint positions after root-aligning each person independently, i.e. subtracting the pelvis joint from all joints of each skeleton before computing the error. This follows the standard convention in single-person pose estimation and removes global translation ambiguity:

$$\text{MPJPE} = \frac{1}{|\mathcal{R}|} \sum_{r \in \mathcal{R}} \frac{1}{K} \sum_{k=1}^{K} \|\hat{J}_{r,k} - J_{r,k}^{\text{GT}}\|_2,$$

where $\hat{J}_{r,k}$ and $J_{r,k}^{\text{GT}}$ are the root-centered predicted and ground-truth positions of joint k for person r, and K is the number of joints. MPJPE is reported in millimetres.

PA-MPJPE$_{\text{per}}$. Procrustes-aligned MPJPE applies a rigid alignment between each predicted and ground-truth skeleton before computing joint error, removing the effect of global rotation and scale. This makes it a measure of pure body shape and pose accuracy, independent of how the person is oriented in the scene.

PA-MPJPE$_{\text{2p}}$. To assess the accuracy of the relative interaction configuration rather than individual poses in isolation, we follow BUDDI [26] and compute a two-person PA-MPJPE by concatenating the joint positions of both people in their original coordinates into a single vector, then applying Procrustes alignment jointly to the concatenated set:

$$\text{PA-MPJPE}_{\text{2p}} = \text{PA-MPJPE}\Big([\hat{J}_{\text{t}}; \hat{J}_{\text{c}}],\ [J_{\text{t}}^{\text{GT}}; J_{\text{c}}^{\text{GT}}]\Big).$$

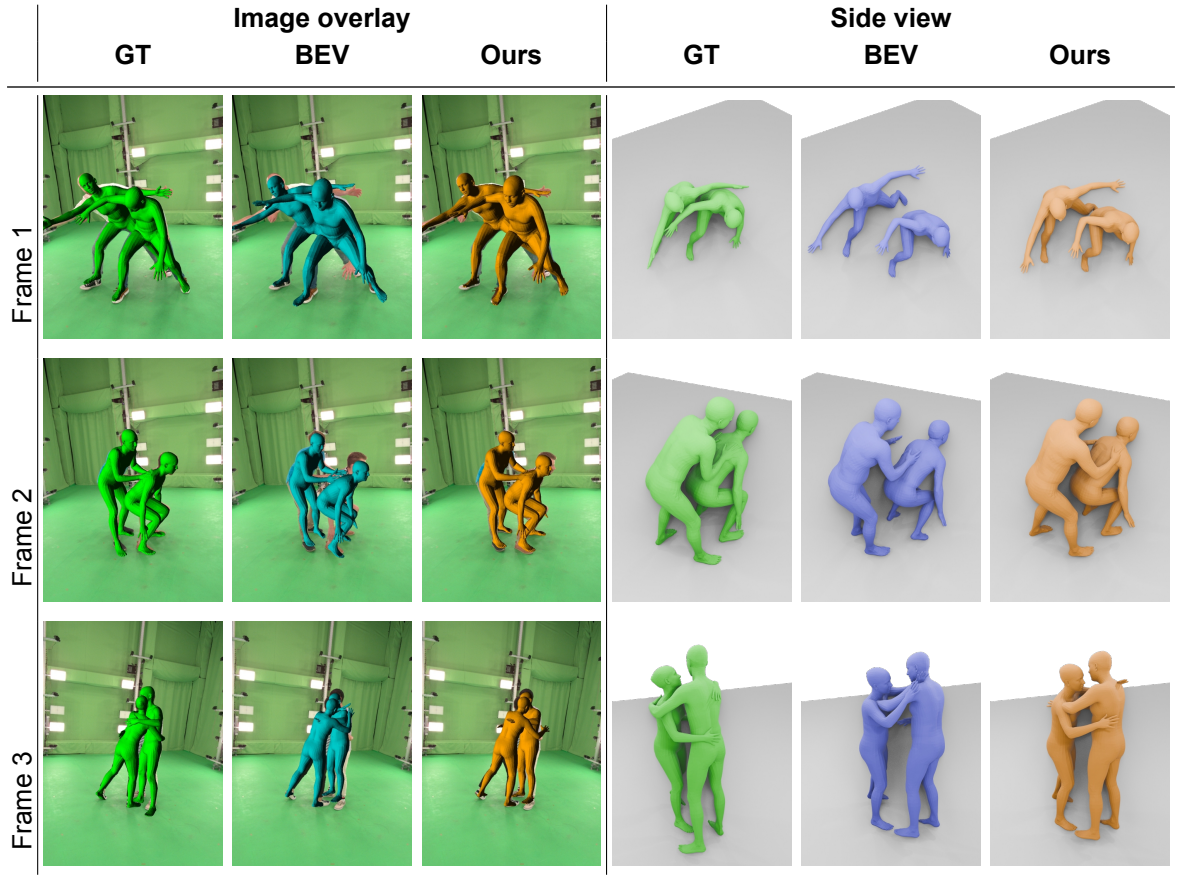


Figure 4.5: Qualitative comparison of image-to-mesh refinement on three Hi4D test frames (rows) where the interaction prior has a visible effect on the recovered spatial configuration. *Left group:* SMPL-X mesh overlays on the input image for the ground truth (GT), BEV initialization, and BEV refined with OpenPose and our interaction prior. *Right group:* corresponding side-view Blender renders, which show the difference in inter-person distance and relative orientation that are not visible in the frontal view.

No root alignment is applied before the joint Procrustes step, so the metric keeps information about the relative placement of the two people. Therefore, it penalizes errors in inter-person distance and relative orientation that per-person metrics cannot capture, such as two people reconstructed in correct individual poses but placed at an incorrect distance or facing direction.

Contact PCC. MPJPE-based metrics do not directly measure whether the reconstructed bodies are in contact where the ground truth indicates they should be. We therefore report precision, recall, and F1 score based on surface proximity, computed over the contact vertices of both bodies combined. We define a binary contact mask over the SMPL-X vertices of both people: a vertex is considered to be in contact if its distance to the other person's surface is below $r = 2$ cm. Applying the same threshold to the predicted meshes gives a predicted contact mask, and we measure how well the two masks agree using precision, recall, and F1:

$$\text{Precision} = \frac{|C_{GT} \cap C_{pred}|}{|C_{pred}|}, \quad \text{Recall} = \frac{|C_{GT} \cap C_{pred}|}{|C_{GT}|},$$

$$\text{F1} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}.$$

where C_{GT} and C_{pred} denote the sets of contact vertices in the ground truth and prediction respectively.

4.3.1. Ablation: Prior Without a Strong Initialization

The main comparison above starts from a BEV estimate, which already provides a reasonable initial pose and placement for both people. To test whether the interaction prior contributes beyond leveraging

a good initialization, we run an additional ablation in which optimization starts from a deliberately weak, uninformative state.

Template initialization. Both people are placed at the origin with zero relative translation between them, each initialized to a neutral T-pose. The only scene-aware information used at this stage is a coarse estimate of body orientation: we inspect the OpenPose shoulder detections to determine whether each person's left and right shoulders appear in the expected image order, and flip the global yaw by 180° if the order suggests the person is facing away from the camera. Starting from this initialization, both the global translations and body poses of both people are highly ambiguous: both people start at the same position in 3D space and there is no prior on their relative placement. This setting therefore isolates the contribution of the interaction prior from the quality of the initialization, and tests whether the prior can recover plausible two-person configurations even when depth and relative placement are very under-constrained from the 2D evidence alone.

We compare two configurations in this setting:

- **OpenPose + VPoser:** SMPL-X parameters are optimized to minimize 2D keypoint reprojection error with an additional single-person VPoser regularization term E_{vp}, enforcing anatomical plausibility per person independently. No interaction constraint is applied, so relative placement is determined entirely by the 2D evidence.
- **OpenPose + prior (ours):** optimization with the full interaction prior E_{prior} and VPoser regularization, providing both single-person plausibility and two-person interaction constraints.

5 Discussion

The central question this thesis investigates is whether a conditional normalizing flow can serve as a practical interaction prior for close human-human interaction: one that captures the multimodal structure of the pose space, provides an exact and tractable likelihood for plausibility scoring, and acts as a differentiable regularizer in image-to-mesh optimization. The three experiments address each of these roles in turn. The results confirm the feasibility of the approach while also revealing its limits, most notably the expressiveness gap relative to diffusion models in generation and the dependence on training-data coverage in plausibility scoring. We discuss the results of each role, examine how well the prior transfers across datasets, and then address broader limitations.

5.1. Generative Quality and Multimodality

The primary question this experiment asks is whether a normalizing flow can capture the multimodal structure of human-human interaction, producing qualitatively distinct reactions for the same reference pose rather than collapsing to a single averaged output. The results (Table 4.1) confirm that it can: our CNF achieves the highest sample diversity of all models (APD 0.326 vs. 0.165 for DH-like, 0.167 for BUDDI), and Figure 4.1 shows that samples span qualitatively distinct spatial configurations and poses for the same conditioning pose.

By contrast, the DH-like baseline maps each reference pose to a single Gaussian, so repeated sampling produces a tight cluster around one mean reaction (APD 0.165, Pose KID 0.262). The CNF is not constrained to a single parametric family per input and can therefore place probability mass in multiple disconnected regions of the distribution.

BUDDI achieves better realism on both KID metrics: Pose KID 0.052 vs. 0.211 for CNF, and Interaction KID 0.106 vs. 0.269 (the latter only comparable between CNF and BUDDI, as the other models do not predict spatial layout). This means BUDDI samples sit closer to the data distribution in both pose and spatial arrangement. Diffusion models compose many small denoising steps and can represent highly complex distributions, making them among the most expressive generative model families. An affine-coupling flow, by contrast, learns a single bijective mapping from base distribution to data, and how well it approximates the data manifold is limited by the capacity of the coupling networks. The practical consequence is that the CNF fits the training distribution less tightly than BUDDI. The tradeoff is the exact likelihood: diffusion models do not provide a closed-form density, whereas the CNF evaluates one in a single forward pass.

BUDDI's low diversity also comes partly from how we sample it, not just from the model itself. To draw a conditional reaction we clamp person A back to ground truth after every denoising step, which keeps pulling person B's trajectory toward configurations that fit a fixed A and narrows the range of poses it explores. The CNF is conditional by construction and never faces this pull, so the comparison is tilted in our favour. Inpainting is still the reasonable thing to do here, since it is the standard way to get conditional samples from a jointly trained diffusion model without retraining it into something else, but a diffusion model trained directly to predict person B from person A would avoid the clamping and show

how much of BUDDI's diversity gap is real. We leave that comparison to future work.

The pose-only CNF ablation (APD 0.224, Pose KID 0.072) helps separate two effects. It already outperforms the DH-like baseline on diversity (APD 0.224 vs. 0.165) without any spatial modeling, showing that the CNF architecture itself is the primary driver of the diversity gain. Adding the spatial components $(t_{\mathrm{rel}}, r_{\mathrm{rel}})$ increases diversity further (APD 0.224 to 0.326) but substantially worsens Pose KID (0.072 to 0.211), even though both models share the same CNF architecture.

5.2. Plausibility Scoring

The plausibility scoring results (Table 4.2) show that the learned density is a meaningful discriminative signal, and reveal important differences between the models in how they handle pose versus spatial plausibility.

BUDDI scores near chance on every task (AUC 0.530 on mismatched, 0.498–0.536 across noise levels), and is excluded from further comparison below. BUDDI provides no closed-form likelihood and was not designed for direct plausibility scoring; the denoiser-based surrogate used here is a rough approximation, and how to best adapt diffusion models for this use case remains an open question.

Noisy pose discrimination. The noisy task is the primary controlled comparison: spatial configuration is held fixed, so performance differences reflect only the quality of the learned conditional pose distribution. The most discriminative model on this task is the pose-only CNF ablation, which achieves AUC 0.711 / pairwise 0.901 at $\sigma{=}0.02$ and reaches AUC 0.998 / pairwise 1.000 at $\sigma{=}0.15$ (Table 4.2). Our full CNF trails the pose-only ablation at low noise (AUC 0.635 vs. 0.711 at $\sigma{=}0.02$), and the two converge at higher noise levels where both approach perfect discrimination.

The noisy task perturbs only body pose while leaving $(t_{\mathrm{rel}}, r_{\mathrm{rel}})$ unchanged. The full CNF evaluates a joint log-density over pose and spatial configuration; because the spatial dimensions are identical in real and noisy pairs, they carry no discriminative information and dilute the pose signal. The pose-only model concentrates all its discriminative capacity on the pose dimensions alone, which is precisely what the task requires.

The DH-like baseline improves more gradually with noise (AUC 0.510, 0.563, 0.677, 0.796 at $\sigma \in \{0.02, 0.05, 0.10, 0.15\}$), confirming that a Gaussian-to-Gaussian mapping can distinguish clean from corrupted poses at high noise but struggles when the corruption is subtle. Pairwise accuracy measures something different: given two poses side by side, which is cleaner. At $\sigma{=}0.15$ the DH-like baseline reaches 0.917, showing that even a coarser model can make the right call when the corruption is large enough. AUC captures the harder ability: judging whether a single pair is realistic, not just picking the better of two options. The CNF models consistently outperform the baseline across all noise levels, as shown in the ROC curves in Figure 4.2.

Mismatch discrimination. On mismatched pairs, our full CNF achieves AUC 0.722 and pairwise accuracy 0.727, ahead of both pose-only models. The pose-only ablation reaches AUC 0.678 and pairwise 0.667, and the DH-like baseline performs worse still, at AUC 0.623 and pairwise 0.633. Mismatched pairs differ in both body pose and spatial placement, so a pose-only model cannot detect a spatial mismatch and both score below the full CNF. The pose-only ablation still beats the DH-like baseline, consistent with the noisy task. This reflects the CNF's greater expressive power. The DH-like baseline maps one person's pose to a single Gaussian over the other's, whereas the CNF can represent a multimodal density. Since a reference pose allows several plausible reactions, a unimodal Gaussian must average over them, placing mass on implausible intermediate poses and blurring the real-versus-corrupted boundary that the CNF keeps sharp. The two gaps, full CNF to pose-only and pose-only to DH-like, are similar in size, so the full CNF's advantage comes from two roughly equal contributions: the added spatial signal, and the CNF's improved pose modelling already visible in the pose-only comparison.

Qualitative behaviour and failure cases. Figure 4.3 illustrates the expected behaviour: ground-truth pairs consistently score higher than their noisy counterparts, and scores decay as noise increases. Frame 3 in that figure exposes a failure case. The ground-truth interaction, one person leaning over

another who is seated on the floor, receives a lower log-probability than both mismatched pairs. This happens because that type of pose is rare in the training data: the model has seen few similar sequences and assigns low probability to it, while the mismatched pair happens to resemble more common interaction patterns. The model scores what it was trained on, so interactions underrepresented in the training data will fall outside the model's learned distribution and receive low scores.

Figure 4.4 provides complementary qualitative evidence that the learned density is geometrically meaningful. Holding the reference person fixed and rotating person B by increasing angles around the vertical axis, the model assigns the highest log-probability near the ground-truth orientation and lower probabilities as the rotation departs from it, forming a smooth density peak around the plausible interaction configuration. This confirms that the prior encodes not just which individual poses are plausible, but how the two people should be oriented relative to each other.

5.3. Application to Image-to-Mesh Optimization

The image-to-mesh results (Table 4.3) show that the interaction prior consistently improves relative placement accuracy, while its effect on per-person pose is limited. The placement improvement is the expected outcome. The limited effect on per-person pose was less expected: one motivation for using an interaction prior was that interaction semantics could help resolve occluded poses, but this did not produce measurable gains.

Relative placement. The two-person PA-MPJPE measures Procrustes alignment over the concatenated skeletons of both people without root-centering, penalizing errors in inter-person distance and relative orientation that per-person metrics cannot capture. It improves across refinement steps: 93.16 (BEV) $\rightarrow$ 89.65 (BEV + OpenPose) $\rightarrow$ 87.25 (BEV + OpenPose + prior). The consistent improvement at each stage confirms that 2D keypoint guidance first corrects relative placement, and the interaction prior then refines the residual configuration errors that the keypoints cannot resolve due to depth ambiguity and occlusion. Figure 4.5 shows selected frames where the effect is most visible; the changes are subtle, but the side views confirm that the prior shifts relative depth and orientation in the right direction.

Individual pose accuracy. PA-MPJPE$_{\text{per}}$ shows no meaningful change across conditions (50.41 $\rightarrow$ 52.73 $\rightarrow$ 51.83). BEV already provides individual pose estimates competitive with other monocular baselines on this benchmark, leaving little room for improvement, and the prior's primary contribution is relative placement rather than individual poses.

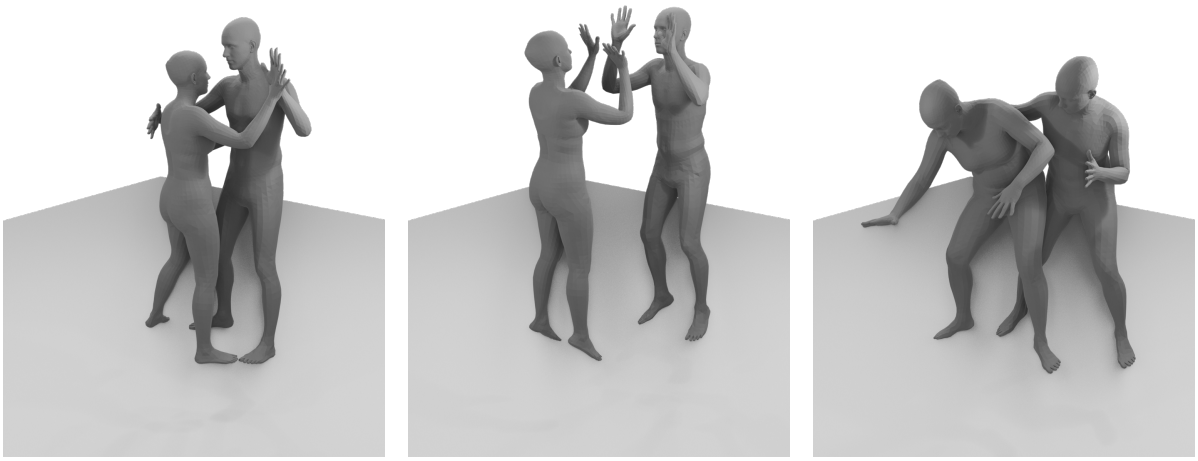
Contact metrics. Contact precision, recall, and F1 are low across all methods and differences are small. The main recall improvement comes from the OpenPose refinement step (0.134 $\rightarrow$ 0.185) rather than from adding the interaction prior (0.185 $\rightarrow$ 0.187), suggesting that 2D keypoint guidance is the primary driver of contact recovery. The prior operates in VPoser latent space and models global relative placement, which is too coarse to resolve precise contact geometry. Explicitly incorporating contact into the interaction prior, for example by conditioning on or jointly modeling contact regions, could be a useful direction for future work.

The role of initialization. The template-initialization ablation isolates the prior's contribution from the quality of the BEV starting point. Starting from a zero-relative-translation T-pose, adding the prior reduces per-person PA-MPJPE from 104.90 to 92.58 and two-person PA-MPJPE from 179.58 to 165.17 (Table 4.3). These are larger gains than in the BEV-initialized setting (two-person PA-MPJPE improvement of 89.65 $\rightarrow$ 87.25), confirming that the prior contributes more when the starting point is weak and relative placement is highly ambiguous. That said, absolute errors from the template start remain far above the BEV-initialized results, highlighting that initialization is critical in gradient-based optimization: the prior can guide refinement but cannot recover from a poor starting point.

5.4. Cross-Dataset Generalization

To test how well the learned prior transfers to a different dataset, we trained the CNF on Inter-X and evaluated it on Hi4D by computing log-likelihoods for ground-truth Hi4D interaction pairs. Despite these

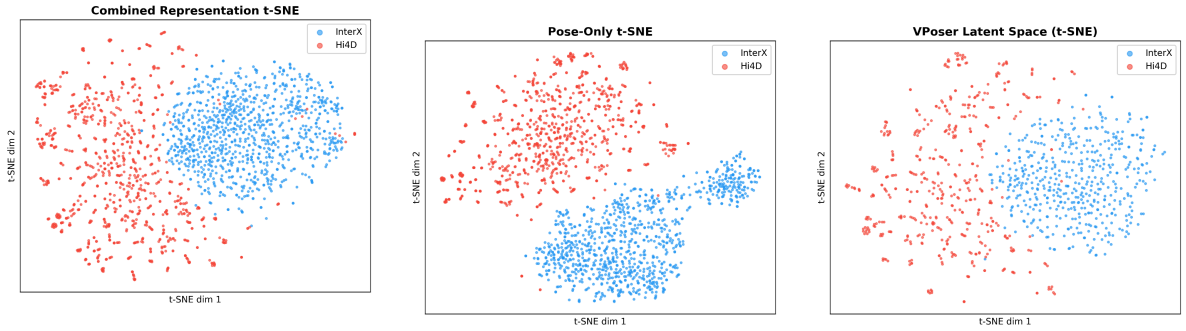
being real, plausible interactions, the model assigned them extremely low probability, well below the scores the same model assigns to held-out Inter-X test frames, as illustrated in Figure 5.1.



(a) Frame 1. Log-prob: -107.4 **(b)** Frame 2. Log-prob: -45.8 **(c)** Frame 3. Log-prob: -42.1

Figure 5.1: Three randomly selected Hi4D ground-truth frames with their log-likelihoods assigned by the Inter-X-trained CNF. Despite being real interactions, all three receive extremely low scores, indicating that the model does not generalize across the domain gap between the two datasets.

To investigate whether this reflects overfitting to Inter-X or a deeper difference in the data, we applied t-SNE [24] to our full interaction representation (VPoser latents of both people combined with relative translation and rotation) and found that the two datasets form two almost perfectly separable clusters (Figure 5.2a). This separation persists even when reducing the representation to raw joint angles alone (Figure 5.2b) and carries through into VPoser latent space (Figure 5.2c). This points to a systematic difference in the underlying pose data rather than the model simply memorizing Inter-X.



(a) Combined representation (VPoser + relative pose). **(b)** Raw joint angles only. **(c)** VPoser latent space.

Figure 5.2: t-SNE projections of Inter-X and Hi4D samples across three representations. In all three cases the datasets form near-perfectly separable clusters, indicating a systematic domain gap in SMPL-X pose parameters.

We believe the primary source of this gap is the difference in SMPL-X fitting pipelines between the two datasets. Recovering SMPL-X parameters from joint positions is a one-to-many problem: rotations at a parent joint in the kinematic chain can be cancelled by compensating rotations at child joints while keeping 3D joint locations identical. Because this mapping is one-to-many, the particular solution an optimizer converges to depends heavily on its initialization, regularization and objective function, meaning two fitting pipelines can produce very different SMPL-X parameters for the same 3D joint positions. We observed that some fitted poses contain joint rotations along axes that are anatomically implausible, introduced as a byproduct of the optimization rather than reflecting actual body motion. Since VPoser operates on SMPL-X parameters rather than joint positions directly, these systematic fitting differences

place the two datasets in distinct regions of the VPoser latent space, making them appear far apart to the CNF even when the underlying interactions are visually similar.

5.5. Limitations

Interpenetration and contact. Our proposed method has no explicit geometric constraint preventing body mesh overlap in generated samples. Training on real data reduces this in practice but does not eliminate it, a differentiable interpenetration penalty during training might be a good direction to explore. More broadly, the prior operates over VPoser latents and relative transforms, capturing global pose and placement but not fine-grained surface contact. It can bring bodies into proximity but cannot reason about which specific surfaces should be touching, which is a meaningful gap for close interactions where contact is a defining feature. Explicitly incorporating contact into the prior representation is a potential direction for future work.

Representation. The prior is conditioned on VPoser latents, which encode pose but not body shape. Differences in height and build affect how people physically interact, and these are not captured by the current representation. Conditioning on shape parameters alongside pose latents would be a natural extension, but is difficult in practice since the existing HHI datasets have limited body shape diversity, making it hard to learn a shape-aware interaction prior from data alone.

Interaction semantics. The model treats all interaction types uniformly: the learned distribution averages over hugs, handshakes, dances, and other categories in the training data. A class-conditional normalizing flow could produce more precise distributions per interaction type. More broadly, both Inter-X and Hi4D cover a finite set of interaction actions recorded in controlled studio settings, and how well the prior generalizes to unseen interaction categories or in-the-wild poses remains untested. Real-world interactions involve greater variability in clothing, lighting, occlusion, and body configuration than what is present in either dataset, and the prior may not transfer reliably to such conditions.

6

Reflection

This chapter situates the work in two broader contexts: the long-standing tension between expressiveness and tractability in generative modeling, and the practical and ethical implications of deploying the system in real-world settings.

6.1. Expressiveness vs Tractability in Broader CS and ML Field

Generative modeling in machine learning has long been shaped by the tension between the flexibility of a model and its ability to support exact probabilistic inference. Early probabilistic models such as graphical models [19], mixture models [20], and probabilistic circuits [5, 18, 25] prioritized tractability, enabling exact computation of marginals, conditionals, and likelihoods, but at the cost of limited expressiveness in high-dimensional spaces [32]. The rise of deep learning shifted this balance: GANs [10], VAEs [16], and most recently diffusion models [12] achieve impressive generative quality by abandoning exact inference, instead relying on adversarial objectives, variational lower bounds, or score-based training objectives where the model learns to predict the gradient of the log-density rather than the density itself [33]. Normalizing flows take a different approach, using invertible neural transformations and the change-of-variables formula to provide exact likelihoods while retaining much of the expressive capacity of deep networks [30, 6].

A key limitation of many high-fidelity deep generative models used in practice, particularly GANs and diffusion models, is that they do not provide tractable exact likelihoods, and thus cannot easily quantify how certain they are about a specific input or output. A diffusion model can generate a plausible-looking reconstruction, but has no direct way to tell whether that reconstruction is highly likely or just one of many equally plausible solutions. Exact likelihoods make this possible: a model that assigns a concrete probability to each configuration can, in principle, rank hypotheses, flag uncertain predictions, and detect when an input falls outside the training distribution. In this work we rely on these properties. Plausibility scoring relies on exactly this property, requiring the model to evaluate how probable a given pose pair is directly, rather than approximating it through sampling. Optimization regularization does not strictly require exact likelihoods, but having them makes the regularization term directly interpretable as a log-probability rather than a surrogate score.

This need for tractable inference is not unique to human pose modeling. In population genomics, GANs and Restricted Boltzmann Machines define joint distributions but lack tractable mechanisms to efficiently and exactly compute the probability of a missing variant given observed ones, making high-quality genotype imputation difficult in practice [4]. Probabilistic circuits, which support exact marginalization, close this gap and produce more accurate artificial genomes than deep generative models on several standard population-genetic metrics [4]. The same tension is now actively driving research in language modeling and image generation. Zhang et al. [41] showed that standard autoregressive language models struggle to enforce global lexical constraints, and addressed this by distilling a tractable probabilistic model that can exactly compute the probability that future continuations satisfy those constraints under the distilled model. In image generation, Liu et al. [21] used probabilistic circuits to steer

diffusion model outputs toward regions of higher likelihood under a tractable model of images, providing exact constrained posterior computations in the probabilistic circuit and using them to guide a diffusion model that cannot perform such inference on its own. In each of these domains, high-capacity generative models produce strong outputs but cannot reason precisely about their own distributions. Bridging generation and inference requires tractability, and that is why the normalizing flow was the right choice for this work.

6.2. Real-World Impact and Responsible Use

6.2.1. Research Prototype and Path to Deployment

The system presented in this thesis is a research prototype. Hyperparameters were selected based on reasonable defaults and limited ablation rather than systematic tuning, the model was trained and evaluated exclusively on studio-captured motion capture data, and no in-the-wild evaluation was performed. This means that while the results are promising within the controlled experimental setting, the model's behavior on real-world video, diverse body types, or unusual interaction geometries is unknown. Any deployment in a downstream application would require retraining on broader data, systematic hyperparameter optimization, and careful evaluation on the target population before conclusions about reliability could be made.

6.2.2. Applications and Beneficiaries

The introduction outlined the primary application domains at a high level, but it is worth reflecting on them more carefully here. In AR/VR telepresence and social avatar animation, the model enables coherent relative placement of two avatars in a shared virtual space, which matters for the sense of presence that makes such applications useful. The stakes are relatively low and the main beneficiaries are the end users, though an interaction prior trained on a limited population may produce configurations that feel spatially unnatural for users whose movement patterns fall outside the training distribution. In motion capture for film, games, and sports analysis, the model addresses a technical bottleneck where close interactions cause severe occlusion and independent fitting fails. The beneficiaries are studios, athletes, and coaches, and the application is unlikely to cause direct harm, though decisions made on the basis of sports analysis (selection, training loads, performance evaluation) do affect athletes who have limited visibility into how automated systems contribute to those decisions.

Physical rehabilitation is a more sensitive domain. A plausibility scorer over two-person interactions could help therapists detect when a patient's movement during an assisted exercise deviates from expected patterns, providing an additional signal where clinical assessment is otherwise subjective. HPE-based scoring systems for single-person rehabilitation have been explored in research [37], demonstrating the potential for automated movement assessment in clinical settings. The interaction prior presented in this thesis extends that concept to two-person settings. This could serve as a useful supporting tool, but over-reliance on it carries real risk. The model produces a statistical score derived from a dataset of healthy, able-bodied subjects performing interactions in controlled studio conditions. A patient in rehabilitation may have movement patterns that are atypical precisely because of their condition, and a system that flags these as implausible could discourage or override clinical judgment that correctly identifies the deviation as acceptable or expected for that individual. Scoring is useful for flagging deviations, but the final call should always rest with the clinician, and that boundary needs to be clearly enforced in any deployment. Responsibility for this lies with the deploying institution and the clinical teams integrating it.

6.2.3. Risks and Stakeholder Considerations

Bias and unequal performance. The model was trained on Inter-X and Hi4D, both of which were captured in controlled studio settings with a limited number of subjects. The population diversity in these datasets is constrained in terms of age, body size, and physical ability, and the range of interaction types and cultural contexts is similarly narrow. A model trained on such data reflects the interaction patterns present there and will generalize less reliably to populations and contexts that are underrepresented. In a rehabilitation or sports analysis context, this means the system may perform worse for exactly the populations that could benefit most from assistive technology, such as elderly patients or people with physical disabilities. This is a well-known problem in human pose estimation more broadly, and it is

more severe in two-person modeling due to the additional scarcity of diverse interaction data.

Privacy. Compared to raw video or image data, operating on 3D body pose representations provides a meaningful privacy benefit: pose parameters capture body configuration and movement rather than facial appearance, skin tone, or other directly identifying visual features. This makes the representation less sensitive than many alternatives, which is an advantage when the goal is to build systems that reason about interaction geometry without requiring access to raw imagery. That said, body pose in context is not fully anonymized. Consider a workplace wellness program where employees participate in motion capture sessions and are told that data is anonymized: if pose sequences and interaction geometry are recorded repeatedly over time, the resulting data can become distinctive enough to re-identify individuals, particularly when combined with metadata such as location or timestamp. Any application that stores or transmits pose data should ensure that participants have consented to the specific use case rather than to data collection in general, and that re-identification risks are assessed before data is shared.

Potential misuse of the plausibility scorer. The most substantive risk concerns the plausibility scoring capability specifically. Within this thesis, it is used to separate real interaction pairs from mismatched or noisy ones and to guide image-to-mesh optimization toward more plausible reconstructions. These are specific, contained uses with clear evaluation criteria. However, a function that assigns a probability score to any observed pair of people based on their pose and relative spatial configuration could be repurposed to flag interactions deemed statistically unusual in a surveillance context. A system that automatically identifies when two people's configuration deviates from the learned distribution could be deployed for crowd monitoring, access control, or institutional oversight, framed as a safety tool while functioning as a behavioral classifier. The people whose interactions are scored in such a deployment would typically have no knowledge that such a system is in use and no ability to contest its outputs. This dynamic is not unique to this work: anomaly detection using learned priors is already deployed in security contexts, and supplementing such a prior with labeled data on interaction types would make the step to a behavioral classifier a short one. Any deployment where scores influence decisions about individuals should require human review before consequential action, clear disclosure to those being assessed, and a process for contesting outputs.

7

Conclusion

We presented a conditional normalizing flow over VPoser latents and relative spatial transforms as an interaction prior for human-human interaction. Unlike prior work that produces implicit likelihoods or collapses to a single predicted reaction, the CNF provides a closed-form, tractable log-density that can be both sampled from and evaluated for any pose pair.

We evaluated the model in three roles. As a generative model, it produces diverse, multimodal reactions to a given reference pose, achieving nearly twice the sample diversity of a Gaussian-to-Gaussian baseline. A diffusion-based model (BUDDI) achieves better sample realism by KID, reflecting the higher expressiveness of diffusion models, but does not provide a tractable likelihood. As a plausibility scorer, the learned density reliably separates real from corrupted interactions, both in pairwise comparisons and globally across thresholds; the spatial components improve detection of mismatched pairs, while pose-only scoring is more sensitive to subtle pose corruptions. As a differentiable prior in image-to-mesh optimization, it consistently improves the relative placement of two-person reconstructions, though the effect on per-person pose accuracy is limited.

The results show that an explicit conditional density over a position-aware latent representation is a practical and versatile tool for reasoning about close human-human interactions that can be used in different downstream tasks without modification. A key limitation is generalization: the prior is sensitive to the SMPL-X fitting pipeline used to produce training data, and does not transfer across datasets with different fitting conventions. Open directions include incorporating contact and body shape into the prior representation, extending to semantic interaction types, and evaluating generalization to in-the-wild settings.

References

[1] Mikołaj Bińkowski et al. “Demystifying MMD GANs”. In: *International Conference on Learning Representations*. 2018. URL: https://openreview.net/forum?id=r1lUOzWCW.

[2] Federica Bogo et al. “Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image”. In: *European conference on computer vision*. Springer. 2016, pp. 561–578.

[3] Zhe Cao et al. “Openpose: Realtime multi-person 2d pose estimation using part affinity fields”. In: *IEEE transactions on pattern analysis and machine intelligence* 43.1 (2019), pp. 172–186.

[4] Meihua Dang et al. “Tractable and expressive generative models of genetic variation data”. In: *bioRxiv* (2023).

[5] Adnan Darwiche. “A differential approach to inference in Bayesian networks”. In: *Journal of the ACM (JACM)* 50.3 (2003), pp. 280–305.

[6] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. “Density estimation using real nvp”. In: *arXiv preprint arXiv:1605.08803* (2016).

[7] Hao-Shu Fang et al. “Alphapose: Whole-body regional multi-person pose estimation and tracking in real-time”. In: *IEEE transactions on pattern analysis and machine intelligence* 45.6 (2022), pp. 7157–7173.

[8] Qi Fang et al. “Capturing closely interacted two-person motions with reaction priors”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024, pp. 655–665.

[9] Mihai Fieraru et al. “Reconstructing three-dimensional models of interacting humans”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2025).

[10] Ian J Goodfellow et al. “Generative adversarial nets”. In: *Advances in neural information processing systems* 27 (2014).

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

[12] Jonathan Ho, Ajay Jain, and Pieter Abbeel. “Denoising diffusion probabilistic models”. In: *Advances in neural information processing systems* 33 (2020), pp. 6840–6851.

[13] Wen Jiang et al. “Coherent Reconstruction of Multiple Humans From a Single Image”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2020.

[14] Angjoo Kanazawa et al. “End-to-end recovery of human shape and pose”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 7122–7131.

[15] Diederik P Kingma and Jimmy Ba. “Adam: A method for stochastic optimization”. In: *arXiv preprint arXiv:1412.6980* (2014).

[16] Diederik P. Kingma and Max Welling. “Auto-Encoding Variational Bayes”. In: *CoRR* abs/1312.6114 (2013). URL: https://api.semanticscholar.org/CorpusID:216078090.

[17] Durk P Kingma and Prafulla Dhariwal. “Glow: Generative flow with invertible 1x1 convolutions”. In: *Advances in neural information processing systems* 31 (2018).

[18] Doga Kisa et al. “Probabilistic sentential decision diagrams”. In: *Proceedings of the Fourteenth International Conference on Principles of Knowledge Representation and Reasoning*. KR’14. Vienna, Austria: AAAI Press, 2014, pp. 558–567. ISBN: 1577356578.

[19] Daphne Koller and Nir Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

[20] B.G. Lindsay. *Mixture Models: Theory, Geometry, and Applications*. NSF-CBMS regional conference series in probability and statistics. Institute of Mathematical Statistics, 1995. ISBN: 9780940600324. URL: https://books.google.nl/books?id=VFDzNhikFbQC.

[21] Anji Liu, Mathias Niepert, and Guy Van den Broeck. “Image Inpainting via Tractable Steering of Diffusion Models”. In: *The Twelfth International Conference on Learning Representations*. 2024. URL: https://openreview.net/forum?id=NSIVHTbZBR.

[22] Matthew Loper et al. “SMPL: A skinned multi-person linear model”. In: *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*. 2023, pp. 851–866.

[23] Junzhe Lu et al. “Dposer: Diffusion model as robust 3d human pose prior”. In: *arXiv preprint arXiv:2312.05541* (2023).

[24] Laurens van der Maaten and Geoffrey Hinton. “Visualizing Data using t-SNE”. In: *Journal of Machine Learning Research* 9.86 (2008), pp. 2579–2605. URL: http://jmlr.org/papers/v9/vandermaaten08a.html.

[25] Radu Marinescu and Rina Dechter. “AND/OR branch-and-bound for graphical models”. In: *Proceedings of the 19th International Joint Conference on Artificial Intelligence*. IJCAI’05. Edinburgh, Scotland: Morgan Kaufmann Publishers Inc., 2005, pp. 224–229.

[26] Lea Müller et al. “Generative proxemics: A prior for 3d social interaction from images”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024, pp. 9687–9697.

[27] Georgios Pavlakos et al. “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*. 2019, pp. 10975–10985.

[28] Georgios Pavlakos et al. “Learning to estimate 3D human pose and shape from a single color image”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 459–468.

[29] Davis Rempe et al. “Humor: 3d human motion model for robust pose estimation”. In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2021, pp. 11488–11499.

[30] Danilo Rezende and Shakir Mohamed. “Variational inference with normalizing flows”. In: *International conference on machine learning*. PMLR. 2015, pp. 1530–1538.

[31] Qing Shuai et al. “Reconstructing Close Human Interactions from Multiple Views”. In: *ACM Transactions on Graphics* 42.6 (Dec. 2023), pp. 1–14. ISSN: 1557-7368. DOI: 10.1145/3618336. URL: http://dx.doi.org/10.1145/3618336.

[32] Sahil Sidheekh and Sriraam Natarajan. *Building Expressive and Tractable Probabilistic Generative Models: A Review*. 2024. arXiv: 2402.00759 [cs.LG]. URL: https://arxiv.org/abs/2402.00759.

[33] Yang Song and Stefano Ermon. “Generative Modeling by Estimating Gradients of the Data Distribution”. In: *Advances in Neural Information Processing Systems*. 2019, pp. 11895–11907.

[34] Yu Sun et al. “Monocular, one-stage, regression of multiple 3d people”. In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2021, pp. 11179–11188.

[35] Yu Sun et al. “Putting people in their place: Monocular regression of 3d people in depth”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 13243–13252.

[36] Garvita Tiwari et al. “Pose-ndf: Modeling human pose manifolds with neural distance fields”. In: *European Conference on Computer Vision*. Springer. 2022, pp. 572–589.

[37] Rahmat Ullah et al. “A real time action scoring system for movement analysis and feedback in physical therapy using human pose estimation”. In: *Scientific Reports* 15.1 (2025), p. 44784.

[38] Christina Winkler et al. “Learning likelihoods with conditional normalizing flows”. In: *arXiv preprint arXiv:1912.00042* (2019).

[39] Liang Xu et al. “Inter-x: Towards versatile human-human interaction analysis”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2024, pp. 22260–22271.

[40] Yifei Yin et al. “Hi4d: 4d instance segmentation of close human interaction”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, pp. 17016–17027.

[41] Honghua Zhang et al. “Tractable control for autoregressive language generation”. In: *Proceedings of the 40th International Conference on Machine Learning*. ICML’23. Honolulu, Hawaii, USA: JMLR.org, 2023.

[42] Yi Zhou et al. “On the continuity of rotation representations in neural networks”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019, pp. 5745–5753.

A

Supplementary Material

A.1. Use of AI tools

I used two AI tools during this project: Claude (Anthropic) and Codex (OpenAI). I used Claude to get quick summaries of papers during the literature search, which helped me decide whether a paper was worth reading in full. I also used it to help find relevant work by suggesting search terms and related topics. When writing, I used Claude to improve the clarity and structure of my text, rewrite awkward sentences, and format tables and figures in LaTeX. I used Codex to help write and debug Python code throughout the implementation. I reviewed the output of all these tools, and all decisions and conclusions throughout the research were still made by me.

A.2. Additional Qualitative Results

A.2.1. Learned Likelihood in Noisy and Mismatched Cases

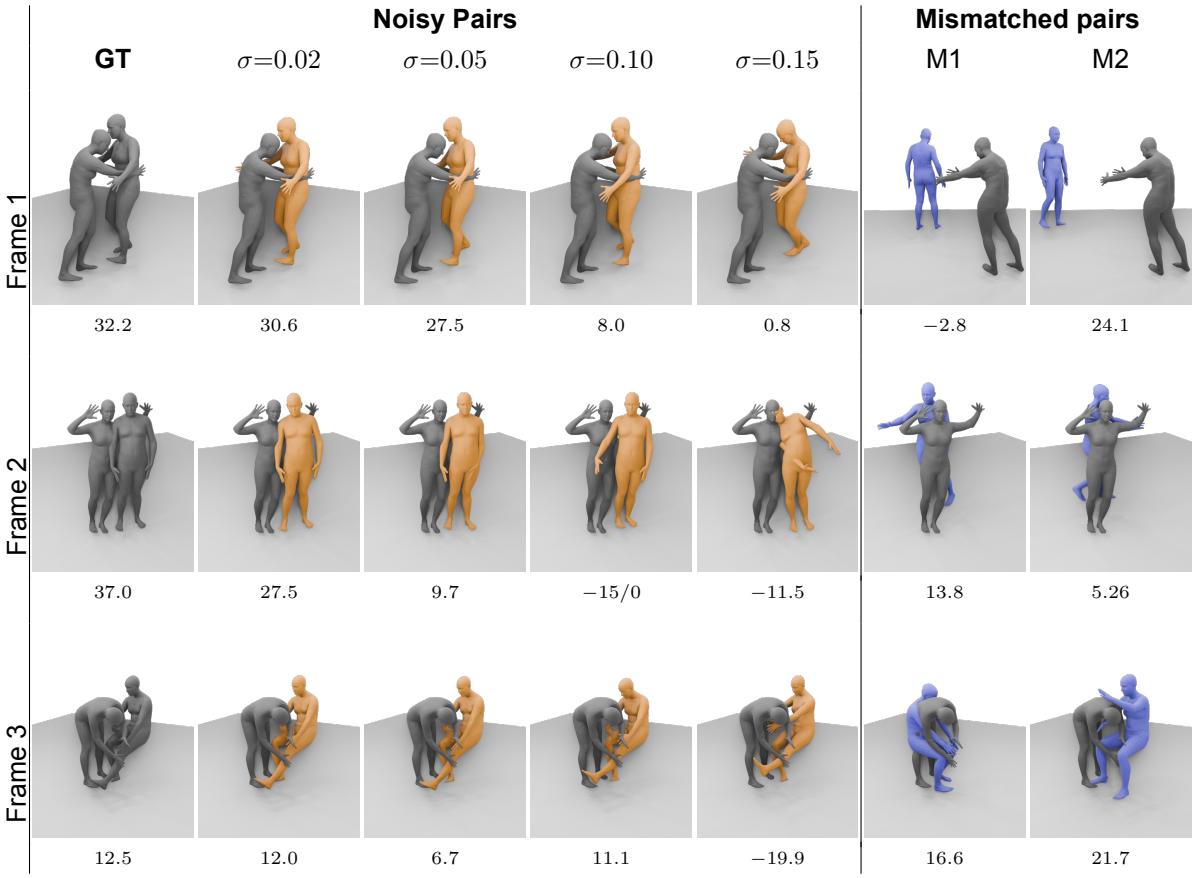
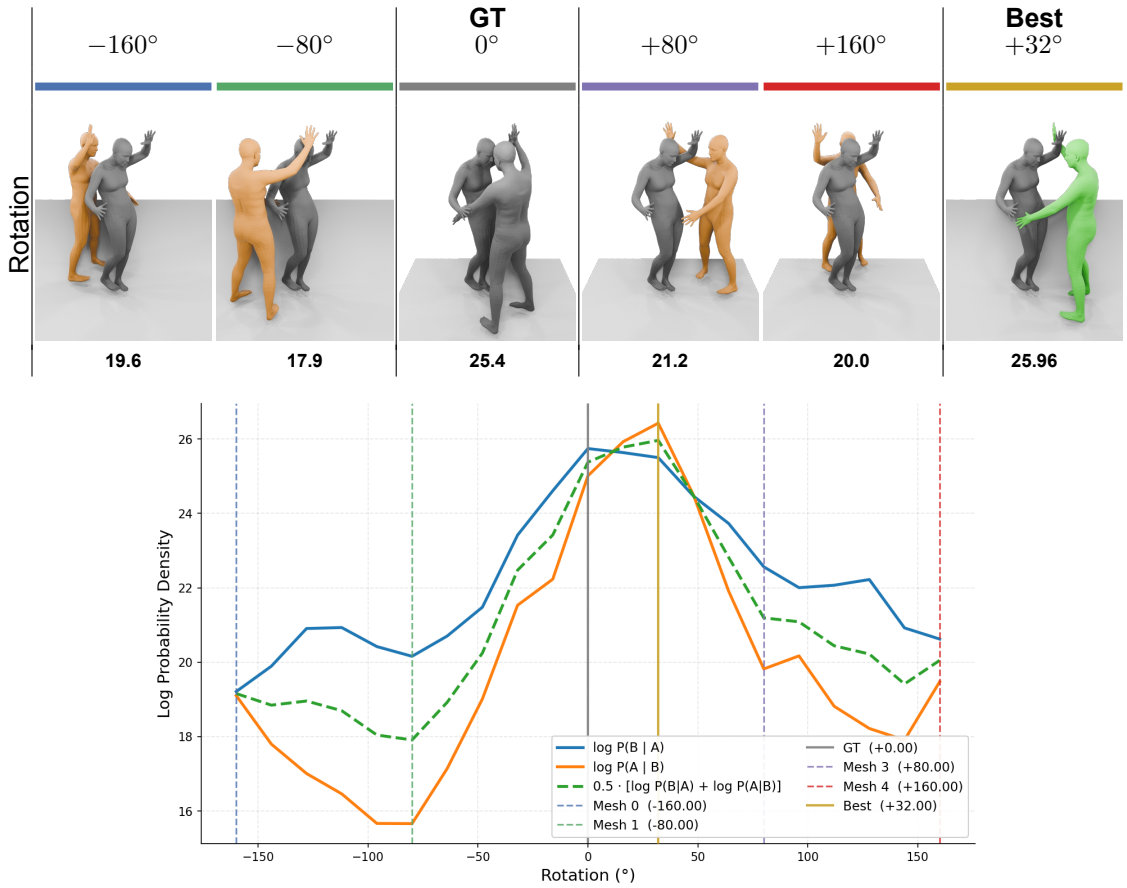


Figure A.1: Additional plausibility scoring examples using our CNF prior (three further test frames). Ground truth is followed by the same pair with increasing Gaussian pose noise added to person B ($\sigma \in \{0.02, 0.05, 0.10, 0.15\}$); the two right-most columns show mismatched pairs. Values below each rendering are the symmetric log-probability $\frac{1}{2}[\log p(B|A) + \log p(A|B)]$ assigned by the model.

A.2.2. Learned Likelihood Under Spatial Transformation



Symmetric log-probability vs. rotation angle

Figure A.2: Additional rotation sweep example. Symmetric log-probability $\frac{1}{2}[\log p(B|A) + \log p(A|B)]$ as person B orbits around person A in equal angular steps. Columns show two orientations before the ground truth (GT) and two after, with the highest-scoring orientation ($+32°$) shown on the right (Best). The bottom panel shows the full probability curve over 20 equally spaced steps.