Robotics: Science and Systems 2026
Sydney, Australia, July 13–17, 2026

HumanFlow – Diffusion-Driven MAV Navigation Among Humans via Tightly-Coupled Motion Tracking, Forecasting, and Control

Simon Schaefer[1,2,3,†] Joshua Näf[4] Stefan Leutenegger[4]
[1]Technical University of Munich [2]MCML [3]MIRMI [4]ETH Zurich

***Abstract*—Robust and accurate perception of humans in their 3D scene context is essential for integrating robots into everyday environments. Existing approaches, however, often fail to predict plausible and accurate human motion estimates that are consistent with the surrounding scene, especially in the presence of heavy occlusions or partial visibility. This can limit both safety and efficiency for robotic operations. We introduce HumanFlow, a latent diffusion model that unifies human motion tracking and forecasting, conditioned on the 3D scene context. We show that our human motion model produces smooth and accurate predictions under challenging conditions, including heavy occlusions, and outperforms state-of-the-art methods in tracking accuracy while being significantly more efficient. Furthermore, we show how HumanFlow's latent space can be tightly coupled with control by conditioning a flow-matching-based, approximate MPC policy on these representations. We validate our policy in simulation with real human trajectories for Micro Air Vehicle (MAV) social navigation, demonstrating superior navigation performance and remaining collision-free, even under partial observability of the human.**

I. Introduction

Robots operating in human-populated environments must navigate safely and efficiently despite the inherent complexity and uncertainty of human behavior. Ensuring safety requires the ability to continuously track and forecast human motion, even under partial observability. While such a task is effortless for humans, it remains a fundamental challenge for robotic systems due to the highly dynamic, articulated, and context-dependent nature of human motion.

Consider a micro aerial vehicle navigating a crowded corridor: a person steps out from behind a pillar, only their shoulder and arm momentarily visible before the robot's own motion sweeps them toward the frame edge. Within milliseconds the planner must infer a full-body state, anticipate where this person is heading, and commit to an avoidance maneuver, all from a partial, rapidly changing observation. Much of the existing literature on perceiving human motion struggles to address conditions that are common in robotic applications, including dynamic camera motion, severe occlusions, and limited human visibility from the robot's viewpoint. Regression-based approaches offer real-time performance but often produce inaccurate or physically implausible motion estimates. In contrast, optimization-based methods are typically more accurate and robust to occlusions, yet are computationally too demanding for onboard deployment.

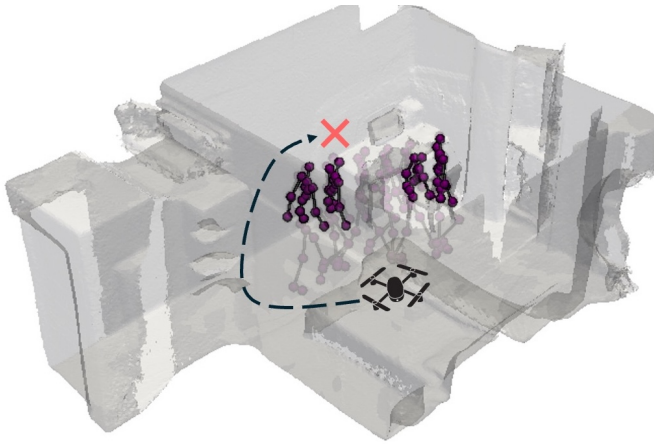

Fig. 1. An MAV navigates a 3D environment containing a human whose motion is partially occluded by furniture. HumanFlow efficiently tracks and forecasts human motion from noisy and heavily occluded observations while conditioning on the surrounding 3D scene using a latent diffusion model. Additionally, HumanFlow compresses human and scene observations into a rich latent representation, enabling a computationally efficient flow-matching-based control policy. The resulting MAV policy is collision-free and enables efficient navigation in cluttered environments despite severe occlusions.

Neither approach incorporates the 3D scene context or has the ability to predict future motion, requiring the use of a separate forecasting model in a robotic use case.

In this work, we address these challenges by introducing HumanFlow, a latent diffusion model for human perception in robotic settings. The model jointly tracks human motion from noisy and occluded observations and forecasts future motion, while being conditioned on the surrounding 3D scene context and operating in real time. To demonstrate the effectiveness of the proposed approach, we apply HumanFlow to the task of social navigation in which an MAV must safely and efficiently navigate among humans. Social navigation in indoor environments exhibits many of the challenges discussed above, including dynamic human motion, rapidly changing viewpoints, and frequent occlusions (see Fig. 1). We show that HumanFlow is not only effective for human motion tracking and forecasting, but also learns a rich latent representation that facilitates learning robust MAV policies. This enables safe social navigation under severe human occlusions while

[†]Corresponding author: simon.k.schaefer@tum.de

remaining computationally efficient for onboard deployment.

Our contributions are the following:

- We propose HumanFlow, a 3D context-conditioned latent diffusion model that jointly performs 3D human motion tracking and forecasting, capturing complex, multi-modal dynamics while remaining robust to noise and occlusions and efficiently representing long-horizon motion in a compact latent space.
- We develop a flow-matching-based MAV policy that leverages the learned latent representation of HumanFlow for proactive, human- and context-aware navigation, enabling real-time control while ensuring safety.
- We show the efficacy of HumanFlow in several benchmarks. It produces smooth, realistic trajectories that align well with the 3D environment and achieves state-of-the-art accuracy. Furthermore, it is robust to high levels of noise and occlusions and is significantly more efficient than existing baselines.
- We demonstrate that using the latent representations of HumanFlow enables our flow-matching-based MAV policy to improve social navigation performance and to remain collision-free, even under challenging conditions with noisy or occluded observations.

II. Related Work

A. Human Motion Tracking and Forecasting

Prior work on 3D human pose and motion recovery largely falls into two camps: regression-based and optimization-based approaches. Regression methods [26, 27, 30, 60, 62] predict pose or motion directly from images or video, and recent work has improved robustness to occlusion and noise [62, 27, 30]. Nevertheless, many regression models struggle with long-term occlusions and typically recover only local motion, which can produce jittery or implausible global trajectories. Optimization-based methods [4, 41, 46, 58, 59] fit parametric body models [33] using temporal priors ranging from hand-crafted smoothness terms to learned motion models. These methods better match observations but are computationally expensive and sensitive to initialization, and rely on short-horizon or pairwise transition priors that break down under long occlusions.

Generative models have recently emerged as a powerful alternative. In particular, diffusion models [7, 23, 24, 50, 56] excel in synthesis [50, 23, 24], pose estimation [15, 18, 17], and forecasting [10] because they can capture long-range spatio-temporal dependencies. RoHM [61] applies coupled diffusion processes to iteratively model global trajectory and local motion, reconstructing globally consistent, complete motions from partial RGB(-D) observations under noise and occlusion. However, its iterative optimization over the full sequence is slow. By contrast, we compress motion into a latent space and employ a single diffusion model in a canonical human-centric space, yielding much faster inference while supporting simultaneous tracking and forecasting.

Scene-conditioned human motion models have mostly focused on forecasting [5, 47, 57, 22], but typically assume a complete scene representation, which is unrealistic in real-time robotic settings. Xing et al. [57] enforce whole-body human–scene coherence by modeling mutual distances between the human and scene and predict future human–scene distance constraints to guide forecasting. This embeds motion into the scene but assumes noise-free input motion and full scene geometry and, as a discriminative model, produces a single deterministic trajectory despite the inherently multi-modal distribution of future motions. In contrast, our approach is generative, models the full distribution of human–scene interactions, explicitly distinguishes occupied, unoccupied, and unobserved scene regions to condition on incomplete scenes, and tolerates noisy past frames.

B. Safe Navigation Among Humans

Safe navigation among humans has been extensively studied for ground robots and, more recently, for aerial platforms. A central challenge is balancing accurate human motion reasoning with computational efficiency, particularly under uncertainty and partial observability.

Early work emphasized modeling human cooperation and intent to avoid overly conservative behavior. Trautman et al. [51] used interacting Gaussian processes to capture cooperative collision avoidance, mitigating the "freezing robot problem" in dense crowds. Sun et al. [48] extended this idea by reasoning over distributions of human preferences rather than single trajectories, enabling richer representations of cooperative behavior. Bai et al. [2] formulated pedestrian interactions as a Partially Observable Markov Decision Process (POMDP), accounting for uncertainty in human intent, while Dragan et al. [12] argued for planning with both physical and cognitive models of humans.

Despite these advances, most cooperative or intention-aware methods assume simplified trajectory models or become computationally prohibitive in highly dynamic, multi-human environments. Safety-critical methods based on Hamilton–Jacobi (HJ) reachability [37, 16] provide formal guarantees but scale poorly, often requiring simplified dynamics. Efficient multi-agent approaches, such as Optimal Reciprocal Collision Avoidance (ORCA) [53, 43], rely on simplified human motion assumptions and are less effective when accurate individual predictions matter. Hybrid strategies that combine learned forecasting with reachability constraints [44] improve realism but remain computationally expensive and are typically restricted to 2D navigation.

Learning-based approaches, including Reinforcement Learning (RL) and inverse RL [8, 13, 29], can implicitly capture human motion patterns and produce socially compliant policies. However, their safety guarantees are probabilistic, they generalize poorly under out-of-distribution conditions, and they are sample-inefficient. Applications to aerial robots [49, 55] have explored drone navigation for human tracking or obstacle avoidance, but these approaches typically assume stationary humans, perfect observations, or static environments, limiting robustness in realistic, cluttered, and dynamic settings.

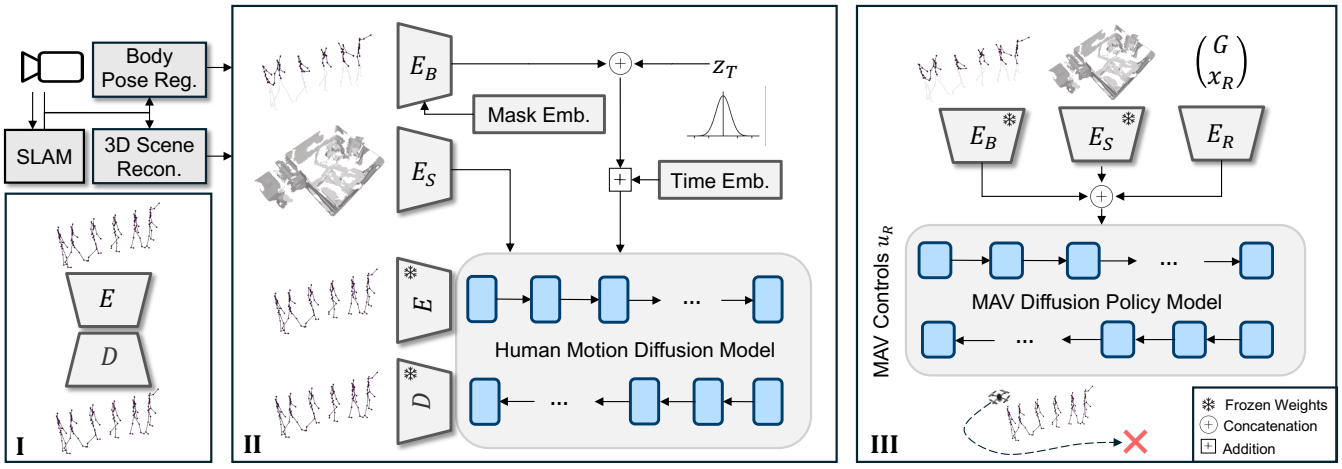

Fig. 2. Overview of HumanFlow. First, clean human motion is encoded into a compact latent space (I). Subsequently, a latent diffusion model is trained to denoise, complete, and forecast human motion sequences, conditioned on scene context and an initial estimate of visible human body frames obtained from video using SLAM and a per-frame human regression model (II). Finally, the latent human and scene embeddings, along with the MAV state and goal, condition a flow-matching policy that outputs collision-free, goal-directed control commands (III).

Recently, HumanHalo [45] proposed an MPC framework for human-aware MAV navigation. Like our approach, it constrains only the initial control input $\mathbf{u}_0$ to prevent the MAV reachable set from fully overlapping with the human reachable set. This avoids overly conservative trajectories while still guaranteeing safety under the assumption of perfect knowledge of the human's current state.

In contrast to relying on expensive optimization or simplified human dynamics models, our work focuses on tightly integrating predictive human motion latents into the control policy via a flow-matching-based formulation. Leveraging these latent representations enables the policy to reason about complex, scene-specific 3D human dynamics without requiring the policy itself to learn these dynamics explicitly, thereby reducing the required model capacity. This enables real-time MAV control, achieving safe, smooth, and efficient navigation without assuming perfect knowledge of human motion or the environment.

III. Preliminaries

A. Coordinate Frames, Transformations, and Notation

Reference coordinate frames are written as $\underrightarrow{\mathcal{F}}_A$; points expressed in frame $\underrightarrow{\mathcal{F}}_A$ are ${}_A\mathbf{r}$. The homogeneous transform from $\underrightarrow{\mathcal{F}}_B$ to $\underrightarrow{\mathcal{F}}_A$ is $\boldsymbol{T}_{AB}$, parameterized by position ${}_A\mathbf{r}_{AB}$ and orientation quaternion $\mathbf{q}_{AB}$. We use two primary frames: the world frame $\underrightarrow{\mathcal{F}}_W$ and the MAV body frame $\underrightarrow{\mathcal{F}}_B$.

B. MAV Model

The MAV state comprises its world-frame position ${}_W\mathbf{r}_B$, orientation $\mathbf{q}_{WB}$, and linear velocity ${}_W\mathbf{v}$:

$$\mathbf{x} = [{}_W\mathbf{r}_B, \mathbf{q}_{WB}, {}_W\mathbf{v}]. \tag{1}$$

The navigation frame $\underrightarrow{\mathcal{F}}_N$, used for control, is obtained by rotating $\underrightarrow{\mathcal{F}}_W$ about its z–axis by yaw ψ and translating its origin to the MAV position, giving:

$${}_N\boldsymbol{r} = \boldsymbol{T}_{NB}\,{}_B\boldsymbol{r} = [x\ y\ z\ 1]^\top, \tag{2}$$

where ${}_N\boldsymbol{r}$ are homogeneous coordinates in $\underrightarrow{\mathcal{F}}_N$.

Following [52, 11], with thrust τ and under the ZYX Euler-angle convention (yaw $\psi \in [-\pi, \pi]$, pitch $\theta \in [-\pi/2, \pi/2]$, roll $\phi \in [-\pi, \pi]$), the MAV dynamics are approximated by a second-order system:

$$\ddot{x} = g\theta - c_x\dot{x}, \tag{3a}$$
$$\ddot{y} = -g\phi - c_y\dot{y}, \tag{3b}$$
$$\ddot{z} = \tau - c_z\dot{z}, \tag{3c}$$
$$\ddot{\theta} = -b_1\theta + b_2\theta^r, \tag{3d}$$
$$\ddot{\phi} = -b_3\phi + b_4\phi^r, \tag{3e}$$

with continuous-time control inputs $\mathbf{u}_t = [\tau, \theta^r, \phi^r]$, reference orientations θ^r, ϕ^r, aerodynamic friction c_x, c_y, c_z, gravity g, and system-identification gains c_i, b_i. Under assumptions of symmetry, small attitudes, and negligible aerodynamic coupling (see [52, 14]), the system is effectively linear and yaw is controlled separately. We assume yaw can track the observed human body independently. Discretizing (3) with sampling time T_s and horizon T yields a time-invariant state-space form:

$$\mathbf{x}_{k+1} = \mathbf{A}\mathbf{x}_k + \mathbf{B}\mathbf{u}_k, \tag{4}$$
$$\mathbf{x}_{0:T} = \boldsymbol{\Phi}\mathbf{x}_0 + \boldsymbol{\Gamma}\mathbf{u}_{0:T-1}, \tag{5}$$

where $\mathbf{A}, \mathbf{B}$ are the state-space matrices and $\boldsymbol{\Phi}, \boldsymbol{\Gamma}$ their stacked forms.

IV. Context-Conditioned Human Motion Model

Inspired by prior work on human motion modeling [50, 19], we represent human motion using a diffusion model, which naturally captures multi-modal future behaviors, supports rich conditioning, and is robust to noisy and incomplete input data. Let $\mathbf{h}_{1:T} \in \mathbb{R}^{T\times J\times 3}$ denote a clean and complete sequence of J 3D body joints, and let $\tilde{\mathbf{h}}_{1:T_{\text{obs}}}$ be a noisy and partially observed version with corresponding binary mask $\mathbf{m}_{1:T_{\text{obs}}} \in \{0,1\}^{T_{\text{obs}}\times J}$ indicating missing observations. Noisy

human joints can be extracted efficiently using regression-based methods such as Multi-HMR [3] and visibility masks from standard 2D keypoint detectors. Scene context can be obtained via 3D reconstruction methods such as [54]. Note that our approach does not rely on completeness or accuracy of these inputs, making it robust to noisy and partial observations.

Given scene context $\mathcal{S}$, our objective is to model the conditional distribution:

$$p_\theta\Big(\mathbf{h}_{1:T} \mid \tilde{\mathbf{h}}_{1:T_{\text{obs}}}, \mathbf{m}_{1:T_{\text{obs}}}, \mathcal{S}\Big), \tag{6}$$

which jointly denoises and completes the observed motion while forecasting plausible future trajectories for $T > T_{\text{obs}}$.

Autoencoder. To efficiently model long temporal horizons, we perform diffusion in a learned latent space. A temporal autoencoder consisting of an encoder $E(\cdot)$ and decoder $D(\cdot)$ maps clean motion sequences to a compact latent representation $\mathbf{z} \in \mathbb{R}^{d \times T'}$ with $T' \ll T$:

$$\mathbf{z} = E(\mathbf{h}_{1:T}), \qquad \hat{\mathbf{h}}_{1:T} = D(\mathbf{z}). \tag{7}$$

In addition to the standard L2 reconstruction loss on joint positions, we apply several regularization terms. A velocity loss encourages temporal smoothness by penalizing differences in predicted and ground-truth joint velocities. A bone length regularization ensures consistent skeletal structure by discouraging changes in bone lengths (see Eqs. (12)). The overall training loss is a weighted sum of all components:

$$\mathcal{L} = \mathcal{L}_{\text{MSE}} + \lambda_{\text{vel}}\mathcal{L}_{\text{vel}} + \lambda_{\text{bone}}\mathcal{L}_{\text{bone}}, \tag{8}$$

where $\lambda_{(\cdot)}$ are weighting hyperparameters.

Diffusion Model. Extending prior work, we adopt a single unified model that performs both human motion tracking and forecasting. This is achieved through the use of mask embeddings, which indicate observed and unobserved portions of the input sequence and allow the same diffusion model to handle denoising, inpainting, and prediction. During training, random temporal masking is applied so that the same model learns to inpaint missing segments and to extrapolate future motion.

Noisy and incomplete inputs are encoded by a separate encoder $E_B(\cdot)$, which shares the same architecture as E. This design encourages the noisy-input encoder to map corrupted observations into a latent space aligned with the latent space of clean data, while allowing it to specialize in handling noise and missing joints during training. Missing joints are padded with the mean SMPL [33] joint configuration, and temporally masked segments are filled using zero-order hold along the sequence. The latent $\mathbf{z}_{\text{noisy}}$ is used as conditioning for the diffusion process by conditioning it on the noise, similar to [25].

$$\mathbf{z}_{\text{noisy}} = E_B(\tilde{\mathbf{h}}_{1:T_{\text{obs}}}). \tag{9}$$

Scene-level context is incorporated through an additional encoder that operates on a human-centric occupancy grid representation of the environment. In contrast to prior approaches that assume a fully known scene [47, 5, 57], we explicitly represent free, occupied, and unknown regions, following modern occupancy mapping frameworks [54]. A scene encoder $E_{\mathcal{S}}(\cdot)$ produces a global scene latent:

$$\mathbf{z}_{\mathcal{S}} = E_{\mathcal{S}}(\mathcal{S}). \tag{10}$$

During training, we obtain incomplete scenes by randomly masking contiguous parts of the input scene. The encoded scene information is injected into the latent diffusion network via Feature-Wise Linear Modulation (FiLM) [40], providing global scene conditioning that biases the generated motion according to the surrounding geometry and uncertainty.

We train the diffusion model using clean state predictions (*x0*-prediction) using DDPM [21] with 100 denoising steps. During inference, we reduce the number of denoising steps to 10. The overall training objective combines a standard diffusion denoising loss $\mathcal{L}_{\text{diff}}$ with several regularization terms similar to those used for the autoencoder. A velocity loss encourages temporal smoothness in the predicted motion. A foot-contact (skating) regularization penalizes unrealistic foot sliding by comparing predicted foot motion to contact labels. Finally, a bone length regularization ensures consistent skeletal structure. The total loss is a weighted sum of all components:

$$\mathcal{L} = \mathcal{L}_{\text{diff}} + \lambda_{\text{vel}}\mathcal{L}_{\text{vel}} + \lambda_{\text{foot}}\mathcal{L}_{\text{foot}} + \lambda_{\text{bone}}\mathcal{L}_{\text{bone}}, \tag{11}$$

$$\mathcal{L}_{\text{vel}} = \frac{1}{J(T-1)} \sum_{k}^{T-1} \sum_{j}^{J} ||\hat{\mathbf{v}}_{j,k} - \mathbf{v}_{j,k}||^2, \tag{12a}$$

$$\mathcal{L}_{\text{foot}} = \frac{1}{J_{\text{foot}}(T-1)} \sum_{k}^{T-1} \sum_{j}^{J_{\text{foot}}} f_{jk}||\hat{\mathbf{v}}_{j,k}||^2, \tag{12b}$$

$$\mathcal{L}_{\text{bone}} = \frac{1}{E(T-1)} \sum_{k}^{T-1} \sum_{j}^{J} \sum_{i}^{J} \delta_{ij}||e_{ij,k} - e_{ij,k-1}||^2, \tag{12c}$$

where $\lambda_{(\cdot)}$ are weighting hyperparameters, $\mathbf{v}_j$ the joint velocity of joint j from numeric differentiation, f_{jk} a foot contact label indicating whether joint j is in contact with the ground at time k, e_{ij} the bone length between joint i and j, δ_{ij} indicating whether joint i is the parent of joint j in the skeleton, and E the number of bones. The motion encoders E_B and E, the decoder D, and the denoising model are implemented as transformer encoders, while the scene encoder $E_{\mathcal{S}}$ consists of 3D convolutions followed by average pooling. For more details about the training data and model architectures, we refer to Section VI and the supplementary materials.

V. Human-Aware MAV Policy

HumanFlow learns a rich latent representation that can be directly leveraged for robot control. To demonstrate its effectiveness, we address the problem of human-aware navigation for an MAV, where the objective is to compute a control policy that drives the MAV toward a goal while avoiding collisions with both humans and the surrounding 3D environment. Formally, at each time step k, the policy Π produces control inputs:

$$\mathbf{u}_k = \Pi(\mathbf{x}_0, \mathbf{g}, \mathcal{H}, \mathcal{S}), \tag{13}$$

where $\mathbf{x}_0$ denotes the initial MAV state, $\mathbf{g}$ the goal position, $\mathcal{H}$ the human motion, and $\mathcal{S}$ the scene. In contrast to prior approaches that rely on explicit parametric representations of human motion and scene geometry, we employ an implicit representation that captures complex human dynamics and human–scene interactions, by using the pretrained encoders E_B and $E_{\mathcal{S}}$ to condition a flow-matching-based MAV control policy.

To generate supervision for policy learning, we utilize the trained human motion diffusion model to sample plausible future human trajectories and solve an offline nonlinear trajectory optimization problem for the MAV. This optimization yields collision-free, goal-directed trajectories that account for both the predicted distribution of human motion and the 3D scene structure.

Finally, the control policy is trained via flow matching to reproduce the distribution of optimized trajectories, conditioned on the MAV state, goal, and latent human and scene representations. Operating in the learned latent space allows the policy to exploit rich information about human behavior and scene structure, enabling robust handling of complex human dynamics and severe occlusions while reducing computational complexity and simplifying policy learning.

The full system architecture, including both the human motion model and the MAV policy, is illustrated in Fig. 2.

Policy Formulation. Given an observed human motion sequence and scene context, we first extract latent representations using the pretrained human and scene encoders to produce $\mathbf{z}_{\text{noisy}}$ and $\mathbf{z}_{\mathcal{S}}$, respectively. Then, these latent features are flattened over time and concatenated with the goal position expressed in the MAV body frame, along with the MAV's initial velocity ${}_B\mathbf{v}_0$, roll ϕ_0, and pitch θ_0. The resulting policy input is:

$$\mathbf{o} = \Big[\text{vec}(\mathbf{z}_{\text{noisy}}),\ \text{vec}(\mathbf{z}_{\mathcal{S}}),\ {}_B\mathbf{g},\ {}_B\mathbf{v}_0\ ,\ \theta_0,\ \phi_0\Big], \tag{14}$$

where $\text{vec}(\cdot)$ flattens the input over time. By operating on these frozen latent features, the policy avoids explicit high-dimensional reasoning over human joint configurations and scene geometry. This enables implementation as a lightweight multilayer perceptron trained using flow matching, supporting efficient onboard inference while maintaining rich contextual awareness. More details on the model architecture can be found in the supplementary materials.

Offline Trajectory Generation. Training data is generated offline using our human motion forecasting model. We sample real human motion sequences from the AMASS [35] and GIMO [63] datasets. We use our human motion model to generate multiple plausible future human trajectories by masking out the last T frames. For each rollout, we randomly sample an initial MAV state $\mathbf{x}_0$ and a goal position $\mathbf{g}$ in the vicinity of the human.

We extend the idea presented in HumanHalo [45] to compute safe but not overly conservative MAV controls by only guaranteeing safety for the initial control input $\mathbf{u}_0$ instead of the full horizon $\mathbf{u}_{0:T-1}$ (see Fig. 3). While HumanHalo [45], however, uses over-approximating reachable sets to represent the distribution of human motion $\mathcal{H}$, we use a set of 10 samples $\{\mathbf{h}^0_{1:T}, \mathbf{h}^1_{1:T}, \mathbf{h}^2_{1:T}, ...\}$ from the human motion model to capture the conditional distribution of future human motion, based on past motion and the scene. Each rollout i defines a scenario. We use a scenario MPC formulation to compute dynamically feasible and collision-free MAV trajectories with respect to each scenario $\mathbf{h}^i_{1:T}$, where we enforce that the initial control action $\mathbf{u}_0$ is equal across all scenarios. Since the optimized initial control action $\mathbf{u}^*_0$ is shared across all scenarios, and each optimized control sequence $\mathbf{u}^{*,i}$ is guaranteed to be collision-free in each scenario, $\mathbf{u}^*_0$ is guaranteed to be safe with respect to the full conditional human motion distribution.

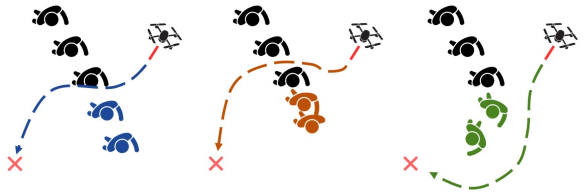

Fig. 3. Exemplary offline ground-truth trajectory generation with three scenarios, each defined by a sampled human motion rollout. The initial control input (red segment) is shared across all scenarios, while the remaining control sequence is collision-free only within its respective scenario.

For all scenarios we minimize the distance between the MAV's state $\mathbf{x}_k$ to the goal $\mathcal{G}$:

$$f(\mathbf{u}^i_{0:T-1}, \mathbf{x}^i_{0:T}) = \sum_{k=0}^{T} \|{}_W\mathbf{r}_{Bk} - \mathbf{g}\|_2^2 . \tag{15}$$

We minimize the objective's expected value across all scenarios i. Formally, we solve the following optimization problem:

$$\begin{aligned}
\min_{\mathbf{u}^i_{0:T-1}} \quad & \mathbb{E}_{\mathcal{H}}[f(\mathbf{u}^i_{0:T-1}, \mathbf{x}^i_{0:T})] \\
\text{s.t.} \quad & \mathbf{x}^i_{k+1} = \mathbf{A}\mathbf{x}^i_k + \mathbf{B}\mathbf{u}^i_k, \quad \forall i \\
& \left\| {}_W\mathbf{r}^{\,i}_{Bk} - \mathbf{h}^i_{j,k} \right\|_2 \geq d, \quad \forall j,\ \forall k,\ \forall i, \\
& \mathbf{u}^i_0 = \mathbf{u}^j_0, \quad \forall i, j \\
& \text{ESDF}(\mathcal{S}, {}_W\mathbf{r}^{\,i}_{Bk}) \geq d, \quad \forall i, \forall k \\
& \mathbf{u}_k \in \mathcal{U}, \quad \mathbf{x}_k \in \mathcal{X},
\end{aligned} \tag{16}$$

where $\mathbf{h}^i_{j,k}$ denotes the position of the j-th human joint at time k in the i-th scenario, d is a predefined safety distance, and $\mathcal{U}$ and $\mathcal{X}$ denote control and state constraints. Whenever 3D context is given, we define $\text{ESDF}(\mathcal{S}, {}_W\mathbf{r}^{\,i}_{Bk})$ to be the euclidean signed distance function of the scene $\mathcal{S}$ at the position ${}_W\mathbf{r}^{\,i}_{Bk}$, and we constrain it to be positive with safety margin d across all scenarios. The constrained non-linear optimization problem is solved using SLSQP [28].

Policy Learning via Flow Matching. Given the optimized MAV trajectories, we train the policy using flow matching [32], inspired by [36]. Let $\mathbf{y}(t)$ denote a continuous-time interpolation between a base distribution $\mathbf{y}(0)$ and the target distribution $\mathbf{y}(1)$, parameterized by time $t \in [0, 1]$. Flow

matching learns a vector field $v_\theta(\mathbf{y}, t, \mathbf{o})$ such that:

$$\frac{d\mathbf{y}(t)}{dt} = v_\theta(\mathbf{y}(t), t, \mathbf{o}), \tag{17}$$

and minimizes the objective:

$$\mathcal{L}_{\text{FM}} = \mathbb{E}_{t,\mathbf{y}(t)} \left[\left\| v_\theta(\mathbf{y}(t), t, \mathbf{o}) - \frac{d\mathbf{y}(t)}{dt} \right\|_2^2 \right]. \tag{18}$$

In practice, $\mathbf{y}(1)$ corresponds to the optimized MAV control sequence $\mathbf{u}^*_{0:T-1} \sim \mathbf{U}^*$. Since each offline scenario shares the same initial control $\mathbf{u}^*_0$, every optimized control sequence $\mathbf{u}^{*,i}_{0:T-1}$ from scenario i is a sample from $\mathbf{U}^*$ and can be used to train the policy.

The policy is trained solely using the flow-matching loss $\mathcal{L}_{\text{FM}}$ on the generated ground-truth control sequences. At inference time, the learned policy produces control commands in real time. We found that 10 denoising steps are sufficient, enabling human-aware MAV navigation that accounts for multi-modal human behavior while remaining computationally efficient for onboard deployment.

VI. Experiments

All experiments within this work have been conducted using the PyTorch framework [38] on a single NVIDIA RTX A4000 GPU. We used the AMASS [35] and GIMO [63] datasets for both training and evaluation.

AMASS. The AMASS dataset [35] provides large-scale, high-quality 3D human motion capture data with pose and shape annotations. We use the SMPL-X neutral body model and downsample all sequences to 10 fps. Following [61], TCD HandMocap, TotalCapture, and SFU are used for evaluation, with the remaining subsets reserved for training.

GIMO. The GIMO dataset [63] is a real-world human–scene interaction dataset. Human motion is captured using an IMU-based motion capture system, while 3D scenes are scanned with a LiDAR-equipped smartphone. GIMO uses the SMPL-X [39] body model. It includes 11 subjects performing diverse actions, such as opening windows or lying in bed, across 19 scenes, resulting in 127 motion sequences comprising 192k frames. We use a random train-test split, with 10% of the data for testing and the remaining 90% for training.

Training Details. All models are trained with the Adam-W optimizer [34]. We use a learning rate of 5e-4 and batch size 256 for the human motion autoencoder and diffusion model, and 4e-5 and batch size 64 for the policy model. All models are trained on sequences of 48 frames, which empirically balance sufficient context and computational cost. Across model training we use $\lambda_{\text{vel}} = 1$, and $\lambda_{\text{foot}} = \lambda_{\text{bone}} = 0.1$.

We adopt a DDPM [21] noise schedule with 100 training timesteps and 10 inference timesteps. The variance schedule is linear with β values ranging from 10^{-4} to 2×10^{-2}. The model predicts denoised samples directly. Masked frames are replaced with the last valid frame before noise. When there is no last valid frame available, it is filled with the mean SMPL [33] joint configuration. This avoids introducing artificial discontinuities and serves as a good a priori estimate.

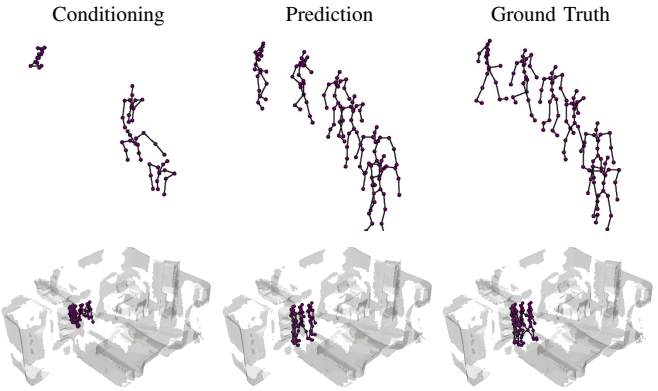

Fig. 4. Qualitative examples of denoised and in-filled body motion with and without scene conditioning. (Upper) The human motion model is conditioned only on noisy observations of the upper body, while the lower body is occluded and several frames are entirely missing. (Lower) The model is conditioned on an incomplete scene. Only the upper body is observed and noisy.

Training Augmentations. During training, we corrupt ground-truth SMPL-X [39] motion sequences by applying noise and masking. Gaussian noise is applied independently to body pose, global orientation, and translation parameters. Rotational parameters are converted to Euler angles, perturbed with zero-mean Gaussian noise with standard deviation 5°, and converted back to rotation vectors. Translation parameters are perturbed using zero-mean Gaussian noise with standard deviation $5cm$. To simulate missing observations, we apply stochastic dropout augmentations. Entire frames are removed with probability 30%, and body-part-specific dropout is applied with probability 20%, masking contiguous body regions rather than individual joints. Additionally, with probability 80%, forecasting augmentation is applied by masking frames at the end of the sequence. The augmentation strength follows a cosine schedule over training iterations, gradually increasing the corruption difficulty.

For the MAV policy, each simulation episode involves random initial and goal states, so the outcome distribution itself is informative. The 25th, 50th, and 75th percentiles of time-to-goal are reported in Table IV. Our method remains fully safe with tightly bounded trajectories across noise levels.

A. Human Motion Tracking and Forecasting

Metrics. We evaluate predicted body joint accuracy using mean-per-joint-position-error without alignment in global frame (GMPJPE). Metrics are reported separately for all, visible, and occluded joints, considering the 22 SMPL [33] body joints. In addition, we assess the physical plausibility of the body motion. We compute the acceleration error (Accel) as the difference in 3D joint accelerations between predictions and ground truth, measured in m/s². Foot skating (Skating) quantifies instances of foot sliding, defined following [59] as cases where all foot joint velocities exceed 10cm/s, toe joints are less than 10 cm above the ground, and ankle joints are below 15 cm. Mean ground penetration (Dist) measures the extent of toe joint penetration into the ground, reported in millimeters, following [41, 46, 61].

Scene-Free Evaluation. To evaluate the robustness to noisy

Input	Noise	Method	GMPJPE ↓ -vis	-occ	-all	Accel↓ [m/s^2]	Skating↓ [·]	Dist↓ [mm]	Runtimes↓ [s]
Occ-L	3	VPoser-t [39]	33.0	242.6	109.2	5.1	0.219	25.91	150
		HuMor [41]	42.4	167.9	88.0	3.3	0.230	2.59	1800
		RoHM [61]	**21.8**	**57.4**	**34.8**	2.3	**0.078**	**0.69**	59
		Ours	27.6	72.1	43.78	**0.58**	0.21	1.55	**0.22**
	5	VPoser-t [39]	43.0	243.1	115.7	7.2	0.179	22.50	150
		HuMor [41]	46.1	163.9	88.9	4.3	0.257	1.81	1800
		RoHM [61]	**31.3**	**66.1**	**44.0**	3.0	**0.105**	**0.69**	59
		Ours	34.3	80.2	50.99	**0.60**	0.17	1.42	**0.22**
	7	VPoser-t [39]	55.1	247.6	125.1	9.4	0.116	18.93	150
		HuMor [41]	70.7	186.2	112.7	5.9	0.269	2.56	1800
		RoHM [61]	**45.6**	**88.9**	**61.3**	4.1	**0.150**	**0.76**	59
		Ours	50.1	92.2	65.0	**0.64**	0.14	1.00	**0.22**
Occ-10%	3	VPoser-t [39]	58.9	136.4	66.4	3.4	0.379	3.12	150
		HuMor [41]	50.0	109.0	55.7	2.6	0.192	0.66	1800
		RoHM [61]	26.3	56.3	29.2	2.3	**0.085**	0.62	59
		Ours	**17.5**	**40.6**	**19.8**	**0.48**	0.10	**0.52**	**0.22**

TABLE I
SCENE-FREE HUMAN MOTION TRACKING EVALUATION ON AMASS [35]. OUR MODEL YIELDS STATE-OF-THE-ART TRACKING ACCURACY AND PHYSICAL PLAUSIBILITY WHILE BEING SIGNIFICANTLY MORE COMPUTATIONALLY EFFICIENT THAN BASELINES.

and occluded data, following [61], we conduct two experiments: (1) motion denoising with lower-body infilling (Occ-L.) and (2) motion denoising with in-betweening (Occ-10%). For a test sequence, (1) masks all lower-body joints to simulate common occlusions, while (2) masks a contiguous 10% subsequence, requiring in-between motion generation. In both cases, Gaussian noise is added to SMPL-X [39] parameters. Noise levels are defined as standard deviations of (k°, k°, k cm) for body pose, root orientation and translation, respectively, which accumulate along the kinematic tree.

We compare our method against several baselines specialized in recovering human motion from noisy and incomplete inputs (tracking scenarios). As shown in Table I, our approach achieves state-of-the-art performance across all tests while being approximately 268× more efficient, as it requires no optimization in the loop and acts in a compressed latent space. This largely reduces the transformer's context length and therefore complexity. Notably, for Occ-10%, our model is 47% more accurate than RoHM [61], despite RoHM's costly dual diffusion and optimization setup, highlighting our model's ability to reason over in-between motion, which is a critical capability for safe MAV navigation when humans are temporarily occluded and non-visible due to the MAV's ego-motion.

Moreover, by predicting body motion in a temporally compressed latent space, our model produces significantly smoother motion, substantially reducing joint acceleration errors across all test sets, by up to 6-fold. This also benefits the MAV policy by preventing sudden jumps in predicted human motion, supporting safer and more stable navigation.

Scene-Conditioned Evaluation. Since the only prior method in this area, HMD2 [19], is not publicly available, we ablate our model's performance with and without scene conditioning on the GIMO [63] dataset. As shown in Table II, incorporating scene information improves tracking accuracy and reduces foot skating.

Method	GMPJPE	Accel	Skat
Ours wo/ Scene Encoder	57.8	**0.5**	0.20
Ours	**51.9**	0.6	**0.12**

TABLE II
SCENE-CONDITIONED MOTION TRACKING EVALUATION ON GIMO [63] DATASET. CONDITIONING ON THE SCENE BOOSTS MODEL ACCURACY AS WELL AS PHYSICAL PLAUSIBILITY.

Qualitative Examples. Fig. 4 and Fig. 5 show qualitative results of our human motion denoising and infilling pipeline across diverse motions, including walking, jumping, sitting, and eating. Despite the variation in motion types, the generated sequences remain diverse while satisfying the imposed constraints, recovering smooth and temporally coherent full-body motion from noisy or occluded inputs, including missing-joint infilling and in-between motion generation.

B. MAV Policy Evaluation

As commonly done in social navigation, we evaluate our method on goal-directed MAV navigation, where it must plan a safe trajectory from an initial state to a goal while avoiding human bodies as dynamic obstacles. We replay 50 AMASS [35] sequences in simulation as ground-truth human motion. Initial states and goals are sampled uniformly from feasible regions, and a planning horizon of 20 steps with a time delta of $0.1s$ is used. For training the policy, we use 10,000 ground-truth trajectories, that have been computed as described in Section V.

Metrics. We assess performance in terms of safety (Collision Avoid.) and efficiency (Time-To-Goal). Safety is measured as the percentage of trajectories maintaining a minimum 0.5 m distance from any human joint. Efficiency captures the elapsed time to reach the goal, penalizing overly conservative behavior.

Baselines. We compare against two classes of baselines. Distance-constraint (DC) methods formulate an MPC-like op-

Method	Collision Avoid. ↑ [%]	Time To Goal ↓ [s]	Collision Avoid. ↑ [%]	Time To Goal ↓ [s]
	No Noise		Noise Level = 5	
DC + Static Human Body	86	4.12	84	5.25
DC + Constant Velocity Model	88	4.14	86	5.29
DC + Forecast-Based Model	86	4.01	96	5.43
RSC + Forward Reachable Set	**100**	4.50	90	5.61
HumanHalo [45]	**100**	4.46	92	5.47
Ours + Past Joints	87	4.30	86	5.34
Ours + Past & Future Joints	92	4.21	91	5.41
Ours + Full Avoidance	**100**	4.45	**100**	5.20
Ours	**100**	**3.80**	**100**	**3.91**

TABLE III
QUANTITATIVE EVALUATION IN SIMULATION OF THE MAV POLICY. OUR POLICY REMAINS SAFE ACROSS ALL SEQUENCES, EVEN IN THE PRESENCE OF HIGH NOISE AND OCCLUSIONS, WHILE YIELDING MORE EFFICIENT TRAJECTORIES COMPARED TO THE BASELINES.

Fig. 5. Qualitative evaluation of the motion tracking model across diverse motion types. The left column shows incomplete and noisy input observations, while the middle and right columns show samples of the reconstructed motion sequences predicted by our model.

timization problem that minimizes a goal-reaching objective while enforcing minimum-distance constraints between the MAV and human joints along a rolled-out trajectory. These methods differ in their assumptions about human motion: (i) *Static*, which fixes human joints to their last observed positions; (ii) *Constant-velocity*, which extrapolates joint motion using an ORCA-inspired instantaneous velocity model [53]; and (iii) *Forecast-based*, which uses predicted future human poses from our motion forecasting model.

Set-based approaches first estimate reachable sets for both the human and the MAV. The *Forward Reachable Set* baseline enforces non-overlap between the two reachable sets over the planning horizon, resulting in conservative behavior.

Finally, we evaluate ablations of our approach. *Ours + Past Joints* replaces the pretrained encoder with past human joint positions expressed in the robot frame while keeping the same flow-matching policy capacity. *Ours + Past & Future Joints* first forecasts a single future human motion rollout and provides it as additional input to the policy network. *Ours + Full Avoidance* follows the same encoder-based conditioning as our method, but instead of constraining a collision-free trajectory per sample during ground-truth generation, only a single $\mathbf{u}_{opt}$ is optimized to avoid all body motion forecasts.

Results. Table III summarizes navigation performance under both noiseless and noisy human motion inputs (noise level 5, as defined in Section VI-A, e.g., regressed by a lightweight human pose estimation model).

Distance-constraint (DC) methods often produce time-efficient trajectories but do not guarantee collision-free navigation. Set-based approaches improve safety by accounting for uncertainty in future human motion, but typically rely on conservative over-approximations that lead to less time-efficient trajectories. Both distance- and set-based methods are particularly sensitive to noisy human motion estimates and may not converge in time, resulting in increased time-to-goal and higher collision rates.

Our ablation studies show that conditioning the policy directly on past joint observations or on a single forecasted trajectory improves navigation efficiency but remains insufficient to ensure safety. In contrast, conditioning the policy on our human motion model's latent representation consistently achieves 100% collision avoidance under both clean and noisy conditions, while maintaining the lowest time-to-goal among

all safe methods. Additionally, learning a policy that constrains only the initial control input for safety further improves time efficiency without compromising collision avoidance.

Inference Details. We follow standard practice in diffusion-based motion (e.g., [50], [59]) and policy models (e.g., [6, 9, 36]). Reducing denoising steps at inference significantly lowers runtime while maintaining quality, which is well-suited to real-time settings where runtime matters more than generation diversity. In the noiseless setting, 10 steps achieve nearly identical performance to 20 steps (100% collision avoidance, 3.80s vs. 3.79s time-to-goal), while fewer steps degrade safety (98% collision avoidance at 7 steps). At 3 steps, we get 100% collision avoidance but with a longer 4.45s time-to-goal. We observe similar trends under noisy conditions (noise level = 5), with 10 steps maintaining 100% collision avoidance. Based on this trade-off between efficiency and consistently safe behavior, we use 10 steps across all experiments in the final setting.

Distribution Shift. We evaluate under complete distribution shift using dancing motion from the AIST++ dataset [31], which neither the human motion model nor the control policy has seen during training. As shown in Table IV, the system achieves 100% collision avoidance with a median time-to-goal of 3.15s on AIST++ dance sequences, demonstrating effective generalization to unseen scenarios without online adaptation. The compressed latent representation captures motion primitives that transfer across motion types.

Method	Collision Avoid. ↑ [%]	Time-to-Goal ↓ [s]		
		25th	50th	75th
	No Noise			
HumanHalo [45]	**100**	3.53	4.46	5.33
Ours	**100**	**3.10**	**3.80**	**5.10**
	Noise Level = 5			
HumanHalo [45]	92	3.46	5.47	6.41
Ours	**100**	**3.05**	**3.91**	**4.80**
	AIST++ Dataset			
Ours	100	1.13	3.15	6.08

TABLE IV
QUANTITATIVE SIMULATION RESULTS OF THE MAV POLICY INCLUDING TIME-TO-GOAL DISTRIBUTION (25TH, 50TH, AND 75TH PERCENTILES) AND OUT-OF-DISTRIBUTION DATASET AIST++ [31].

VII. LIMITATIONS

Human Motion Modeling. While our model captures full-body motion and human–scene interactions, it does not explicitly account for social interactions between multiple humans, where individuals influence each other's motion. Additionally, hand motion is not modeled; incorporating hand dynamics could further improve reasoning about fine-grained scene interactions such as object manipulation or support contacts.

Modeling Human–Robot Interactions. The MAV policy is trained primarily for collision avoidance rather than explicit human–robot interaction, due to the limited availability of data involving close human–MAV encounters. Also, our framework assumes that human motion is unaffected by the MAV's behavior. While this assumption simplifies planning, humans may react to nearby robots and adapt their motion. Modeling such bidirectional human–robot interactions has only been studied for 2D root trajectory generation [42] and remains an open challenge for full 3D body motion. Lastly, we use a simplified linear drone dynamics model for control. Although this enables efficient optimization for generating ground-truth MAV trajectories, our framework allows for efficient generation of controls independent of the underlying MAV dynamics. Thus, extending the policy to more realistic or higher-order dynamics could improve flight agility.

Data Scale and Diversity. Although our method generalizes well across the evaluated benchmarks, datasets capturing human–scene interactions remain limited in scale and diversity. In particular, GIMO [63] includes only a small number of indoor scenes. Expanding training data to more varied environments, including outdoor settings, would likely further improve robustness.

VIII. CONCLUSION

We introduced HumanFlow, a latent diffusion model for 3D scene-conditioned joint human motion tracking and forecasting. We further demonstrated its effectiveness in robotic applications by integrating it into a 3D MAV social navigation policy. In particular, we showed how the latent representations learned by HumanFlow can be leveraged within a flow-matching-based control policy, enabling collision-free and computationally efficient control, even under noisy and occluded observations.

Extensive experiments demonstrate that our human motion model produces smooth and accurate predictions, outperforming state-of-the-art tracking methods at a fraction of their computational cost. When integrated into the navigation pipeline, our approach enables consistently safe and efficient MAV control. The system remains collision-free even under severe noise and partial observability, outperforming strong optimization- and reachability-based baselines.

Our results highlight the importance of jointly reasoning about human motion uncertainty and robot control within a shared latent representation. Future work will focus on extending the framework to richer human–robot interaction scenarios and scaling to more diverse and unstructured environments. While we demonstrate HumanFlow in the context of MAV navigation, the underlying formulation is general and applicable to a broader range of robotic systems, representing a step toward robust, human-aware robotic operation in real-world environments.

ACKNOWLEDGMENTS

This work was supported by TUM Georg Nemetschek Institute under the VAULT-AI project, by TUM AGENDA 2030, funded by the Federal Ministry of Education and Research and the Free State of Bavaria, and by the ETH RobotX research grant funded through the ETH Zurich Foundation. We thank anonymous reviewers for their constructive comments.

References

[1] Abien Fred Agarap. Deep learning using rectified linear units (relu), 2019. URL https://arxiv.org/abs/1803.08375.

[2] Haoyu Bai, Shaojun Cai, Nan Ye, David Hsu, and Wee Sun Lee. Intention-aware online pomdp planning for autonomous driving in a crowd. In *2015 IEEE International Conference on Robotics and Automation (ICRA)*, pages 454–460, 2015. doi: 10.1109/ICRA.2015.7139219.

[3] Fabien Baradel*, Matthieu Armando, Salma Galaaoui, Romain Brégier, Philippe Weinzaepfel, Grégory Rogez, and Thomas Lucas*. Multi-HMR: Multi-person whole-body human mesh recovery in a single shot. In *ECCV*, 2024.

[4] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J. Black. Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In *Computer Vision – ECCV 2016*, Lecture Notes in Computer Science. Springer International Publishing, October 2016.

[5] Zhe Cao, Hang Gao, Karttikeya Mangalam, Qizhi Cai, Minh Vo, and Jitendra Malik. Long-term human motion prediction with scene context. In *ECCV*, 2020.

[6] Hanzhi Chen, Boyang Sun, Anran Zhang, Marc Pollefeys, and Stefan Leutenegger. VidBot: Learning generalizable 3d actions from in-the-wild 2d human videos for zero-shot robotic manipulation. In *CVPR*, 2025.

[7] Xin Chen, Biao Jiang, Wen Liu, Zilong Huang, Bin Fu, Tao Chen, and Gang Yu. Executing your commands via motion diffusion in latent space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18000–18010, 2023.

[8] Richard Cheng, Gabor Orosz, Richard M. Murray, and Joel W. Burdick. End-to-end safe reinforcement learning through barrier functions for safety-critical continuous control tasks, 2019.

[9] Cheng Chi, Siyuan Feng, Yilun Du, Zhenjia Xu, Eric Cousineau, Benjamin Burchfiel, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. In *RSS*, 2023.

[10] Cecilia Curreli, Dominik Muhle, Abhishek Saroha, Zhenzhang Ye, Riccardo Marin, and Daniel Cremers. Non-isotropic gaussian diffusion for realistic 3d human motion prediction. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pages 1871–1882, June 2025.

[11] Georgios Darivianakis, Kostas Alexis, Michael Burri, and Roland Siegwart. Hybrid predictive control for aerial robotic physical interaction towards inspection operations. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 53–58, 2014. doi: 10.1109/ICRA.2014.6906589.

[12] Anca D. Dragan. Robot planning with mathematical models of human state and action, 2017. URL https://arxiv.org/abs/1705.04226.

[13] Michael Everett, Yu Fan Chen, and Jonathan P. How. Collision avoidance in pedestrian-rich environments with deep reinforcement learning. *IEEE Access*, 9: 10357–10377, 2021. ISSN 2169-3536.

[14] Davide Falanga, Philipp Foehn, Peng Lu, and Davide Scaramuzza. PAMPC: Perception-aware model predictive control for quadrotors. In *IEEE/RSJ Int. Conf. Intell. Robot. Syst. (IROS)*, 2018.

[15] Lin Geng Foo, Jia Gong, Hossein Rahmani, and Jun Liu. Distribution-aligned diffusion for human mesh recovery. *arXiv preprint arXiv:2211.16940*, 2023.

[16] David Fridovich-Keil, Sylvia L. Herbert, Jaime F. Fisac, Sampada Deglurkar, and Claire J. Tomlin. Planning, fast and slow: A framework for adaptive real-time safe trajectory planning. *IEEE International Conference on Robotics and Automation*, 2018.

[17] Shubham Goel, Georgios Pavlakos, Jathushan Rajasegaran, Angjoo Kanazawa*, and Jitendra Malik*. Humans in 4D: Reconstructing and tracking humans with transformers. In *International Conference on Computer Vision (ICCV)*, 2023.

[18] Jia Gong, Lin Geng Foo, Zhipeng Fan, Qiuhong Ke, Hossein Rahmani, and Jun Liu. Diffpose: Toward more reliable 3d pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023.

[19] Vladimir Guzov, Yifeng Jiang, Fangzhou Hong, Gerard Pons-Moll, Richard Newcombe, C. Karen Liu, Yuting Ye, and Lingni Ma. HMD2: Environment-aware motion generation from single egocentric head-mounted device. In *International Conference on 3D Vision (3DV)*, March 2025.

[20] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016.

[21] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *arXiv preprint arxiv:2006.11239*, 2020.

[22] Zhiming Hu, Zheming Yin, Daniel Haeufle, Syn Schmitt, and Andreas Bulling. Hoimotion: Forecasting human motion during human-object interactions using egocentric 3d object bounding boxes. *IEEE Transactions on Visualization and Computer Graphics (TVCG)*, pages 1–11, 2024.

[23] Biao Jiang, Xin Chen, Wen Liu, Jingyi Yu, Gang Yu, and Tao Chen. Motiongpt: Human motion as a foreign language. *Advances in Neural Information Processing Systems*, 36, 2024.

[24] Korrawe Karunratanakul, Konpat Preechakul, Supasorn Suwajanakorn, and Siyu Tang. Guided motion diffusion for controllable human motion synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2151–2162, 2023.

[25] Bingxin Ke, Kevin Qu, Tianfu Wang, Nando Metzger, Shengyu Huang, Bo Li, Anton Obukhov, and Konrad Schindler. Marigold: Affordable adaptation of diffusion-based image generators for image analysis, 2025.

[26] Rawal Khirodkar, Shashank Tripathi, and Kris Kitani.

Occluded human mesh recovery. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2022)*, pages 1705–1715, Piscataway, NJ, June 2022. IEEE. doi: 10.1109/CVPR52688.2022.00176.

[27] Muhammed Kocabas, Chun-Hao P. Huang, Otmar Hilliges, and Michael J. Black. PARE: Part attention regressor for 3D human body estimation. In *Proc. International Conference on Computer Vision (ICCV)*, pages 11127–11137, October 2021.

[28] Dieter Kraft. A software package for sequential quadratic programming. Technical Report DFVLR-FB 88-28, Institut für Dynamik der Flugsysteme, Oberpfaffenhofen, July 1988.

[29] Henrik Kretzschmar, Markus Spies, Christoph Sprunk, and Wolfram Burgard. Socially compliant mobile robot navigation via inverse reinforcement learning. *The International Journal of Robotics Research*, 35(11):1289–1307, 2016. doi: 10.1177/0278364915619772.

[30] Jiahao Li, Zongxin Yang, Xiaohan Wang, Jianxin Ma, Chang Zhou, and Yi Yang. Jotr: 3d joint contrastive learning with transformers for occluded human mesh recovery. *ICCV*, 2023.

[31] Ruilong Li, Shan Yang, David A. Ross, and Angjoo Kanazawa. Learn to dance with AIST++: Music conditioned 3d dance generation. *ICCV*, 2021.

[32] Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, and Maximilian Nickel Matt Le. Flow matching for generative modeling. *ICLR*, 2023.

[33] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16, October 2015.

[34] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

[35] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. AMASS: Archive of motion capture as surface shapes. In *International Conference on Computer Vision*, pages 5442–5451, October 2019.

[36] Pau Marquez Julbe, Julian Nubert, Henrik Hose, Sebastian Trimpe, and Katherine J. Kuchenbecker. Diffusion-based approximate MPC: Fast and consistent imitation of multi-modal action distributions. In *IROS*, 2025.

[37] I.M. Mitchell, A.M. Bayen, and C.J. Tomlin. A time-dependent hamilton-jacobi formulation of reachable sets for continuous dynamic games. *IEEE Transactions on Automatic Control*, 50(7):947–957, 2005. doi: 10.1109/TAC.2005.851439.

[38] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: an imperative style, high-performance deep learning library. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, 2019. Curran Associates Inc.

[39] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[40] Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron C. Courville. Film: Visual reasoning with a general conditioning layer. In *AAAI*, 2018.

[41] Davis Rempe, Tolga Birdal, Aaron Hertzmann, Jimei Yang, Srinath Sridhar, and Leonidas J. Guibas. HuMoR: 3d human motion model for robust pose estimation. In *International Conference on Computer Vision (ICCV)*, 2021.

[42] Tim Salzmann, Boris Ivanovic, Punarjay Chakravarty, and Marco Pavone. Trajectron++: Dynamically-feasible trajectory forecasting with heterogeneous data. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pages 683–700, Cham, 2020. Springer International Publishing.

[43] Sepehr Samavi, James R. Han, Florian Shkurti, and Angela P. Schoellig. Sicnav: Safe and interactive crowd navigation using model predictive control and bilevel optimization. *IEEE Transactions on Robotics*, 41:801–818, 2025.

[44] Simon Schaefer, Karen Leung, Boris Ivanovic, and Marco Pavone. Leveraging neural network gradients within trajectory optimization for proactive human-robot interactions. *IEEE International Conference on Robotics and Automation*, 2020.

[45] Simon Schaefer, Helen Oleynikova, Sandra Hirche, and Stefan Leutenegger. HumanHalo - safe and efficient 3d navigation among humans via minimally conservative mpc. In *arxiv*, 2024.

[46] Mingyi Shi, Sebastian Starke, Yuting Ye, Taku Komura, and Jungdam Won. Phasemp: Robust 3d pose estimation via phase-conditioned human motion prior. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14679–14691, 2023. doi: 10.1109/ICCV51070.2023.01353.

[47] Caiyi Sun, Yujing Sun, Xiao Han, Zemin Yang, Jiawei Liu, Xinge Zhu, Siu Ming Yiu, and Yuexin Ma. HUMOF: Human motion forecasting in interactive social scenes. In *The Fourteenth International Conference on Learning Representations*, 2026.

[48] Muchen Sun, Francesca Baldini, Peter Trautman, and Todd Murphey. Move beyond trajectories: Distribution space coupling for crowd navigation. *Robotics: Science and Systems*, 2021.

[49] Rahul Tallamraju, Nitin Saini, Elia Bonetto, Michael Pabst, Yu Tang Liu, Michael Black, and Aamir Ahmad. Aircaprl: Autonomous aerial human motion capture us-

ing deep reinforcement learning. *IEEE Robotics and Automation Letters*, 5(4):6678–6685, October 2020. URL https://ieeexplore.ieee.org/document/9158379.

[50] Guy Tevet, Sigal Raab, Brian Gordon, Yoni Shafir, Daniel Cohen-or, and Amit Haim Bermano. Human motion diffusion model. In *ICLR*, 2023.

[51] Pete Trautman, Jeremy Ma, Richard M. Murray, and Andreas Krause. Robot navigation in dense human crowds: Statistical models and experimental studies of human–robot cooperation. *The International Journal of Robotics Research*, 34(3):335–356, 2015. doi: 10.1177/0278364914557874.

[52] Dimos Tzoumanikas, Wenbin Li, Marius Grimm, Ketao Zhang, Mirko Kovac, and Stefan Leutenegger. Fully autonomous micro air vehicle flight and landing on a moving target using visual–inertial estimation and model-predictive control. *Journal of Field Robotics*, 36 (1):49–77, 2019. doi: https://doi.org/10.1002/rob.21821. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/rob.21821.

[53] Jur van den Berg, Stephen J. Guy, Ming Lin, and Dinesh Manocha. Reciprocal n-body collision avoidance. In Cédric Pradalier, Roland Siegwart, and Gerhard Hirzinger, editors, *Robotics Research*, pages 3–19, Berlin, Heidelberg, 2011. Springer Berlin Heidelberg. ISBN 978-3-642-19457-3.

[54] Emanuele Vespa, Nils Funk, Paul H. J. Kelly, and Stefan Leutenegger. Adaptive-resolution octree-based volumetric SLAM. In *International Conference on 3D Vision (3DV)*, pages 654–662, Québec City, QC, Canada, September 2019. doi: 10.1109/3DV.2019.00077.

[55] Rodolfo Villalobos-Salazar, Abimael Contreras-Carlos, Carlos A. Toro-Arcila, and America Morales-Diaz. Hierarchical drone navigation control for ensuring safety with stationary humans. In *2024 21st International Conference on Electrical Engineering, Computing Science and Automatic Control (CCE)*, pages 1–6, 2024. doi: 10.1109/CCE62852.2024.10771000.

[56] Yin Wang, Zhiying Leng, Frederick W. B. Li, Shun-Cheng Wu, and Xiaohui Liang. Fg-t2m: Fine-grained text-driven human motion generation via diffusion model. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 21978–21987, 2023. doi: 10.1109/ICCV51070.2023.02014.

[57] Chaoyue Xing, Wei Mao, and Miaomiao Liu. Scene-aware human motion forecasting via mutual distance prediction. In *ECCV*, 2024.

[58] Vickie Ye, Georgios Pavlakos, Jitendra Malik, and Angjoo Kanazawa. Decoupling human and camera motion from videos in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023.

[59] Siwei Zhang, Yan Zhang, Federica Bogo, Marc Pollefeys, and Siyu Tang. Learning motion priors for 4d human body capture in 3d scenes. In *International Conference on Computer Vision (ICCV)*, October 2021.

[60] Siwei Zhang, Qianli Ma, Yan Zhang, Sadegh Aliakbarian, Darren Cosker, and Siyu Tang. Probabilistic human mesh recovery in 3d scenes from egocentric views. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

[61] Siwei Zhang, Bharat Lal Bhatnagar, Yuanlu Xu, Alexander Winkler, Petr Kadlecek, Siyu Tang, and Federica Bogo. RoHM: Robust human motion reconstruction via diffusion. In *CVPR*, 2024.

[62] Yi Zhang, Pengliang Ji, Angtian Wang, Jieru Mei, Adam Kortylewski, and Alan L Yuille. 3d-aware neural body fitting for occlusion robust 3d human pose estimation. *ICCV*, 2023.

[63] Yang Zheng, Yanchao Yang, Kaichun Mo, Jiaman Li, Tao Yu, Yebin Liu, Karen Liu, and Leonidas J Guibas. GIMO: Gaze-informed human motion prediction in context. *arXiv preprint arXiv:2204.09443*, 2022.

HumanFlow – Supplementary Material

I. Model Architectures

A. Human Motion Model

Autoencoder. Before encoding, the input is flattened along the joint dimension and processed by two temporal downsampling convolutional blocks. Each block consists of a 1D convolution with kernel size 3 and stride 2, followed by ReLU [1], and a 1D convolution with kernel size 1 followed by ReLU [1]. This reduces the temporal resolution by a factor of four and produces latent features with dimensionality 128. Sinusoidal positional encodings are added and the sequence is processed by a 6-layer transformer encoder with 8 attention heads, feed-forward dimension four times the model dimension, and dropout 0.1. The noisy encoder $E_B(\cdot)$ follows a similar architecture. The latent representation is processed by the decoder D, a transformer encoder with the same configuration as in E. Temporal resolution is restored using two upsampling convolutional blocks. Each block performs nearest-neighbor upsampling by a factor of two, followed by a 1D convolution with kernel size 3 and ReLU [1], and a 1D convolution with kernel size 1 and ReLU [1]. A final upsampling layer followed by a 1D convolution projects the features to the output dimension, producing reconstructed motion with full temporal length.

Scene Encoder. We encode the scene occupancy grid using a 3D convolutional encoder. The encoder consists of three 3D convolutional layers with channel dimensions $1 \to 8 \to 16 \to 32$, each using kernel size 3, padding 1, and ReLU activation. The resulting features are aggregated using global average pooling to obtain a single feature vector per grid. A linear layer then projects the features to a latent embedding of dimension 128.

Diffusion Model. We use a Transformer-based denoising network operating on latent motion representations. The model conditions on diffusion timesteps and encoded context features. Diffusion timesteps are embedded using a learned timestep embedding and added together with sinusoidal positional encodings. The input latent representation is first projected to a hidden dimension of 256 using a linear layer. The denoising network consists of a 10-layer Transformer encoder with FiLM [40] conditioning for the scene, where noisy body latents are concatenated to the input noise. Each layer uses 8 attention heads, feed-forward dimension four times the hidden dimension, GELU [20] activation, and dropout 0.1. The output is projected back to the latent dimensionality using a linear layer.

B. MAV Policy Model

The flow-matching-based MAV policy model is implemented as a Multi-Layer Perceptron (MLP) conditioned on encoded context features. Conditioning sequences are processed by a three-layer MLP with dimensions input_dim $\to 64 \to 32 \to 8$ with ReLU [1] activations, and flattened across time. The processed latents are concatenated with the input noise, the encoded features for the goal position and initial state as well as the denoising timestep. The final prediction network consists of an MLP with layer dimensions inputs $\to 256 \to 128 \to 64 \to 3T$, where T is the prediction horizon. Each hidden layer is followed by a ReLU [1] activation. The network outputs control sequences matching the prediction horizon.

II. Uncertainty Estimates of Reported Metrics

We report mean $\pm$ standard deviation for the AMASS [35] Occ-10% tracking benchmark in Table V, computed over the AMASS [35] evaluation sequences. Our method achieves both the lowest error and tightest bounds across all metrics, confirming robustness.

Input	Noise	GMPJPE-all ↓	Accel ↓	Skating ↓	Dist ↓
Occ-L	3	43.78 ± 4.4	0.58 ± 0.01	0.21 ± 0.09	1.55 ± 0.5
	5	50.99 ± 6.1	0.60 ± 0.02	0.17 ± 0.09	1.42 ± 0.6
	7	65.0 ± 7.8	0.64 ± 0.04	0.14 ± 0.07	1.00 ± 0.5
Occ-10%	3	19.8 ± 5.7	0.48 ± 0.01	0.10 ± 0.04	0.52 ± 0.2

TABLE V
Scene-Free human motion tracking evaluation on AMASS [35] with uncertainty estimates.

III. Model and Loss Function Hyperparameters.

The hyperparameters for the human motion model architecture and the associated loss weights were selected via grid search over a held-out validation split. As an example, Table VI reports an ablation study quantifying the effect of different loss weight settings on model performance, demonstrating the robustness of our chosen configuration.

λ_{vel}	λ_{bone}	GMPJPE-all ↓	Accel ↓	Skating ↓	Dist ↓
0.1	0.01	19.8 ± 5.7	0.48 ± 0.01	0.10 ± 0.04	0.52 ± 0.20
0.1	0.1	22.1 ± 8.8	0.42 ± 0.02	0.11 ± 0.05	0.69 ± 0.29
0.01	0.001	23.2 ± 9.2	0.54 ± 0.01	0.15 ± 0.04	1.40 ± 0.40

TABLE VI
Ablation of human motion tracking on AMASS [35] for the Occ-10% setting across loss weight parameters.