WANDR: Intention-guided Human Motion Generation

Markos Diomataris [1,2] Nikos Athanasiou[1] Omid Taheri[1] Xi Wang[2]
Otmar Hilliges[2] Michael J. Black[1]
[1]Max Planck Institute for Intelligent Systems, Tübingen, Germany [2]ETH Zürich, Switzerland

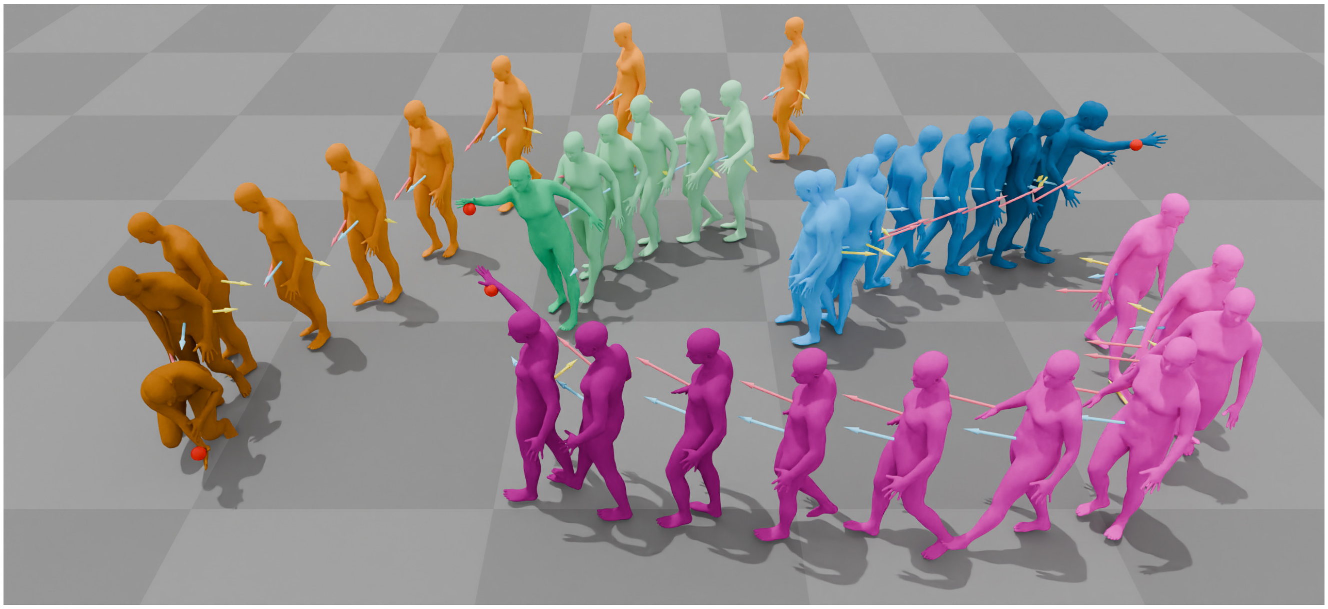

Figure 1. WANDR starts from an arbitrary body pose and generates precise and realistic human motions that reach a specified 3D goal (depicted as a red sphere). Employing a purely data-driven approach, WANDR is a conditional Variational Autoencoder guided by *intention* features (depicted arrows) that steer the human's orientation (yellow), position (cyan) and wrist (pink) towards the goal. WANDR is able to reach a wide range of goals even if they deviate significantly from the training data.

Abstract

Synthesizing natural human motions that enable a 3D human avatar to walk and reach for arbitrary goals in 3D space remains an unsolved problem with many applications. Existing methods (data-driven or using reinforcement learning) are limited in terms of generalization and motion naturalness. A primary obstacle is the scarcity of training data that combines locomotion with goal reaching. To address this, we introduce WANDR, a data-driven model that takes an avatar's initial pose and a goal's 3D position and generates natural human motions that place the end effector (wrist) on the goal location. To solve this, we introduce novel intention features that drive rich goal-oriented movement. Intention guides the agent to the goal, and interactively adapts the generation to novel situations without needing to define sub-goals or the entire motion path. Crucially, intention allows training on datasets that have goal-oriented motions as well as those that do not. WANDR is a conditional Variational Auto-Encoder (c-VAE), which we train using the AMASS and CIRCLE datasets. We evaluate our method extensively and demonstrate its ability to generate natural and long-term motions that reach 3D goals and generalize to unseen goal locations. Our models and code are available for research purposes at wandr.is.tue.mpg.de.

1. Introduction

Goals drive our motions. Even the simplest goal can give rise to intricate motions. Consider reaching for a coffee cup – it can be as straightforward as an arm extension or can involve the coordinated action of our entire body. Actions like bending down, extending our arm, and walking must come together to achieve the goal. At a granular level, we continuously make subtle adjustments to maintain balance and stay on course towards our objective. The result is a fluid motion that seamlessly integrates numerous smaller movements, all converging toward a common and simple goal: placing our hand on the cup. Generating this hierarchy of motions, from the overarching goal to the moment-to-moment individual actions, remains a longstanding challenge in computer vision, graphics, and robotics.

Here we focus on a representative task, illustrated in Fig. 1: given a goal location in space and a starting pose, a humanoid agent must place an end effector (wrist joint) on the goal location while moving in a natural human-like way. To solve the task, the agent needs to be able to approach the goal, orient itself towards it, and reach out such that its wrist makes contact with the goal. Our primary emphasis is on ensuring autonomy for human agents. Consequently, we strive to minimize the guidance information provided, limiting it only to the human's initial pose and the goal's position. Diverging from prior data-driven approaches [4, 17], we choose to refrain from evaluating the model solely on limited labeled data. Instead, we devise an evaluation pipeline that requires agents to reach goals positioned in diverse locations around them. Considering the arbitrary selection of the goal during evaluation, and the minimal guidance information provided, tackling this task is challenging, demanding an approach with the capacity to generalize beyond the distribution of the training dataset.

Existing methods approach this problem either using reinforcement learning (RL) [11, 15, 24, 43] or by capturing task-specific datasets [4, 8, 32]. While RL provides a principled way to explore the solution space, it comes with considerable shortcomings. The "trial and error" of exploratory learning, in combination with the high dimensionality of human motion result in policies requiring an enormous amount of training even to achieve simple tasks such as walking to a waypoint [5, 43]. In addition, since motion naturalness is better captured by data and not reward functions, RL approaches tend to produce motions that lack naturalness and expressiveness. Data-driven approaches on the other hand, rely on plentiful training motions that are acquired through motion capture and carefully curated for the downstream tasks [4, 10]. Such approaches do not scale and do not generalize well to out-of-distribution tasks.

In prioritizing both motion realism and training efficiency, we adopt a data-driven approach. However, current data-driven methods lack the ability to learn both from smaller datasets that provide high-quality human reaching motions with goal labels, and from unlabeled larger scale datasets that contain necessary motion skills such as navigating to a goal position. This raises two key challenges. First, how do we model human motion in a way that generated motions can combine skills from different datasets? Second, what should the training objective be in the cases where goal labels are absent?

To address these challenges, we propose WANDR (Wrist-driven Autonomous Navigation for Data-based goal Reaching). We observe that by modeling human motion generation as an autoregressive stochastic process that produces motions frame by frame, WANDR is able to combine pieces of different dataset distributions when generating a motion sequence. Each generation step is conditioned on goal-related information that we call *intention* (visualized arrows in Fig. 1). We carefully design *intention* in a way that strikes a balance between being informative enough to guide the avatar to reach the goal, while also being abstract enough to promote generalization to unseen goals. This allows our generated motions to reach goals that were never encountered during the training phase in a completely zero-shot evaluation scenario. By generating the motion in an autoregressive way, we disentangle the spatial and temporal dimensions of motion. This is necessary as it allows our model to generate novel long-term sequences while being realistic in terms of local dynamic details.

In more detail, our method is based on a conditional Variational Auto-Encoder (c-VAE) that learns to model motion as a frame-by-frame generation process by auto-encoding the pose difference between two adjacent frames. The condition signal consists of the human's current pose and dynamics along with the *intention* information. *Intention* is a function of both the current pose and the goal location and therefore actively guides the avatar during the motion generation in a closed loop manner. Through training, the c-VAE learns the distribution of potential subsequent poses conditioned on the current dynamic state of the human and its *intention* towards a specific goal. We train WANDR using two datasets: AMASS [18], which captures a wide range of motions including locomotion, and CIRCLE [4], which captures reaching motions.

Although AMASS is large, it lacks any explicit label of goals or intentions. To address this, inspired by the Hindsight Experience Replay paradigm in robotics [3], we define *intention* using a hallucinated goal derived from the ground-truth wrist position in a future frame. This approach allows us to establish a unified training objective spanning AMASS and CIRCLE. Consequently, our model learns to combine motions from both datasets, enabling it to effectively reach arbitrary goals during testing.

In summary, we present WANDR, a data-driven method that combines an autoregressive motion prior with a novel *intention* guiding mechanism and is able to generate avatars that realistically move in space and reach arbitrary goals. We experimentally evaluate our approach, including the benefit of combining multiple datasets as well as the generalization capabilities of our motion generator. Our results underscore the efficacy of the intention mechanism as an elegant way of guiding the motion generation process while also enabling the incorporation of pseudo goal labels for datasets lacking explicit goal annotations. The model and code are available for research purposes.

2. Related Work

Early research in motion generation focuses on tasks like motion prediction [1, 2, 7, 9, 19, 27] and unconstrained motion generation [6, 16, 21, 23, 28, 29, 36–38, 40]. More recently, significant effort has been devoted into improving controllability, with a focus on motion generation conditioned on different types of goals [25, 39], enabling interactions with scenes [12, 13, 20] and objects [31, 33, 42, 45]. Methods that attempt goal-driven motion generation can be broadly divided into reinforcement learning or data-driven approaches.

2.1. Reinforcement Learning for Motion Synthesis

Many existing works employ Reinforcement Learning (RL) for the generation of task-specific long motion sequences. Representative work includes MotionVAE [15] and AMP [24]. MotionVAE [15] employs a two-step process where it initially leverages an autoregressive conditional Variational Autoencoder (VAE) to construct a latent space that encapsulates possible human movements. Subsequently, it utilizes RL to sample from this action space to reach a designated target location while avoiding obstacles by monitoring the area ahead. Similar to MotionVAE, GAMMA [43] learns a policy to extract samples from a latent space and then employs a tree-based search algorithm to find viable motions that steer clear of obstacles by considering the environmental geometric constraints. DIMOS [44] further extends the GAMMA framework by introducing two specialized policy networks: one for locomotion and one for interaction. Together these networks generate goal-conditioned motion sequences that dynamically interact with objects and the environment. AMP [24] learns an adversarial motion prior from unstructured datasets and then applies goal-conditioned RL. This approach involves the formulation of a style reward to encourage the resemblance of the generated sequences to those in the dataset, complemented by a task-specific reward aimed at achieving a particular objective. Hassan et al. [11] extend AMP to produce motions that facilitate interactions with the scene, by conditioning both the discriminator and the policy network on the scene context. However, RL requires significant computation and struggles to generate natural and expressive motion sequences.

2.2. Data-driven Approaches for Motion Synthesis

Most data-driven approaches use existing motion capture (MoCap) datasets [18, 35] to train their models through supervised learning. The pioneering Neural State Machine method [30] is a data-driven technique for generating motion with character-scene interactions, focusing on scenarios with a limited number of objects and interactions. HuMoR [26] proposes a robust model for 3D human shape and temporal pose estimation, yet it falls short of generating motions that are conditioned on specific goals. The SAMP [10] method, designed for real-time stochastic motion synthesis, generates diverse human-scene interaction movements by breaking down the process into predicting goals, planning paths, and generating motion along a predefined route. The GOAL [32] method, trained on the GRAB dataset [31], produces motion sequences in which humans walk towards and grasp 3D objects. However, the generated motions exhibit minimal movements, especially in the feet. To address these constraints, the newly introduced CIRCLE [4] dataset provides a collection of reaching motion data. This dataset is used to train a neural network that generates diverse scene-aware reaching motions. Recently, diffusion models have seen the most success at generating motions conditioned on textural input [34, 41] and spatial data [14]. This advancement enables the synthesized motion to accurately reach specified target locations or navigate around obstacles. Nevertheless, the effectiveness of data-driven approaches is constrained by the amount of training data and they lack generalization to out-of-distribution scenarios.

3. Method

Our goal is to have a virtual human that can autonomously and realistically move from an initial pose to an arbitrary goal position and accurately place its right hand on the target. This challenge requires a nuanced understanding of human motion and the intricate dynamics involved in goal-oriented motions. For example, when the human tries to reach a distant goal, the motions are mostly focused on the legs and navigating the body to approach the object, but when it gets close to the object, the focus will be on moving the arms and upper body to reach the target. Using these observations, we develop a method named WANDR, which, although trained in a supervised setting on motion capture data, exhibits generalization in reaching unseen goal locations during test time. WANDR is designed to generate human motion in an autoregressive frame-by-frame fashion, conditioned on novel *intention* features. During training, these features are extracted by picking a future frame as the goal for the wrist. During inference, the *intention* features are dynamically computed based on the goal's position in a feedback loop, guiding the virtual human to reach the goal. See Fig. 2 for the network overview.

In this section, we first consider the different distributions of the datasets we will be using (section 3.1). Following this, we detail the components of the *intention* features and how they are computed during both the training and inference phases (Section 3.2). Finally, we define the motion representation and motion generator network (Section 3.3).

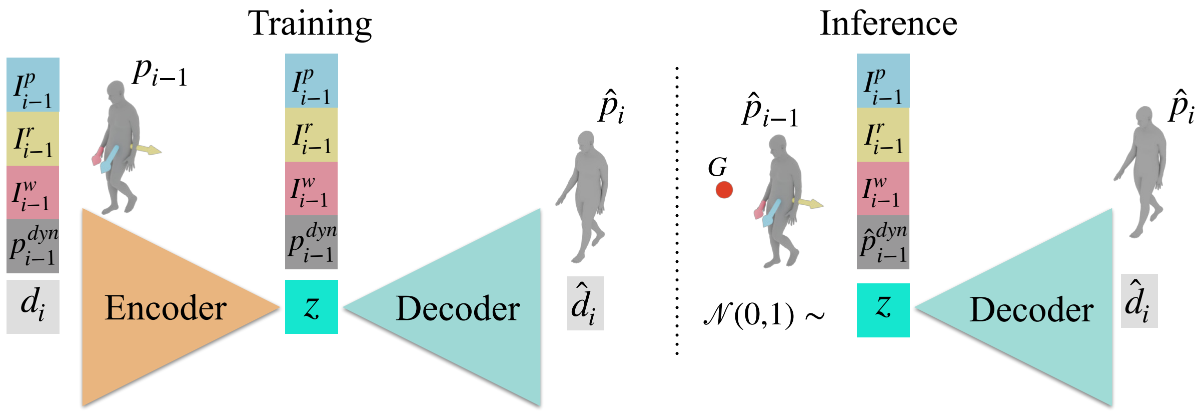

Figure 2. **WANDR architecture.** During training, our model conditions on the intention vectors I^p, I^r and I^w, learning to associate them with actions that result into reaching goals realistically. When the training data has no defined goal, we create a goal based on the wrist location in future frames; see Sec. 3.2. The state of the avatar, p_i^{dyn} expresses the SMPL-X local pose parameters p_i, as well as the deltas d_{i-1} the body parameters have in frame $i-1$. During inference, WANDR takes the intention features, the state, and random noise and returns the change in pose, $\hat{d}_i$. The next pose, $\hat{p}_i$ is obtained by integrating the $\hat{d}_i$ with the previous pose $\hat{p}_{i-1}$.

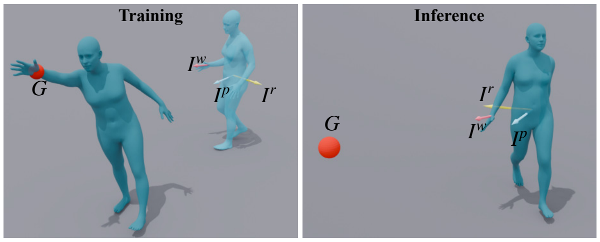

Figure 3. In training, if goals are not specified, they are determined by the future wrist location at a randomly selected future timestep, compensating for the lack of paired ground-truth data in AMASS and direct human motion through intention vectors. During inference, target locations are used as goals with intention vectors calculated based on these specific locations.

3.1. Two Complementing Datasets

For the development of WANDR, we use two key datasets: AMASS and CIRCLE. AMASS is a large-scale dataset that offers a broad range of general human motions but lacks a specific focus on goal-reaching tasks. Its diverse collection of movements provides a solid base for understanding human locomotion and body movement when the person is far away from the goal location. In contrast to AMASS, CIRCLE is tailored towards movements involving reaching specific target positions, particularly capturing the nuances of upper body and arm movements. By integrating AMASS's general motion diversity with CIRCLE's targeted goal-reaching data, we equip WANDR with the ability to generate the entire process of reaching a distant goal, from the initial navigation to the precise target-reaching actions.

3.2. Intention Features

We represent 3D human motion as a sequence of SMPL-X [22] body poses $\mathbf{p} = \{p_1, ..., p_N\}$. Each pose $p_i \in \mathbb{R}^{135}$ consists of three concatenated components: the body's translation $t_i \in \mathbb{R}^3$, root orientation $r_i \in \mathbb{R}^6$, and the body pose $\theta_i \in \mathbb{R}^{21\times6}$, both in 6D format [46].

For the avatar to reach the goal, it is important to be informed about the spatial relation of the goal location with respect to its current pose, as well as to have a sense of time to reach the goal promptly. We achieve this by introducing the *intention* features, which are central to our approach.

To define the *intention* features I_i at timestep i, it is crucial to first establish the selection criteria for the goal $G \in \mathbb{R}^3$, within a motion sequence. Our training involves both label-available scenarios (e.g., CIRCLE dataset) and label-absent scenarios (e.g., AMASS dataset). In label-available cases, where the goal position G and the frame index t_G which it is reached are known, calculating I_i is straightforward. Conversely, for label-absent scenarios like in AMASS, we pick a random future frame as the t_G and define the human's right wrist joint location as the goal G (Fig. 3).

In both scenarios, the *intention* features are defined as:

$$I_i = I_i(p_i, G, t_G, i).$$

These features are essentially a function of the current body pose, the goal position, and time, offering both spatial and temporal insights required for the motion to reach the goal. They are designed to provide sufficient information to reach goals while also enabling test-time generalization. We define them as three distinct components:

$$I_i = (I_i^w, I_i^r, I_i^p).$$

These components represent wrist intention, body orientation intention, and pelvis intention, respectively.

Wrist Intention: This is the main time-dependent component that guides the wrist to reach the goal. It is calculated as the necessary average velocity for the wrist to be at the goal location in time, defined as:

$$I_i^w = \frac{G - W_i}{t_G - i}$$

where t_G is the frame when the goal should be reached. During training, we know the goal frame and the time to reach it. At test time, we know the goal location but need an externally-supplied time to reach it. An emergent behavior of this formulation is that, at inference, the model is able to adjust its movement speed and reach the goal just in time.

Orientation Intention: This component captures the body orientation when reaching the goal location. By conditioning on this, we ensure that the human model orients towards the goal and smoothly navigates towards it, preventing unnatural motions during inference, like walking backward. During training, this is defined as the difference between the forward direction of the current body frame, H_i^{xy}, and the goal body, $H_{t_G}^{xy}$ where xy signifies removing the z component. During inference, since we do not have the goal body, we use the pelvis position P_i to calculate the pelvis-to-goal direction as the desired orientation. This feature is formulated as:

$$I_i^r = \begin{cases} H_{t_G}^{xy} - H_i^{xy} & \text{during training} \\ (G - P_i)^{xy} - H_i^{xy} & \text{during inference.} \end{cases}$$

Pelvis Intention: This feature captures information about the position of the goal relative to the body. It is the difference between the goal and the pelvis joint, excluding the z (height) component. Following the approach in [32], we scale this distance by an exponential function that saturates this vector to have a maximum norm of 2. This formulation helps the method generalize to navigating towards the goal during longer motions and helps the model learn since the distance from the goal does not grow indefinitely in extreme scenarios. This intention is defined by the equation:

$$I_i^p = 2 \times (1 - e^{||G^{xy} - P_i^{xy}||_2}) \times \frac{G^{xy} - P_i^{xy}}{||G^{xy} - P_i^{xy}||_2}.$$

In the experimental section, we delve into the significance of each of these intention features and discuss the rationale behind our design choices, illustrating their impact on the effectiveness of our model.

3.3. Motion Network (WANDR)

WANDR is designed as a conditional Variational Auto-Encoder (c-VAE) network, operating in an autoregressive manner to generate sequential motion frames. This framework is pivotal in predicting the subsequent pose in a motion sequence, emphasizing an incremental, frame-by-frame approach.

Central to our approach is the training of the c-VAE to autoencode pose deltas, denoted as $d_i \in \mathbb{R}^{135}$. These deltas represent the difference between two consecutive poses, p_i and p_{i-1}. By focusing on pose deltas rather than absolute pose values, our model benefits from an important inductive bias, enhancing its learning efficiency and performance, as supported by prior research [15, 26]. We separate rotational differences into: body orientation (d_i^r) and body pose (d_i^θ), each expressed in a 6-D rotational format. Translation deltas are denoted as $d_i^t = t_i - t_{i-1}$.

To enhance the motion representation's invariance, we remove information related to the global z-orientation. This is accomplished by subtracting the global z Euler angle of r_{i-1} from both the translational (d_i^t) and rotational (d_i^r) deltas. The resulting deltas, d_i^{t-z} for translation and d_i^{r-z} for orientation, provide a more robust and consistent representation of motion, irrespective of global direction. Consequently, the delta pose features for any given frame i are composed as follows:

$$d_i = (d_i^{t-z}, d_i^{r-z}, d_i^\theta).$$

An advantage of this representation is its consistency across different motion global orientations. For instance, in the scenario of a person walking, the delta representation remains agnostic to the walking direction. This attribute underscores the efficacy of our method in capturing the essence of motion without being biased towards any specific orientation or direction.

Condition Inputs: For each motion frame, the decoder is conditioned on a combination of state and intention features. Specifically, this condition signal is formulated as $c_i = (p_i^{dyn}, I_i)$. The state features, p_i^{dyn}, encapsulate the avatar's current local pose, focusing on the z-component of translation and a modified orientation that excludes the global z Euler angle, along with the pose deltas $di - 1$ of the previous step. That combination ensures that the generated motion at each step is informed by both the local pose configuration of the avatar, its dynamics and its directional intention towards the set goal, vital for producing realistic, goal-oriented human motions. An overview of the network architecture is shown in Fig. 2.

3.4. Training Losses

Our training objective is a composite of three distinct loss functions:

$$\mathcal{L} = \mathcal{L}_{rec} + \alpha \mathcal{L}_{KL} + \mathcal{L}_J.$$

The reconstruction loss, $\mathcal{L}_{rec}$, measures the accuracy of the motion reconstruction, quantified as the mean square error (MSE) between the input pose delta, d_i, and its reconstructed counterpart, $\hat{d}_i$. It ensures the network's ability to faithfully replicate the input motion.

The KL Divergence Loss, ($\mathcal{L}_{KL}$), evaluates the deviation of the encoded distribution from a standard normal distribution. It is formulated as:

$$\mathcal{L}_{KL} = \mathcal{KL}(\mathcal{N}(0, I)||\mathcal{N}(\mu_i, \sigma_i)).$$

Here, μ_i and σ_i represent the mean and variance of the Gaussian distribution predicted by the encoder. We balance this term with $\alpha = 10^{-2}$ to prevent the over-dominance of $\mathcal{L}_{KL}$, thereby aiding the decoder in avoiding collapse to mean predictions.

Finally, we use a Joint Error Loss ($\mathcal{L}_J$), to ensure perceptual accuracy, by integrating the predicted $\hat{d}_i$ to get the predicted next pose $\hat{p}_i$, which is then fed into the SMPL-X model to obtain the predicted joint positions, $\hat{J}$. The loss $\mathcal{L}_J$ is the MSE between these predicted joints $\hat{J}$ and the ground truth joints J, addressing errors that might not be apparent in parameter space but are perceptually significant, such as incorrect body orientation.

Notably, our approach does not incorporate any explicit loss functions directly related to reaching a goal. This omission is a deliberate choice, aligning with our method's emphasis on generalizing to diverse goal-reaching scenarios without being constrained by goal-specific training losses.

3.5. Motion Generation

In the inference phase of WANDR, our primary objective is to generate human motion that is driven towards a specific goal. Using the decoder of the WANDR c-VAE, we iteratively generate and integrate pose deltas. This process is initiated from the starting pose and progressively builds upon each subsequent pose.

The intention features are recalculated at each step based on the current predicted pose and the goal location. They serve as a guiding mechanism, ensuring that the generated motion is consistently oriented towards placing the human's right wrist on the target.

The user can control the motion's pace by specifying the time to reach the goal t_G. This directly affects the wrist intention feature, enabling adjustments from fast to slow motions to suit various scenarios and constraints.

4. Experiments

In this section, we outline the datasets used for training and evaluation and benchmark how each dataset affects the goal-reaching ability and the quality of the generated motions. Furthermore, we compare our approach with several baselines and ablate the effect of the different components of our *intention* vector.

4.1. Datasets & Processing

Our model is trained on two datasets: AMASS [18] and CIRCLE [4]. AMASS is a collection of 17k sequences, containing a wide range of motion types including long-term navigational skills like walking and turning. CIRCLE, on the other hand, contains 7.2k shorter sequences, each marked with a specific goal reached by a hand. For training, we refine AMASS by excluding sequences where feet are more than $20cm$ above the ground, resulting in a combined dataset of nearly 20k sequences. This dataset is split into 80% training, 10% validation, and 10% test sets. All motions are re-sampled to 30 frames per second (fps).

4.2. Evaluation Strategy

Our evaluation procedure aims at testing the degree which WANDR can generalize to generating reaching motions that start from unseen poses and reach the whole range of 3D space around the starting pose. This is why we choose not to evaluate on held-out motion-goal pairs from the training data. Instead, we only hold out initial poses. During evaluation, starting from these unseen poses, we generate motions that attempt to reach goals that uniformly cover the volume of a cylinder centered on the human, including completely out-of-distribution goal locations (see Sup. Mat. Sec. 5).

In particular, the set of evaluation goals is defined in a cylindrical coordinate frame by taking all the combinations of (1) 5 angles equally separating the 360 degrees around the human, (2) 5 different goal heights ranging from 0 to 1.8 meters and (3) 5 distances from 0.5 to 5 meters. We generate motions from 6 different initial poses, with an 8-second duration specified for reaching each goal. Five motions are sampled for each pose-goal combination, resulting in $5 \times 5 \times 5 \times 6 \times 5 = 3750$ unique motion sequences from which our metrics are computed. This setup allows us to thoroughly test our model in diverse scenarios, including long-term movements, navigational skills, and reaching motions at various heights and distances.

4.3. Evaluation Metrics

To accurately assess the effectiveness of our approach in generating realistic, goal-oriented human motion, we employ a set of metrics focused on both the ability to successfully reach the intended goal and the naturalness of the motion. These metrics are:

- **Success Rate (SR)**: This quantifies the percentage of motions where the right wrist reaches within 10cm of the goal, indicating successful goal attainment. The criterion for success aligns with that used in [4].
- **Foot Skating (FS)**: FS evaluates the naturalness of the motion based on foot skating, where a frame is considered as having foot skating if the lowest vertex of the human mesh moves more than $0.66cm$ between consecutive

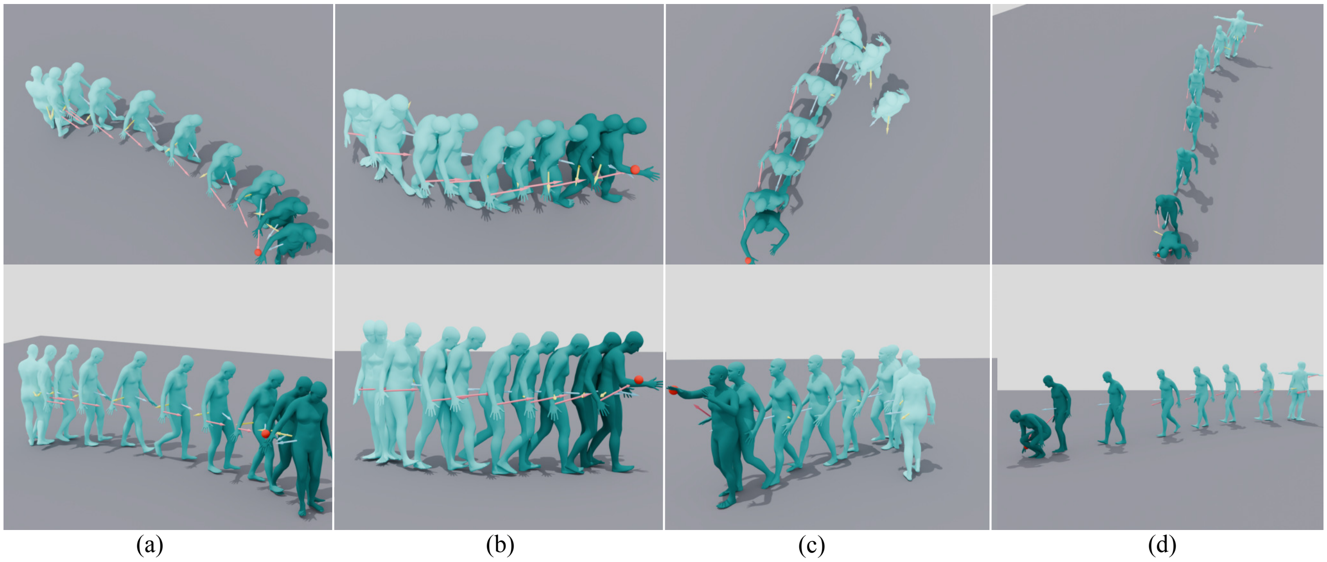

Figure 4. Diverse motion generated with WANDR: Displaying a range of motions generated by WANDR from various initial poses towards arbitrary goals. Examples include navigating towards goals from initial orientations not facing the goal (a, b, c, d), elevating the right hand to reach higher targets (c), and bending down to access goals near the floor (d), showcasing the model's ability to adapt to novel goal locations.

frames (adjusted from the 1cm threshold used in [4] to accommodate our 30fps motion generation).

- **Distance to Goal (DTG)**: DTG records the closest distance in cm that the right wrist gets to the goal during a motion. This metric offers a nuanced view of the model's capability to guide the motion towards the goal.

4.4. Results

4.4.1 Quantitative Results

Combining AMASS and CIRCLE: Our evaluation in Table 1 validates our hypothesis about the benefits of training with both the AMASS and CIRCLE datasets. On the one hand, training only on CIRCLE (line 1) is not sufficient for the model to learn necessary navigational skills, such as walking, due to the dataset's narrow focus or reaching motions. This is apparent from the very high foot-skating. On the other hand, training only on AMASS (line 2) results in high-quality motion generation with low foot-skating, but a relatively low success rate in goal-reaching. The Distance to Goal (DTG) metric, suggests that the model is able to navigate close to the target, but it lacks the precise movement needed to successfully reach for the goal. Using both datasets (line 3) illustrates how our approach effectively merges the broad motion vocabulary of AMASS with the goal-oriented precision of CIRCLE, leading to both high-quality motion and improved goal-reaching capability.

We also compare with GOAL [32], a method that generates human motions that reach and grasp objects. GOAL is trained on GRAB [31], a dataset purely consisting of motions of humans grasping and manipulating objects. Since the relative positioning of the human and the object as well as the motions in GRAB have very small variations, GOAL does not succeed in any of the evaluation configurations (line 4).

Train Set	**SR** ↑	**FS** ↓	**DTG (cm)** ↓
WANDR/Circle	0%	56%	205.4
WANDR/AMASS	16%	19%	48.0
WANDR	**32%**	**16%**	**24.8**
GOAL [32]	0%	29%	149.2

Table 1. We evaluate WANDR trained on different datasets and compare with GOAL [32]. Training solely on CIRCLE results in unrealistic motions, whereas AMASS excels in motion quality but struggles with finer goal-reaching skills. WANDR, leveraging both of what these datasets offer, demonstrates realistic motions as well as better ability to reach goals compared to baselines and existing methods.

Ablation of Intention Features: In order to demonstrate the contribution of each component of the *intention* feature we conduct an ablation study (Table 2). Using only wrist intention (line 1) results in the lowest foot skating, due to the minimal constraints applied to the motion, allowing for more adaptable motion planning. But wrist intention features, are time-dependent and do not carry information about the absolute distance to the goal. This is why it can lead to the avatar over- or under-shooting and thus achieving a low success ratio (SR). The addition of pelvis intention (line 2) enables the avatar to sense the distance to the goal while the orientation intention (line 3) properly aligns the body to face the goal. Since pelvis and orientation intention add more constraints to the motion, they can sometimes cause more challenging body dynamics and produce motions with higher foot-skating (FS) (e.g. turning around in place instead of walking in a U-turn). We also try removing the *intention* from the motion prior and optimizing the latent space of a randomly generated initial motion to minimize the distance between the wrist and the goal (VAE + opt). This approach fails since the result is heavily dependent on the initialization of the latent variables of the motion. Our results confirm that each component of the intention feature is essential to achieving the overall performance of the model. For more details on the optimization see Sup. Mat.

Train Set	**SR** ↑	**FS** ↓	**DTG (cm)** ↓
WANDR (I^w)	15%	**13%**	62.9
WANDR ($I^w + I^p$)	18%	17%	44.9
WANDR ($I^w + I^r$)	19%	19%	36.0
WANDR (full *intention*)	**32%**	16%	**24.8**
VAE + opt	3%	4%	217.0

Table 2. **Ablation Study.** We evaluate the impact of each component of the intention vector. We also compare with an optimization baseline that does not use any condition signals. The results highlight the effectiveness of all of the components of intention as well as the fact that the complexity of the task makes "brute-forcing" with optimization unsuccessful.

Success Ratio Distribution: Our decomposition of the model's success ratio, presented in Fig. 5, offers insights into how the model's performance varies with respect to different goal positions. WANDR demonstrates a consistent ability to reach goals across various distances (blue) and directions (green). It is more capable at reaching goals that are closer to the natural position of the wrist and do not require extensive bending or stretching (yellow). This trend likely results from the abundance of standing or upright motion sequences in the training data, as opposed to motions involving bending or crouching. This analysis provides valuable information for future improvements and dataset balancing.

4.4.2 Qualitative Results

In Fig. 4, we show a variety of motion sequences generated with our network featuring reaching goals located at varying distances and heights, highlighting the model's ability to realistically and smoothly orient, navigate, and reach for goals. These goals require actions such as bending down, turning, or stretching upwards. A critical aspect observed is the model's ability to decelerate as it approaches the goal, seamlessly coordinating body and arm movements to achieve a natural-looking reaching motion. Overall, the qualitative results show that our network generalizes well to novel goal locations while generating realistic motions. For more results please see Sup. Mat. and the video.

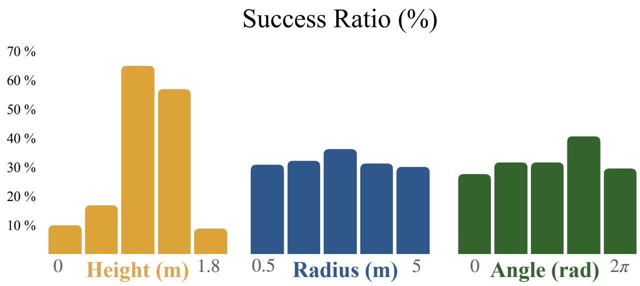

Figure 5. We show the success rates of reaching goals at various heights, angles, and distances from the initial human pose. It highlights how goal position affects the model in accurately navigating and achieving the goals.

5. Conclusion

In conclusion, our research presents a novel data-driven approach to human motion generation, focusing on the task of reaching arbitrary goals in space. We introduce novel *intention* features that enable learning both general navigational skills from AMASS and goal-reaching skills from the CIRCLE dataset under the same distribution. We evaluate our model's ability to reach unseen goals that cover the whole space an avatar should be able to reach around it. The autoregressive design of WANDR demonstrates generalizability in generating realistic human motions that reach unseen goals without requiring any extra guidance information such as a pre-defined trajectory.

Limitations and Future Work: Our approach is not without its limitations. Currently, error accumulation can sometimes bring the avatar to states where it can no longer recover. Additionally, our model shows less proficiency in reaching extremely low or high goals, reflecting a need for more diverse training data encompassing a wider range of body movements. Future work could focus on incorporating realistic grasping mechanisms and interactions with objects, as well as including scene navigation capabilities. This could involve integrating more complex datasets or developing advanced algorithms capable of understanding and interacting with varied environmental contexts, thereby pushing the boundaries of realistic human motion simulation.

Acknowledgments: Markos Diomataris was supported in part by the Max Planck ETH Center for Learning Systems. We thank P. Ghosh, O. Ben Dov S. Tripathi, for the fruitful discussions and A. Cseke, T. Niewiadomski, T. Alexiadis, T. McConnell, for conducting the user studies.

Conflicts of Interest:
https://files.is.tue.mpg.de/black/CoI_CVPR_2024.txt

References

[1] Emre Aksan, Manuel Kaufmann, Peng Cao, and Otmar Hilliges. A spatio-temporal transformer for 3d human motion prediction. In *2021 International Conference on 3D Vision (3DV)*, pages 565–574. IEEE, 2021. 3

[2] Sadegh Aliakbarian, Fatemeh Sadat Saleh, Mathieu Salzmann, Lars Petersson, and Stephen Gould. A stochastic conditioning scheme for diverse human motion prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5223–5232, 2020. 3

[3] Marcin Andrychowicz, Filip Wolski, Alex Ray, Jonas Schneider, Rachel Fong, Peter Welinder, Bob McGrew, Josh Tobin, OpenAI Pieter Abbeel, and Wojciech Zaremba. Hindsight experience replay. In *Advances in Neural Information Processing Systems*, 2017. 2

[4] Joao Pedro Araújo, Jiaman Li, Karthik Vetrivel, Rishi Agarwal, Jiajun Wu, Deepak Gopinath, Alexander William Clegg, and Karen Liu. Circle: Capture in rich contextual environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21211–21221, 2023. 2, 3, 6, 7

[5] Jona Braun, Sammy Christen, Muhammed Kocabas, Emre Aksan, and Otmar Hilliges. Physically plausible full-body hand-object interaction synthesis. In *International Conference on 3D Vision (3DV)*, 2024. 2

[6] Yujun Cai, Yiwei Wang, Yiheng Zhu, Tat-Jen Cham, Jianfei Cai, Junsong Yuan, Jun Liu, Chuanxia Zheng, Sijie Yan, Henghui Ding, et al. A unified 3D human motion synthesis model via conditional variational auto-encoder. In *Computer Vision and Pattern Recognition (CVPR)*, pages 11645–11655, 2021. 3

[7] Zhe Cao, Hang Gao, Karttikeya Mangalam, Qi-Zhi Cai, Minh Vo, and Jitendra Malik. Long-term human motion prediction with scene context. In *European Conference on Computer Vision (ECCV)*, pages 387–404. Springer, 2020. 3

[8] Zicong Fan, Omid Taheri, Dimitrios Tzionas, Muhammed Kocabas, Manuel Kaufmann, Michael J. Black, and Otmar Hilliges. ARCTIC: A dataset for dexterous bimanual hand-object manipulation. In *Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[9] Partha Ghosh, Jie Song, Emre Aksan, and Otmar Hilliges. Learning human motion models for long-term predictions. In *2017 International Conference on 3D Vision (3DV)*, pages 458–466. IEEE, 2017. 3

[10] Mohamed Hassan, Duygu Ceylan, Ruben Villegas, Jun Saito, Jimei Yang, Yi Zhou, and Michael Black. Stochastic scene-aware motion prediction. *arXiv:2108.08284*, 2021. 2, 3

[11] Mohamed Hassan, Yunrong Guo, Tingwu Wang, Michael Black, Sanja Fidler, and Xue Bin Peng. Synthesizing physical character-scene interactions. In *International Conference on Computer Graphics and Interactive Techniques (SIGGRAPH)*, 2023. 2, 3

[12] Yana Hasson, Gül Varol, Dimitrios Tzionas, Igor Kalevatykh, Michael J Black, Ivan Laptev, and Cordelia Schmid. Learning joint reconstruction of hands and manipulated objects. *arXiv:1904.05767*, 2019. 3

[13] Chun-Hao P Huang, Hongwei Yi, Markus Höschle, Matvey Safroshkin, Tsvetelina Alexiadis, Senya Polikovsky, Daniel Scharstein, and Michael J Black. Capturing and inferring dense full-body human-scene contact. In *Computer Vision and Pattern Recognition (CVPR)*, pages 13274–13285, 2022. 3

[14] Korrawe Karunratanakul, Konpat Preechakul, Supasorn Suwajanakorn, and Siyu Tang. Gmd: Controllable human motion synthesis via guided diffusion models. In *Computer Vision and Pattern Recognition (CVPR)*, 2023. 3

[15] Hung Yu Ling, Fabio Zinno, George Cheng, and Michiel Van De Panne. Character controllers using motion VAEs. *Transactions on Graphics (TOG)*, 2020. 2, 3, 5

[16] Hung Yu Ling, Fabio Zinno, George Cheng, and Michiel Van De Panne. Character controllers using motion vaes. *Transactions on Graphics (TOG)*, 2020. 3

[17] Matthew Loper, Naureen Mahmood, and Michael J Black. Mosh: motion and shape capture from sparse markers. *International Conference on Computer Graphics and Interactive Techniques (SIGGRAPH)*, 33(6):220–1, 2014. 2

[18] 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 (ICCV)*, 2019. 2, 3, 6

[19] Julieta Martinez, Michael J Black, and Javier Romero. On human motion prediction using recurrent neural networks. In *Computer Vision and Pattern Recognition (CVPR)*, 2017. 3

[20] Aymen Mir, Xavier Puig, Angjoo Kanazawa, and Gerard Pons-Moll. Generating continual human motion in diverse 3d scenes. In *International Conference on 3D Vision (3DV)*, 2024. 3

[21] Dirk Ormoneit, Hedvig Sidenbladh, Michael Black, and Trevor Hastie. Learning and tracking cyclic human motion. *Conference on Neural Information Processing Systems (NeurIPS)*, 13, 2000. 3

[22] 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 *Computer Vision and Pattern Recognition (CVPR)*, pages 10975–10985, 2019. 4

[23] Vladimir Pavlovic, James M Rehg, and John MacCormick. Learning switching linear models of human motion. *Conference on Neural Information Processing Systems (NeurIPS)*, 13, 2000. 3

[24] Xue Bin Peng, Ze Ma, Pieter Abbeel, Sergey Levine, and Angjoo Kanazawa. Amp: adversarial motion priors for stylized physics-based character control. *Transactions on Graphics (TOG)*, 2021. 2, 3

[25] Mathis Petrovich, Michael J Black, and Gül Varol. Action-conditioned 3d human motion synthesis with transformer vae. In *Computer Vision and Pattern Recognition (CVPR)*, 2021. 3

[26] 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. 3, 5

[27] Xiangbo Shu, Liyan Zhang, Guo-Jun Qi, Wei Liu, and Jinhui Tang. Spatiotemporal co-attention recurrent neural networks for human-skeleton motion prediction. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 44 (6):3300–3315, 2021. 3

[28] Hedvig Sidenbladh, Michael J Black, and David J Fleet. Stochastic tracking of 3d human figures using 2d image motion. In *European Conference on Computer Vision (ECCV)*. Springer, 2000. 3

[29] Leonid Sigal, Alexandru O Balan, and Michael J Black. Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. *International Journal of Computer Vision (IJCV)*, 87 (1-2):4–27, 2010. 3

[30] Sebastian Starke, He Zhang, Taku Komura, and Jun Saito. Neural state machine for character-scene interactions. *Transactions on Graphics (TOG)*, 2019. 3

[31] Omid Taheri, Nima Ghorbani, Michael J. Black, and Dimitrios Tzionas. GRAB: A dataset of whole-body human grasping of objects. In *European Conference on Computer Vision (ECCV)*, 2020. 3, 7

[32] Omid Taheri, Vasileios Choutas, Michael J. Black, and Dimitrios Tzionas. GOAL: Generating 4D whole-body motion for hand-object grasping. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2, 3, 5, 7

[33] Omid Taheri, Yi Zhou, Dimitrios Tzionas, Yang Zhou, Duygu Ceylan, Soren Pirk, and Michael J. Black. Grip: Generating interaction poses using latent consistency and spatial cues, 2023. 3

[34] Guy Tevet, Sigal Raab, Brian Gordon, Yoni Shafir, Daniel Cohen-or, and Amit Haim Bermano. Human motion diffusion model. In *The Eleventh International Conference on Learning Representations*, 2023. 3

[35] Matthew Trumble, Andrew Gilbert, Charles Malleson, Adrian Hilton, and John Collomosse. Total capture: 3d human pose estimation fusing video and inertial sensors. In *British Machine Vision Conference (BMVC)*, pages 1–13, 2017. 3

[36] Raquel Urtasun, David J Fleet, and Pascal Fua. Temporal motion models for monocular and multiview 3d human body tracking. *Computer Vision and Image Understanding (CVIU)*, 104(2-3):157–177, 2006. 3

[37] Qi Wang, Thierry Artières, Mickael Chen, and Ludovic Denoyer. Adversarial learning for modeling human motion. *The Visual Computer*, 36(1):141–160, 2020.

[38] Zhenyi Wang, Ping Yu, Yang Zhao, Ruiyi Zhang, Yufan Zhou, Junsong Yuan, and Changyou Chen. Learning diverse stochastic human-action generators by learning smooth latent transitions. In *Proceedings of the AAAI conference on artificial intelligence*, pages 12281–12288, 2020. 3

[39] Liang Xu, Ziyang Song, Dongliang Wang, Jing Su, Zhicheng Fang, Chenjing Ding, Weihao Gan, Yichao Yan, Xin Jin, Xiaokang Yang, et al. Actformer: A gan-based transformer towards general action-conditioned 3d human motion generation. In *International Conference on Computer Vision (ICCV)*, pages 2228–2238, 2023. 3

[40] Ping Yu, Yang Zhao, Chunyuan Li, Junsong Yuan, and Changyou Chen. Structure-aware human-action generation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16*, pages 18–34. Springer, 2020. 3

[41] Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. Motiondiffuse: Text-driven human motion generation with diffusion model. *arXiv preprint arXiv:2208.15001*, 2022. 3

[42] Xiaohan Zhang, Bharat Lal Bhatnagar, Sebastian Starke, Vladimir Guzov, and Gerard Pons-Moll. Couch: Towards controllable human-chair interactions. In *European Conference on Computer Vision*, pages 518–535. Springer, 2022. 3

[43] Yan Zhang and Siyu Tang. The wanderings of odysseus in 3D scenes. In *Computer Vision and Pattern Recognition (CVPR)*, 2022. 2, 3

[44] Kaifeng Zhao, Yan Zhang, Shaofei Wang, Thabo Beeler, , and Siyu Tang. Synthesizing diverse human motions in 3d indoor scenes. In *International conference on computer vision (ICCV)*, 2023. 3

[45] Yang Zheng, Yanchao Yang, Kaichun Mo, Jiaman Li, Tao Yu, Yebin Liu, C Karen Liu, and Leonidas J Guibas. Gimo: Gaze-informed human motion prediction in context. In *European Conference on Computer Vision (ECCV)*, pages 676–694. Springer, 2022. 3

[46] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *Computer Vision and Pattern Recognition (CVPR)*, pages 5745–5753, 2019. 4