Prime and Reach: Synthesising Body Motion for Gaze-Primed Object Reach

Masashi Hatano[1*] Saptarshi Sinha[2*] Jacob Chalk[2] Wei-Hong Li[2]
Hideo Saito[1] Dima Damen[2]
[1]Keio University, Japan [2]University of Bristol, UK *: Equal Contribution
https://masashi-hatano.github.io/prime-and-reach/

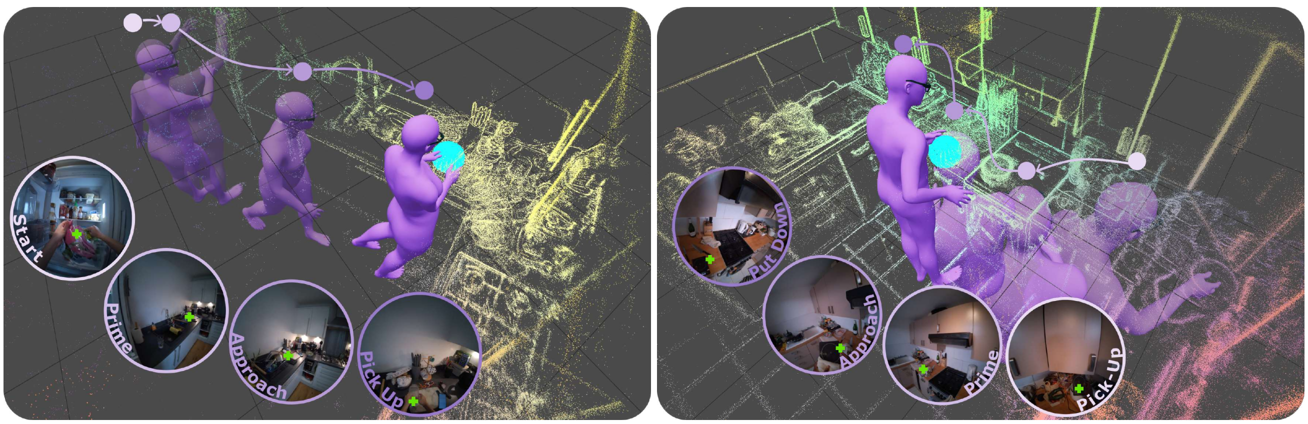

Figure 1. Prime & Reach sequences from HD-EPIC [65], using full-body pose from EgoAllo [91]. **(Left)** A sequence starting with the intention to reach the container (cyan sphere). Gaze priming is evident (gaze intersecting the object) during the approach before reaching the object. **(Right)** Similar behaviour is noted for priming and picking up the scale (cyan sphere). [darker colors indicate later time].

Abstract

Human motion generation is a challenging task that aims to create realistic motion imitating natural human behaviour. We focus on the well-studied behaviour of priming an object/location for pick up or put down – that is, the spotting of an object/location from a distance, known as gaze priming, followed by the motion of approaching and reaching the target location. To that end, we curate, for the first time, 23.7K gaze-primed human motion sequences for reaching target object locations from five publicly available datasets, i.e., *HD-EPIC, MoGaze, HOT3D, ADT, and GIMO.*

We pre-train a text-conditioned diffusion-based motion generation model, then fine-tune it conditioned on goal pose or location, on our curated sequences. Importantly, we evaluate the ability of the generated motion to imitate natural human movement through several metrics, including the 'Reach Success' and a newly introduced 'Prime Success' metric. On the largest dataset, HD-EPIC (Fig. 1), our model achieves 60% prime success and 89% reach success when conditioned on the goal object location.

1. Introduction

Text-guided human motion generation has seen significant advancements, generating increasingly realistic body motions [22, 25, 80]. These works excelled at translating textual descriptions into a wide array of human movements, ranging from simple actions like walking and running to more complex motions such as dancing. The scope has then expanded to include navigating virtual environments towards a spatial target location [14, 37, 88]. More recently, models are designed to generate human motions of interactions with various objects [13, 61, 78, 79, 89], such as grasping small objects or interacting with furniture. However, these works heavily rely on synthetic datasets [4, 79] or datasets collected in lab settings [35, 54, 77]. As a result, they do not capture or model the prime and reach (P&R) behaviour, synthesising motions that are less natural and thus less useful as a digital replica of human behaviour.

On the other hand, egocentric data collection provides a compelling alternative solution, enabling the capture of more natural interactions directly from the first-person perspective, accompanied by a gaze-mounted camera. Several datasets capture natural interactions in diverse daily activities [20, 55] and human-human interactions [96], as observed from an egocentric perspective. Nevertheless, most egocentric datasets [12, 19, 65] do not capture full-body motion, as pairing egocentric and 3D sensors is costly and challenging. Several recent works [9, 26, 46, 91] have explored estimating full-body motion conditioned on the egocentric camera pose and viewpoint, training on diverse data [20, 55]. Although all these egocentric datasets capture diverse human behaviour, they have not been explored for modelling the preparatory priming and reaching motion (see Fig 1 for sample sequences).

Human visuomotor coordination is fundamentally anticipatory, continuously integrating sensory information to facilitate fluid, goal-directed actions [31]. The anticipatory mechanism heavily relies on gaze. By directing visual attention to relevant objects or areas before reaching, gaze offers crucial predictive signals that enable the motor system to prepare and execute actions efficiently and naturally [28, 31, 36, 43]. The ability to prime and reach has recently been investigated in various real-world applications, including robotics [38, 70, 73]. For example, in a manipulation involving a towel, the robot would visually fixate on the towel and the grasp point before extending its arm [38]. However, such behaviour has not been explored in training or evaluating motion generation.

In this work, we enable the synthesis (or generation) of priming and reaching behaviour by combining egocentric datasets that offer gaze priming and full-body motion with conditioned diffusion models.

We list our contributions as follows:

- We explore, for the first time, synthesising full-body motion to replicate humans' natural ability to first gaze-prime then approach and reach objects for interactions.
- We curate five datasets, totalling 23.7K prime and reach sequences from five public datasets.
- We train goal-conditioned motion generation models on our curated datasets and evaluate their ability to generate prime and reach motion. We compare conditioning on the goal pose vs the goal object location.
- We introduce the 'Prime Success' metric to particularly evaluate the ability to synthesise gaze priming behaviour.
- Our results show that using goal pose as a condition, our model can boost priming ability by up to 18x compared to previous methods, and can similarly improve reach success by up to 8x. When conditioning on the target object location only, our model improves the priming ability by up to 5x.

2. Related Work

2.1. Human Motion Generation

Human motion generation aims to create realistic, continuous human movements that simulate or animate natural human motion. The majority of work in this field generates human motion from a single modality such as text [5, 22, 67, 68, 99], action [8, 21, 51, 66], speech [3, 100], and music [48, 76, 84], with a few recent works tackle motion generation from multiple modalities [7, 44, 47, 52].

Text-to-motion Generation. Initial efforts [1, 18] in text-to-motion were deterministic, converging to an averaged motion given an input text. After the advent of the denoising diffusion models [29, 74], these models are nowadays a common practice to generate text-conditioned human motion [39, 80, 94, 95]. MDM [80], MotionDiffuse [95], and FLAME [39] are conditioned on features extracted by a pre-trained text encoder. Another common approach is to disentangle motion representation from generation [24, 25, 32, 93]. This is a two-stage process: first, a VQ-VAE [85] is trained to create a discrete codebook that tokenises continuous motion sequences. Then, a Transformer-based autoregressive model is trained on these discrete tokens to learn the motion primitives by predicting the next token in a sequence. Although the text can serve as a strong signal for conditioning semantic motion, these methods often lack precise control over body positions.

Location-Conditioned Motion Generation. Our work is related to a line of research that generates human motion conditioned on a target location. Several works extend the text-to-motion model to enable various spatial controls (*e.g.*, entire root trajectory [37], keyframe root location [37], obstacle avoidance [37], temporally and spatially sparse joints [69, 88], or goal location [81]). Recently, WANDR [14] introduced a data-driven model conditioned on the initial pose and the goal location of the right wrist to generate avatars that walk and reach the goal in 3D space. Other works [15, 45] address the goal-reaching human motion generation via egocentric perception and reinforcement learning. Despite these advancements, these methods rely on synthetic datasets [4] or MoCap-based datasets [56], which limit their ability to generate natural interactions in real-world scenarios. Additionally, these works do not address or evaluate priming. In this work, we curate the first set of datasets that include full-body, priming, and reaching, with a focus on replicating this human priming-then-reaching motion through generation.

Ego-body Pose Generation. Recent research has explored estimating [9, 26, 33, 34, 53] or forecasting [17, 27, 62, 92] human motion from an egocentric perspective. These methods typically adopt a generative approach as the human body is largely invisible from an egocentric view, unlike ego-body pose estimation from a downward-facing camera [2, 11, 59, 83, 86, 97]. EgoEgo [46] is the first work to propose head pose (camera pose) conditioned human motion generation, but was mainly evaluated on synthetic datasets. Subsequently, EgoAllo [91] proposes a head-centric representation (*i.e.*, canonicalisation) to achieve spatial and temporal invariance, and also enables the integration of in-view 3D hand poses for better prediction. We utilise the ego-body pose estimation method [91] to generate human motion on gaze-primed and reach sequences curated from egocentric datasets.

2.2. Eye-gaze in Motion

Eye-gaze is an important predictive signal that directs attention and primes the processing of future movements [28, 36, 43]. Recognising the critical role of gaze, recent research has been focusing on estimating gaze/saliency [10, 41, 42] or explicitly leveraging this cue for various problems, such as video understanding [58, 64] or human-robot interactions [72, 75]. Several works focus on future motion prediction following gaze priming [30, 31, 50, 87, 90]. Tian *et al.* [82] generate hand-object interactions but only in table-top settings. Different from these works, we wish to synthesise both the gaze priming and the reach motion, for the full body, conditioned on the goal.

3. Prime and Reach Data Curation

We first introduce the principle of curating 'Prime and Reach' (P&R) sequences from longer videos. We then detail the steps we carried out to curate these sequences from five public datasets. We note the statistics of these sequences, which we use for training and evaluation.

3.1. P&R sequence Curation

Interaction datasets include multiple and frequent object reaching and manipulation behaviours. However, a critical aspect largely unexplored is the role of gaze in priming or "spotting" objects prior to the reaching motion. We take this missed opportunity and curate for the first time P&R motion sequences from datasets capturing wearable gaze and object interactions. We are inspired by the "*gaze priming*" discussed in [65] where objects were annotated in 3D and then used to identify the fixation that occurs prior to the physical action, signalling intent of interaction.

Starting from long videos, we extract timestamps for object pick-up or put-down events. We note that priming takes place also during put-down where the future location of an object is primed before the action. Given the known pick/put event at time t_e, we analyse a temporal window of duration w immediately preceding it to find a moment $t_p \in [t_e - w, t_e]$. We wish to identify when the user's gaze first attends to or primes the relevant location for the pick-up/put-down event. For pick-up events, we associate the event with 3D location of the object. This location will be

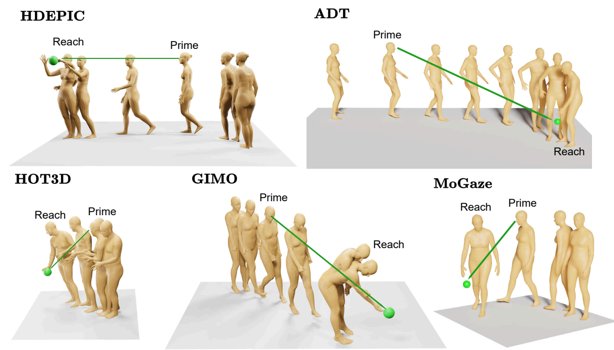

Figure 2. Examples of curated P&R motion sequences from five different datasets.

used to identify the priming event. Importantly, for put-down events, we instead use the future 3D location of the object (which at the start of the motion is an empty part of the 3D space) to search for the priming – *i.e.*, we track the intersection of the gaze of the camera wearer with this empty space, priming the location where the object is going to be placed.

Specifically, we project the user's gaze into the 3D environment to form a ray. First, the gaze direction provided by the eye-tracker in the camera's local coordinate system at any time t, $\mathbf{p}^t_{\text{gaze\_cam}}$, is transformed using the camera-to-world transformation matrix $\mathbf{T}^t_{\text{c2w}}$. The final normalised gaze direction vector, $\hat{\mathbf{d}}^t_{\text{gaze}}$, is then computed as the vector from the camera's world position, $\mathbf{o}^t_{\text{cam}}$, to this new world-space gaze point $\mathbf{p}^t_{\text{gaze\_world}}$ as shown in Equation 1.

$$\begin{aligned} \mathbf{p}^t_{\text{gaze\_world}} &= (\mathbf{T}^t_{\text{c2w}} \mathbf{p}^t_{\text{gaze\_cam}}) \\ \hat{\mathbf{d}}^t_{\text{gaze}} &= \frac{\mathbf{p}^t_{\text{gaze\_world}} - \mathbf{o}^t_{\text{cam}}}{||\mathbf{p}^t_{\text{gaze\_world}} - \mathbf{o}^t_{\text{cam}}||} \end{aligned} \tag{1}$$

We register a relevant location as primed if the gaze ray, originating from the camera's position, intersects with the corresponding 3D bounding box or 3D location o_{3D}. Therefore, we define the prime time t_p as

$$\begin{aligned} T_{\text{int}} &= \{t | t \in [t_e - w, t_e], \mathbb{I}(\text{intersect}(\hat{\mathbf{d}}^t_{\text{gaze}}, o_{\text{3D}})) = 1\} \\ t_p &= \min_{t \in T_{\text{int}}} t, \end{aligned} \tag{2}$$

where T_{int} is the set of all timestamps within the temporal window $[t_e - w, t_e]$, where the gaze ray intersects the 3D location and t_p is the first moment where the intersection happens. We discard sequences where $T_{\text{int}} = \emptyset$. Following [65], we use $w = 10$ secs. To compute the intersection *i.e.* intersect$(\hat{\mathbf{d}}^t_{\text{gaze}}, o_{\text{3D}})$, we use the slab test method [57], details of which are available in the supplementary. At the end of this process, we get P&R sequences each defined by a prime time t_p and reach time t_e.

Table 1. **Curated Dataset Statistics**. We report statistics on curated P&R sequences across five publicly available datasets, ordering them by the size of curated sequences. We report the number of P&R sequences, duration between prime time and reach time *i.e.* $t_e - t_r$ (Prime Gap), body pose type, the distance/movement of body and hand. * indicates that body poses are estimated using [91].

Dataset	#P&R Seq.	Sequence Duration (s)	Prime Gap (s)	Body Pose Type	Body Movement (m)	Hand Movement (m)
HD-EPIC [65]	18,134	5.49 ± 2.76	3.55 ± 2.79	SMPL-H*	0.72 ± 0.67	0.45 ± 0.22
MoGaze [40]	2,637	3.64 ± 0.94	1.53 ± 0.92	3D Skeleton	1.07 ± 0.62	0.75 ± 0.25
HOT3D [6]	2,416	4.31 ± 1.54	2.37 ± 1.57	SMPL-H*	0.20 ± 0.15	0.38 ± 0.19
ADT [60]	411	7.44 ± 2.47	4.51 ± 2.74	3D Skeleton	1.23 ± 1.28	0.56 ± 0.24
GIMO [98]	130	7.11 ± 2.49	4.47 ± 1.49	SMPL-X [63]	3.09 ± 1.20	0.64 ± 0.19

3.2. Datasets

As explained in Sec. 3.1, we formulate how P&R sequences can be curated from long video sequences. We consider five publicly available human-object interaction datasets, all of which contain 2D gaze tracked from wearable gaze trackers along with camera poses [6, 40, 60, 65, 98].

HD-EPIC [65] is an egocentric video dataset capturing diverse human-object interactions in the kitchen using the Aria Device [16]. The dataset provides timestamp annotations of every object's pick/put events, along with 3D location and bounding boxes around the objects at the pick and put locations. Using the 3D annotations with the gaze and camera pose in Eqs. (1) and (2), we determine the priming time t_p for each pick and put interaction in the dataset. We prepend these sequences by a fixed 2-second duration, so the start of the sequence is ahead of the priming, resulting in the sequence $[t_p - 2 \text{ secs}, t_e]$. We accordingly curate 18,134 P&R sequences collected from 156 different videos. We use EgoAllo [91] to estimate full-body motion for our P&R sequences as SMPL-H[49] parameters.

MoGaze [40] is human motion data designed explicitly for human-object interactions, with a particular focus on 'pick' and 'put'. The dataset includes synchronised full-body motion captured using motion capture markers, 3D object models, and 6-DoF and eye-gaze data. The dataset contains 180 minutes of motion capture data from seven participants performing pick-and-put actions along with temporal segment annotations of these actions. Using the gaze data and the object locations, we determine priming timestamps (t_p) for each pick and put event. We slice the motion data for $t \in [t_p - 2, t_e]$ constructing $2,637$ P&R sequences.

HOT3D [6] captures 3D hand-object interactions. The dataset offers 198 ARIA recordings featuring 14 subjects interacting with 33 diverse objects. We only use the ARIA videos as these provide gaze information. As pick/put timestamp annotations are not provided, we extract temporal segments where an object is in-hand by thresholding (< 5 cm) the distance between the nearest hand vertices and the object locations. After identifying these in-hand segments, we refine the segment boundaries by detecting the object state change from stationary to in-hand or vice-versa. This gives us pick/put events t_e. We use gaze and camera pose to estimate the prime time for these events, resulting in $2,416$ P&R sequences. We estimate full-body motion for these sequences using EgoAllo [91].

Aria Digital Twin (ADT) [60] provides a rich collection of synchronised data, including images, eye-tracking data, 6-DoF object data, and 3D human poses. 72 videos in the dataset capture indoor activities and interactions involving 398 unique objects and provide paired eye gaze and 3D body motion data. We curate P&R sequences from these videos. Same as HOT3D, we find temporal segments when objects are in-hand by thresholding the distance between object locations and nearest wrist locations from the corresponding body poses. We identify pick and put events near the temporal boundaries of these segments based on how the object state changed. These events were then primed to determine t_p, resulting in 411 P&R sequences with full-body motion data.

GIMO [98] is a benchmark that focuses on intent-guided human motion segments. It provides 217 trimmed segments, along with corresponding SMPL-X fitted IMU-captured body poses and egocentric views with eye gaze data captured by the HoloLens 2. We discard all segments that include resting activities *e.g.*, sitting or lying on the bed. Gimo segments include object pick-up but no put-down. However, they do not provide the 3D locations or timestamps of object pick-up. Therefore, we manually annotate these event timestamps (t_e) from RGB videos and use the relevant wrist location at the timestamp as our object locations. We determine t_p for each of these events following the same method as before. We get 130 P&R SMPL-X body motion sequences.

In total, we curate $23,728$ P&R sequences from the five datasets. Statistics are provided in Tab. 1. We showcase sample P&R sequences in Fig. 2. Importantly, we unify body pose formats across datasets by representing them using the canonicalised 22-joint body motion used in HumanML3D [23]. Following [80], we convert the 22 joint positions to 263-dim vector representation that combines local pose, rotation and velocity of each joint. The curated sequences from each dataset are split 70%-30% into train and test sets. Details are provided in supplementary.

4. Method

4.1. Objective

Here, we address the task of goal-conditioned human motion generation with the ability to prime and reach a given object. Specifically, the task aims to generate human motion sequences $\{x^i\}_{i=1}^N$ of length N, where $x^i \in R^{J\times3}$ represents 3D positions of J body joints, guided either by desired **goal pose** or target **goal object location** as a condition. We consider and compare the two conditions. As sequences can be of varying length, we opt to predict a fixed-length sequence by uniformly up-/down-sampling N frames. Predicted sequences can then be up- or down-sampled temporally so that the motion length is the same as the ground truth.

4.2. Prime & Reach Motion Diffusion Model

Conditioning. Diffusion models have demonstrated exceptional capability for text-conditioned motion generation [80, 94]. Motivated by this, we use a diffusion generative model, as in [80] for our task. We present our architecture in Fig. 3. Starting from pure noise at $t = T$, the transformer decoder generates motion through iterative denoising over multiple diffusion timesteps $t = \{T, ..., 0\}$ where $t = 0$ produces the predicted motion. This generation is guided through a set of conditions injected into the decoder. We condition our prime and reach motion generation on –

- **Text prompt**: this allows the model to benefit from text-to-motion pre-training. We describe the action as *e.g.*, ‘The person moves across and picks/puts an object.’ We use the knowledge of the action (*i.e.*, whether it is a pick up or a put down) in both training and inference to guide the synthesis. We refer to this conditioning text as $\mathbf{c}$.
- **Initial state** of the body describing where and how the motion initiates. As P&R sequences do not start from a static or neutral pose, but are sampled from within a longer sequence, it is important to feed the initial pose and the velocity at the first frame, as this impacts the guided motion. We represent this by the starting pose $(\hat{x}^1) \in R^{J\times3}$ and the joint velocities *i.e.* $(\hat{x}^1 - \hat{x}^0)$ where $\hat{x}^0$ is the preceding frame before the curated P&Rsequence. We flatten and concatenate the start pose and velocities into a single vector, which forms our initial state.
- **Goal.** We evaluate two possible goal formulations. The first is the complete goal/target pose at the end of the motion $(\hat{x}^N)$. The **goal pose** not only guides the motion to reach the object but additionally guides where the agent would stand relative to the object (through the full pose at the end of the motion) as well as which hand would be reaching the object (guided through the position of the hand joints). Second, we use the more challenging goal of only specifying the **object location** $(o_{3D} \in R^3)$ as a condition. For this goal, the model needs to estimate where the body should be relative to the object and which hand to be the one reaching for the object.

Architecture. Next, we explain how we use these conditions for the generation process. First, we encode the input text prompt $\mathbf{c}$ using a pre-trained text encoder [71]. Both the diffusion noise time step t and $\mathbf{c}$ are then projected to a latent space and summed to get a token $\mathbf{z_t}$. Note that for generating motion conditioned on text, $\mathbf{z_t}$ is directly injected into the transformer decoder’s layers. We pre-train this model for text-conditioned motion generation. This pre-training enables the model to learn prior knowledge of fine-grained full-body motion involved in everyday activities.

For learning P&R motion generation, we initialise our model with the pre-trained text-to-motion conditioned weights and fine-tune it with all three conditions. To add the initial state and goal condition, we first flatten and concatenate them into a single 1D vector. The resultant vector is linearly projected to a latent space to give $\mathbf{p}$. $\mathbf{z_t}$ is modified by adding condition $\mathbf{p}$ to $\mathbf{z_t}$.

$$\tilde{\mathbf{z}}_\mathbf{t} = \mathbf{z_t} + \mathbf{p} \tag{3}$$

This modified condition $\tilde{\mathbf{z}}_\mathbf{t}$ is now injected into the transformer decoder’s layers through cross-attention blocks. This guides the motion sequence generation over multiple diffusion timesteps. We ablate other options of injecting $\mathbf{p}$, including cross-attention, in the supplementary.

Once de-noised, the decoder produces 263-dim representations of body joints $\{v^n\}_{n=1}^N$ where $v^n \in R^{263}$ combines local position, local rotation and local velocity of all 22 body joints. This is post-processed as $\{x^n\}_{n=1}^N = g(\{v^n\}_{n=1}^N)$ to get the predicted 22 joint positions. We use the same g as in [37, 80].

Training. During training, following [80], the model is optimised to reduce the reconstruction error between the 263-dim representations of generated and ground truth motion sequence *i.e.* $\mathcal{L} = \sum_{n=1}^N ||\hat{v}^n - v^n||_2^2$, where $\{\hat{v}^n\}_{n=1}^N$ is the 263-dim representation from the ground truth motion $v^n = g^{-1}(\hat{x}^n)$. We add a joint reconstruction loss as $\mathcal{L}_{joint} = \sum_{n=1}^N ||\hat{x}^n - x^n||_2^2$ which acts on the original 22 joint pose. Our total loss is $\mathcal{L} + \mathcal{L}_{joint}$.

5. Experiments

Here we explain implementation details (Sec. 5.1), evaluation metrics (Sec. 5.2), baselines (Sec. 5.3), experimental results (Sec. 5.4) and ablations (Sec. 5.5).

5.1. Implementation Details

We pre-train our P&R motion diffusion model for text-conditioned motion generation on the large-scale Nymeria [55] dataset to learn the motion prior of everyday activities. Nymeria provides large-scale full-body motion data of

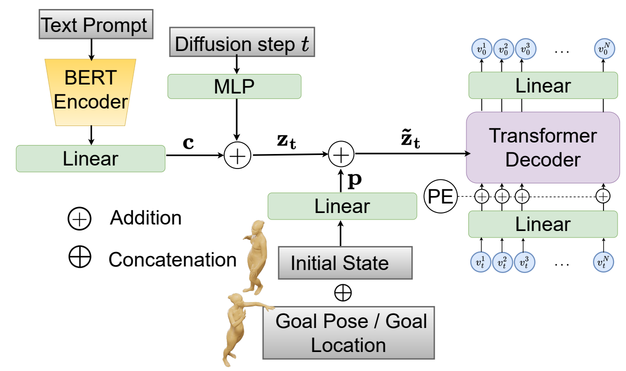

Figure 3. P&R motion diffusion model for goal-conditioned motion generation. We concatenate the initial state of the human body and the goal pose/goal object as conditions, along with a text condition describing the type of action the motion is expected to perform. This accumulated condition is injected into the transformer decoder layers, which then outputs an N-length motion sequence over multiple diffusion steps.

participants performing diverse actions captured by Xsens mocap sensors, accompanied by atomic narrations describing the actions. The narrations are used as text guidance. For pre-training, an initial learning rate of $1e-4$ is used, for a maximum of 600K steps. We use a motion length of $N = 150$ and classifier-free guidance with a probability of 0.2. The pre-training takes $\sim$ 36 hours on one H200 GPU.

Initialised with the pre-trained weights, we fine-tune our Prime and Reach model on the curated P&R sequences from each dataset separately. This is because various datasets capture different types of activities. For fine-tuning, we use a learning rate of $5e-5$ for 250K steps, taking around 15-20 hours on HD-EPIC and HOT3D. For MoGaze, ADT and GIMO, we use a learning rate of $1e-5$ for $\sim 300K$, $\sim 150K$, and $\sim 170K$ steps respectively. We use $T = 50$ diffusion steps in pre-training, finetuning, and inference, following [80].

5.2. Evaluation Metrics

We report results on six metrics: two to directly evaluate our ability to prime and reach, two to evaluate the body pose at the goal, and two to evaluate the entire generated motion.

(1) Prime Success. This is evaluating whether the generated motion is priming the object. Importantly, we wish to evaluate if the object is primed at roughly the same time in the synthesised motion compared to the ground-truth motion, allowing some error in the metric. Specifically, we calculate the angular error between the head forward vector of predicted motion ($\mathbf{H}$) and actual gaze of ground truth motion ($\hat{\mathbf{H}}$) at frame (t_p) as follows:

$$\frac{1}{B}\sum_{b=1}^{B}\mathbb{I}\{\min_{t=t_p-\sigma}^{t_p+\sigma}(\cos^{-1}(\hat{\mathbf{H}}_{\mathbf{t_p}},\mathbf{H_t}))\leq\theta\}, \tag{4}$$

where B is the total number of motion sequences, $\mathbb{I}$ is the indicator function and σ, θ are hyperparameters. Unless mentioned otherwise, we use $\theta = 16\,\text{deg}$ and $\sigma = 0.2$ sec. Prime Success gives the percentage of sequences where this angular error is $\leq$ threshold θ. Having σ relaxes the metric to allow predicted motion to prime the object in a temporal window of 2σ around t_p.

(2) Reach Success. We follow [14] and evaluate our reaching ability. This calculates the percentage of sequences where either wrist in the predicted motion reaches within 10 cm of the goal at N.

(3) Location Error. Also used in [37], we calculate the distance between the final pelvis locations of the ground truth and predicted motions and report the percentage of sequences where the distance is $\geq$ (50 cm).

(4) Goal MPJPE, which calculates the error between the final full body pose of the ground truth and the predicted motion. This includes the error in all joints including the hand reaching the object. Notice that this error assumes the same hand is reaching out to the object as the ground truth.

(5) Mean Per Joint Position Error (MPJPE), which averages the joint position error in Euclidean distance over all generation frames.

(6) Foot Skating, a common evaluation of the generated human motion [37], which measures the proportion of frames where either foot slides with velocity ($>$ 50 cm/s) between consecutive frames while maintaining contact with the ground (foot height $<$ 5 cm).

5.3. Baselines

To assess the ability of generated motion to replicate humans' P&R behaviour, we benchmark five methods on our curated datasets: one naive baseline and three previous works (one from the text-to-motion generation and two from location-conditioned human motion generation):

- **Static** is a naive baseline that uses the average full-body pose of training data and keeps it static for the entire motion sequence. It showcases the difficulty of the dataset.
- **MDM** [80]. We evaluate the checkpoint trained on HumanML3D [23] vs our pre-training on Nymeria [55]. We also fine-tune this model, pre-trained on Nymeria, on each dataset using only text conditioning.
- **GMD** [37], a guided motion diffusion trained on HumanML3D based on UNet architecture, is a two-stage process designed for controllable human motion synthesis with text. The first stage generates a root trajectory that guides full-body motion generation in the second stage.
- **WANDR** [14] is a data-driven, autoregressive method that uses a conditional Variational Auto-Encoder (c-VAE) to generate motion frame-by-frame, trained on the combined dataset of AMASS [56] and CIRCLE [4]. The model is guided by *intention features* that encode the goal's position and the time remaining to reach it.

Table 2. **Comparison of motion generation baselines** on our curated P&R sequences using different metrics. We show results for HD-EPIC, MoGaze, HOT3D, ADT and GIMO separately. The baselines are grouped based on the kind of condition they use for generation. † denotes that the model is trained on HumanML3D. * denotes the model is trained on Nymeria.[55].

Condition	Method	HD-EPIC Prime Success ↑	HD-EPIC Reach Success ↑	HD-EPIC Goal MPJPE ↓	HD-EPIC Loc Err ↓	HD-EPIC MPJPE ↓	HD-EPIC Foot Skating ↓	MoGaze Prime Success ↑	MoGaze Reach Success ↑	MoGaze Goal MPJPE ↓	MoGaze Loc Err ↓	MoGaze MPJPE ↓	MoGaze Foot Skating ↓
No condition	Static	30.15	15.02	0.82	50.92	0.45	–	1.64	4.17	1.03	74.26	0.64	–
Text	MDM †	4.33	0.50	0.94	71.52	0.66	0.21	10.19	3.20	2.03	93.01	1.45	0.29
	MDM *	5.82	1.66	0.83	58.82	0.58	0.03	1.86	1.79	1.18	78.72	0.74	0.07
	MDM	30.52	18.76	0.84	58.73	0.54	0.04	14.43	7.74	1.24	85.42	0.72	0.27
+ Initial State & Goal Pose	GMD †[37]	29.23	26.30	0.30	4.00	0.32	0.17	2.01	50.24	0.35	2.23	0.54	**0.11**
	P&R	**66.31**	**91.23**	**0.09**	**0.20**	**0.17**	**0.13**	**35.34**	**82.96**	**0.09**	**0**	**0.24**	0.21
+ Initial State & Object Loc.	WANDR †[14]	10.21	**95.66**	0.56	54.85	0.58	0.21	11.26	75.62	0.60	64.91	0.56	0.30
	P&R	**59.06**	89.48	**0.30**	**16.28**	**0.24**	**0.12**	**48.51**	**92.78**	**0.39**	**22.32**	**0.33**	**0.25**

Condition	Method	HOT3D Prime Success ↑	HOT3D Reach Success ↑	HOT3D Goal MPJPE ↓	HOT3D Loc Err ↓	HOT3D MPJPE ↓	HOT3D Foot Skating ↓	ADT Prime Success ↑	ADT Reach Success ↑	ADT Goal MPJPE ↓	ADT Loc Err ↓	ADT MPJPE ↓	ADT Foot Skating ↓
No condition	Static	42.13	31.38	0.30	6.96	0.27	–	7.29	11.46	1.98	72.92	1.12	–
Text	MDM †	19.10	2.45	0.49	38.01	0.44	0.10	14.58	3.65	2.72	94.79	1.85	0.24
	MDM *	0.89	0.35	0.54	42.34	0.46	0.01	7.81	6.25	2.15	89.06	1.25	0.07
	MDM	36.44	30.32	0.36	14.45	0.32	0.00	8.85	8.85	2.23	90.63	1.23	0.17
+ Initial State & Goal Pose	GMD †[37]	34.04	11.70	0.41	28.73	0.36	0.02	20.69	55.17	0.42	15.79	0.48	**0.14**
	P&R	**75.26**	**90.69**	**0.12**	**0.53**	**0.11**	**0**	**28.65**	**67.71**	**0.22**	**7.81**	**0.31**	0.18
+ Initial State & Object Loc.	WANDR †[14]	12.26	90.41	0.58	59.70	0.52	0.07	14.94	**85.71**	0.73	79.87	0.66	0.32
	P&R	**67.50**	**94.50**	**0.16**	**2.35**	**0.12**	**0**	**25.00**	75.52	**0.56**	**53.15**	**0.47**	**0.22**

Condition	Method	GIMO Prime Success ↑	GIMO Reach Success ↑	GIMO Goal MPJPE ↓	GIMO Loc Err ↓	GIMO MPJPE ↓	GIMO Foot Skating ↓
No condition	Static	14.28	0	2.85	100	1.84	–
Text	MDM †	0	0	5.26	100	2.23	0.17
	MDM *	4.76	0	3.95	100	2.10	0.04
	MDM	0	0	2.96	100	1.64	0.14
+ Initial State & Goal Pose	GMD †[37]	0	4.76	0.44	**11.90**	0.62	0.43
	P&R	**42.12**	**38.12**	**0.41**	15.32	**0.54**	**0.15**
+ Initial State & Object Loc.	WANDR †[14]	19.04	**71.42**	0.75	85.71	0.77	0.21
	P&R	**47.61**	**71.42**	**0.54**	**52.38**	**0.62**	**0.16**

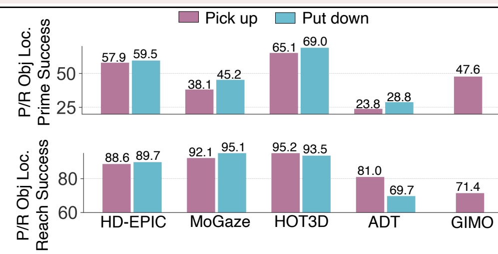

Figure 4. **P&R performance for pick v/s put.**

5.4. Results

In Tab. 2, we compare our proposed P&R diffusion model against baselines on five different datasets. The naive static baseline performs poorly on all the metrics for all datasets. Its poor prime success (20% on average) and reach success (12% on average) highlight the difficulty of the task on our curated datasets.

Text-conditioned baselines rely solely on the knowledge of the pick/put action to generate motion and have no information about the target location to prime or reach. We report the performance of three variants of MDM [80]. These baselines perform poorly on all the metrics as they lacks sufficient guidance. Finetuning MDM on the target datasets helps generate better prime and reach motions with a maximum gain of +24.7% in prime success and +17.1% in reach success on HD-EPIC. The fine-tuned MDM only marginally improves over the static baseline (average of 20.0% prime success and 13.1% reach success), showing that text is not a sufficient condition for prime and reach.

Conditioning with Goal Pose. GMD [37] achieves low location error for all datasets with a best of 2.23% on MoGaze, but it fails to improve prime success, scoring an average of only 17%. Note that GMD gets a high location error in HOT3D. This is because HOT3D has a lot of sitting poses, and GMD tends to always generate standing poses. Our goal-pose conditioned P&R model achieves significant improvements on all the metrics, with maximum gains in prime and reach success. P&R achieves a maximum boost of +41.2% prime success and +79.0% reach success on HOT3D. It is also interesting to note that P&R achieves lower location errors than GMD with an improvement of −28.2% on HOT3D.

Conditioning with Object Location. WANDR [14] achieves strong success scores over all datasets, as that the model's reward focuses on the reaching of goal. However, it underperforms on all other metrics, including the metrics that evaluate the motion between the initial state and target, and it is very poor at priming, with an average of only 16% over all datasets. P&R trained using initial state and object location as condition achieves significant improvements in prime success with gains ranging from +11.9% on ADT to

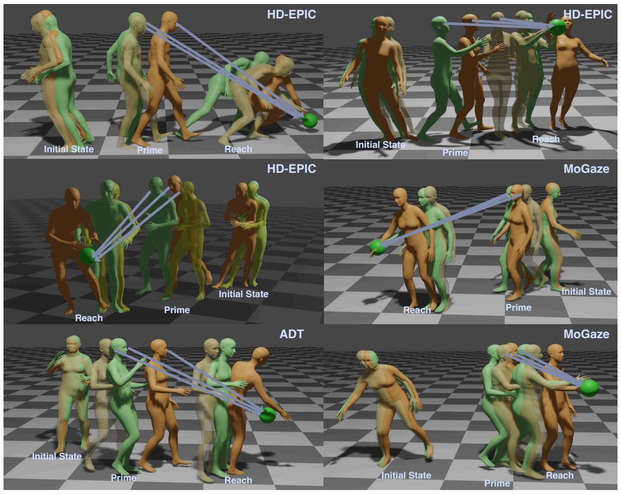

Figure 5. Qualitative results on 3 datasets: Ground truth sequence in light green, goal-pose conditioned prediction in translucent yellow, and target location conditioned generation in brown. We show the pose at the initial, prime, and reach timesteps. Prime direction for both ground truth and predictions are shown using arrows, and target object location is shown in sphere.

+48.8% on HD-EPIC. It reaches or outperforms WANDR in reach success on three out of the five datasets. It also outperforms WANDR by a big margin on all other metrics.

Importantly, when comparing the object location condition with the goal pose, we note that metrics that evaluate the goal location (Goal MPJPE and Loc Error) surely benefit from the guidance on where the person is exactly standing at the end. However, when focusing on the prime and reach success, object location can outperform the stronger goal pose on MoGaze and GIMO. We find that due to limited training data, performance on small-scaled ADT and GIMO are evidently worse.

Results for pick and put. We analyse the performance of the P&R model separated by the action (i.e. pick up or put down) in Fig. 4. P&R performs well on both, but pick-up motions are relatively more challenging than put-down actions, especially for priming ability on MoGaze.

Qualitative Results. We demonstrate qualitative examples of our P&R generated motions in Fig. 5. Generated P&R motions appear natural. Starting with an initial pose and velocity, our generated motion first primes the target object (see arrow) and then reaches to it with one of the hands. Evidently, using the goal location matches better the ground-truth. However, using the object location condition solely successfully synthesises reach but positions the body in a different location at the goal. We demonstrate this ability on various target locations including challenging ones, where the location can be located low or high.

Table 3. **Impact of condition**: We show how each of our modified condition impact P&R model's performance (last row).

				HD-EPIC				MoGaze			
Object Loc.	Initial Pose	Initial Vel.	Text	Prime Success ↑	Reach Success ↑	Loc Err ↓	MPJPE ↓	Prime Success ↑	Reach Success ↑	Loc Err ↓	MPJPE ↓
✗	✓	✓	✓	40.14	30.62	45.50	0.38	18.45	10.94	73.21	0.49
✓	✗	✗	✓	38.71	83.44	27.30	0.37	41.07	87.05	28.57	0.46
✓	✗	✓	✓	42.84	88.15	22.41	0.34	36.83	87.57	24.33	0.35
✓	✓	✗	✓	56.76	86.31	16.90	0.31	38.39	**94.57**	23.59	0.38
✓	✓	✓	✗	56.23	87.66	17.57	**0.24**	41.00	91.52	23.14	**0.33**
✓	✓	✓	✓	**59.06**	**89.48**	**16.28**	**0.24**	**41.44**	93.53	**22.92**	**0.33**

Table 4. **Impact of pre-training**. To validate the effectiveness of pre-training on Nymeria, we show the P&R model's performance trained from scratch or pre-trained on HumanML3D.

	HD-EPIC				MoGaze			
Pre-train	Prime Success ↑	Reach Success ↑	Loc Err ↓	MPJPE ↓	Prime Success ↑	Reach Success ↑	Loc Err ↓	MPJPE ↓
No pre-train	49.76	83.86	22.18	0.30	37.20	87.50	24.11	**0.33**
HumanML3D	56.23	87.56	18.96	0.28	38.39	93.45	24.85	0.34
Nymeria (Ours)	**59.06**	**89.48**	**16.28**	**0.24**	**41.44**	**93.53**	**22.92**	**0.33**

5.5. Ablation and Analysis

We ablate the proposed P&R model our two largest datasets: HD-EPIC and MoGaze. These cover both estimated motion (using EgoAllo for HD-EPIC) and MoCap data (in MoGaze), to ensure our ablation covers both cases.

Condition Ablation. As explained in Sec. 4.2, the proposed P&R method uses text, initial state (initial pose and velocity), and target location as conditions to generate motion. We ablate the impact of each of these conditions in Tab. 3. Using the object location condition gives a significant boost in all metrics, with maximum gains of 82.6% for reach success and 50.3% in location error on MoGaze. This showcases the difficulty of the task, and that it is not trivial to synthesise P&R motions without knowledge of the target. Using the initial state of the body as a condition helps to improve the priming ability of the generated motion, leading to a gain of 20.3% in prime success on HD-EPIC. We find that having both initial pose and initial velocity as initial state conditions is important, especially for the prime success, with drops of at least 2.3% when either is removed on HD-EPIC. Finally, the ablations show that having action knowledge via text (*i.e.*, drop or pick) also improves P&R motion generation on most metrics.

Impact of pre-training. We find that the Nymeria pre-trained model gives a better initialisation (Tab. 4). This is probably because Nymeria has fine-grained motion for everyday activities; therefore, it infuses the knowledge that humans naturally prime and reach objects for everyday actions, which serves better for the downstream task.

6. Conclusion and Future Work

Humans naturally spot or prime an object before reaching it. Previous motion synthesis benchmarks or methods have failed to explore the role of priming for object reaching. To that end, we curate Prime and Reach (P&R) sequences from five datasets using gaze information and object locations. We propose a P&R motion diffusion model that generates full-body motion using goal pose or target location as a condition, along with initial state and text conditioning. We show that the P&R model can generate better prime and reach motion than previous baselines.

Limitation The current method, in line with similar methods [14], does not model hand pose (only body up to wrist). Generating hand motion is an interesting future direction due to its relevance to grasping objects upon reach.

Acknowledgements: This work uses publicly available datasets and annotations to curate P&R sequences. Research at the University of Bristol is supported by EPSRC UMPIRE (EP/T004991/1). M Hatano is supported by JST BOOST, Japan Grant Number JPMJBS2409, and Amano Institute of Technology. S Sinha and J Chalk are supported by EPSRC DTP studentships.

At Keio University, we used ABCI 3.0 provided by AIST and AIST Solutions.

At the University of Bristol, we acknowledge the use of resources provided by the Isambard-AI National AI Research Resource (AIRR), funded by the UK Government's Department for Science, Innovation and Technology (DSIT) via UK Research and Innovation; and the Science and Technology Facilities Council [ST/AIRR/I-A-I/1023]. In particular, we acknowledge the usage of GPU Node hours granted as part of the Sovereign AI Unit call project "Gen Model in Ego-sensed World" (Aug-Nov 2025) as well as the usage of GPU Node hours granted by AIRR Early Access Project ANON-BYYG-VXU6-M (March-May 2025).

References

[1] Chaitanya Ahuja and Louis-Philippe Morency. Language2pose: Natural language grounded pose forecasting. In *3DV*, 2019. 2

[2] Hiroyasu Akada, Jian Wang, Soshi Shimada, Masaki Takahashi, Christian Theobalt, and Vladislav Golyanik. Unrealego: A new dataset for robust egocentric 3d human motion capture. In *ECCV*, 2022. 3

[3] Simon Alexanderson, Rajmund Nagy, Jonas Beskow, and Gustav Eje Henter. Listen, denoise, action! audio-driven motion synthesis with diffusion models. *ACM Transactions on Graphics (TOG)*, 42(4), 2023. 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 *CVPR*, 2023. 1, 2, 6

[5] Nikos Athanasiou, Mathis Petrovich, Michael J Black, and Gül Varol. Teach: Temporal action composition for 3d humans. In *3DV*, 2022. 2

[6] Prithviraj Banerjee, Sindi Shkodrani, Pierre Moulon, Shreyas Hampali, Shangchen Han, Fan Zhang, Linguang Zhang, Jade Fountain, Edward Miller, Selen Basol, Richard Newcombe, Robert Wang, Jakob Julian Engel, and Tomas Hodan. Hot3d: Hand and object tracking in 3d from egocentric multi-view videos. In *CVPR*, 2025. 4

[7] Yuxuan Bian, Ailing Zeng, Xuan Ju, Xian Liu, Zhaoyang Zhang, Wei Liu, and Qiang Xu. Motioncraft: Crafting whole-body motion with plug-and-play multimodal controls. In *AAAI*, 2025. 2

[8] Pablo Cervantes, Yusuke Sekikawa, Ikuro Sato, and Koichi Shinoda. Implicit neural representations for variable length human motion generation. In *ECCV*, 2022. 2

[9] Seunggeun Chi, Pin-Hao Huang, Enna Sachdeva, Hengbo Ma, Karthik Ramani, and Kwonjoon Lee. Estimating egobody pose from doubly sparse egocentric video data. In *NeurIPS*, 2024. 2

[10] Eunji Chong, Nataniel Ruiz, Yongxin Wang, Yun Zhang, Agata Rozga, and James M Rehg. Connecting gaze, scene, and attention: Generalized attention estimation via joint modeling of gaze and scene saliency. In *ECCV*, 2018. 3

[11] Hanz Cuevas-Velasquez, Charlie Hewitt, Sadegh Aliakbarian, and Tadas Baltrušaitis. SimpleEgo: Predicting probabilistic body pose from egocentric cameras. In *3DV*, 2024. 3

[12] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Jian Ma, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100. *International Journal of Computer Vision (IJCV)*, 130(1), 2022. 2

[13] Christian Diller and Angela Dai. Cg-hoi: Contact-guided 3d human-object interaction generation. In *CVPR*, 2024. 1

[14] Markos Diomataris, Nikos Athanasiou, Omid Taheri, Xi Wang, Otmar Hilliges, and Michael J. Black. WANDR: Intention-guided human motion generation. In *CVPR*, 2024. 1, 2, 6, 7, 9

[15] Markos Diomataris, Berat Mert Albaba, Giorgio Becherini, Partha Ghosh, Omid Taheri, and Michael J. Black. Moving by looking: Towards vision-driven avatar motion generation. *arXiv preprint arXiv:2509.19259*, 2025. 2

[16] Jakob Engel, Kiran Somasundaram, Michael Goesele, Albert Sun, Alexander Gamino, Andrew Turner, Arjang Talattof, Arnie Yuan, Bilal Souti, Brighid Meredith, Cheng Peng, Chris Sweeney, Cole Wilson, Dan Barnes, Daniel DeTone, David Caruso, Derek Valleroy, Dinesh Ginjupalli, Duncan Frost, Edward Miller, et al. Project aria: A new tool for egocentric multi-modal ai research. *arXiv preprint arXiv:2308.13561*, 2023. 4

[17] Maria Escobar, Juanita Puentes, Cristhian Forigua, Jordi Pont-Tuset, Kevis-Kokitsi Maninis, and Pablo Arbeláez. Egocast: Forecasting egocentric human pose in the wild. In *WACV*, 2025. 2

[18] Anindita Ghosh, Noshaba Cheema, Cennet Oguz, Christian Theobalt, and Philipp Slusallek. Synthesis of compositional animations from textual descriptions. In *ICCV*, 2021. 2

[19] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, Miguel Martin, Tushar Nagarajan, Ilija Radosavovic, Santhosh Kumar Ramakrishnan, Fiona Ryan, Jayant Sharma, Michael Wray, Mengmeng Xu, Eric Zhongcong Xu, Chen Zhao, Siddhant Bansal, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *CVPR*, 2022. 2

[20] Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, Eugene Byrne, Zach Chavis, Joya Chen, Feng Cheng, Fu-Jen Chu, Sean Crane, Avijit Dasgupta, Jing Dong, Maria Escobar, Cristhian Forigua, Abrham Gebreselasie, et al. Ego-exo4d: Understanding skilled human activity from first- and third-person perspectives. In *CVPR*, 2024. 2

[21] Chuan Guo, Xinxin Zuo, Sen Wang, Shihao Zou, Qingyao Sun, Annan Deng, Minglun Gong, and Li Cheng. Action2motion: Conditioned generation of 3d human motions. In *ACMMM*, 2020. 2

[22] Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating diverse and natural 3d human motions from text. In *CVPR*, 2022. 1, 2

[23] Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating diverse and natural 3d human motions from text. In *CVPR*, 2022. 4, 6

[24] Chuan Guo, Xinxin Zuo, Sen Wang, and Li Cheng. Tm2t: Stochastic and tokenized modeling for the reciprocal generation of 3d human motions and texts. In *ECCV*, 2022. 2

[25] Chuan Guo, Yuxuan Mu, Muhammad Gohar Javed, Sen Wang, and Li Cheng. Momask: Generative masked modeling of 3d human motions. In *CVPR*, 2024. 1, 2

[26] Vladimir Guzov, Yifeng Jiang, Fangzhou Hong, Gerard Pons-Moll, Richard Newcombe, C. Karen Liu, Yuting Ye, and Lingni Ma. Hmd^2: Environment-aware motion generation from single egocentric head-mounted device. In *3DV*, 2025. 2

[27] Masashi Hatano, Zhifan Zhu, Hideo Saito, and Dima Damen. The invisible egohand: 3d hand forecasting through egobody pose estimation. *arXiv preprint arXiv:2504.08654*, 2025. 2

[28] Mary Hayhoe and Dana Ballard. Eye movements in natural behavior. *Trends in cognitive sciences*, 9(4), 2005. 2, 3

[29] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020. 2

[30] Zhiming Hu, Daniel Haeufle, Syn Schmitt, and Andreas Bulling. Hoigaze: Gaze estimation during hand-object interactions in extended reality exploiting eye-hand-head coordination. In *SIGGRAPH*, 2025. 3

[31] Wenqi Jia, Bolin Lai, Miao Liu, Danfei Xu, and James M Rehg. Learning predictive visuomotor coordination. *arXiv preprint arXiv:2503.23300*, 2025. 2, 3

[32] Biao Jiang, Xin Chen, Wen Liu, Jingyi Yu, Gang Yu, and Tao Chen. Motiongpt: Human motion as a foreign language. In *NeurIPS*, 2024. 2

[33] Jiaxi Jiang, Paul Streli, Huajian Qiu, Andreas Fender, Larissa Laich, Patrick Snape, and Christian Holz. Avatarposer: Articulated full-body pose tracking from sparse motion sensing. In *ECCV*, 2022. 2

[34] Jiaxi Jiang, Paul Streli, Manuel Meier, and Christian Holz. Egoposer: Robust real-time egocentric pose estimation from sparse and intermittent observations everywhere. In *ECCV*, 2024. 2

[35] Nan Jiang, Tengyu Liu, Zhexuan Cao, Jieming Cui, Zhiyuan Zhang, Yixin Chen, He Wang, Yixin Zhu, and Siyuan Huang. Full-body articulated human-object interaction. In *ICCV*, 2023. 1

[36] Roland S Johansson, Göran Westling, Anders Bäckström, and J Randall Flanagan. Eye–hand coordination in object manipulation. *Journal of neuroscience*, 21(17), 2001. 2, 3

[37] Korrawe Karunratanakul, Konpat Preechakul, Supasorn Suwajanakorn, and Siyu Tang. Guided motion diffusion for controllable human motion synthesis. In *ICCV*, 2023. 1, 2, 5, 6, 7

[38] Justin Kerr, Kush Hari, Ethan Weber, Chung Min Kim, Brent Yi, Tyler Bonnen, Ken Goldberg, and Angjoo Kanazawa. Eye, robot: Learning to look to act with a bc-rl perception-action loop. In *CoRL*, 2025. 2

[39] Jihoon Kim, Jiseob Kim, and Sungjoon Choi. Flame: Free-form language-based motion synthesis & editing. In *AAAI*, 2023. 2

[40] Philipp Kratzer, Simon Bihlmaier, Niteesh Balachandra Midlagajni, Rohit Prakash, Marc Toussaint, and Jim Mainprice. Mogaze: A dataset of full-body motions that includes workspace geometry and eye-gaze. *IEEE Robotics and Automation Letters (RA-L)*, 6(2), 2020. 4

[41] Bolin Lai, Miao Liu, Fiona Ryan, and James Rehg. In the eye of transformer: Global-local correlation for egocentric gaze estimation. In *BMVC*, 2022. 3

[42] Bolin Lai, Miao Liu, Fiona Ryan, and James M Rehg. In the eye of transformer: Global–local correlation for egocentric gaze estimation and beyond. *International Journal of Computer Vision (IJCV)*, 132(3), 2024. 3

[43] Michael Land, Neil Mennie, and Jennifer Rusted. The roles of vision and eye movements in the control of activities of daily living. *Perception*, 28(11), 1999. 2, 3

[44] Chuqiao Li, Julian Chibane, Yannan He, Naama Pearl, Andreas Geiger, and Gerard Pons-Moll. Unimotion: Unifying 3d human motion synthesis and understanding. In *3DV*, 2025. 2

[45] Gen Li, Kaifeng Zhao, Siwei Zhang, Xiaozhong Lyu, Mihai Dusmanu, Yan Zhang, Marc Pollefeys, and Siyu Tang. Egogen: An egocentric synthetic data generator. In *CVPR*, 2024. 2

[46] Jiaman Li, Karen Liu, and Jiajun Wu. Ego-body pose estimation via ego-head pose estimation. In *CVPR*, 2023. 2, 3

[47] Jiefeng Li, Jinkun Cao, Haotian Zhang, Davis Rempe, Jan Kautz, Umar Iqbal, and Ye Yuan. Genmo: Generative models for human motion synthesis. In *ICCV*, 2025. 2

[48] Ruilong Li, Shan Yang, David A. Ross, and Angjoo Kanazawa. Ai choreographer: Music conditioned 3d dance generation with aist++. In *ICCV*, 2021. 2

[49] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Transactions on Graphics (TOG)*, 34(6), 2015. 4

[50] Zhenyu Lou, Qiongjie Cui, Haofan Wang, Xu Tang, and Hong Zhou. Multimodal sense-informed forecasting of 3d human motions. In *CVPR*, 2024. 3

[51] Thomas Lucas, Fabien Baradel, Philippe Weinzaepfel, and Grégory Rogez. Posegpt: Quantization-based 3d human motion generation and forecasting. In *ECCV*, 2022. 2

[52] Mingshuang Luo, Ruibing Hou, Zhuo Li, Hong Chang, Zimo Liu, Yaowei Wang, and Shiguang Shan. M^3gpt: An advanced multimodal, multitask framework for motion comprehension and generation. In *NeurIPS*, 2024. 2

[53] Zhengyi Luo, Ryo Hachiuma, Ye Yuan, and Kris Kitani. Dynamics-regulated kinematic policy for egocentric pose estimation. In *NeurIPS*, 2021. 2

[54] Xintao Lv, Liang Xu, Yichao Yan, Xin Jin, Congsheng Xu, Shuwen Wu, Yifan Liu, Lincheng Li, Mengxiao Bi, Wenjun Zeng, and Xiaokang Yang. Himo: A new benchmark for full-body human interacting with multiple objects. In *ECCV*, 2024. 1

[55] Lingni Ma, Yuting Ye, Fangzhou Hong, Vladimir Guzov, Yifeng Jiang, Rowan Postyeni, Luis Pesqueira, Alexander Gamino, Vijay Baiyya, Hyo Jin Kim, Kevin Bailey, David S. Fosas, C. Karen Liu, Ziwei Liu, Jakob Engel, Renzo De Nardi, and Richard Newcombe. Nymeria: A massive collection of multimodal egocentric daily motion in the wild. In *ECCV*, 2024. 2, 5, 6, 7

[56] Naureen Mahmood, Nima Ghorbani, Nikolaus F Troje, Gerard Pons-Moll, and Michael J Black. Amass: Archive of motion capture as surface shapes. In *ICCV*, 2019. 2, 6

[57] Alexander Majercik, Cyril Crassin, Peter Shirley, and Morgan McGuire. A ray-box intersection algorithm and efficient dynamic voxel rendering. *Journal of Computer Graphics Techniques (JCGT)*, 7(3), 2018. 3

[58] Michele Mazzamuto, Antonino Furnari, Yoichi Sato, and Giovanni Maria Farinella. Gazing into missteps: Leveraging eye-gaze for unsupervised mistake detection in egocentric videos of skilled human activities. In *CVPR*, 2025. 3

[59] Christen Millerdurai, Hiroyasu Akada, Jian Wang, Diogo Luvizon, Christian Theobalt, and Vladislav Golyanik. Eventego3d: 3d human motion capture from egocentric event streams. In *CVPR*, 2024. 3

[60] Xiaqing Pan, Nicholas Charron, Yongqian Yang, Scott Peters, Thomas Whelan, Chen Kong, Omkar Parkhi, Richard Newcombe, and Yuheng Carl Ren. Aria digital twin: A new benchmark dataset for egocentric 3d machine perception. In *ICCV*, 2023. 4

[61] Georgios Paschalidis, Romana Wilschut, Dimitrije Antić, Omid Taheri, and Dimitrios Tzionas. 3d whole-body grasp synthesis with directional controllability. In *3DV*, 2025. 1

[62] Chaitanya Patel, Hiroki Nakamura, Yuta Kyuragi, Kazuki Kozuka, Juan Carlos Niebles, and Ehsan Adeli. Uniegomotion: A unified model for egocentric motion reconstruction, forecasting, and generation. In *ICCV*, 2025. 2

[63] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. Expressive body capture: 3d hands, face, and body from a single image. In *CVPR*, 2019. 4

[64] Taiying Peng, Jiacheng Hua, Miao Liu, and Feng Lu. In the eye of mllm: Benchmarking egocentric video intent understanding with gaze-guided prompting. In *NeurIPS*, 2025. 3

[65] Toby Perrett, Ahmad Darkhalil, Saptarshi Sinha, Omar Emara, Sam Pollard, Kranti Parida, Kaiting Liu, Prajwal Gatti, Siddhant Bansal, Kevin Flanagan, Jacob Chalk, Zhifan Zhu, Rhodri Guerrier, Fahd Abdelazim, Bin Zhu, Davide Moltisanti, Michael Wray, Hazel Doughty, and Dima Damen. Hd-epic: A highly-detailed egocentric video dataset. In *CVPR*, 2025. 1, 2, 3, 4

[66] Mathis Petrovich, Michael J. Black, and Gül Varol. Action-conditioned 3D human motion synthesis with transformer VAE. In *ICCV*, 2021. 2

[67] Mathis Petrovich, Michael J. Black, and Gül Varol. TEMOS: Generating diverse human motions from textual descriptions. In *ECCV*, 2022. 2

[68] Mathis Petrovich, Michael J. Black, and Gül Varol. TMR: Text-to-motion retrieval using contrastive 3D human motion synthesis. In *ICCV*, 2023. 2

[69] Ekkasit Pinyoanuntapong, Muhammad Usama Saleem, Korrawe Karunratanakul, Pu Wang, Hongfei Xue, Chen Chen, Chuan Guo, Junli Cao, Jian Ren, and Sergey Tulyakov. Maskcontrol: Spatio-temporal control for masked motion synthesis. In *ICCV*, 2025. 2

[70] Nima Rahrakhshan, Matthias Kerzel, Philipp Allgeuer, Nicolas Duczek, and Stefan Wermter. Learning to autonomously reach objects with nico and grow-when-required networks. In *Humanoids*, 2022. 2

[71] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019. 5

[72] Akanksha Saran, Srinjoy Majumdar, Elaine Schaertl Short, Andrea Thomaz, and Scott Niekum. Human gaze following for human-robot interaction. In *IROS*, 2018. 3

[73] Ali Shafti, Pavel Orlov, and A Aldo Faisal. Gaze-based, context-aware robotic system for assisted reaching and grasping. In *ICRA*, 2019. 2

[74] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*, 2021. 2

[75] Janis Stolzenwald and Walterio W Mayol-Cuevas. I can see your aim: Estimating user attention from gaze for handheld robot collaboration. In *IROS*, 2018. 3

[76] Jiangxin Sun, Chunyu Wang, Huang Hu, Hanjiang Lai, Zhi Jin, and Jian-Fang Hu. You never stop dancing: Non-freezing dance generation via bank-constrained manifold projection. In *NeurIPS*, 2022. 2

[77] Omid Taheri, Nima Ghorbani, Michael J Black, and Dimitrios Tzionas. Grab: A dataset of whole-body human grasping of objects. In *ECCV*, 2020. 1

[78] Omid Taheri, Vasileios Choutas, Michael J Black, and Dimitrios Tzionas. Goal: Generating 4d whole-body motion for hand-object grasping. In *CVPR*, 2022. 1

[79] Purva Tendulkar, Dídac Surís, and Carl Vondrick. Flex: Full-body grasping without full-body grasps. In *CVPR*, 2023. 1

[80] Guy Tevet, Sigal Raab, Brian Gordon, Yoni Shafir, Daniel Cohen-or, and Amit Haim Bermano. Human motion diffusion model. In *ICLR*, 2023. 1, 2, 4, 5, 6, 7

[81] Guy Tevet, Sigal Raab, Setareh Cohan, Daniele Reda, Zhengyi Luo, Xue Bin Peng, Amit Haim Bermano, and Michiel van de Panne. CLoSD: Closing the loop between simulation and diffusion for multi-task character control. In *ICLR*, 2025. 2

[82] Jie Tian, Ran Ji, Lingxiao Yang, Suting Ni, Yuexin Ma, Lan Xu, Jingyi Yu, Ye Shi, and Jingya Wang. Gaze-guided hand-object interaction synthesis: Dataset and method. *arXiv preprint arXiv:2403.16169*, 2024. 3

[83] Denis Tome, Patrick Peluse, Lourdes Agapito, and Hernan Badino. xr-egopose: Egocentric 3d human pose from an hmd camera. In *ICCV*, 2019. 3

[84] Jonathan Tseng, Rodrigo Castellon, and C Karen Liu. Edge: Editable dance generation from music. In *CVPR*, 2023. 2

[85] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. In *NeurIPS*, 2017. 2

[86] Jian Wang, Zhe Cao, Diogo Luvizon, Lingjie Liu, Kripasindhu Sarkar, Danhang Tang, Thabo Beeler, and Christian Theobalt. Egocentric whole-body motion capture with fisheyevit and diffusion-based motion refinement. In *CVPR*, 2024. 3

[87] Ping Wei, Yang Liu, Tianmin Shu, Nanning Zheng, and Song-Chun Zhu. Where and why are they looking? jointly inferring human attention and intentions in complex tasks. In *CVPR*, 2018. 3

[88] Yiming Xie, Varun Jampani, Lei Zhong, Deqing Sun, and Huaizu Jiang. Omnicontrol: Control any joint at any time for human motion generation. In *ICLR*, 2024. 1, 2

[89] Sirui Xu, Zhengyuan Li, Yu-Xiong Wang, and Liang-Yan Gui. Interdiff: Generating 3d human-object interactions with physics-informed diffusion. In *ICCV*, 2023. 1

[90] Haodong Yan, Zhiming Hu, Syn Schmitt, and Andreas Bulling. Gazemodiff: Gaze-guided diffusion model for stochastic human motion prediction. *arXiv preprint arXiv:2312.12090*, 2023. 3

[91] Brent Yi, Vickie Ye, Maya Zheng, Yunqi Li, Lea Müller, Georgios Pavlakos, Yi Ma, Jitendra Malik, and Angjoo Kanazawa. Estimating body and hand motion in an ego-sensed world. In *CVPR*, 2025. 1, 2, 3, 4

[92] Ye Yuan and Kris Kitani. Ego-pose estimation and forecasting as real-time pd control. In *ICCV*, 2019. 2

[93] Jianrong Zhang, Yangsong Zhang, Xiaodong Cun, Shaoli Huang, Yong Zhang, Hongwei Zhao, Hongtao Lu, and Xi Shen. T2m-gpt: Generating human motion from textual descriptions with discrete representations. In *CVPR*, 2023. 2

[94] Mingyuan Zhang, Xinying Guo, Liang Pan, Zhongang Cai, Fangzhou Hong, Huirong Li, Lei Yang, and Ziwei Liu. Remodiffuse: Retrieval-augmented motion diffusion model. In *ICCV*, 2023. 2, 5

[95] Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. Motiondiffuse: Text-driven human motion generation with diffusion model. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 46(6), 2024. 2

[96] Siwei Zhang, Qianli Ma, Yan Zhang, Zhiyin Qian, Taein Kwon, Marc Pollefeys, Federica Bogo, and Siyu Tang. Egobody: Human body shape and motion of interacting people from head-mounted devices. In *ECCV*, 2022. 2

[97] Dongxu Zhao, Zhen Wei, Jisan Mahmud, and Jan-Michael Frahm. Egoglass: Egocentric-view human pose estimation from an eyeglass frame. In *3DV*, 2021. 3

[98] 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 *ECCV*, 2022. 4

[99] Chongyang Zhong, Lei Hu, Zihao Zhang, and Shihong Xia. Attt2m: Text-driven human motion generation with multi-perspective attention mechanism. In *CVPR*, 2023. 2

[100] Lingting Zhu, Xian Liu, Xuanyu Liu, Rui Qian, Ziwei Liu, and Lequan Yu. Taming diffusion models for audio-driven co-speech gesture generation. In *CVPR*, 2023. 2