

Document Version

Final published version

Licence

Dutch Copyright Act (Article 25fa)

Citation (APA)

He, Y., Zhang, X., & Stienen, A. H. A. (2025). Gaze-Guided 3D Hand Motion Prediction for Detecting Intent in Egocentric Grasping Tasks. In C. Laugier, & A. Renzaglia (Eds.), *Proceedings of the 2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (pp. 14580-14586). (IEEE International Conference on Intelligent Robots and Systems). IEEE. <https://doi.org/10.1109/IROS60139.2025.11246573>

Important note

To cite this publication, please use the final published version (if applicable).
Please check the document version above.

Copyright

In case the licence states "Dutch Copyright Act (Article 25fa)", this publication was made available Green Open Access via the TU Delft Institutional Repository pursuant to Dutch Copyright Act (Article 25fa, the Taverne amendment). This provision does not affect copyright ownership.
Unless copyright is transferred by contract or statute, it remains with the copyright holder.

Sharing and reuse

Other than for strictly personal use, it is not permitted to download, forward or distribute the text or part of it, without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license such as Creative Commons.

Takedown policy

Please contact us and provide details if you believe this document breaches copyrights.
We will remove access to the work immediately and investigate your claim.

Gaze-Guided 3D Hand Motion Prediction for Detecting Intent in Egocentric Grasping Tasks

Yufei He¹, Xucong Zhang², and Arno H. A. Stienen¹

Abstract—Human intention detection with hand motion prediction is critical to drive the upper-extremity assistive robots in neurorehabilitation applications. However, the traditional methods relying on physiological signal measurement are restrictive and often lack environmental context. We propose a novel approach that predicts future sequences of both hand poses and joint positions. This method integrates gaze information, historical hand motion sequences, and environmental object data, adapting dynamically to the assistive needs of the patient without prior knowledge of the intended object for grasping. Specifically, we use a vector-quantized variational autoencoder for robust hand pose encoding with an autoregressive generative transformer for effective hand motion sequence prediction. We demonstrate the usability of these novel techniques in a pilot study with healthy subjects. To train and evaluate the proposed method, we collect a dataset consisting of various types of grasp actions on different objects from multiple subjects. Through extensive experiments, we demonstrate that the proposed method can successfully predict sequential hand movement. Especially, the gaze information shows significant enhancements in prediction capabilities, particularly with fewer input frames, highlighting the potential of the proposed method for real-world applications.

I. INTRODUCTION

Upper extremity movement disorders caused by stroke or traumatic brain injury often limit daily task performance [1], [2], driving the development of assistive robots that enhance re-acquisition of motor function through adaptive exercises [3]. Intention detection is crucial for robots to assist patients, as it allows the robot to understand the desired actions from users and provide customized assistance [4]. The transition from clinic-based rehabilitation robots to home-assistive robots that support everyday tasks highlights the importance of quickly understanding and assisting with user intentions in domestic settings. Conventional methods such as surface electromyography (sEMG) and electroencephalography (EEG) have been used for intention detection, which can directly measure physiological signals [5]. However, these methods usually restrict movement, require frequent re-calibrations, and lack the perception of environmental context. Additionally, these signals in post-stroke conditions are disturbed [6] and become hard to correlate with movement.

Recent developments in computer vision have enhanced the utility of vision signals for intention interpretation. This progress enables robots to learn from natural human behaviors and observe environmental interactions, thus significantly improving their ability to assist in a manner that aligns

closely with human needs [7]. A variety of datasets capturing both hand and object interactions have been developed [8]–[14], facilitating studies into grasp generation [15]–[17]. For example, Jiang et al. [15] refined grasping gestures using contact maps on objects. However, these studies have been limited to generating static gestures.

Gaze information, which reveals user intent by identifying focus areas before physical actions, enhances human motion prediction for assistive robots [18], [19]. Nevertheless, the application of gaze data often focuses on classification tasks and may suffer from inaccuracies due to false positives [7]. Although research has utilized egocentric signals [20]–[22], or combined them with gaze data [23], [24], the focus has generally been on semantic predictions while the capability of explicit future hand motion prediction is missing. Developing this capability is essential for assistive robots to offer effective assistance throughout training. Studies for hand motion prediction often depend on explicit conditions such as the geometry or position of the object, initial or final hand positions [25]–[27], trajectories [28], or textual descriptions [29]. For example, Christen et al. [30] introduced a method to synthesize diverse hand movements based on the start and end poses of an object.

In response to this challenge, we propose a novel task for intention detection. Given a set of potential grasping objects and initial hand movements, we want to predict future intended hand motions, including hand poses and joint positions. This task focuses on two fundamental aspects: 1) utilizing only implicit environmental context, and 2) producing explicit hand motion outputs represented by 21 hand key points. To tackle this task, we have developed a method that employs gaze- and egocentric-view visual signals to predict future hand motions. This setting is practical for assistive robot applications because the user can operate the robot with the head-mounted device.

Specifically for hand motion prediction, we develop a method consisting of a Vector-Quantized Variational AutoEncoder (VQ-VAE) and an auto-regressive generative transformer. The VQ-VAE encodes hand poses to capture in-distribution features for precise motion prediction, while the transformer generates future hand motion sequences from any input frame. Additionally, our feature fusion architecture, consisting of linear layers, integrates gaze and object features to enhance accuracy.

We validate the generalizability of our model in our self-collected dataset across different subjects and motions, and we explore the impact of various gaze fusion methods on model performance. Our findings indicate robust generaliza-

¹ Department of Biomechanical Engineering, Faculty of Mechanical Engineering, Delft University of Technology

² Department of Intelligent Systems, Delft University of Technology

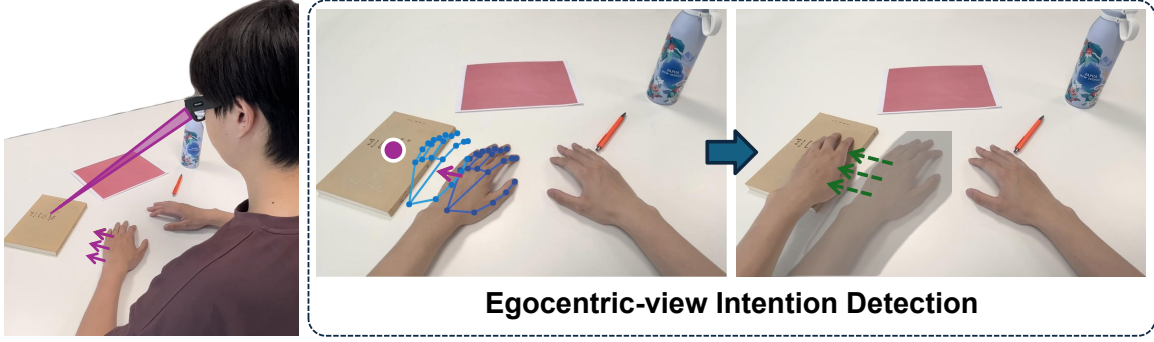


Fig. 1: **Overview of gaze-guided human intention detection.** Left: The user wears eye-tracking glasses to capture gaze fixation points (purple ray), initial hand motion (purple arrow), and object locations as input. Right: Using egocentric-view data, including eye fixation points (purple dot) and hand motion (blue hand skeletons), the system predicts a sequence of hand motions leading to the final grasping action (green arrow).

tion across diverse validation settings, particularly in distance accuracy. The gaze-integrated model significantly outperforms the no-gaze model, especially with fewer input frames, highlighting the value of gaze information when historical data is limited. The results show that, compared with the no-gaze model, our model has the potential to provide accurate and timely predictions in real-time situations.

In summary, our paper introduces a novel approach to hand motion prediction that enhances hand movements in interactive tasks. The main contributions are:

- 1) We propose a new task of hand motion sequence prediction with the goal of driving upper extremity assistive robots.
- 2) We combine gaze data with egocentric visual signals for hand motion prediction.
- 3) We validate that our model generalizes effectively to grasping behaviors, illustrating its broad applicability.

II. METHOD

Our method performs intention detection via hand motion prediction from the user eye gaze, historical hand motions, and object information. To perform the task, our method includes two modules: the hand motion VQ-VAE for discrete hand pose codebook learning and the hand motion generator to predict the sequence of hand motions.

A. Problem Formulation

The intention detection can be formulated as a model M that predicts a sequence of future hand motion $\hat{H}$ based on an initial sequence of hand motion H , a corresponding sequence of eye gazes G , and the representation of the possible interactive objects O in the first frame. The task is formally defined as:

$$\hat{H} = M(H, G, O). \quad (1)$$

$H = \{h_t\}_{t=1}^{\tau}$ is a sequence of input hand motion, where τ is the input frame number. The hand pose at frame t , denoted by $h_t \in \mathbb{R}^{126}$, is defined by the positions of 21 3D hand joint positions (x, y, z) for both hands. This configuration

includes 20 finger joints and one wrist position per hand, according to the Mediapipe framework [31], resulting in a total dimension of $126 = 21 \times 2 \times 3$. Similarly, $G = \{g_t\}_{t=1}^{\tau}$ is a sequence of eye gaze fixation points, represented as $g_t \in \mathbb{R}^3$, is characterized by the 3D eye fixation point (x, y, z) in the world coordinate system. Objects in the scene are represented by maximum of four 3D points $O = \{o^k\}$, where $o_k \in \mathbb{R}^{12}$. For instance, a sheet of paper is described using the positions of its four corners, while a pen is represented by the positions of its tip and bottom, reflecting their distinct shapes. The predicted sequence hand motion $\hat{H} = \{\hat{h}_t\}_{t=\tau+1}^T$ consists of hand pose at frame t , $\hat{h}_t \in \mathbb{R}^{126}$, maintains the same dimension as the input hand pose h_t starting from frame $\tau + 1$ until the end frame T .

B. Hand Motion VQ-VAE

Hand poses have a large space of movements that is difficult to model. A similar problem exists in the human body pose modeling, where the VQ-VAE [32] has been proposed to encode the continuous body movements into discrete classes within a latent space [33]. We utilize the VQ-VAE to learn multiple hand poses, which can be represented as discrete classes in the motion generation phase. An overview of the Hand Motion VQ-VAE model is presented in Fig. 2 (a). The encoded features serve as inputs for the hand motion generation network. The codebook is defined as $C = \{c_i\}_{i=1}^K$, where each c_i belongs to $\mathbb{R}^{D_c}$, K represents the size of the discrete latent space, and D_c is the dimensionality of each embedding vector. The sequence is encoded as $E(H) = \{e_t\}_{t=1}^{\lfloor T/l \rfloor}$, with each embedding $e \in \mathbb{R}^{D_c}$ and l denoting the downsampling scale. The discrete embeddings $Q = \{q_t\}_{t=1}^{\lfloor T/l \rfloor}$ and indices $S = \{s_t\}_{t=1}^{\lfloor T/l \rfloor}$ for each frame are computed as:

$$q_t = \arg \min_{c_i \in C} \|e_t - c_i\|_2 \quad (2)$$

$$s_t = \arg \min_i \|e_t - c_i\|_2 \quad (3)$$

1) *Network Architecture:* Inspired by previous work [34], the encoder utilizes two convolutional layers with a stride of two for temporal downsampling, reducing the temporal

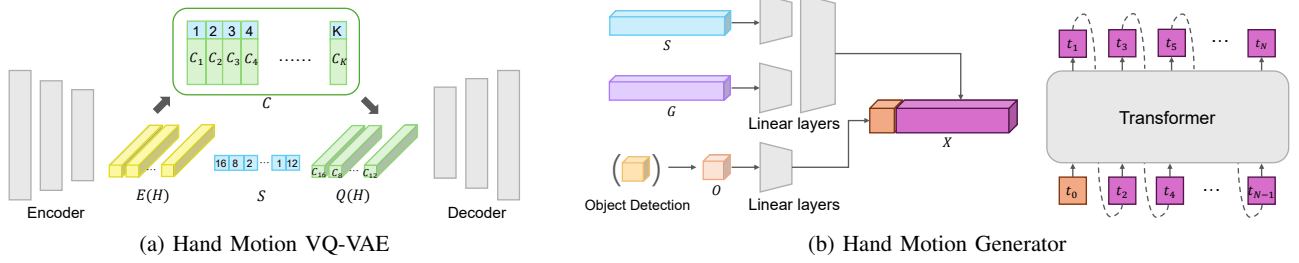


Fig. 2: **Overview of our framework for hand motion prediction.** The proposed method consists of two main components: (a) Hand-Motion VQ-VAE, which encodes hand motion into codebook C with indices S ; and (b) Hand Motion Generator, which contains feature fusion layers and a transformer. In feature fusion layers, the encoded hand motion S is integrated with eye-gaze and object features G and O , together forming fused feature X . The transformer predicts future hand motion indices in an auto-regressive manner using a transformer architecture. These indices are subsequently decoded using the VQ-VAE decoder to obtain the predicted hand motions.

length by a factor of four. This approach not only minimizes computational demands but also reduces noise within the input data. In contrast, the decoder employs nearest-neighbor interpolation for upsampling, facilitating the reconstruction of the complete hand motion sequence.

2) *Optimization Strategy*: To optimize the VQ-VAE model, the loss function L_{vq} consists of reconstruction loss, embedding loss, and commitment loss, detailed as follows:

$$L_{recon} = \begin{cases} 0.5(h_t - \hat{h}_t)^2 / \beta, & \text{if } |h_t - \hat{h}_t| < \beta \\ |h_t - \hat{h}_t| - 0.5\beta, & \text{otherwise} \end{cases} \quad (4)$$

$$L_{embed} = \|\text{sg}[e_t] - q_t\|_2^2 \quad (5)$$

$$L_{commit} = \gamma \|e_t - \text{sg}[q_t]\|_2^2 \quad (6)$$

where β and γ are hyper-parameters that influence reconstruction loss and commitment loss, respectively. “sg” represents the stop-gradient operator, which prevents the back-propagation of gradient, treating the variable as a constant during the optimization process. The total loss is written as $L_{vq} = L_{recon} + L_{embed} + L_{commit}$.

C. Hand Motion Generator

With a trained hand-motion VQ-VAE model, the input hand-motion sequence $H = \{h_t\}_{t=1}^{T_g}$ is encoded into a sequence of quantized indices $S = \{s_t\}_{t=1}^{T_d}$, where $T_d = \lfloor T_g/l \rfloor$. As demonstrated in Fig. 2 (b), these indices are fused with gaze features G and conditioned on object features O , serving as inputs X to the hand-motion generator. This generator operates in an autoregressive manner, producing predicted hand motion indices $\hat{S} = \{\hat{s}_t\}_{t=1}^{T_d+1}$. Given the combined features up to the previous frame X_{t-1} , the probability of each code book index being selected for the next frame hand-motion is calculated as $p_i(s_t|X_{t-1})$. The next-frame hand-motion index is determined by:

$$\hat{s}_t = \arg \max_i p_i(s_t|X_{t-1}) \quad (7)$$

The sequence of indices is mapped to the learned codebook embeddings, forming $\hat{Q} = \{\hat{q}_t\}_{t=1}^{T_d+1}$, where each $\hat{q}_t$ corresponds to $c_{\hat{s}_t}$ from the codebook. This encoded sequence

$\hat{Q}$ is then processed by the decoder D , which reconstructs the predicted hand motion sequence $\hat{H} = \{\hat{h}_t\}_{t=1}^{T_p}$.

1) *Feature Combination*: To align the dimensional differences between gaze features and hand-motion token embeddings, the gaze features are expanded using a linear layer, producing $G' = \{g'_t\}_{t=1}^{T_d}$ where each $g'_t \in D_g$. These embeddings are then concatenated and passed through a linear feature fusion layer followed by a ReLU function, resulting in the combined hand-eye embeddings $F(S, G) = \{f_t\}_{t=1}^{T_d}$, where each $f_t \in \mathbb{R}^{D_x}$. Notably, we do not apply off-the-shelf object detection here, due to the existence of well-established methods for accurate real-time object detection, allowing us to focus on other aspects of our study. Instead, we manually extract object positions from the first frame and transform them via a linear layer to match D_x , forming $O' \in \mathbb{R}^{D_x}$, which acts as a conditioning input. The object embeddings are concatenated at the start of the sequence to create $X = \text{Concat}(O', F) = \{x_t\}_{t=0}^{T_d}$, with each $x_t \in \mathbb{R}^{D_x}$.

2) *Decoder-only Transformer Architecture*: We employ decoder-only transformers with masked self-attention layers similar to [34] for human pose generation, enabling the model to learn input tokens sequentially.

The masked self-attention is calculated as follows:

$$Q = XW^Q; K = XW^K; V = XW^V \quad (8)$$

$$\text{Att}(Q, K, V) = \text{softmax}\left(\frac{QK^T - M}{\sqrt{D_x}}\right)V \quad (9)$$

$$M_{i,j} = \begin{cases} 0 & \text{if } i \geq j, \\ -\infty & \text{if } i < j. \end{cases} \quad (10)$$

W^Q , W^K , and $W^V \in \mathbb{R}^{D_x \times D_x}$ represent the linear projection weights for queries, keys, and values, respectively. Att is the soft attention operation. M is the mask ensuring predictions for a position do not depend on the following positions.

3) *Optimization Strategy*: The loss for the transformer model is computed as a classification task as follows:

$$L_{transformer} = -\sum_{t=1}^N w_t \cdot \log(p(\hat{s}_t|X_{t-1})), \quad (11)$$

TABLE I: Summary of Motions and Interactions

Motion	Grasping Type [36]	Involved Object	Num. Hands
Pick up a bottle	Type A	Bottle	1
Move a piece of paper	Type B	Paper	1
Pick up a book	Type C	Book	1
Pick up a phone	Type C	Phone	1
Pick up a pen	Type D	Pen	1
Pick (an) earphone(s)	Type D	Earphone(s)	1 or 2
Write on paper	Type B, D	Paper, Pen	2

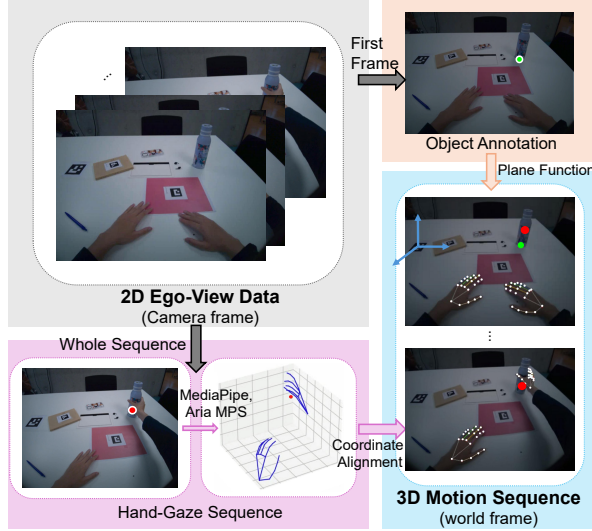


Fig. 3: **Data processing pipeline.** This figure illustrates the sequence of steps applied to process egocentric video data for analysis: (a) Raw 2D images are captured from an egocentric-view video. (b) Throughout the entire sequence, the Mediapipe framework and Aria MPS are utilized to extract 3D hand motion, while Aria MPS extracts 3D gaze points. (c) The object representation is manually annotated on the first frame of the video. (d) A world coordinate is employed to integrate the hand-gaze sequence with the object representation into a unified 3D world frame.

where w_t is the weight assigned to each index, and N is the length of the learned sequence. Specifically, w_N , the weight for the last index, is greater than the weights assigned to other indices to enhance the final hand pose prediction.

III. EXPERIMENTS AND RESULTS

A. Dataset Collection

To evaluate the developed method, we utilize the Aria Glasses from Meta [35] to capture eye-tracking data and the egocentric view of grasping procedures. In this study, 15 volunteers were recruited. Prior to the experiments, participants receive comprehensive instructions detailing the tasks and procedures, and each object is associated with a specific grasping type [36], as shown in Table I. The dataset collection is approved by the Human Research Ethics Committee at TU Delft with case ID 146041.

Participants, already equipped with eye-tracking glasses, are seated at a table with their hands placed palms down.

An instructor randomly positions a target object on the table for grasping alongside other objects to simulate a real-life scenario. Participants are instructed to pick up the object with their preferred hand. Each object is grasped five times by each participant from randomly determined positions, totaling thirty grasping attempts. With the exception of earphones—which may be placed singly or in pairs, requiring bimanual coordination if paired—all objects can be grasped with one hand. Additionally, a bimanual task—writing on paper—is included and performed once per participant.

1) *3D Eye-Hand-Object Data Acquisition:* The data acquisition procedure is illustrated in Fig. 3. We extract relative 3D hand joints from the 2D video using MediaPipe [31]. Subsequently, using the Project Aria Machine Perception Service (MPS), we obtain 3D wrist positions from an egocentric viewpoint. By assuming uniform joint-to-wrist lengths across participants, we project these relative 3D joints into a global 3D space based on the known 3D wrist positions. For object positioning, we manually identify their exact locations in the video and map these into 3D space using the plane function defined by the table surface. The 3D eye gaze point is provided by the MPS. Lastly, we synchronize the hand motion, gaze, and object data within a unified coordinate system origin at 1.15 meters to the right and 0.5 meters from the bottom of the table plane.

B. Evaluation Metric and Baseline

We calculate the Euclidean distance as the position error between the predicted motion and the ground truth as

$$e_{\text{position}} = \frac{1}{T} \sum_{t=1}^T \|p_t - \hat{p}_t\|_2, \quad (12)$$

where the 3-dimensional p_t and $\hat{p}_t$ are the ground truth and predicted palm positions at each frame t , and T is the total number of frames. We compute the position error across all frames and refer to it as *Average Position Error*. In addition, for the upper extremity rehabilitation robot applications, we are concerned about the prediction of the final grabbing pose given the current input. It has practical usage in giving signals to rehabilitation robots early. To reflect the ability of early prediction, we compute the position error only on the final grabbing pose as *End-Pose Position Error*, which we consider only the last frame ($t = T$), where $p_t = p_T$, $\hat{p}_t = \hat{p}_T$.

We establish a baseline model that only takes hand motion sequences and object embeddings as input, i.e. without the eye-gaze feature. The model architecture remains the same as our proposed model. With this baseline, we want to investigate the effectiveness of the eye gaze feature for the human intention detection task.

C. Cross-Subject and Motion Generalization

To evaluate the generalization capability of the proposed method in terms of subjects and motions, we design three evaluation settings. Two specific actions, “pick up a book” and “write on a piece of paper”, are selected for motion validation that have always been excluded from any training procedure. The first evaluation setting is cross-subject (CS),

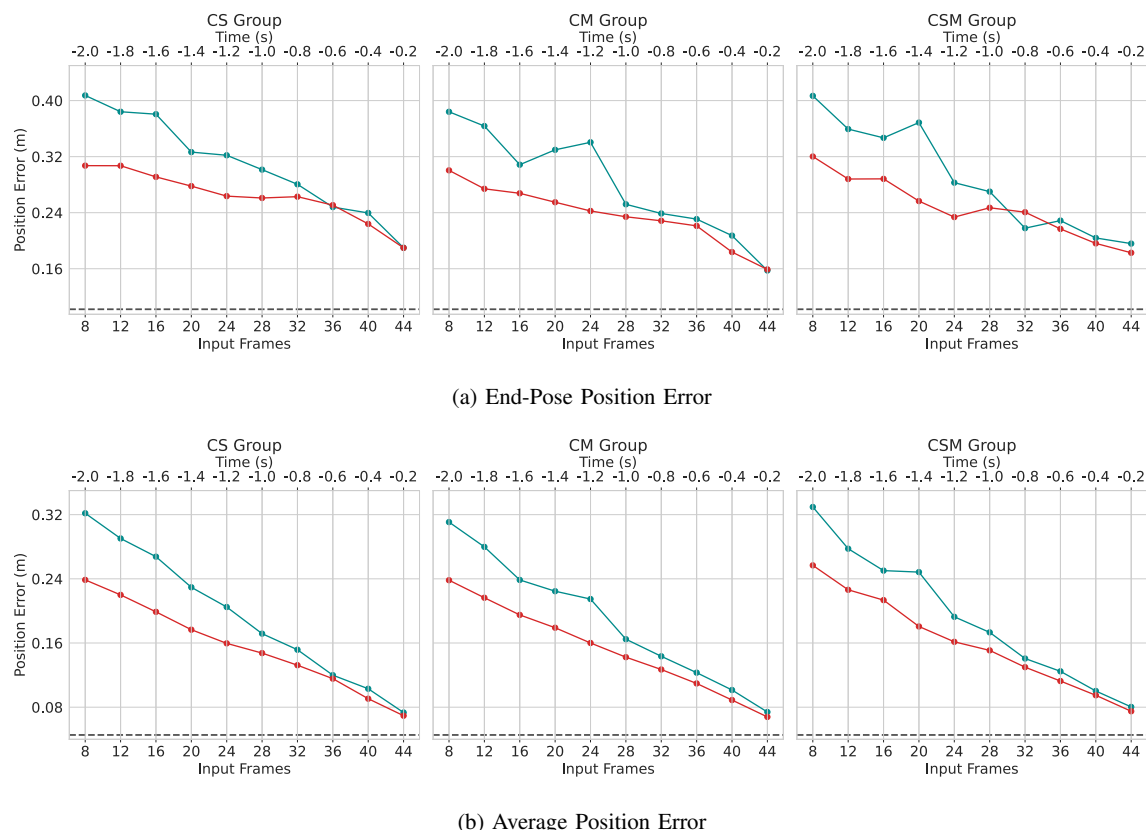


Fig. 4: **Position Errors (in m) across Various Input Frames and Time (in s).** This figure displays the end-pose (first row) and average (second row) *position* errors within the CS, CM, and CSM groups across different numbers of input frames and time before contact. Red lines represent results with gaze, and green lines represent results without gaze. Gray dashed lines at the bottom represent the position error calculated directly through the encoder and decoder of the hand-pose VQ-VAE.

where we perform the five-fold cross-subject evaluation on the 15 subjects from our self-collected dataset. Note that the subjects are different in the training and test sets, while hand actions are the same. The second evaluation setting is cross-motion (CM), where we train and test on the same groups of subjects while testing only on the “pick up a book” and “write on a piece of paper” actions that are not presented in the training set. Note that the subjects are the same in the training and test sets, while the hand actions are different. The third evaluation setting is across both subjects and motions, where we perform the five-fold cross-subject evaluation while only testing on the “pick up a book” and “write on a piece of paper” actions. Note that both subjects and hand actions are different. All results reported are derived from this comprehensive cross-validation strategy. Position errors are evaluated across a range of input frames from 8 to 44, increasing in increments of four to show the early prediction of the proposed method.

We present the results of end-pose error and average position error in Fig. 4. Across all evaluation settings, the position errors demonstrated a decreasing trend as the number of input frames increased for both end-pose and

average position errors. Models integrating gaze information generally exhibited lower errors across CS, CM, and CSM settings, although there are exceptions. All three settings exhibited similar position error patterns, indicating that the model generalizes well across different settings. Notably, the disparity in position errors between models with and without gaze became more pronounced with fewer input frames. Gaze-enhanced models showed smaller errors, suggesting the potential of gaze-enhanced models to provide more accurate and immediate corrections in real-time applications where rapid response is crucial.

The average error throughout the entire grasping process is evaluated similarly to the end-pose error, as depicted in Fig. 4 (b). Across all experimental groups—CS, CM, and CSM—demonstrate consistent trends where models with gaze information outperformed those without. This indicates that gaze information plays a critical role in guiding the movement process. The performance gap between the gaze and no-gaze models becomes more pronounced as the number of input frames decreases, suggesting that gaze information is particularly beneficial in the early stages of input where less historical data is available to aid prediction.

TABLE II: **Position error Comparison of End-Pose by Fusion Type.** This table displays the position error (in m) for end-pose across different input frames, comparing various fusion types within the CS, CM, and CSM groups.

Validation Type	Fusion Type	Input Frames									
		8	12	16	20	24	28	32	36	40	44
CS	Linear	0.3071	0.3070	0.2911	0.2779	0.2636	0.2610	0.2629	0.2508	0.2240	0.1898
	Convolution	0.3320	0.3195	0.3417	0.3044	0.2765	0.2819	0.2705	0.2635	0.2344	0.1954
	Summation	0.4149	0.3265	0.3171	0.3199	0.2829	0.2715	0.2611	0.2425	0.2260	0.1967
CM	Linear	0.3004	0.2741	0.2677	0.2549	0.2425	0.2341	0.2284	0.2212	0.1838	0.1590
	Convolution	0.3427	0.2988	0.3041	0.2934	0.2810	0.2719	0.2629	0.2475	0.2222	0.1822
	Summation	0.3494	0.3223	0.3146	0.2988	0.2656	0.2561	0.2295	0.2007	0.1882	0.1624
CSM	Linear	0.3202	0.2880	0.2883	0.2566	0.2337	0.2470	0.2407	0.2168	0.1962	0.1828
	Convolution	0.3479	0.3233	0.3309	0.3108	0.2681	0.2665	0.2458	0.2349	0.2210	0.1863
	Summation	0.3315	0.3067	0.2980	0.2805	0.2783	0.2706	0.2245	0.2146	0.2139	0.1713

D. Ablation Study on Gaze Fusion Techniques

To investigate the optimal way of integrating gaze information into the model, we conduct comparisons between our standard linear feature integration method and two simple yet effective gaze-fusion methods: convolutional fusion and direct summation. For the convolutional fusion method, we incorporate convolution layers equipped with 1×1 kernels. In the direct summation method, we first expand the gaze feature to match the dimensionality of the hand motion features before adding them together.

The results based on the end-pose error across the CS, CM and CSM groups are shown Tab. II. From the table, we can see that the linear combination method in our model generally surpasses other methods in reducing position error across the CS, CM, and CSM groups, although there are a few exceptions. Specifically, for shorter input sequences ranging from 4 to 28 frames, the linear combination consistently delivers superior performance compared to other methods. As the length of the input sequences increases, providing a longer historical context, the performance of the other methods becomes comparable.

E. Visualization of hand motion prediction process

In practical applications, the proposed method could be used to predict possible hand positions in the near future, in order to guide the robot to assist individuals in planning the next actions. We take one example from our method and illustrate the progressive process in Fig. 5.

In the figure, we show the predictions demonstrate a trend of movement of both hands towards the target or the designed target zone. The method aims to bring the hand closer to the target, with human intervention taking over once within the target zone. The prediction sequences, visualized by dots linked by arrows, show the hand movements from an initial position progressing towards the target over a period of 0.9 seconds, with updates every 0.3 seconds. Given that human movements naturally refine and adjust positioning, the network can utilize these human-corrected positions to further forecast future positions with greater precision. This capability demonstrates the usability of the system

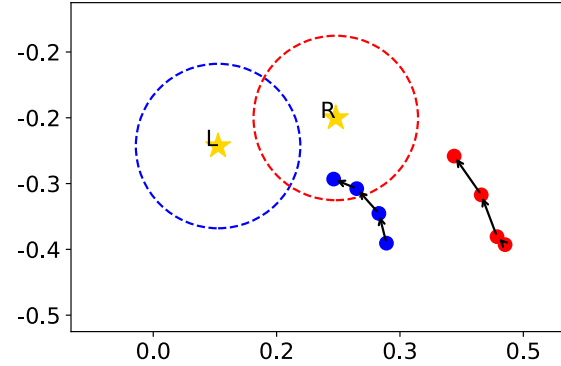


Fig. 5: **An example of hand sequential position predictions from top view.** The red and blue dots indicate the predicted right and left hand positions respectively, linked by arrows to be the prediction sequences. The stars mark the final targets for both hands, surrounded by dashed circles denoted as the ‘target zone’. The axis has a unit of meters.

in dynamic real-world environments and shows promising research toward robust hand motion prediction.

IV. DISCUSSION AND CONCLUSION

This study presents a gaze-guided intention detection method for predicting hand motion in egocentric grasping tasks. By integrating gaze and visual cues with a VQ-VAE and a transformer, our approach achieves accurate and robust hand pose prediction. Our method demonstrates real-time capability, predicting 40 frames with only 40 ms latency. Our method performs well with fewer input frames, particularly when gaze is included, supporting early and responsive intent detection. These results indicate that our system is well-suited for real-time assistive scenarios.

Compared to traditional EMG or EEG-based methods, our non-intrusive setup offers greater usability and environmental awareness, which is crucial for assistive applications. Although no direct comparisons were made, previous studies

highlight that physiological signals are often noisy after stroke and restrict natural movement.

ACKNOWLEDGMENT

We thank Meta for the Aria Glasses in this study.

REFERENCES

- [1] A. C. Lo, P. D. Guarino, L. G. Richards, J. K. Haselkorn, G. F. Wittenberg, D. G. Federman, R. J. Ringer, T. H. Wagner, H. I. Krebs, B. T. Volpe, C. T. Bever, D. M. Bravata, P. W. Duncan, B. H. Corn, A. D. Maffucci, S. E. Nadeau, S. S. Conroy, J. M. Powell, G. D. Huang, and P. Peduzzi, "Robot-assisted therapy for long-term upper-limb impairment after stroke," *New England Journal of Medicine*, vol. 362, no. 19, pp. 1772–1783, 2010.
- [2] S. Luo, Q. Meng, S. Li, and H. Yu, "Research of intent recognition in rehabilitation robots: a systematic review," *Disability and Rehabilitation: Assistive Technology*, vol. 19, no. 4, pp. 1307–1318, 2024.
- [3] A. Basteris, S. M. Nijenhuis, A. H. Stienen, J. H. Buurke, G. B. Prange, and F. Amirabdollahian, "Training modalities in robot-mediated upper limb rehabilitation in stroke: a framework for classification based on a systematic review," *Journal of neuroengineering and rehabilitation*, vol. 11, pp. 1–15, 2014.
- [4] D. Wang, X. Gu, and H. Yu, "Sensors and algorithms for locomotion intention detection of lower limb exoskeletons," *Medical Engineering & Physics*, p. 103960, 2023.
- [5] Y. Gu, Y. Xu, Y. Shen, H. Huang, T. Liu, L. Jin, H. Ren, and J. Wang, "A review of hand function rehabilitation systems based on hand motion recognition devices and artificial intelligence," *Brain Sciences*, vol. 12, no. 8, p. 1079, 2022.
- [6] M. S. N. Ag Lamat, M. S. H. Abd Rahman, W. A. Wan Zaidi, W. N. N. W. Yahya, C. S. Khoo, R. Hod, and H. J. Tan, "Qualitative electroencephalogram and its predictors in the diagnosis of stroke," *Frontiers in Neurology*, vol. 14, p. 1118903, 2023.
- [7] Y. Guo, X. Gu, and G.-Z. Yang, *Human-robot interaction for rehabilitation robotics*. Springer, 2021.
- [8] O. Taheri, N. Ghorbani, M. J. Black, and D. Tzionas, "Grab: A dataset of whole-body human grasping of objects," in *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*, pp. 581–600, Springer, 2020.
- [9] S. Hampali, M. Rad, M. Oberweger, and V. Lepetit, "Honnotate: A method for 3d annotation of hand and object poses," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3196–3206, 2020.
- [10] Y.-W. Chao, W. Yang, Y. Xiang, P. Molchanov, A. Handa, J. Tremblay, Y. S. Narang, K. Van Wyk, U. Iqbal, S. Birchfield, et al., "Dexycb: A benchmark for capturing hand grasping of objects," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9044–9053, 2021.
- [11] T. Kwon, B. Tekin, J. Stühmer, F. Bogo, and M. Pollefeys, "H2o: Two hands manipulating objects for first person interaction recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10138–10148, 2021.
- [12] L. Yang, K. Li, X. Zhan, F. Wu, A. Xu, L. Liu, and C. Lu, "Oakink: A large-scale knowledge repository for understanding hand-object interaction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20953–20962, 2022.
- [13] Y. Liu, Y. Liu, C. Jiang, K. Lyu, W. Wan, H. Shen, B. Liang, Z. Fu, H. Wang, and L. Yi, "Hoi4d: A 4d egocentric dataset for category-level human-object interaction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21013–21022, 2022.
- [14] Z. Fan, O. Taheri, D. Tzionas, M. Kocabas, M. Kaufmann, M. J. Black, and O. Hilliges, "Arctic: A dataset for dexterous bimanual hand-object manipulation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12943–12954, 2023.
- [15] H. Jiang, S. Liu, J. Wang, and X. Wang, "Hand-object contact consistency reasoning for human grasps generation," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 11107–11116, 2021.
- [16] K. Karunratanakul, J. Yang, Y. Zhang, M. J. Black, K. Muandet, and S. Tang, "Grasping field: Learning implicit representations for human grasps," in *2020 International Conference on 3D Vision (3DV)*, pp. 333–344, IEEE, 2020.
- [17] K. Karunratanakul, A. Spurr, Z. Fan, O. Hilliges, and S. Tang, "A skeleton-driven neural occupancy representation for articulated hands," in *2021 International Conference on 3D Vision (3DV)*, pp. 11–21, IEEE, 2021.
- [18] C. Lin, C. Zhang, J. Xu, R. Liu, Y. Leng, and C. Fu, "Neural correlation of eeg and eye movement in natural grasping intention estimation," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 31, pp. 4329–4337, 2023.
- [19] V. Crocher, R. Singh, J. Newn, and D. Oetomo, "Towards a gaze-informed movement intention model for robot-assisted upper-limb rehabilitation," in *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pp. 6155–6158, IEEE, 2021.
- [20] R. Girdhar and K. Grauman, "Anticipative video transformer," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 13505–13515, 2021.
- [21] B. Soran, A. Farhadi, and L. Shapiro, "Generating notifications for missing actions: Don't forget to turn the lights off!," in *2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 4669–4677, 2015.
- [22] A. Furnari and G. M. Farinella, "Rolling-unrolling lstms for action anticipation from first-person video," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 11, pp. 4021–4036, 2021.
- [23] Y. Li, M. Liu, and J. M. Rehg, "In the eye of the beholder: Gaze and actions in first person video," *IEEE transactions on pattern analysis and machine intelligence*, vol. 45, no. 6, pp. 6731–6747, 2021.
- [24] S. Adebayo, S. McLoone, and J. C. Dessing, "Hand-eye-object tracking for human intention inference," *IFAC-PapersOnLine*, vol. 55, no. 15, pp. 174–179, 2022.
- [25] Q. Li, J. Wang, C. C. Loy, and B. Dai, "Task-oriented human-object interactions generation with implicit neural representations," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 3035–3044, 2024.
- [26] Y. Wu, J. Wang, Y. Zhang, S. Zhang, O. Hilliges, F. Yu, and S. Tang, "Saga: Stochastic whole-body grasping with contact," in *European Conference on Computer Vision*, pp. 257–274, Springer, 2022.
- [27] J. Zheng, Q. Zheng, L. Fang, Y. Liu, and L. Yi, "Cams: Canonicalized manipulation spaces for category-level functional hand-object manipulation synthesis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 585–594, 2023.
- [28] M. Zhang, Y. Ye, T. Shiratori, and T. Komura, "Manipnet: neural manipulation synthesis with a hand-object spatial representation," 2021.
- [29] A. Ghosh, R. Dabral, V. Golyanik, C. Theobalt, and P. Slusallek, "Imos: Intent-driven full-body motion synthesis for human-object interactions," in *Computer Graphics Forum*, vol. 42, pp. 1–12, Wiley Online Library, 2023.
- [30] S. Christen, M. Kocabas, E. Aksan, J. Hwangbo, J. Song, and O. Hilliges, "D-grasp: Physically plausible dynamic grasp synthesis for hand-object interactions," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20577–20586, 2022.
- [31] C. Lugaresi, J. Tang, H. Nash, C. McClanahan, E. Uboweja, M. Hays, F. Zhang, C.-L. Chang, M. G. Yong, J. Lee, et al., "Mediapipe: A framework for building perception pipelines," *arXiv preprint arXiv:1906.08172*, 2019.
- [32] A. Van Den Oord, O. Vinyals, et al., "Neural discrete representation learning," *Advances in neural information processing systems*, vol. 30, 2017.
- [33] M. S. Yasar and T. Iqbal, "Vader: Vector-quantized generative adversarial network for motion prediction," in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 3827–3834, IEEE, 2023.
- [34] J. Zhang, Y. Zhang, X. Cun, S. Huang, Y. Zhang, H. Zhao, H. Lu, and X. Shen, "T2m-gpt: Generating human motion from textual descriptions with discrete representations," *arXiv preprint arXiv:2301.06052*, 2023.
- [35] Jakob Engel et al., "Project aria: A new tool for egocentric multi-modal ai research," *arXiv preprint arXiv:2308.13561*, 2023.
- [36] G. Prange, L. Smulders, J. van Wijngaarden, G. Lijbers, S. Nijenhuis, P. Veltink, J. Buurke, and A. Stienen, "User requirements for assistance of the supporting hand in bimanual daily activities via a robotic glove for severely affected stroke patients," in *2015 IEEE International Conference on Rehabilitation Robotics (ICORR)*, pp. 357–361, 2015.