2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

SignBot: Learning Human-to-Humanoid Sign Language Interaction

Guanren Qiao[1], Sixu Lin[1], Ronglai Zuo[2], Zhizheng Wu[1], Kui Jia[1,3], Guiliang Liu[1,†]

***Abstract*— Sign language is a natural and visual form of language that uses movements and expressions to convey meaning, serving as a crucial means of communication for individuals who are deaf or hard-of-hearing (DHH). However, the number of people proficient in sign language remains limited, highlighting the need for technological advancements to bridge communication gaps and foster interactions with minorities. Based on recent advancements in embodied humanoid robots, we propose SignBot, a novel framework for human-robot sign language interaction. SignBot integrates a cerebellum-inspired motion control component and a cerebral-oriented module for comprehension and interaction. Specifically, SignBot consists of: 1) Motion Retargeting, which converts human sign language datasets into robot-compatible kinematics; 2) Motion Control, which leverages a learning-based paradigm to develop a robust humanoid control policy for tracking sign language gestures; and 3) Generative Interaction, which incorporates translator, responder, and generator of sign language, thereby enabling natural and effective communication between robots and humans. Simulation and real-world experimental results demonstrate that SignBot can effectively facilitate human-robot interaction and perform sign language motions with diverse robots and datasets. SignBot represents a significant advancement in automatic sign language interaction on embodied humanoid robot platforms, providing a promising solution to improve communication accessibility for the DHH community. Please refer to our webpage: https://qiaoguanren.github.io/SignBot-demo/**

I. INTRODUCTION

Sign language, as the primary linguistic medium for the deaf and hard-of-hearing (DHH) communities, plays a vital role in bridging the communication barriers between these communities and others. Recent advancements in computer vision and large language models (LLMs) have significantly enhanced sign language applications, including generation, translation, and recognition [1], [2]. These advancements enable effective translation between sign language text, videos, and mesh representations. However, despite these advancements, their real-world impact on assisting individuals with disabilities remains limited. A key reason is that these systems are primarily demonstrated in models and cannot facilitate physical interaction with people in real-world scenarios.

To address this gap, Embodied Artificial Intelligence (EAI) integrates AI models into physical agents, enabling real-world interaction, task execution, and continuous learning. Recent progress in humanoid robots [3] highlights their potential, as human-like structures allow seamless integration into daily environments for tasks such as housekeeping, cooking, and navigation, thereby fostering natural human–robot interaction [4], [5]. However, no previous study has explored some methods for sign language applications. Teleoperation-based approaches [6], [7] typically rely on human manipulation, preventing robots from autonomously performing sign language. Learning-based control methods [8], [9] primarily focus on the robot's body without addressing the complexities of dexterous hand movements. Additionally, many dexterous robotic hands have limited degrees of freedom (DoFs), and the lack of wrist flexibility further restricts the accurate expression of the rich and diverse movements required for sign language.

Fig. 1: Motivation of SignBot: Human-Robot Sign Language Interaction.

To overcome these challenges, we introduce SignBot, an expressive robotic sign language framework designed for seamless interaction with sign language users. Figure 1 illustrates the motivation of our work. SignBot mainly consists of three key components: 1) *Motion Retargeting*, which maps the action sequences from human sign language datasets into a format compatible with robotic kinematics [8], [10]. 2) *Policy Training (SignBot's Cerebellum)*, which enables humanoid robots to first learn diverse sign language motions in a simulated environment with a decoupled policy [7], [8]. Specifically, we utilize decoupled body policies to learn the entire sign language gesture. The upper body, including the hands, learns to track the target sign language poses through imitation learning, while the lower body maintains a stable standing posture using a reinforcement learning (RL) policy [11], [12]. 3) *Sign Language Interaction (SignBot's Cerebral)* integrates a sign language translator [13], a sign language responder [14], and a sign language generator, enabling the robot to understand user expressions and respond appropriately in sign language. By combining these elements, our framework enhances real-time human-robot interaction, bridging the communication gap between sign language users and embodied robot systems.

We design various experiments to verify the performance of SignBot. Experimental results show that SignBot exhibits *accuracy*, *generalization*, *naturalness*, and *interactivity* with various datasets and robots. Overall, the main contributions

† Corresponding Author
1 School of Data Science, the Chinese University of Hong Kong, Shenzhen (E-mail: guanrenqiao1@link.cuhk.edu.cn; liuguiliang@cuhk.edu.cn)
2 Imperial College London
3 DexForce Technology

of our paper are as follows:

(1) **Human-Robot Interaction for Minority.** We propose an interactive sign language framework that enables seamless communication between robots and the DHH community.
(2) **Precise Sign Language Execution.** Our robot control policy robustly adapts to a diverse range of human sign language motions, ensuring stable and accurate execution.
(3) **Embodiment and Domain Adaptation.** Signbot can be transferred to different robots, achieving sign language interaction in Sim-to-Real scenarios.

II. RELATED WORK

Humanoid Robotic Imitation of Human Behavior. For the imitating human behavior problem of humanoid robots, researchers often adopt a whole-body control learning paradigm [3]. This paradigm consists mainly of two approaches. One approach is to decouple the upper and lower body policies, which are responsible for managing different parts of a humanoid robot. Representative works include Exbody [8], OmniH2O [7], etc. Although upper and lower body policies are decoupled, they can still be integrated into a whole-body control paradigm. The alternative approach involves providing reference motions for the humanoid robot. Given the physical similarities to humans, a promising reference is the collection of human movements from motion datasets, such as Exbody [8], ASAP [9], H2O [15], HWC-loco [16], etc. We simultaneously leverage the advantages of both methods to perform sign language. These reference motions provide rich signals for humanoid robots to imitate human-like motions.

Sign Language Processing. The field encompasses two primary research directions: sign language translation (SLT), and sign language generation (SLG). SLT and SLG form complementary pathways for bidirectional communication between deaf and hearing populations, specializing in sign-to-text and text-to-sign conversion, respectively. Some studies successfully incorporated language models (LMs) pre-trained on extensive natural language corpora into SLT frameworks, yielding substantial accuracy enhancements [1]. Recent state-of-the art SLG works can be categorized into two classes: the first group of methods [17] employs diffusion models to generate sign motions conditioned on text inputs; the second group of methods [2] considers the linguistic nature of sign languages and adopts a tokenizer-LM autoregressive generation approach.

III. PROBLEM DEFINITION

Robot Learning Environment. We formulate the task of tracking human sign language motions as a Partially Observable Markov Decision Process (POMDP) defined by the tuple $\mathcal{M} = (\mathcal{O}, \mathcal{S}, \mathcal{A}, P_{\mathcal{T}}, \mathcal{R}, \mu_0, \gamma)$, where: 1) Within the observation space $\mathcal{O}$, each observation $o_t \in \mathcal{O}$ consists of two components: proprioception (o_t^p) and goal imitation (o_t^y). The proprioception o_t^p includes essential motion-related information such as the root state, joint positions, and joint velocities. Meanwhile, the goal imitation o_t^y represents a unified encoding of the whole-body sign language pose that must be tracked during RL training. 2) $s_t \in \mathcal{S}$ records the complete information and environment of the robot. We summarize a state as $s_t = [o_t, o_{t-1}, ..., o_{t-H}]$ and each $o_t = [o_t^p, o_t^y]$. 3) $a_t \in \mathcal{A}$ denotes the action space, and action $a \in \mathcal{A}$ denotes the target joint positions that a PD controller uses to actuate the DoF. 4) $r_t = \mathcal{R}(s_t, a_t)$ denotes the reward functions, which typically consist of penalty, regularization, and task rewards. These reward signals determine the level of optimality in the control policy. 5) $P_{\mathcal{T}} \in \Delta_{\mathcal{S}\times\mathcal{A}}^{\mathcal{S}}$ denotes the transition function as a mapping from state-action pairs to a distribution of future states. 6) $\mu_0 \in \Delta^{\mathcal{S}}$ denotes the initial state distribution. 7) $\gamma \in (0, 1]$ denotes the discounting factor. Under this POMDP, our goal is learning a control policy $\pi : \mathcal{S} \to \mathcal{A}$ that can maximize the discounted cumulative rewards $\sum_{t=0}^{T-1} \gamma^t r_t$.

Human-Robot Sign Language Interaction. The problem of sign Language Interaction can be modelled as a closed-loop system. Firstly, the robot should observe a sequence of the user's sign motions $\boldsymbol{o}^{y_{\text{human}}} = [o_1^{y_{\text{human}}}, \ldots, o_K^{y_{\text{human}}}]$. Then, the robot translates sign language $\boldsymbol{o}^{y_{\text{human}}}$ into a sequence of text description $\boldsymbol{x} \sim \mathcal{X}$ ($\mathcal{X}$ denotes the space of text sequence) with the translation function $f_{\mathcal{T}} : \mathcal{O}^K \to \mathcal{X}$. To response, the system must understand the intention of $\boldsymbol{x}_t$ and answer $\boldsymbol{x}'_t$ with the responding function $f_{\mathcal{R}} : \mathcal{X} \to \mathcal{X}$. Based on the text response $\boldsymbol{x}'_t$, the system should generate sign language $\boldsymbol{o}_t^{y_{robot}}$ with the generation function $f_{\mathcal{G}} : \mathcal{X} \to \mathcal{O}^K$, which can be used as imitation goals for robot controller. Specifically, we formulate the sign language generation problem as a conditional sequence generation task, where the goal is to generate a sign language SMPL-X sequence from input semantic information. The output sign language sequence $\boldsymbol{o}^y$ is the SMPL-X representation space of sign language and K denotes the length of the motion sequence. Typically, this is modeled by the conditional probability distribution$P_{\mathcal{G}}(\boldsymbol{o}^y|\boldsymbol{x}) = \prod_{k=1}^{K} P_{\mathcal{G}}(o_k^y \mid o_{<k}^y, \boldsymbol{x})$.

IV. SIGNBOT

In this section, we introduce the pipeline of SignBot, which is divided into three parts: 1) **Motion Retargeting** of the body and hands, 2) **Policy Training** to control the robot's movements as *"SignBot's cerebellum"*, and 3) **Sign Language Reasoning** for comprehensive and responding users' sign languages as *"SignBot's cerebral"*. Figure 2 illustrates the SignBot pipeline.

A. Motion Retargeting

As shown in the first stage of Figure 2, we extract the motion from the video mesh for subsequent data processing. To mitigate differences in body shape between humans and the humanoid robot, we perform retargeting separately for the human body and hands.

Body Retargeting. Our humanoid body retargeting system is based on the [18], [19]. By establishing a mapping relationship between the source keypoints and the target keypoints, we follow a dual T-Pose (the standard poses of the source and target skeletons) as a spatial alignment reference. We convert the local quaternion of each joint to the form of axis angles. It is important to note that to make the robot's

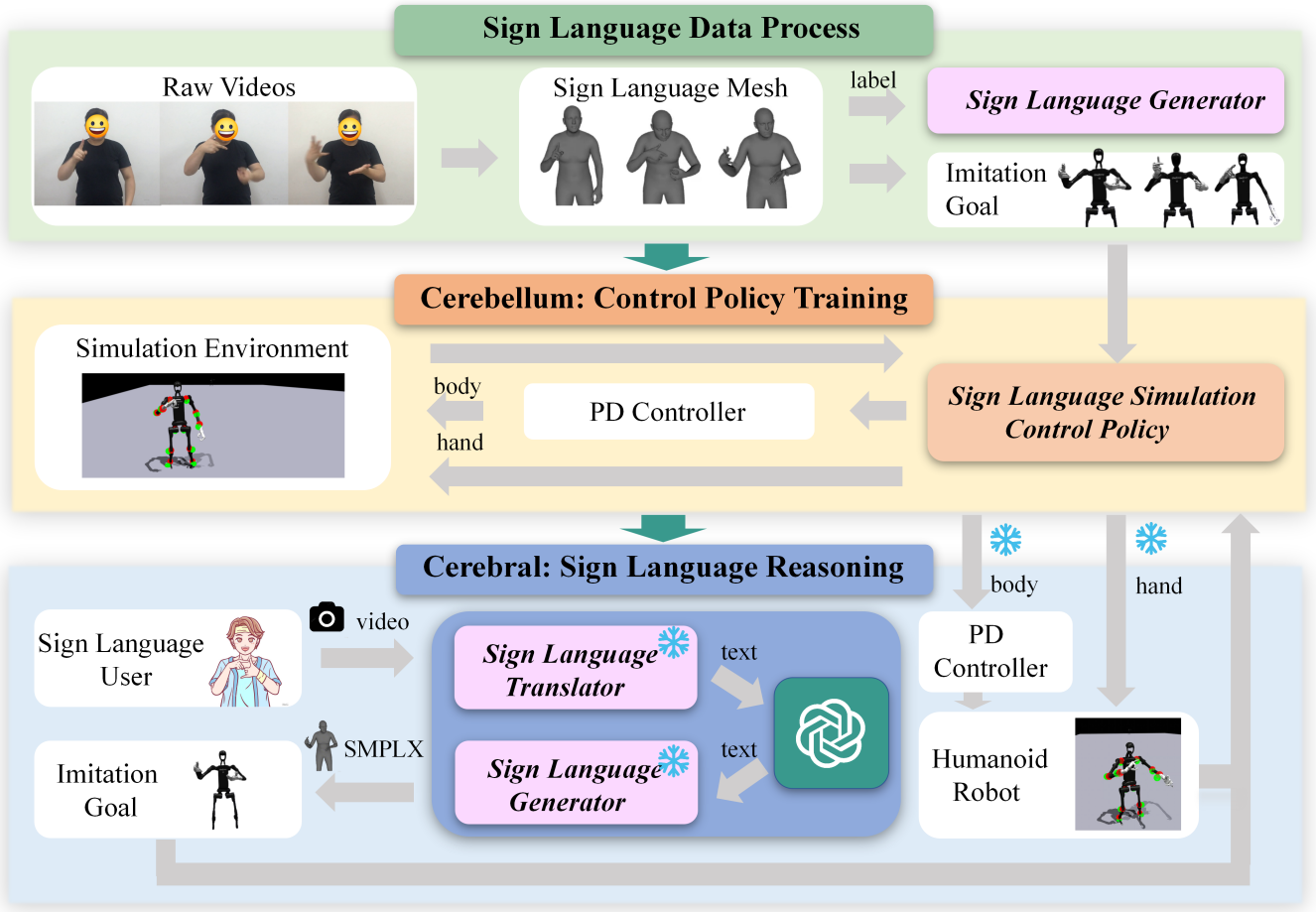

Fig. 2: Overview of SignBot: The framework consists of three stages: (1) *Motion Retargeting* aligns human sign language gestures with the body structure of humanoid robots (Section IV-A). In addition, we use the processed mesh along with text labels to train the sign language generator model. (2) *Cerebellum* performs Sim2Real policy training that enables the robot to track various sign language gestures in the simulated environment and deploy the policy to real-world (Section IV-B). (3) *Cerebral* conducts sign language reasoning to facilitate communication with sign language users through the sign language translator, response, and generator within the cerebral (Section IV-C).

sign language movements more natural, we add two additional degrees of freedom at the robot's wrist and transform the wrists and shoulders represented by a 1D joint to a 3D joint.

Hand Retargeting. To map human hand pose data to the joint positions of the Linker hand (the robot hands used by SignBot), we apply the preprocessing method from [10], adapting it specifically for the Linker hand. This process is often formulated as an optimization problem, where the difference between the keypoint vectors of the human hand model and the dexterous hand is minimized. Since the linker hand is larger than a typical dexterous hand, we adjust the scale factor in the optimization process. Additionally, we modify the regularization term to smooth the sign language movements between consecutive frames.

B. SignBot's Cerebellar: Control Policy Training

Within our SignBot humanoid agent, the cerebellum controls the low-level movements for performing sign language. As shown in the second stage of Figure 2, we train a control policy to enable the humanoid robot to track and imitate sign language gestures in a simulated environment. In this section, we discuss our approach from three key perspectives: observation space, decoupled policy, and reward design.

Observation Space. SignBot's observation $o_t \in \mathcal{O}$ consists of proprioception (o_t^p) and goal imitation (o_t^y) (Section III). Our proprioception is defined as $o_t^p \triangleq [\mathbf{q}_t, \dot{\mathbf{q}}_t, \mathbf{v}_t, \mathbf{w}_t, \mathbf{g}_t, \mathbf{a}_{t-1}]$, which includes joint position $\mathbf{q}_t \in \mathbb{R}^{55}$ (DoF position), joint velocity $\dot{\mathbf{q}}_t \in \mathbb{R}^{55}$ (DoF velocity), root linear velocity $\mathbf{v}_t \in \mathbb{R}^3$, root angular velocity $\mathbf{w}_t \in \mathbb{R}^3$, root projected gravity $\mathbf{g}_t \in \mathbb{R}^3$, and the previous action $\mathbf{a}_{t-1} \in \mathbb{R}^{55}$. The goal observation is $\mathbf{o_t^y} \triangleq \left[\hat{\mathbf{q}}^{kp}, \hat{\mathbf{q}}, \dot{\hat{\mathbf{q}}}\right]$, where $\hat{\mathbf{q}}^{kp} \in \mathbb{R}^{14\times3}$ are the positions of 14 selected reference keypoints [8] to ensure that the humanoid robot and the imitation goal are oriented in the same direction, $\hat{\mathbf{q}} \in \mathbb{R}^{55\times3}$ are the positions of all reference joints, and $\dot{\hat{\mathbf{q}}}$ is the linear velocity of the reference joints. At the same time, we generally perform domain randomization [7] in the simulation environment to ensure robustness.

Decoupled Policy. Given that sign language involves precise coordination of both hand poses and full-body motion with high DoFs, learning a unified control policy is inherently challenging. More importantly, while the dual arms in the upper body can often be governed by a shared control strategy,

the control approach for the lower body may vary significantly depending on the physical design of the humanoid robot. For instance, bipedal humanoid robots typically utilize RL controller, whereas wheeled robots often employ model predictive control (MPC) techniques.

Motivated by recent advances in whole-body humanoid control [8], we adopt a decoupled architecture that separates the control policies of the upper and lower body.The primary objective of the upper-body policy π^{upper} is to track the retargeted actions, while the lower-body policy π^{lower} ensures balance in the robot's default standing pose while adapting to the movements of π^{upper}.

$$\pi^{upper} = \arg\min_{\pi} \mathbb{E}_{(s,a_*^{up})} \Big[\mathcal{D}_f[\delta(a_*^{up}) \| \pi(a^{up}|s)] \Big] \tag{1}$$

$$\pi^{lower} = \arg\max_{\pi} \mathbb{E}_{\pi(a^{low}|s)} \left[\sum_{t=0}^{\infty} \gamma^t r_t(s,a) | a^{lower} \sim \pi^{upper} \right]$$

Where: 1) a_*^{up} represents the retargeted action of the upper body, and δ denotes the Dirac delta function. 2) $\mathcal{D}_f(\cdot\|\cdot)$ refers to the f-divergence (e.g., KL-divergence) between two distributions. 3) The whole-body humanoid action, $a = [a^{low}, a^{up}]$, is the concatenation of upper and lower body actions. 4) The reward from the humanoid control learning environment can be represented by the weighted penalty $r_{\mathfrak{P}}$, task $r_{\mathfrak{T}}$, and regularization $r_{\mathfrak{R}}$ terms: $r_t = \beta_{\mathfrak{T}} r_{\mathfrak{T}} + \beta_{\mathfrak{P}} r_{\mathfrak{P}} + \beta_{\mathfrak{R}} r_{\mathfrak{R}}$. In our task, task rewards ($r_{\mathfrak{T}}$) measure the robot's performance in tracking joint motions and body velocities. Penalty rewards ($r_{\mathfrak{P}}$) serve to discourage undesirable outcomes such as falling and violating dynamic constraints like joint or torque limits. Regularization rewards ($r_{\mathfrak{R}}$) are used to align the humanoid's sign language gestures with human preferences.

Note that to train the lower-body policy π^{lower}, we apply the PPO [20], [21] algorithm for legged humanoid robots and MPC for wheeled humanoid robots. In this way, SignBot can be scaled to multiple embodiments, as we demonstrate in the experiment.

C. SignBot's Cerebral: Sign Language Reasoning

Within our SignBot humanoid agent, the cerebral system controls high-level reasoning skills, enabling it to respond to sign language and generate appropriate responses. An ideal embodied humanoid robot needs to have both cerebellar and cerebral capabilities, enabling it to interact effectively with its surrounding environment and other agents. The third stage of Figure 2 illustrates the cerebral system of the SignBot. This system processes the user's input and generates an appropriate sign language response, which serves as the imitation goal for the control policy ($\mathbf{o_t^c}$ in Section IV-B).

Implementation. SignBot utilizes a camera to observe the motions of sign language users and then stores this as a video input into its cerebral system [22]. For communicating with sign language users in real-time, SignBot's cerebral system is implemented by three models: a *sign language translator* understanding sign language contents, a *sign language responder* interpreting semantics and generating responses, and a *sign language generator* converting texts to the SMPLX format with Transformer-based models. We introduce these two models in the following:

Sign Language Translator is implemented with a LLM [13]. In the pre-training stage, it extracts the sign language features from sign language videos and images, aligns them with the dimensions of the language model, and then inputs them into the model. In the fine-tuning stage, we utilize the translation text from the sign language dataset [23], [24] to construct the supervised labels for fine-tuning our translation model.

Sign Language Responder is implemented using the DeepSeek-Chat API [14], [25] due to its ability to seamlessly comprehend semantic information and facilitate multi-turn conversations. Other chat models can also be used here. To simulate natural conversations with sign language users, we design a suitable prompt template [12]. When generating a response, the system integrates the text-format sign language into the prompt and produces a contextually appropriate reply. We ensure that DeepSeek is aware of the vocabulary and scope covered by the CSL dataset to prevent the output of dangerous or out-of-distribution (OOD) corpus.

Sign Language Generator autoregressively generates sign motions from text input based on a multilingual LM [26], [27]. First, we design a decoupled VQVAE tokenizer to map continuous sign motions to discrete tokens over upper body (UB), left hand (LH), and right hand (RH) movements. Given a K-frame sign sequence, we decompose it into three part-wise motion sequences based on the SMPL-X: $\mathbf{C}^u \in \mathbb{R}^K$, where $u \in \{\text{UB}, \text{LH}, \text{RH}\}$. For each body part, we train a separate VQ-VAE comprising an encoder that projects the sequence into a latent space $\mathbf{C}^u_{f_{enc}} = \{c^u_{f_{enc},k}\}_{k=1}^K \in \mathbb{R}^{K_{f_{enc}} \times C}$, a decoder for reconstruction, and a learnable codebook $\mathbf{Z}_u \in \mathbb{R}^{N_Z^u \times d}$, where N_Z^u represents the number of codes and d denotes the code dimension. Then, for each pose, we can derive a set of discrete tokens $\hat{\boldsymbol{z}}^u_{1,\dots,K} = [\hat{z}^u_1, \dots, \hat{z}^u_K]$, which searches for the nearest neighbor from the codebook Z_u:

$$\hat{z}^u_k = \arg\min_{z_n \in \mathbf{Z}_u} \| s^u_{f_{enc},k} - z_n \|^2, \quad n \in [1, N_Z^u] \tag{2}$$

where $\forall u \in \{\text{UB}, \text{LH}, \text{RH}\}$. Given a text description $\boldsymbol{x}$, the generator retrieves word-level signs based on the $\boldsymbol{x}$ from external dictionaries made by the decoupled tokenizer. These word-level signs are represented with discrete tokens $\hat{\boldsymbol{z}}^u_{1,\dots,K}$. We feed these tokens and text sequence $\boldsymbol{x}$ into the LM encoder at the same time. During decoding, we adopt a multi-head decoding strategy [28]. We design three language modeling heads, implemented as fully connected layers, to predict motion tokens for each body part from m simultaneously at each step. The decoding process can be formulated as:

$$P_{Dec}(\boldsymbol{o}^y|\mathbf{h}) = \prod_{k=1}^{K} \prod_{u} P_{Dec}(o_k^{y^u} | \boldsymbol{o}^{y,\mathbf{h}}_{<k}) \tag{3}$$

where $(\boldsymbol{o}^{y^u}_{<k}, \mathbf{h})$ is the simplification of $\boldsymbol{o}^{y,\mathbf{h}}_{<k}$, $\boldsymbol{o}^y = \{o_1^{y^{\text{UB}}}, o_1^{y^{\text{LH}}}, o_1^{y^{\text{RH}}}, \dots, o_K^{y^{\text{UB}}}, o_K^{y^{\text{LH}}}, o_K^{y^{\text{RH}}}\}$ and $\mathbf{h} = f_{\text{Enc}}(\boldsymbol{x}, \hat{\boldsymbol{z}}^u_{1,\dots,K})$ denotes the output of LM Encoder.

Finally, the derived motion tokens are used to reconstruct sign motions.

Sim-to-Real Deployment. In real-world environments, robot movement from positions from t to $t+1$ is not instantaneous. To ensure smooth transitions, we employ the Ruckig algorithm [29] for online trajectory generation with third-order (jerk) constraints and complete kinematic targets. Ruckig computes time-optimal trajectories between arbitrary states, defined by position, velocity, and acceleration, while respecting velocity, acceleration, and jerk limits. To maintain smooth motion, we interpolate intermediate values between target positions. For safety during deployment, we avoid commanding joint angles near their physical limits, as slight discrepancies between simulation and reality may lead to motor power loss or low-voltage issues, even when using identical limit settings.

V. Environment

Experiment Settings. To conduct a comprehensive evaluation, we quantify the performance of SignBot in simulated (IssacGym [30]) and realistic environments from the following perspectives: 1) *Accuracy*: How to effectively align sign language actions between robots and humans? 2) *Generalization*: How well does SignBot perform with diverse sign language datasets and different robots? 3) *Naturalness*: How effectively does SignBot imitate human-like sign language norms in the real-world interaction scenarios? 4) *Interactivity*: What is the DHH community's genuine evaluation of this work? We use the public CSL-Daily [23] and How2Sign (ASL) [24] Sign Language dataset. Experiments span different embodiments: H1 legged robot, W1 wheeled robot, and Linker hand.

Metrics. We adopt consistent metrics across settings, including errors in 1) DoF positions, body yaw, linear velocity, and roll/pitch. 2) *Cumulative Rewards*: are calculated by all the weighted reward functions. 3) *BLEU-1* calculates the single-word overlap between the generated text and the reference. 4) *BLEU-4* evaluates the match of four-word sequences (4-grams). 5) *ROUGE* assesses how well the generated output covers the essential content of the reference. 6) *DTW-PA-JPE* [17] evaluates sequence-level distances between the generated signs and ground truth.

Comparison Methods. To demonstrate the effectiveness of SignBot, we compare with other baselines based on whole-body control or RL as follows: 1) **SignBot (w/o Lower-Body Tracking** follows SignBot, but allows the lower body to self-adapt and maintain balance. 2) **Whole-Body Tracking + AMP** uses an AMP reward [18] to encourage the transitions of the policy to be similar to the motions of the sign language features. 3) **Whole-Body Tracking** [7] learns the movement of the upper body and lower body simultaneously.

A. Accuracy: Alignment of Sign Language between Human and Robot

Precise execution of sign language is vital for effective communication within the DHH community, as even minor inaccuracies can lead to misunderstandings, in contrast to boxing techniques, where approximate gesture replication suffices for demonstration purposes. To construct a high-fidelity dataset with precise 3D SMPLX annotations, we leverage state-of-the-art methods for 3D hand [31] and body reconstruction [32]. Specifically, for each 2D sign language video, we first utilize OSX [32] to obtain an initial body pose estimation. Recognizing that OSX often struggles with accurate arm and hand pose estimation, we implement a two-stage refinement process. For high-precision hand pose reconstruction, we employ WiLoR [31], a cutting-edge 3D reconstruction pipeline that can robustly detect and reconstruct even challenging hand configurations. The hand pose parameters and global orientation outputs from WiLoR are then directly integrated to replace the corresponding OSX-derived estimates, ensuring high fidelity.

As the datasets contain multiple demonstrators performing the same actions, we select the sign language motions of one demonstrator and segment them by difficulty based on the length of the label (longer labels correspond to motions of longer duration, for example, labels with 10 words, 10-20 words, and 20 words). We divide the CSL-Daily sign language dataset [23] into three difficulty levels: simple (929 sentences), intermediate (4558 sentences), and difficult (1089 sentences). We compare SignBot with the previously mentioned baselines on the data across these three difficulty levels.

Notably, since the Unitree H1's wrist has only one DoF, it can only rotate while maintaining a fixed orientation and cannot bend. To address this, we modify the robot by adding two additional DoFs to the wrist, which allows the wrist to move up/down and left/right, enabling more flexible sign language gestures. Table I illustrates the training performance of each baseline under different difficulty levels. The results indicate that SignBot achieves the lowest error in metrics such as DoF position tracking and yaw angle tracking, while also achieving the highest reward value, demonstrating the effectiveness of the SignBot control policy. The improvement in tracking accuracy and the reduction in training difficulty are attributed to the provision of upper body DoF position data and the tracking of lower body bent standing position. SignBot excels in controlling yaw and roll&pitch angles, indicating strong stability of the robot base. We also test other baselines, such as tracking the upper and whole body DoF positions, but the performance of these is unsatisfactory. This is because sign language differs from other upper-body movements; it is flexible and varied, and the frequency of sign language actions in the dataset is relatively fast. In addition, we observe that the length of the sentences does not significantly affect SignBot's performance. When the sentences are short, SignBot tends to have larger errors in tracking keypoints. This may be due to the shorter episode length, which causes the robot to complete and reset its random learning for the next action quickly. During this period, it needs to adjust the robot's global pose frequently.

B. Generalization: Imitation across diverse datasets and robots

Generalizing SignBot to different embodiments. We retarget the human sign language data to align with the joints

Baseline \ Metric	DoF Pos ↓	Yaw ↓	Linear Velocity ↓	Roll&Pitch ↓	Cumulative Rewards ↑
			Easy		
Whole-Body Tracking	0.97 ± 0.01	0.16 ± 0.01	$\mathbf{0.18 \pm 0.01}$	0.05 ± 0.00	10.41 ± 0.70
Whole-Body Tracking + AMP	1.13 ± 0.01	0.11 ± 0.00	0.22 ± 0.00	0.04 ± 0.01	8.89 ± 0.32
SignBot (w/o Lower-Body Tracking)	0.85 ± 0.01	0.95 ± 0.01	1.86 ± 0.02	0.26 ± 0.01	0.54 ± 0.01
SignBot	$\mathbf{0.59 \pm 0.01}$	$\mathbf{0.04 \pm 0.00}$	0.20 ± 0.01	$\mathbf{0.03 \pm 0.01}$	$\mathbf{12.46 \pm 0.40}$
			Medium		
Whole-Body Tracking	0.95 ± 0.01	0.17 ± 0.00	$\mathbf{0.17 \pm 0.01}$	0.05 ± 0.00	10.63 ± 0.63
Whole-Body Tracking + AMP	1.12 ± 0.01	0.10 ± 0.00	0.22 ± 0.00	$\mathbf{0.04 \pm 0.00}$	9.18 ± 0.32
SignBot (w/o Lower-Body Tracking)	0.87 ± 0.01	0.97 ± 0.00	1.87 ± 0.02	0.27 ± 0.01	0.33 ± 0.01
SignBot	$\mathbf{0.61 \pm 0.00}$	$\mathbf{0.08 \pm 0.00}$	0.20 ± 0.01	$\mathbf{0.04 \pm 0.01}$	$\mathbf{12.24 \pm 0.57}$
			Hard		
Whole-Body Tracking	0.94 ± 0.01	0.17 ± 0.00	$\mathbf{0.17 \pm 0.01}$	0.05 ± 0.00	9.67 ± 0.20
Whole-Body Tracking + AMP	1.11 ± 0.02	0.10 ± 0.00	0.23 ± 0.01	0.04 ± 0.01	9.04 ± 0.24
SignBot (w/o Lower-Body Tracking)	0.85 ± 0.00	0.92 ± 0.01	2.00 ± 0.02	0.04 ± 0.00	0.52 ± 0.03
SignBot	$\mathbf{0.59 \pm 0.01}$	$\mathbf{0.04 \pm 0.00}$	0.20 ± 0.03	$\mathbf{0.03 \pm 0.01}$	$\mathbf{12.56 \pm 0.30}$

TABLE I: Tracking Performance: We compare model performance under three difficulty levels. Bolded results indicate the best performance. The rewards calculation denotes the cumulative returns over a trajectory, while the other metrics are calculated using the mean square error.

of humanoid robots to visualize the precision in imitating sign language poses. Figure 3 illustrates the alignment between human and robot sign language gestures after applying our SignBot method, demonstrated on both the legged Unitree H1 and the wheeled W1 robots. A notable observation is that SignBot maintains a high accuracy in imitating sign language across both robots.

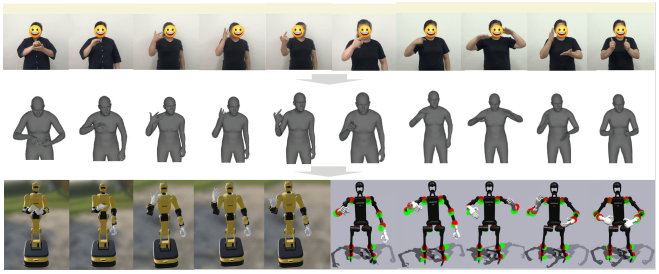

Fig. 3: Sign Language Alignment between Human and Robots: We display the source video of human sign language in the first row, followed by the mesh from video processing, and the last row shows the results of different robots. For the H1 robot, the red nodes represent the robot dof pos, while the green nodes represent the retargeted demonstration nodes.

Generalizing SignBot to different Datasets. To demonstrate the generalization capability of SignBot, we also evaluate SignBot with the How2Sign dataset. Specifically, we preprocess a portion of the 2,286 data entries and evaluate the robustness of the policy for different sign languages following Table I. Table II illustrates SignBot's performance on the How2Sign dataset, demonstrating that SignBot achieves a minimal tracking error.

Baseline \ Metric	DoF Pos ↓	Yaw ↓	Linear Velocity ↓	Roll&Pitch ↓
		Easy		
Whole-Body Tracking	1.05 ± 0.02	0.32 ± 0.00	$\mathbf{0.16 \pm 0.01}$	$\mathbf{0.04 \pm 0.00}$
SignBot	$\mathbf{0.63 \pm 0.02}$	$\mathbf{0.07 \pm 0.02}$	0.19 ± 0.03	0.07 ± 0.01
		Medium		
Whole-Body Tracking	0.88 ± 0.02	0.34 ± 0.01	$\mathbf{0.15 \pm 0.02}$	$\mathbf{0.04 \pm 0.01}$
SignBot	$\mathbf{0.57 \pm 0.02}$	$\mathbf{0.11 \pm 0.01}$	0.17 ± 0.02	0.05 ± 0.00
		Hard		
Whole-Body Tracking	0.90 ± 0.03	0.30 ± 0.03	$\mathbf{0.17 \pm 0.02}$	$\mathbf{0.05 \pm 0.01}$
SignBot	$\mathbf{0.56 \pm 0.01}$	$\mathbf{0.11 \pm 0.02}$	0.18 ± 0.02	$\mathbf{0.05 \pm 0.01}$

TABLE II: Generalization Experiment.

C. *Naturalness: Sim-to-Real Human-Robot Interaction*

To demonstrate the naturalness of SignBot, we choose the W1 robot to demonstrate the performance of the SignBot framework in realistic environments. The W1 robot and Linker Hand are controlled through the ROS system to drive the various joints during real deployment. Figure 4 illustrates several examples of interactions between a human and a robot functioning as a supermarket cashier. The individual shown in the figure is a proficient sign language user from our research team. In this scenario, the robot initiates communication by asking customers about their intended purchases. More examples can be found in the supplementary video material.

We evaluate the performance of the sign language translation and generation modules in interaction. We compare our modules with the current SOTA baselines on the ASL and CSL datasets using the test sets (since the translator is a gloss-free method, we compare it with the gloss-free baseline instead of the gloss-based baseline). Additionally, we test the effectiveness of both modules using real-world sign language videos and dialogue corpus. Since our sign language users are familiar with CSL, the actual performance is only evaluated on CSL. Table III shows that our modules outperform the existing SOTA baseline. "Deployment" refers to the evaluation using our real test data from non-CSL test sets during the actual device deployment phase. Furthermore, since the dialogue content mainly falls within the vocabulary and scope covered by the CSL dataset, the performance degradation is not significant, providing a basic guarantee for sign language interaction.

D. *Interactivity: Human Feedback from the DHH Community*

We conduct a user study and invite members of the DHH community to evaluate this work. Due to constraints of the experimental venue and hardware safety concerns, we do not involve DHH individuals in actual interactions. As an alternative, we provide comprehensive documentation and video recordings of the entire sign language interaction process to facilitate their assessment.

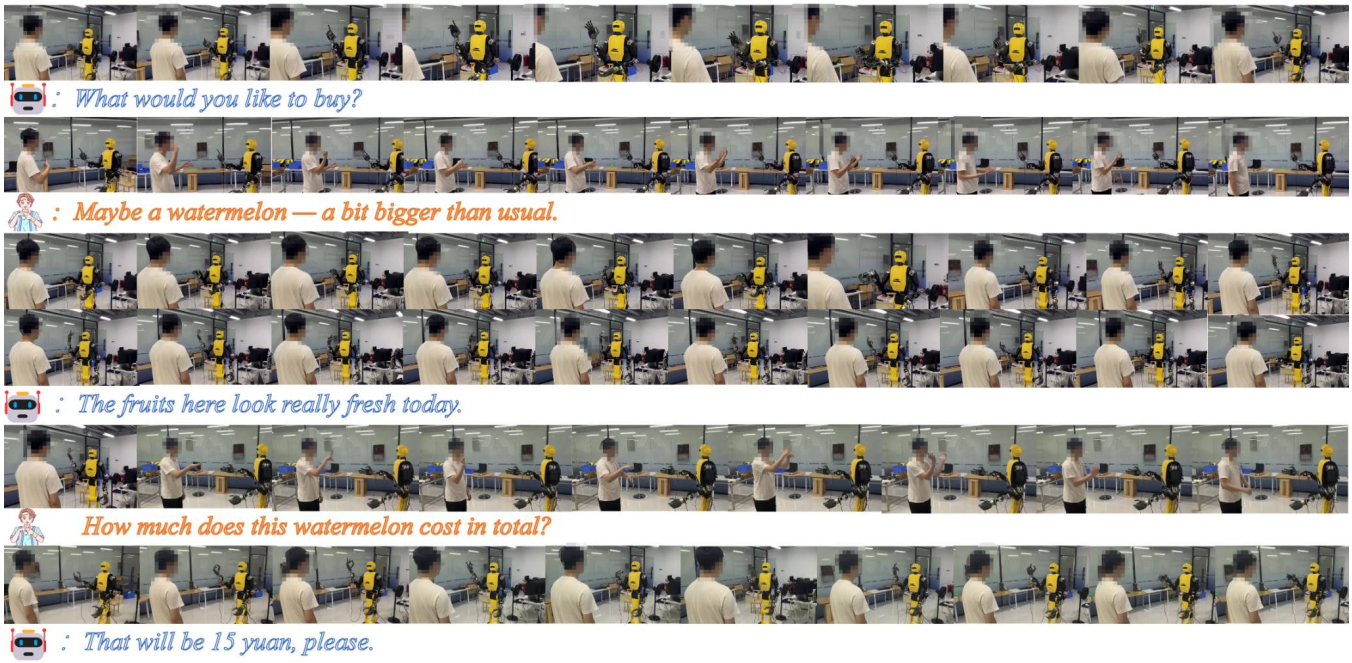

Fig. 4: An example of real-world interaction between the robot and the human customer.

Translation Method	BLEU-1 ↑	BLEU-4 ↑	ROUGE ↑	Generation Method	DTW-PA-JPE (body) ↓	DTW-PA-JPE (hand) ↓
C^2RL (CSL) [33]	49.32	21.61	48.21	NSA (ASL) [17]	7.83	7.33
Translator (ASL)	40.20	14.90	36.01	Generator (ASL)	6.82	2.35
Translator (CSL)	55.08	26.36	56.51	Generator (CSL)	6.24	1.71
Translator (CSL, Deployment)	53.20	20.53	54.48	Generator (CSL, Deployment)	7.63	2.19

TABLE III: Evaluation of Sign Language Translation and Generation Models in the simulation and real environment.

Evaluator	Satisfaction	Enjoyment	Naturalness	Accuracy	Understanding	Convenience
Evaluator 1	6	5	6	5	6	8
Evaluator 2	7	8	7	6	5	5
Evaluator 3	8	7	9	7	7	8
Evaluator 4	7	6	7	7	7	8
Evaluator 5	6	6	6	5	6	8
Evaluator 6	6	9	5	7	6	9
Evaluator 7	9	9	7	8	6	5
Evaluator 8	3	3	3	3	3	3
Evaluator 9	1	1	1	1	5	6
Evaluator 10	8	6	7	7	7	6

TABLE IV: Evaluation Results from the DHH Community

Experiment Setup: Among these people, there are young and elderly individuals, all coming from different places of birth. To ensure fairness, we do not collect any personal or private information from the experiment participants. The assessment covers the following aspects: 1) Overall satisfaction with the system, 2) enjoyment of the interactive experience, 3) naturalness and fluency of responses, 4) accuracy of motion execution, 5) ability to understand human-expressed intentions, and 6) whether the system could bring convenience to their daily lives. Ratings are given on a scale of 1–10, where 10 represents the highest score. A score of 6 represents a baseline level of satisfaction.

Table IV presents the individual ratings from all 10 evaluators. Most evaluators give scores above 6 in the majority of categories, indicating general satisfaction with the system's performance. The convenience criterion received the highest average score (mean: 6.60), suggesting that evaluators think this research is meaningful and has the potential to positively impact their daily lives. If we exclude Evaluators 8 and 9 (likely outliers due to unfamiliar regional signs), the average score across all criteria increases significantly (e.g., Satisfaction average rises from 6.10 to 7.13). Naturalness (from 5.80 to 6.75) and Accuracy (from 5.60 to 6.50) receive moderately high scores, despite the absence of facial expressions or lip movements.

Limitation. Based on feedback from members of the DHH community, we have summarized several current limitations of our framework: 1) *Limitations in Sign Language Translation and Generation.* Current state-of-the-art open-source models for sign language translation and generation still exhibit errors. Although we currently use public sign language datasets, our research has found that sign language can vary across different regions within the same country. 2) *Limitations of hardware support*. Our robot cannot support facial expressions or lip movements, which may affect the understanding of certain sign language content.

Despite these limitations, a significant portion of the DHH community strongly affirms both the technical quality of our implementation and the social value of our research.

VI. Conclusion

We introduce SignBot, a human-robot sign language interaction framework that incorporates an embodied cerebellum + cerebral cooperation mechanism. This framework has been validated across various sign language motions, demonstrating exceptional accuracy, generalization, naturalness, and adaptability across diverse sign language scenarios. In particular, the cerebellum + cerebral cooperation mechanism in SignBot achieves reliable performance in daily communication through

the sign language translator, response, and generator. We think SignBot is a foundational solution for sign language applications, such as daily sign language robots serving the DHH community. In the future, we will focus on enabling the robot to exhibit facial expressions, enhancing the interaction experience, and collecting field data in different provinces to accommodate regional variations in sign language.

ACKNOWLEDGMENTS

This work is supported in part by Shenzhen Science and Technology Program under grant KJZD20240903104008012, Shenzhen Science and Technology Program under grant ZDCY20250901113000001, CUHK-CUHK(SZ)-GDSTC Joint Collaboration Fund No. 2025A0505000053, GuangDong Key Laboratory of Big Data Computing (2021B1212040002) and Guangdong Provincial Key Laboratory of Mathematical Foundations for Artificial Intelligence (2023B1212010001).

REFERENCES

[1] P. Jiao, Y. Min, and X. Chen, "Visual alignment pre-training for sign language translation," in *European Conference on Computer Vision, (ECCV)*, 2024.

[2] Z. Yu, S. Huang, Y. Cheng, and T. Birdal, "Signavatars: A large-scale 3d sign language holistic motion dataset and benchmark," in *European Conference on Computer Vision (ECCV)*, 2024, pp. 1–19.

[3] Z. Gu, J. Li, W. Shen, W. Yu, Z. Xie, S. McCrory, X. Cheng, A. Shamsah, R. Griffin, C. K. Liu, A. Kheddar, X. B. Peng, Y. Zhu, G. Shi, Q. Nguyen, G. Cheng, H. Gao, and Y. Zhao, "Humanoid locomotion and manipulation: Current progress and challenges in control, planning, and learning," *arXiv preprint arXiv:2501.02116*, 2025.

[4] Y. Liu, W. Chen, Y. Bai, G. Li, W. Gao, and L. Lin, "Aligning cyber space with physical world: A comprehensive survey on embodied ai," *arXiv preprint arXiv:2407.06886*, 2024.

[5] S. Nasiriany, A. Maddukuri, L. Zhang, A. Parikh, A. Lo, A. Joshi, A. Mandlekar, and Y. Zhu, "Robocasa: Large-scale simulation of everyday tasks for generalist robots," *arXiv preprint arXiv:2406.02523*, 2024.

[6] Z. Fu, Q. Zhao, Q. Wu, G. Wetzstein, and C. Finn, "Humanplus: Humanoid shadowing and imitation from humans," *Conference on Robot Learning, (CoRL)*, 2024.

[7] T. He, Z. Luo, X. He, W. Xiao, C. Zhang, W. Zhang, K. Kitani, C. Liu, and G. Shi, "Omnih2o: Universal and dexterous human-to-humanoid whole-body teleoperation and learning," in *Conference on Robot Learning, (CoRL)*, 2024.

[8] X. Cheng, Y. Ji, J. Chen, R. Yang, G. Yang, and X. Wang, "Expressive whole-body control for humanoid robots," in *Proceedings of Robotics: Science and Systems, (RSS)*, 2024.

[9] T. He, J. Gao, W. Xiao, Y. Zhang, Z. Wang, J. Wang, Z. Luo, G. He, N. Sobanbab, C. Pan, *et al.*, "Asap: Aligning simulation and real-world physics for learning agile humanoid whole-body skills," *arXiv preprint arXiv:2502.01143*, 2025.

[10] Y. Qin, W. Yang, B. Huang, K. Van Wyk, H. Su, X. Wang, Y.-W. Chao, and D. Fox, "Anyteleop: A general vision-based dexterous robot arm-hand teleoperation system," in *Robotics: Science and Systems, (RSS)*, 2023.

[11] R. S. Sutton, "Reinforcement learning: An introduction," *A Bradford Book*, 2018.

[12] S. Liu, S. Xu, W. Qiu, H. Zhang, and M. Zhu, "Explainable reinforcement learning from human feedback to improve language model alignment," in *Advances in Neural Information Processing Systems, (NeurIPS)*, 2025.

[13] Z. Li, W. Zhou, W. Zhao, K. Wu, H. Hu, and H. Li, "Uni-sign: Toward unified sign language understanding at scale," *International Conference on Learning Representations, (ICLR)*, 2025.

[14] D. G. DeepSeek-AI, D. Yang, H. Zhang, J. Song, R. Zhang, R. Xu, Q. Zhu, S. Ma, P. Wang, X. Bi, *et al.*, "Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning," *arXiv preprint arXiv:2501.12948*, 2025.

[15] T. He, Z. Luo, W. Xiao, C. Zhang, K. Kitani, C. Liu, and G. Shi, "Learning human-to-humanoid real-time whole-body teleoperation," *IEEE/RSJ International Conference on Intelligent Robots and Systems, (IROS)*, 2024.

[16] S. Lin, G. Qiao, Y. Tai, A. Li, K. Jia, and G. Liu, "Hwc-loco: A hierarchical whole-body control approach to robust humanoid locomotion," *arXiv preprint arXiv:2503.00923*, 2025.

[17] V. Baltatzis, R. A. Potamias, E. Ververas, G. Sun, J. Deng, and S. Zafeiriou, "Neural sign actors: A diffusion model for 3d sign language production from text," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, (CVPR)*, 2024, pp. 1985–1995.

[18] X. B. Peng, Y. Guo, L. Halper, S. Levine, and S. Fidler, "Ase: Large-scale reusable adversarial skill embeddings for physically simulated characters," *ACM Transactions On Graphics (TOG)*, vol. 41, no. 4, pp. 1–17, 2022.

[19] J. P. Araujo, Y. Ze, P. Xu, J. Wu, and C. K. Liu, "Retargeting matters: General motion retargeting for humanoid motion tracking," *arXiv preprint arXiv:2510.02252*, 2025.

[20] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *arXiv preprint arXiv:1707.06347*, 2017.

[21] G. Qiao, G. Liu, P. Poupart, and Z. Xu, "Multi-modal inverse constrained reinforcement learning from a mixture of demonstrations," *Advances in Neural Information Processing Systems, (NeurIPS)*, vol. 36, pp. 60 384–60 396, 2023.

[22] H.-C. Dan, Z. Huang, B. Lu, and M. Li, "Image-driven prediction system: Automatic extraction of aggregate gradation of pavement core samples integrating deep learning and interactive image processing framework," *Construction and Building Materials*, vol. 453, p. 139056, 2024.

[23] H. Zhou, W. Zhou, W. Qi, J. Pu, and H. Li, "Improving sign language translation with monolingual data by sign back-translation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, (CVPR)*, 2021.

[24] A. Duarte, S. Palaskar, L. Ventura, D. Ghadiyaram, K. DeHaan, F. Metze, J. Torres, and X. Giro-i Nieto, "How2sign: a large-scale multimodal dataset for continuous american sign language," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, (CVPR)*, 2021, pp. 2735–2744.

[25] J. Zhuang, H. Jin, Y. Zhang, Z. Kang, W. Zhang, G. G. Dagher, and H. Wang, "Exploring the vulnerability of the content moderation guardrail in large language models via intent manipulation," *arXiv preprint arXiv:2505.18556*, 2025.

[26] Y. Liu, J. Gu, N. Goyal, X. Li, S. Edunov, M. Ghazvininejad, M. Lewis, and L. Zettlemoyer, "Multilingual denoising pre-training for neural machine translation," *Transactions of the Association for Computational Linguistics*, vol. 8, pp. 726–742, 2020.

[27] R. Zuo, R. A. Potamias, E. Ververas, J. Deng, and S. Zafeiriou, "Signs as tokens: A retrieval-enhanced multilingual sign language generator," in *International Conference on Computer Vision, (ICCV)*, 2025.

[28] G. Qiao, G. Quan, R. Qu, and G. Liu, "Modelling competitive behaviors in autonomous driving under generative world model," in *European Conference on Computer Vision, (ECCV)*. Springer, 2024, pp. 19–36.

[29] L. Berscheid and T. Kröger, "Jerk-limited real-time trajectory generation with arbitrary target states," *Robotics: Science and Systems XVII, (RSS)*, 2021.

[30] V. Makoviychuk, L. Wawrzyniak, Y. Guo, M. Lu, K. Storey, M. Macklin, D. Hoeller, N. Rudin, A. Allshire, A. Handa, *et al.*, "Isaac gym: High performance gpu-based physics simulation for robot learning," *arXiv preprint arXiv:2108.10470*, 2021.

[31] R. A. Potamias, J. Zhang, J. Deng, and S. Zafeiriou, "Wilor: End-to-end 3d hand localization and reconstruction in-the-wild," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, (CVPR)*, 2025.

[32] J. Lin, A. Zeng, H. Wang, L. Zhang, and Y. Li, "One-stage 3d whole-body mesh recovery with component-aware transformer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, (CVPR)*, 2023.

[33] Z. Chen, B. Zhou, Y. Huang, J. Wan, Y. Hu, H. Shi, Y. Liang, Z. Lei, and D. Zhang, "C^2rl: Content and context representation learning for gloss-free sign language translation and retrieval," *CoRR*, vol. abs/2408.09949, 2024.