Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2024.0429000

Expanding Multilingual Co-Speech Interaction: The Impact of Enhanced Gesture Units in Text-to-Gesture Synthesis for Digital Humans

GHAZANFAR ALI[1], WOOJOO KIM[2], MUHAMMAD SHAHID ANWAR[3], JAE-IN HWANG[1], and AHYOUNG CHOI[3]*

[1][1][2]RC - Intelligence and Interaction Research Center, Korea Institute of Science and Technology (KIST), 5, Hwarang-ro 14-gil, Seongbuk-gu, Seoul, 02792, Republic of Korea
[2]Division of Liberal Studies, Kangwon National University, 1, Gangwondaehak-gil, Chuncheon-si, Gangwon-do, 24341, Republic of Korea
[3]Department of AI, Software, Gachon University, 1342, Seongnam-daero, Sujeong-gu, Seongnam-si, Gyeonggi-do, 13120, Republic of Korea

Corresponding author: Ahyoung Choi (e-mail: aychoi@gachon.ac.kr).

This work was supported by Korea Institute of Science and Technology of the Republic of Korea(Funding 2E33842)

ABSTRACT In this study, we explore the effects of co-speech gesture generation on user experience in 3D digital human interaction by testing two key hypotheses. The first hypothesis posits that increasing the number of gestures enhances the user experience across criteria such as naturalness, human-likeness, temporal consistency, semantic consistency, and social presence. The second hypothesis suggests that language translation does not degrade the user experience across these criteria. To explore these hypotheses, we investigated three conditions using a digital human: voice only with no gestures, limited(56 gestures) co-speech gestures, and full system functionality with over 2000 unique gestures. For the second hypothesis, we used language translation to provide multilingual support, retrieving gestures from an English rule base. We obtained text and pose from English videos and matched the pose with gesture units derived from Korean speakers' motion-capture sequences, enhancing a comprehensive rule base that we used for gesture retrieval for given text input. We used translation of non-English input language to English for text matching. Our novel method utilizes an improved pipeline to extract text, 2D pose data, and 3D gesture units. Incorporating a cutting-edge gesture-pose matching model with deep contrastive learning, we retrieved gestures from a comprehensive rule base containing 210,000 rules. This approach optimizes alignment and generates realistic, semantically consistent co-speech gestures adaptable to various languages. A comprehensive user study evaluated our hypotheses. The results underscored the positive impact of diverse gestures, supporting the first hypothesis. Additionally, multilingual capabilities did not degrade the user experience, confirming the second hypothesis. Highlighting the scalability and flexibility of our method, this study provides valuable insights into cross-lingual data and expert systems for gesture generation, contributing significantly to more engaging and immersive digital human interactions and the broader field of human-computer interaction.

INDEX TERMS Co-speech gestures, Gesture generation, HCI, machine learning, augmented/virtual/mixed realities

I. INTRODUCTION

THE use of digital humans in virtual environments is gaining popularity for various applications, such as acting as guides in virtual museums [1] or as actors in pre-visualizations [2]. For these digital humans to appear natural, they require co-speech gestures that are convincingly synchronized with their speech [3]. To advance the state of the art in this area, our research is guided by two central hypotheses: **(1) a greater diversity and number of available gestures significantly enhances the user experience across criteria such as naturalness, human-likeness, temporal consistency, semantic consistency, and social presence;** and **(2) leveraging language translation to enable multilingual gesture generation does not degrade this user experience.** To investigate these hypotheses, we developed and evaluated a novel system using a cutting-edge gesture-pose matching model with deep contrastive learning and tested it across three conditions: a digital human with no gestures, one with a limited set of gestures, and our full system with a comprehensive gesture library.

Traditional methods for generating co-speech gestures involve labor-intensive motion capture and manual animation mapping. Rule-based animations offer control and flexibility but suffer from subjective expert definitions [4]. On the other hand, data-driven methods use probabilistic models and intensive and elaborate annotation schemes [5]. Recently there has been a significant increase in the use of end-to-end deep-learning techniques to generate co-speech gestures, but given the non-deterministic nature of co-speech gestures, the results are underwhelming so far [6].

In this study, we address the problem of generating co-speech gestures for digital humans by proposing a cross-lingual text-to-gesture synthesis method that leverages both English and Korean data. Our specific aims are to develop a more accurate gesture matching system, diversify the output to create a more natural communication experience, and demonstrate the effectiveness of our approach through a user study.

While numerous studies have explored the generation of gestures using audio-to-gesture methods [7], [8], Audio-only modality systems generate better gesture flow than text-only methods but lack semantic understanding [6]. Text-only systems provide better output control and semantic awareness due to pre-trained LLMs [3]. However, these systems may face challenges in capturing the subtleties of human motion and may not generalize well across languages and cultures. To address these limitations, we focus on text-based input, which provides richer semantic information and aligns with the current trend of text-only chatbot technologies. This shift is particularly relevant as text-to-speech (TTS) systems continue improving, moving away from robotic-like voices to more natural and expressive human-like ones.

The quality of co-speech gestures generated by any method strongly depends on the data used [6]. Methods that use motion capture data have better motion flow [9]–[11] than wild data from videos [12], [13]. However, the quality can decrease for systems trained on motion capture data when the input is semantically different from the training data [11]. Public videos provide diverse text corpus but may have motions of lower quality, resulting in the model learning averaged motions [3]. To address these issues, a hybrid approach was proposed that combines diverse text data from public videos and clean motion data from motion capture, using intermediate models via deep learning for text and motion matching [14].

Regarding language, most co-speech gesture-generation methods use English data [7], [8]. Other languages are rarely explored in this context, presenting further research opportunities. Studies have shown that language and culture influence gesture production [15]–[17], suggesting language-specific models are needed. Our study aims to address this gap by developing a cross-lingual text-to-gesture synthesis method that leverages both English and Korean data.

Ali et al. [3] proposed a method to address the issue of subjectivity in rule-based animations. Their approach involved using transcript and 2D pose data extracted from monologue-style public videos of human presenters. They then used a set of predefined gesture units as sliding windows over the 2D pose sequences and extracted corresponding text for the best matches. This process resulted in a comprehensive text-to-gesture mapping. In addition, the authors used the mean cosine similarity for the pose-to-gesture matching, which performs poorly on temporal misalignment and noise. However, this novel approach can help mitigate the subjectivity problem of rule-based animations. The gesture units were 56, and the rules were 84k. We used this system as a baseline system referred to as 56GestureBase, seeking to improve upon it by incorporating cross-lingual data and developing a more accurate gesture matching system.

Our approach is based on GestureCLR [14] and automatic text-to-gesture rule generation [3] with cross-lingual data, specifically videos of English speakers and motion capture (mocap) data of Korean speakers, to create an automatic rule-map that can be used with input data in both English and Korean. To achieve this, we first needed to identify gesture units and develop a more effective method for matching 2D pose data to 3D mocap gesture units.

To generate gesture units from the mocap data, we designed an expert system that increased the number of gestures to 2035, which enabled us to capture the subtleties of human motion. We then used deep contrastive learning to train our model to match 210k rules from a set of 350k pose-text pairs. The method we propose also integrates translation technology, allowing us to support multiple languages without compromising user experience.

Our user study involved three conditions: No gestures, just voice(NoneKor), a Korean version of the baseline system 56GestureBase (Ali et al. [3]), and our proposed system in English and Korean voices. The results confirmed that using deep contrastive learning and the increased number of gestures positively impacted various user experience criteria. The results also supported our hypotheses that increasing the number of gestures enhances user experience and that language translation does not degrade the user experience.

The findings of this study contribute to the broader field of human-computer interaction by creating more engaging and immersive digital human interactions. The enhanced user experience is achieved through the innovative application of cross-lingual text-to-gesture synthesis methods, expert system design, deep contrastive learning, and comprehensive user study. In the following sections, we will review the related work, explain our approach, detail the implementation, and discuss the user study. We conclude by shedding light on the potential of using cross-lingual data and expert systems for gesture generation, providing valuable insights for future research, including incorporating more languages and understanding cultural nuances in gesture generation.

II. RELATED WORK

Since our work is text-based and we have given its reasons, this section will provide a detailed review of the related work in text-to-co-speech gesture generation, focusing on the two

main components, i.e., data and methods. Although the focus is on text-based methods, other methods will be mentioned if necessary.

A. DATA FOR CO-SPEECH GESTURE GENERATION

Data availability and quality significantly impact the style and quality of the generated co-speech gestures. There are three main types of data used for co-speech gesture generation: (1) Handcrafted data, (2) studio-captured data, also known as motion capture (mocap) data, and (3) videos from open sources. Initial methods used Handcrafted animation data to make gesture units [4], [18]. With the development of motion capture technology, obtaining human motion data became easier. Some researchers converted this motion into gesture units [19] or used the motion and audio for end-to-end learning of gesture generation [9]–[11]. More recently, large-scale public motion capture datasets have been released, further enabling data-driven approaches [20]. Neff M [5] developed a method to use video annotations to build probabilistic models. Another way of using videos was to process them using pose estimation and transcription by speech recognition [13], [21], [22]. This type of data is typically used for 2D gesture generation. For 3D gesture generation, either direct 2D-to-3D estimation is needed, or grounding with approximate 3D motion is required, a technique that has shown promising results in hybrid systems [14], [55].

B. METHODS FOR CO-SPEECH GESTURE GENERATION

Text-to-gesture systems have traditionally been rule-based. Cassell et al. [18] developed a system that animates conversational agents with synchronized speech, intonation, facial expressions, and hand gestures. The Behavior Expression Animation Toolkit (BEAT) [4] generates nonverbal behaviors from text, using rules derived from human conversational behavior research. While foundational, these systems can be labor-intensive and lack adaptability [39]. The field has since shifted towards data-driven and learning-based methods. Early deep learning models often regressed gestures from audio, but this could result in averaged, lifeless motions [21]. To improve realism and diversity, subsequent works introduced adversarial losses [11], probabilistic models [22], and advanced sequence models like Transformers [20], [26]. More recently, the state-of-the-art has advanced with the introduction of diffusion models for generating fluid, audio-driven gestures [58] and the use of Large Language Models (LLMs) to guide gesture selection and scripting from text, significantly enhancing semantic coherence [57], [56], [55]. Our work builds on hybrid approaches that combine the control of rule-based systems with the scalability of data-driven methods. Ali et al. [3] utilized pose data from videos to expand a rule-map of high-quality mocap gestures. Our approach extends this concept by introducing a more robust cross-lingual framework and a superior motion-matching model, GestureCLR [14], positioning our contribution at the intersection of retrieval-based systems and deep learning.

III. APPROACH

Our approach breaks the problem into three parts as shown in Fig.1 and 2, i.e., 1) GestureCLR model for 2D and 3D motion matching 2) Gesture units and clustering, 3) Rulemap generation with GestureCLR. We choose videos on a wide variety of topics with different presenters. The topic ranges from current affairs talk show to lectures in universities.

Algorithm 1: Extracting motion clips from a motion sequence

```
Input: motion sequence
Output: list of extracted motion clips
while motion sequence has more than 3 seconds do
    for each possible starting position s do
        for each possible ending position e do
            if 30 ≤ (e − s) ≤ 45 at 15 fps and
              Euclidean distance between s and e is
              minimal then
                clip ← extract motion clip from s to e;
                if clip has optimal variance then
                    add clip to the list of extracted
                      motion clips;
                end
            end
        end
    end
    Remove the longest clip from the motion
      sequence;
end
```

A. GESTURE UNITS AND CLUSTERING

In this study, we defined a gesture unit as a short, meaningful sequence of upper body motion that naturally occurs during speech. To build a comprehensive and realistic library of these units, we sourced two hours of high-fidelity motion capture (mocap) data. This data features a male and a female participant engaged in unscripted, spontaneous conversations about a wide range of topics, including life experiences and general discussions. We deliberately chose this type of natural, dyadic conversational data over scripted or monologue-style performances because it more accurately reflects the nuances of real-world co-speech gesturing. The spontaneity of the interaction ensures a wide variety of non-emblematic and non-repetitive gesticulations, which are crucial for generating a gesture library that appears authentic and dynamic rather than robotic. While the specific content of their conversation is not used for text mapping, the diversity of topics naturally elicits a broad spectrum of expressive motions, forming an ideal foundation for our gesture unit extraction process.

As manual extraction of gesture units is time-consuming and subjective, we automated this process using the rules defined in Algorithm 1. The first two rules concern temporal length (2-3 seconds, corresponding to 30-45 frames at 15 fps) and pose similarity (minimal Euclidean distance between

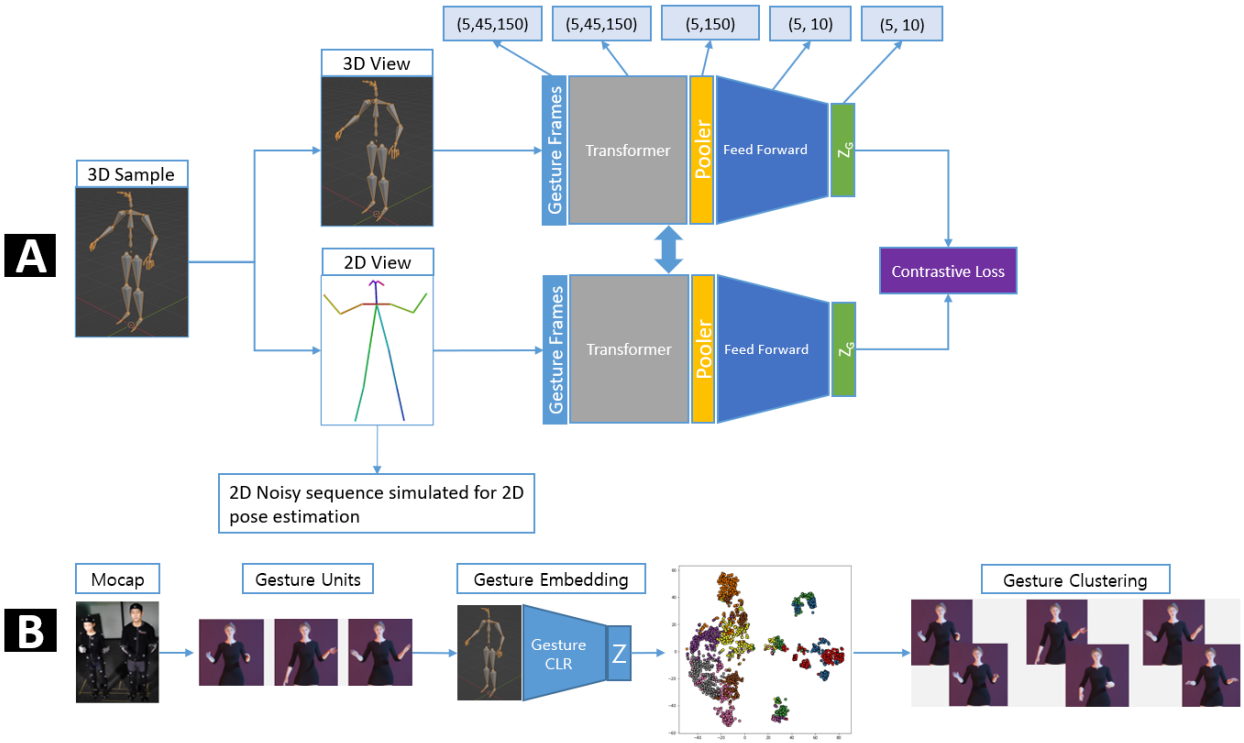

FIGURE 1. A) GestureCLR Training B) Gesture unit embedding with GestureCLR and gesture clustering.

the start and end poses). The third rule, ensuring "optimal variance," was designed to filter out static or non-expressive motions. To calibrate this, we first manually screened the full motion capture sequences to remove any segments with obvious data errors. Then, for all potential clips, we calculated the variance of all upper-body joint positions over the clip's duration. By plotting these variance values in ascending order, we could identify an "elbow point" where the curve began to rise sharply, indicating a transition from low-motion to high-motion clips. We set our variance threshold at this point. This expert-guided calibration process was repeated across different data segments to ensure robustness, yielding gesture units with sufficient dynamism. We extracted a total of 2035 gesture units using these rules (1025 male-style, 1010 female-style). To improve output diversity and avoid repetitive animations, we clustered these units. Our goal was to create semantically meaningful groups, with each cluster containing 10-15 stylistically similar variations of a gesture. Initially, a standard K-means algorithm with K=100 (aiming for 10 variations per 1000 gestures) resulted in highly imbalanced clusters. We also found that quantitative metrics like the elbow method were not informative for this specific task. Consequently, we adopted the Bisecting K-Means algorithm. This hybrid approach recursively splits the largest cluster, leading to more balanced and hierarchically sound groupings. Setting K=100 with Bisecting K-Means successfully produced clusters that were both reasonably balanced and aligned with our goal of having 10-15 gesture variations per cluster. This allowed us to replace a simple Top-K sampling rule with a more robust strategy: for a given text, we identify the best-matching cluster and then randomly sample one of the gesture variations from within that cluster, ensuring both relevance and diversity in the final animation.

B. 2D POSE-TEXT PAIRS EXTRACTION

In the field of natural language processing, grounding refers to the process of connecting language to the real world. Grounding is the process of linking words or phrases in natural language to specific entities or concepts in the real world, such as physical objects, locations, or actions. In the context of analyzing public monologue-based videos with human presenters, we obtained timestamped speech transcripts and 2D human pose data with OpenPose [28]. The pose sequences were then processed to make 3-second chunks (Text, Pose) pairs. However, the pose sequences obtained from OpenPose were usually noisy and not directly usable. Some common noise sources in pose data include incomplete or missing pose data, drift, jitter, and false positives.

To address the limitations of OpenPose, such as the lack of person tracking, we used facial recognition to identify the person of interest in the videos. A dictionary of faces was built to identify the face present in the majority of frames. By tracking the person of interest's head position, we could create segments where the person was continuously present. Speech recognition was then employed to find segments containing speech, and the text was extracted from those segments. Segments without speech were discarded.

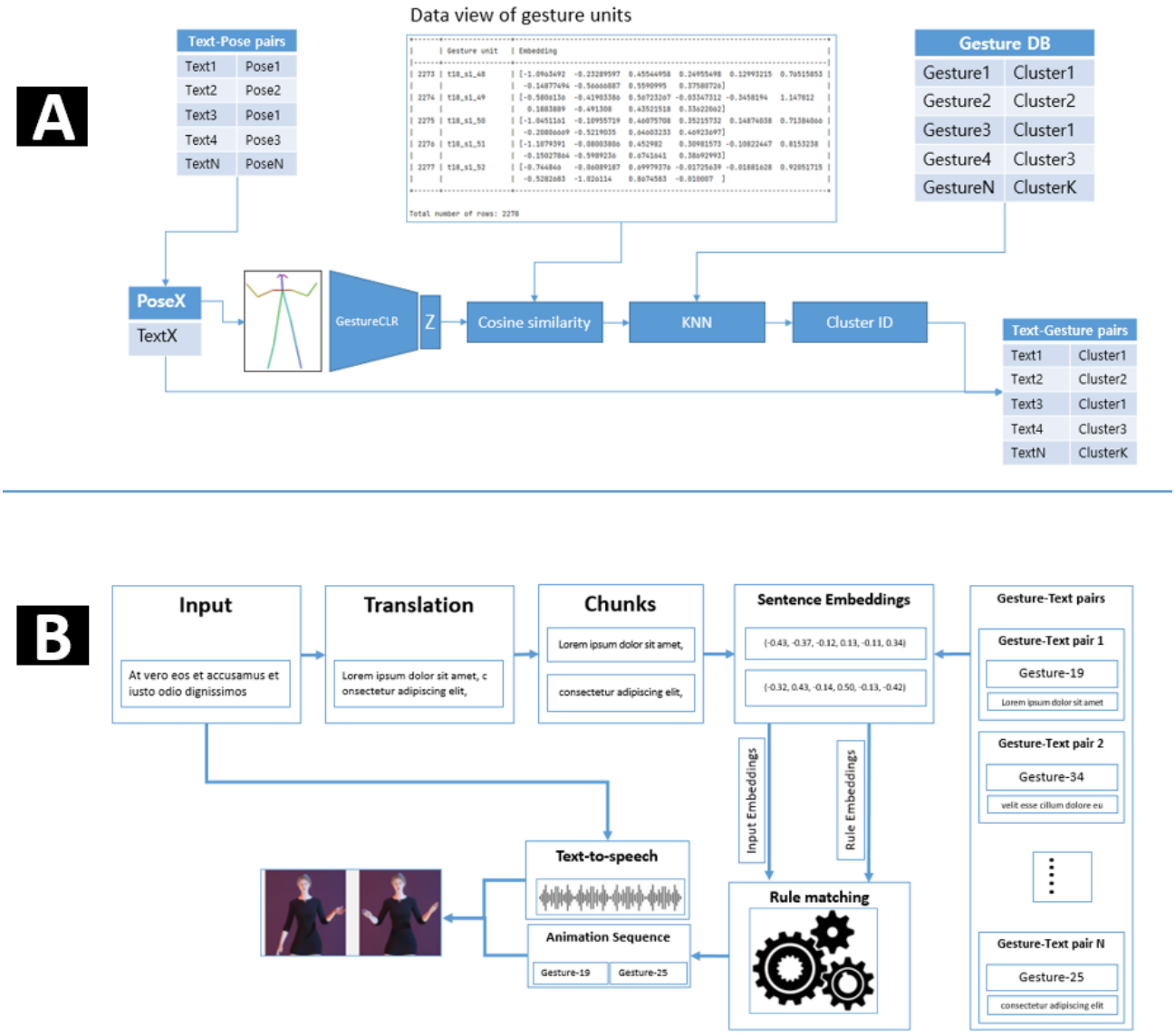

FIGURE 2. **A) Gesture Unit matching with pose B) Gesture retrieval pipeline for given text input.**

Pose estimation was run on the segments obtained from the previous step, and only those poses whose head position matched the head position from facial recognition were selected. The noisy pose sequences were not cleaned up, but instead, a model was trained to link them with predefined animations, as detailed in the section titled "Rulemap Generation with GestureCLR."

The process of extracting the 2D Pose-Text pairs can be summarized as follows:

1) Get a set of videos where ideally one person is the main subject, such as a talk show host or lecturer delivering lectures.
2) Take a video from the set of videos.
3) Run facial recognition on the video and build a dictionary of faces. Identify the face which is present in the majority of frames as the person of interest.
4) Track the person of interest's head position and create segments where the person is continuously present.
5) Run speech recognition to find the segments containing speech. Obtain text from these segments and discard segments without speech.
6) Run pose estimation on the segments obtained from the previous step, selecting only those poses whose head position matched the head position from facial recognition.
7) Save the pose-text pairs from the given segment and continue with other segments.
8) Repeat the process for the remaining videos in the set.

By following this process, we successfully built a dataset of accurate and usable (Text, Pose) pairs extracted from public monologue-based videos.

C. RULEMAP GENERATION WITH GESTURECLR

In this study, we developed GestureCLR, a deep contrastive learning model based on the SimCLR architecture proposed by Chen et al. [29], for recognizing gestures from noisy and varied input data.

We trained our model on motion capture data from the Talking with hands dataset [30]. We first processed the data to create sequence units of 2-3 seconds, or 30-45 frames, as the frames per second are set to 15. The input included a clean 3D unit and a 2D version of the same unit. The 2D version was augmented with Gaussian noise and temporal shifts to mimic the variability and noise present in real-world 2D pose data. Gaussian noise was generated with a variable variance of 0.001, 0.01, and 0.1 for low, medium, and high noise. Temporal shifts were introduced by creating two sequences of 45 frames, one filled with the mean pose of the original sequence and the other with zero poses. Contiguous segments of 30 frames were randomly selected from the original sequence and inserted into the new sequences at a random position between 1 and 15, contiguously occupying the next 30 frames.

Our GestureCLR model is built on top of a Transformer encoder architecture [31]. The model takes as input the 3D and 2D body gesture data and learns to map them to a latent space, where the representations of corresponding 3D and 2D gestures have high similarity, and non-corresponding pairs have low similarity. The main components of the model are a Positional Encoding module, a Transformer Encoder, and a fully connected layer followed by a batch normalization and a leaky ReLU activation function.

The Positional Encoding module generates sinusoidal positional encodings for the input data, enabling the model to capture the sequential nature of the gesture data. The Transformer Encoder consists of 3 layers, each having 5 self-attention heads and a feedforward network with a hidden dimension of 512. The input and output dimensions of the encoder are set to 150. The fully connected layer maps the output of the Transformer Encoder to a latent space of dimension 10. We experimented with latent space dimensions of 10, 16, 32, and 64 and found that a dimension of 10 resulted in the best accuracy.

We trained the model with a contrastive learning approach using the Normalized Temperature-Scaled Cross Entropy (NT-Xent) loss. The loss function is defined as:

$$L_{i,j} = -\log\left(\frac{\exp(\cos\_\text{sim}_{i,j})}{\sum_k \exp(\cos\_\text{sim}_{i,k})}\right) \tag{1}$$

where $\cos\_\text{sim}_{i,j}$ is the cosine similarity between the i-th 2D motion sequence and the j-th 3D motion sequence, computed as:

$$\cos\_\text{sim}_{i,j} = \frac{2D_i \cdot 3D_j}{\|2D_i\|_2 \|3D_j\|_2} / T \tag{2}$$

In these equations: - $2D_i$ and $3D_j$ are the latent representations of the i-th 2D motion sequence and the j-th 3D motion sequence, respectively. - $\cdot$ denotes the dot product. - $\|\cdot\|_2$ denotes the L2 norm. - T is the temperature hyperparameter, controlling the concentration of the distribution. - The index k runs over all possible negative samples.

The final loss is the mean over all pairs (i, j):

$$L = \frac{1}{N^2} \sum_{i,j} L_{i,j} \tag{3}$$

where N is the total number of samples in the batch. We used the AdamW optimizer [32] with a Cosine Annealing learning rate scheduler for training. The model was trained with a learning rate of 5×10^{-4}, a weight decay of 10^{-4}, and a maximum number of epochs set to 1000. The training and validation batch sizes were set to 512.

To implement the GestureCLR model and training pipeline, we used PyTorch Lightning [33], a high-level library for PyTorch that simplifies the training process. The model checkpoints were saved periodically based on the validation loss, and the learning rate was logged during training.

Using the GestureCLR model, we can match 2D poses from OpenPose with 3D poses from motion capture. It can also be used as a motion encoder to encode gesture units and cluster them using the latent representations obtained from the GestureCLR.

Finally, we applied our model to poses from (Text, Pose) pairs and gesture units from our mocap data to obtain (Text, Gesture) pairs. To obtain an embedding for the text input, we used SentenceBERT [34]. This gave us (Text_{Emb}, Gesture) pairs as rulebase.

The selection of the latent space dimension was a critical hyperparameter choice. We conducted experiments with dimensions of 10, 16, 32, and 64 to determine the optimal balance between representational capacity and model generalization. Our evaluation involved both quantitative and qualitative measures. Quantitatively, we assessed the matching accuracy on our most challenging test sets (high noise and temporal shifts). As shown in Table 1, the model with a latent dimension of 10 achieved the highest accuracy. We observed a performance decline with higher dimensions, suggesting potential underfitting or issues related to sparsity in a larger latent space. Qualitatively, we visualized the gesture unit embeddings from each model using t-SNE. The model with a dimension of 10 produced visually distinct and coherent clusters, indicating that it effectively learned a meaningful representational space. In contrast, the visualizations for dimensions 16, 32, and 64 showed less defined and more overlapping clusters. Based on these combined findings, we selected the latent space dimension of 10 for our final model, as it provided the best performance in both our matching task and the semantic clustering of gestures.

IV. IMPLEMENTATION OF ANIMATION SYNTHESIS PIPELINE

Having detailed our three-part approach—developing the GestureCLR model, extracting and clustering gesture units, and generating the rule map—we now describe how these

TABLE 1. GestureCLR Matching Accuracy with Different Latent Space Dimensions.

Latent Dimension	High Noise	Temporal Shift
10	**97%**	**62%**
16	94%	55%
32	80%	53%
64	84%	48%

components were integrated into a practical, real-time animation synthesis pipeline. We employed the server-client model to implement our gesture system, with Unity as the platform for visualizing the gestures. Our gesture units were initially in BVH format, but as described in the section "Gesture units and clustering," we applied an algorithm on the BVH files to extract the start and end of each gesture unit and saved the information in a file. We then used Blender to convert the BVH files to FBX format. The FBX files were loaded into Unity, and a script used the gesture unit information from the file to create clips in the FBX, which were then loaded into the Unity animator.

We chose 6-gram tokenization (6-word chunks) for the text input based on the assumption of 2.5 words per second for an average text-to-speech system. As our gestures have a minimum duration of 2 seconds and a maximum of 3 seconds, this segmentation accurately represents the text. The tokens were sent to the server's gesture system, which converted them into embeddings using SentenceBERT.

To find the best match in the rule map, we used cosine similarity on text embeddings and ranked them to identify the gesture unit ID with the lowest distance. As described in the section "Rulemap Generation with GestureCLR," this process was repeated for each token, and the resulting IDs were returned to the Unity animator. The implementation guarantees that an input text will always match with some text in the rules, even if the matching distance might be low in some cases. A minimum threshold can be applied if needed, and an idle gesture can be played in instances with low matching distances.

We employed clustering to ensure that the generated gestures are diverse and do not become repetitive, as explained in the "Gesture units and clustering" section. The rules have a cluster ID instead of a gesture ID, and given a cluster ID, we randomly pick an animation. This approach balances the relevance of the animation with the input text, creating a slightly different animation sequence each time while maintaining overall similarity.

We also implemented a multi-stage process to address the critical challenge of synchronizing gestures with speech. Our strategy was designed to be robust against variations in speech rate and the presence of natural pauses. First, to prevent timing errors accumulating over long passages, any input text exceeding 30 words is segmented into smaller, sentence-level chunks. For each chunk, the synchronization process is as follows: Initial Pacing: The text is sent to our TTS engine. We use the total duration of the generated audio and the number of retrieved gestures (one per 6-word token) to calculate an initial, average duration for each gesture. This provides a coarse alignment. Timestamp-based Refinement: The TTS engine also provides word-level timestamps. We use these to refine the duration of each gesture, aligning its playback time with the precise start and end times of its corresponding 6-word text segment. This ensures that gestures are tightly coupled with the spoken words. Animation Blending: To avoid unnatural, choppy transitions between consecutive gestures, which can arise from this precise time-scaling, we incorporate a blending mechanism. Each gesture unit is designed with a 5-frame blending buffer at both its start and end. This allows for smooth interpolation from the end of the previous gesture to the start of the new one, maintaining fluid motion across the entire sequence. This combination of coarse pacing, fine-grained timestamp alignment, and animation blending creates a synchronized and natural-looking performance. It is also worth noting that co-speech gestures in human communication are not always perfectly synchronous; slight temporal shifts between speech and gesture are common and can even enhance perceived naturalness. Our system inherently accommodates these minor variations, contributing to a more realistic user experience.

Overall, our gesture system relied on a combination of server-side and client-side technologies to provide a seamless experience for users. Using a rule map and SentenceBERT, we can accurately match text inputs with appropriate gesture units, allowing for clear and effective communication through the Unity platform.

A core component of our system's multilingual capability is the integration of a translation layer for non-English input. For this study, we utilized the Naver Papago API, which is widely recognized for its high-quality and low-latency Korean-to-English translation. Our primary methodological challenge was not technical but conceptual. Given the significant structural differences between Korean (an SOV language) and English (an SVO language), we were concerned that direct translation could lead to semantic mismatches or awkward timing when paired with a gesture rule map derived from English text. Our hypothesis was that by breaking down the input text into short, 6-word chunks, we could mitigate these macro-level structural differences. The idea was to match smaller, localized semantic units to our atomic gesture units, thereby bypassing the complexities of full-sentence syntax. The success of this approach was a key question for our user study. As our results in Section 5.2.4 indicate, the lack of significant degradation in user experience between the OursKor and OursEng conditions validated this design choice, demonstrating that a robust rule map can indeed be effectively leveraged across structurally different languages through this chunking and translation strategy.

V. EVALUATION

A. OBJECTIVE EVALUATION

To assess the effectiveness of our model, we conducted experiments using a test set of 500 randomly selected gesture

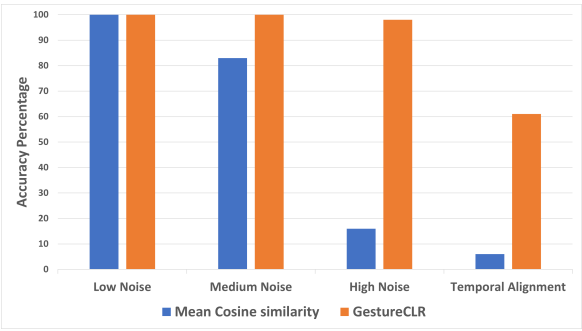

FIGURE 3. Comparative accuracy performance of GestureCLR and Mean Cosine Similarity.

units extracted from 3D motion capture data. We created 2D noisy versions of the gesture units using Gaussian noise with variable variance levels of 0.001, 0.01, and 0.1 for low, medium, and high noise, respectively. Additionally, we created a temporally shifted 2D version by randomly selecting contiguous segments of frames and inserting them into new sequences.

We evaluated our model's output by comparing it with the frame-by-frame mean cosine similarity metric proposed by Ali et al. [3]. While this metric has limitations in that it is not robust for noise and temporal shifts, our model outperformed the metric in all scenarios. Specifically, for low noise, the accuracy of our model and the metric was 100%, whereas, for medium noise, our model achieved 100% accuracy compared to 82% for the metric. For high noise, our model achieved an accuracy of 97% compared to only 18% for the metric. Finally, for temporal shifts, our model achieved an accuracy of 62% compared to only 6% for the metric. The results are presented in Fig. 3.

While our objective evaluation demonstrates the robustness of GestureCLR, we acknowledge certain limitations. The comparison was focused on retrieval accuracy against a single baseline metric, mean cosine similarity. Future work could provide a more comprehensive assessment by incorporating additional evaluation indicators, such as comparing the distribution of generated motions to real motions using metrics like Frechet Gesture Distance (FGD), to better evaluate the naturalness of the retrieved gestures. Furthermore, the test set itself has constraints. The 500 gesture units were randomly sampled from our mocap database, which, while diverse in conversational gestures, may not cover the full spectrum of human motion, such as highly emblematic or action-specific gestures. A potential bias, therefore, is that our model's performance is optimized for the types of gesticulation present in our source data. Our synthetic noise models (Gaussian noise, random temporal shifts) also may not fully capture the complex, structured artifacts seen in "in-the-wild" 2D pose data, such as partial occlusions or inconsistent camera angles. Consequently, while the results are strong within our controlled testbed, the model's performance on truly unconstrained, real-world video footage remains an area for future investigation.

B. USER STUDY

While our objective evaluation confirmed the superior matching accuracy of the GestureCLR model, the ultimate measure of success for a co-speech gesture system is the user's subjective experience. Therefore, we conducted a comprehensive user study to assess how our technical improvements in gesture selection and diversity translate into perceived enhancements in naturalness, human-likeness, and overall social presence.

1) Participants and Experimental Settings

A total of 51 Korean university students (33 males, 18 females) aged 20–27 (M=23.7, SD=1.7) participated in the experiment. All participants were native Korean speakers. To address the cross-lingual aspect of our study, we also considered their English language proficiency. As a prerequisite for admission to their university, all students are required to hold a minimum English certification, such as a TOEIC score of 700 or its equivalent. This requirement ensures that all participants possess at least an intermediate level of English comprehension, making them capable of understanding the speech content and evaluating the appropriateness of the accompanying gestures in both the Korean and English conditions. The study was approved by the Institutional Review Board (KIST-202304-HR-003) and all participants gave written informed consent for their participation.

The experiment took place in a university classroom, where all participants were sitting and watching a projector screen located at the front of the classroom simultaneously while the video stimuli were being presented at the projector screen (Fig. 4). A young female avatar was used to generate gesture videos, as young or female embodied agents have been preferred over old or male agents [35], [36].

FIGURE 4. The experimental settings

2) Experimental Design and Procedure

Participants were asked to watch a total of 20 videos (4 gesture conditions × 5 speech contents) in random order. Five speech contents were retrieved from some of the most popular TED talks of all time [37]. Each speech content consisted of 49–83 words and was around 3–5 sentences long. Four gesture conditions were selected to answer following three research questions:

- Is the digital human with co-speech gestures generated by the proposed method perceived better than the digital human without co-speech gestures?
- Is the digital human with co-speech gestures generated by the proposed method perceived better than the digital human with co-speech gestures generated by the baseline system [3]?
- Does the language affect how the digital human with co-speech gestures generated by the proposed method is perceived?

Therefore, gesture conditions considered in this study were composed of combinations between three gesture generation methods (None, 56GestureBase system [3], and Ours) and two speech languages (Korean and English). Fig. 5 shows the four gesture conditions evaluated in this study: *NoneKor* (None & Korean), *56GestureBase* (Ali et al. (2020) [3] & Korean), *OursKor* (Ours & Korean), and *OursEng* (Ours & English). All 20 videos can be seen from the following link: https://bit.ly/3Ma6vyR

	NoneKor	*56GestureBase*	*OursKor*	*OursEng*
Gesture	None	Ali et al., 2020	Ours	Ours
Language	Korean	Korean	Korean	English

FIGURE 5. Four gesture conditions evaluated in this study

After watching each video, participants were asked to give ratings on a 7-point Likert scale (1=strongly disagree, 7=strongly agree) for five evaluation items to subjectively assess the user experience of each gesture condition. The five subjective evaluation items were naturalness (''Digital human was natural''), human-likeness (''Digital human was human-like''), time consistency (''Gesture timing was matched to speech''), semantic consistency (''Gesture was matched to speech content''), and social presence (''I felt like digital human was 'real' and 'there' ''). These items have been frequently used in evaluating co-speech gestures of embodied agents [38], [39].

3) Data Analysis

The first speech content was considered as practice thus excluded, so a total of 4 gesture conditions × 4 speech contents × 51 participants = 816 trials was arranged for the analysis. After getting the mean ratings of four speech contents for each gesture condition and participant, repeated-measures analysis of variance (RM-ANOVA) and Bonferroni post-hoc tests were conducted. The Shapiro-Wilk test for normality and inspection of Q-Q plots found no measures were severely deviated from the normal distribution. The degree of freedom was corrected with Greenhouse-Geisser correction if the p-value of Mauchly's test was equal to or less than 0.05. R 4.2.2 and ''*rstatix*'' R package were used to conduct all data processing and statistical analyses at a significance level of 0.05, and the effect size in terms of generalized eta-squared (η_G^2) [40] was further calculated to check practical significance.

4) Results and Discussion

Fig. 6 shows the subjective evaluation ratings (mean of four speech contents) for each of five investigated items. The significant main effect of gesture condition was found in all evaluation items: naturalness ($F_{1.73,86.55} = 39.42, p < 0.001, \eta_G^2 = 0.146$), human-likeness ($F_{1.69,84.25} = 24.68, p < 0.001, \eta_G^2 = 0.088$), time consistency ($F_{1.88,93.97} = 56.45, p < 0.001, \eta_G^2 = 0.274$), semantic consistency ($F_{2.24,112.22} = 47.71, p < 0.001, \eta_G^2 = 0.292$), and social presence ($F_{1.89,94.26} = 23.24, p < 0.001, \eta_G^2 = 0.097$).

Regarding the pairwise comparison, we only highlight the results of comparisons between *NoneKor* and *56GestureBase*, *56GestureBase* and *OursKor*, and *OursKor* and *OursEng* here for better clarity. According to the post-hoc test results, the same pattern was found across all evaluation items. *OursKor* had significantly higher ratings than *56GestureBase*, which had significantly higher ratings compared to *NoneKor*, whereas no significant differences were found between *OursKor* and *OursEng*. The mean and standard deviation of subjective evaluation ratings are shown in Table 2.

The results indicate that the digital human with co-speech gestures generated by the proposed method was perceived more natural, more human-like, more temporally and semantically consistent, and more socially present than the digital human without co-speech gestures or with co-speech gestures generated by the previous counterpart [3]. The fact that the existence of co-speech gestures helps digital human to look more alive and engaging has been supported by numerous studies [41]–[44]. Our results align with existing studies, emphasizing the significance of naturalness and human-likeness in perceptions of digital humans. Studies reported that lifelike co-speech gestures that closely mimic human gestures could enhance perception of the digital human's human-like characteristics [45], [46]. Our proposed methods could generate gestures with greater time and semantic consistency, which might have affected other evaluation aspects. Coherence between gestures and verbal timing and content play a significant role in improving the perceived naturalness and effectiveness of digital humans [47]–[49]. The perception of social presence, or the degree to which a virtual agent is perceived as a social entity, was also found to be influenced by co-speech

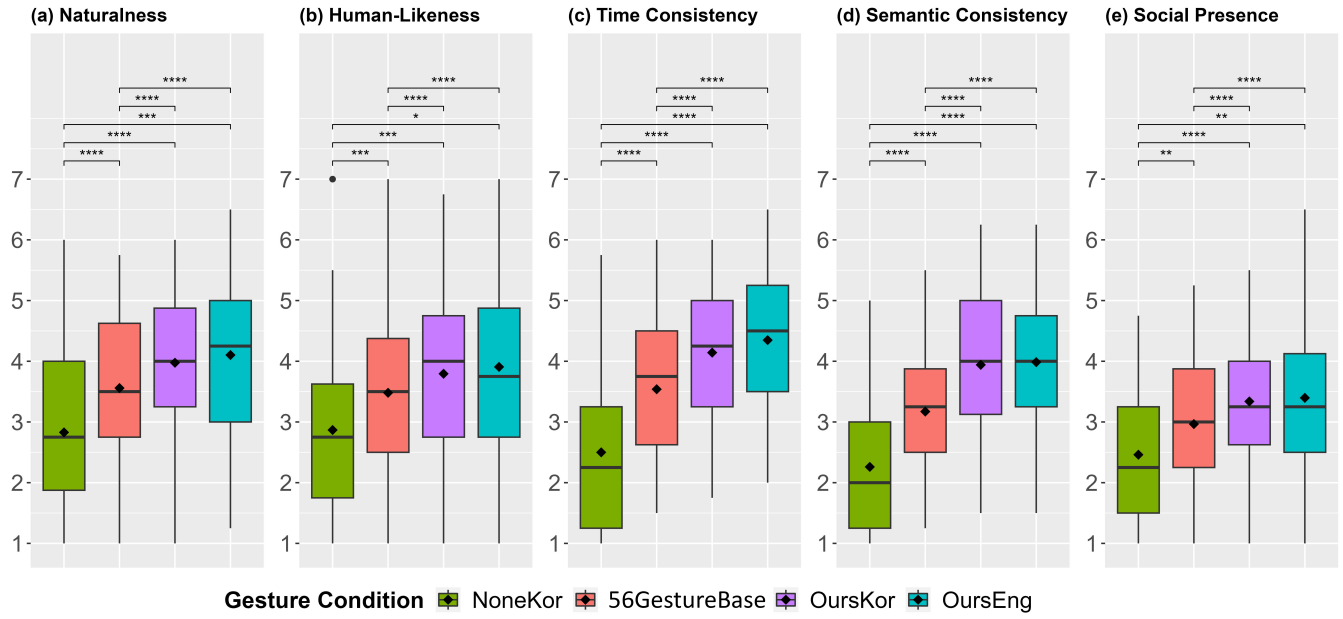

FIGURE 6. **Mean ratings for subjective evaluation by gesture condition for each item. *Note*: Diamond symbols indicate mean ratings, and circular symbols indicate outliers outside of the interquartile range. Asterisks indicate the significance of Bonferroni post-hoc tests ($p < 0.05^{*}, p < 0.01^{**}, p < 0.001^{***}, p < 0.0001^{****}$)**

TABLE 2. **Subjective evaluation ratings ($M \pm SD$)**

Evaluation Item	NoneKor	56GestureBase	OursKor	OursEng
Naturalness	2.83 ± 1.34	3.56 ± 1.15	3.98 ± 1.13	**4.10 ± 1.24**
Human-likeness	2.87 ± 1.36	3.48 ± 1.26	3.79 ± 1.27	**3.91 ± 1.35**
Time consistency	2.50 ± 1.29	3.54 ± 1.16	4.14 ± 1.12	**4.35 ± 1.15**
Semantic consistency	2.26 ± 1.12	3.17 ± 0.99	3.94 ± 1.16	**3.98 ± 1.14**
Social presence	2.46 ± 1.10	2.97 ± 1.07	3.34 ± 1.17	**3.40 ± 1.24**

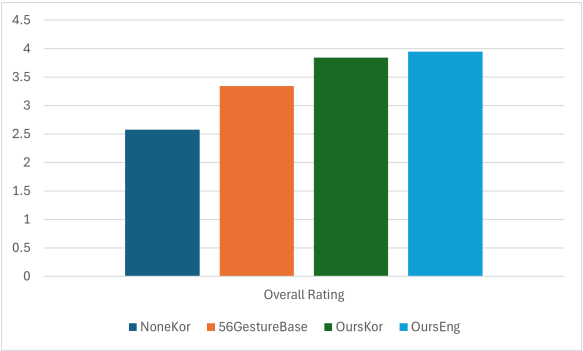

FIGURE 7. **Overall user experience ratings, averaged across all five subjective evaluation items (Naturalness, Human-likeness, Time Consistency, Semantic Consistency, and Social Presence). This chart summarizes the main trend, showing a significant improvement from no gestures to the baseline system, and a further significant improvement with our proposed method. A detailed breakdown for each item is provided in Fig. 6.**

gestures. Our results are in agreement with previous studies suggesting that gestures contribute to the perception of social presence in digital humans [50], [51]. The proposed method (*OursKor*) could generate more suitable co-speech gestures than baselines (*NoneKor* & *56GestureBase*), thus leading to an increased sense of social presence. Results suggest that increasing the number of diverse gesture units has a net positive effect on user experience.

Our results also provide insight into the multifaceted nature of what users perceive as "realistic" co-speech gestures. The term "realism" in this context is not a monolithic concept but rather an emergent property derived from several key attributes, as reflected in our evaluation items. First, our findings confirm that realism is tied to motion diversity and quality. The significant improvement from the 56GestureBase condition to our OursKor condition demonstrates that a larger, more varied gesture library, free from the repetitive motions that often plague simpler systems, is a foundational element. Second, realism is critically dependent on temporal and semantic coherence. The high ratings for Time consistency and Semantic consistency for our system suggest that users perceive gestures as realistic only when they are precisely synchronized with speech and are contextually appropriate. A perfectly rendered gesture that is poorly timed or semantically mismatched shatters the illusion of realism. Finally, our study touches upon cultural nuance as a component of realism. The fact that gestures derived from Korean motion data were perceived positively by Korean participants, even when paired

with translated English, suggests that the underlying motion patterns were culturally congruent. Therefore, a realistic gesture is not merely a physically plausible motion but one that is varied, well-timed, semantically relevant, and culturally resonant.

In addition, the results revealed no significant differences between the evaluations when the speech-language was Korean compared to when it was English. Our results suggest that the perception of co-speech gestures in digital humans can be relatively consistent across different languages, which aligns with some previous research [52], [53]. These studies argued that gestures are a universal aspect of human communication and convey meaning irrespective of the specific language spoken, which could explain the lack of significant differences in our study. However, it is important to note that cultural factors can influence the perception of gestures. Different cultures may have different expectations and preferences for co-speech gestures [16], [17], which could affect the evaluations of digital humans. Our findings might indicate that the cultural background of the participants was not significantly different or diverse enough to reveal any potential effects of language on the evaluation of co-speech gestures. Another factor that might have influenced the result is the participants' language proficiency. People with higher language proficiency may better understand the nuances of the spoken language, enabling them to more accurately assess the appropriateness and naturalness of the accompanying gestures [54]. Nevertheless, it is unclear whether the relatively lower language proficiency in English compared to Korean participants in this study positively or negatively affected the evaluation results.

The findings of this study carry significant practical implications for practitioners developing interactive virtual agents. The demonstrated success of our hybrid approach—combining a large, high-quality motion library with a robust text-matching system—offers a scalable solution for applications where natural and engaging communication is paramount. For instance, in the domain of virtual assistants and customer service avatars, our system can generate non-repetitive, contextually relevant gestures that enhance social presence and user trust. In educational tools and virtual tutoring, the ability to produce semantically consistent gestures can help clarify complex concepts and maintain student engagement. Furthermore, the cross-lingual capabilities shown here are particularly valuable for applications targeting global audiences, such as virtual museum guides or corporate trainers, allowing a single, well-curated gesture system to serve multiple language markets without a discernible loss in quality. By providing a pathway to more human-like and expressive digital characters, our research offers practitioners a validated method to improve user experience and communication effectiveness in a wide array of HCI applications.

C. LIMITATIONS AND FUTURE WORK

While our findings support our hypotheses, we acknowledge several limitations that offer valuable directions for future research. First, the experimental setting, where all participants viewed stimuli simultaneously on a single projector screen, introduces potential confounding variables. Although we confirmed that all participants could perceive the gestures, variations in viewing angles and distances were unavoidable. Furthermore, this group setting could introduce peer influence effects, where participants' responses might be unconsciously influenced by the reactions of others. Future studies could mitigate this by conducting experiments in isolated settings for each participant to ensure independent evaluations. Second, our primary user study was conducted with a homogeneous demographic group: Korean university students. While their verified English proficiency was sufficient for the tasks, this single cultural background could be seen as a limitation on the generalizability of our findings. However, it is important to note that our gesture synthesis system has been successfully deployed in a separate real-world application: powering a humanoid avatar to deliver medical explanations to a cohort of native English-speaking users. In that study, which compared traditional slide-based presentations to an avatar-led presentation using our system, the gesturing avatar significantly enhanced the users' sense of social presence. This positive result from a different linguistic and user group provides preliminary evidence for the cross-lingual robustness of our method. Nevertheless, to fully validate our findings, future work should still aim to include a more diverse participant pool, incorporating multiple cultural backgrounds to formally investigate the cross-cultural applicability of our approach. Finally, our evaluation relied on single-item questions for each metric. While based on established practices, future work would benefit from using more comprehensive, multi-item questionnaires for each evaluated construct (e.g., naturalness, social presence) to capture a more nuanced understanding of user experience.

VI. CONCLUSION

In conclusion, our approach highlights the importance of utilizing text-based input to generate more contextually relevant and meaningful co-speech gestures for digital humans in virtual environments. We have developed an automatic rule map by employing cross-lingual data, expert systems, and deep contrastive learning to generate more natural, human-like gestures in multiple languages. With an increased number of gesture units, the proposed method has outperformed baseline methods in terms of naturalness, human likeness, temporal and semantic consistency, and social presence. Furthermore, our user study indicates that the perception of co-speech gestures remains relatively consistent across different languages, demonstrating the potential for creating a more engaging and immersive experience for users.

Building on our findings, several exciting directions for future research emerge. A primary avenue is to move beyond language translation and investigate the direct impact of cultural context on gesture perception and generation. While our study demonstrated cross-lingual consistency, a more rigorous investigation would involve creating culturally-specific

gesture libraries (e.g., from Italian or Japanese speakers) and evaluating their reception by different cultural groups. This would help untangle the universal aspects of co-speech gesture from the culturally-specific ones, a critical step for creating truly global digital humans. Furthermore, the methodology itself can be advanced. Our GestureCLR model could be enhanced by incorporating multimodal inputs, such as audio prosody, during the matching process to capture even finer nuances of expression. Another promising direction is to explore few-shot style adaptation, where the system could learn a new speaker's unique gesture style from a very short video clip, allowing for rapid personalization. Finally, to address the limitations of our evaluation, future user studies should be conducted in controlled, individual settings and employ a more diverse, multicultural participant base to fully validate the scalability of our approach. These efforts will continue to drive the development of more engaging, natural, and socially intelligent digital humans.

REFERENCES

[1] G. Ali, H.-Q. Le, J. Kim, S.-W. Hwang, and J.-I. Hwang, "Design of seamless multi-modal interaction framework for intelligent virtual agents in wearable mixed reality environment," in *CASA '19*, pp. 47–52, 2019, doi:10.1145/3328756.3328758.

[2] H. Kim, G. Ali, and J.-I. Hwang, "Asap: Auto-generating storyboard and previz with virtual humans," in *2021 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*, pp. 316–320, 2021, doi:10.1109/ISMAR-Adjunct54149.2021.00071.

[3] G. Ali, M. Lee, and J. Hwang, "Automatic text-to-gesture rule generation for embodied conversational agents," *Computer Animation and Virtual Worlds*, vol. 31, no. 4–5, p. 1944, 2020, doi:10.1002/cav.1944.

[4] J. Cassell, H. H. Vilhjálmsson, and T. Bickmore, "Beat: The behavior expression animation toolkit," in *Proc. 28th Annu. Conf. Computer Graphics and Interactive Techniques, SIGGRAPH 2001*, pp. 477–486, 2001, doi:10.1145/383259.383315.

[5] M. Neff, M. Kipp, I. Albrecht, and H. P. Seidel, "Gesture modeling and animation based on a probabilistic re-creation of speaker style," *ACM Trans. Graph.*, 2008, doi:10.1145/1330511.1330516.

[6] S. Nyatsanga, T. Kucherenko, C. Ahuja, G. E. Henter, and M. Neff, "A comprehensive review of data-driven co-speech gesture generation," vol. 42, 2023.

[7] T. Kucherenko, P. Jonell, Y. Yoon, P. Wolfert, and G. E. Henter, "A large, crowdsourced evaluation of gesture generation systems on common data: The genea challenge 2020," in *Proc. Int. Conf. Intelligent User Interfaces (IUI)*, pp. 11–21, 2021, doi:10.1145/3397481.3450692.

[8] Y. Yoon, P. Wolfert, T. Kucherenko, C. Viegas, T. Nikolov, M. Tsakov, and G. E. Henter, "The genea challenge 2022: A large evaluation of data-driven co-speech gesture generation," in *ACM Int. Conf. Proc. Ser.*, pp. 736–747, 2022, doi:10.1145/3536221.3558058.

[9] T. Kucherenko, D. Hasegawa, G. E. Henter, N. Kaneko, and H. Kjellström, "Analyzing input and output representations for speech-driven gesture generation," 2019, doi:10.1145/3308532.3329472.

[10] K. Takeuchi, D. Hasegawa, S. Shirakawa, N. Kaneko, H. Sakuta, and K. Sumi, "Speech-to-gesture generation," in *Proc. ACM Press*, pp. 365–369, 2017, doi:10.1145/3125739.3132594.

[11] Y. Ferstl, M. Neff, and R. McDonnell, "Expressgesture: Expressive gesture generation from speech through database matching," *Computer Animation and Virtual Worlds*, vol. 32, 2021, doi:10.1002/cav.2016.

[12] Y. Yoon, W.-R. Ko, M. Jang, J. Lee, J. Kim, and G. Lee, "Robots learn social skills: End-to-end learning of co-speech gesture generation for humanoid robots," 2018.

[13] C. Ahuja, D. W. Lee, Y. I. Nakano, and L.-P. Morency, "Style transfer for co-speech gesture animation: A multi-speaker conditional-mixture approach," arXiv preprint, pp. 248–265, 2020, doi:10.1007/978-3-030-58523-5_15.

[14] G. Ali and J.-I. Hwang, "Improving co-speech gesture rule-map generation via wild pose matching with gesture units," *SIGGRAPH Asia 2022 Posters*, PartF168147-3, pp. 1575–1585, 2022, doi:10.1145/3550082.

[15] S. Kita, *Pointing: Where Language, Culture, and Cognition Meet*. Psychology Press, 2003.

[16] D. Archer, "Unspoken diversity: Cultural differences in gestures," *Qualitative Sociology*, vol. 20, no. 1, p. 79, 1997.

[17] S. Kita, "Cross-cultural variation of speech-accompanying gesture: A review," *Language and Cognitive Processes*, vol. 24, no. 2, pp. 145–167, 2009.

[18] J. Cassell, C. Pelachaud, N. Badler, M. Steedman, B. Achorn, T. Becket, B. Douville, S. Prevost, and M. Stone, "Animated conversation: Rule-based generation of facial expression, gesture and spoken intonation for multiple conversational agents," in *Proc. SIGGRAPH 1994*, pp. 413–420, 1994, doi:10.1145/192161.192272.

[19] J. Lee and S. Marsella, "Nonverbal behavior generator for embodied conversational agents," in *Lecture Notes in Computer Science*, vol. 4133, pp. 243–255, 2006, doi:10.1007/11821830_20.

[20] H. Liu, Z. Zhu, N. Iwamoto, Y. Peng, Z. Li, Y. Zhou, E. Bozkurt, and B. Zheng, "Beat: A large-scale semantic and emotional multi-modal dataset for conversational gestures synthesis," in *Lecture Notes in Computer Science*, vol. 13667, pp. 612–630, 2022, doi:10.1007/978-3-031-20071-7_36/TABLES/6.

[21] S. Ginosar, A. Bar, G. Kohavi, C. Chan, A. Owens, and J. Malik, "Learning individual styles of conversational gesture," 2019.

[22] Y. Yoon, B. Cha, J.-H. Lee, M. Jang, J. Lee, J. Kim, and G. Lee, "Speech gesture generation from the trimodal context of text, audio, and speaker identity," *ACM Trans. Graph.*, vol. 39, no. 6, 2020.

[23] M. Stone, D. DeCarlo, I. Oh, C. Rodriguez, A. Stere, A. Lees, and C. Bregler, "Speaking with hands: Creating animated conversational characters from recordings of human performance," *ACM Trans. Graph.*, vol. 23, pp. 506–513, 2004, doi:10.1145/1015706.1015753.

[24] S. Levine, P. Krähenbühl, S. Thrun, and V. Koltun, "Gesture controllers," 2010, doi:10.1145/1778765.1778861.

[25] S. Marsella, Y. Xu, M. Lhommet, A. Feng, S. Scherer, and A. Shapiro, "Virtual character performance from speech," in *Proc. SCA 2013: 12th ACM SIGGRAPH/Eurographics Symp. on Computer Animation*, pp. 25–36, 2013, doi:10.1145/2485895.2485900.

[26] U. Bhattacharya, N. Rewkowski, A. Banerjee, P. Guhan, A. Bera, and D. Manocha, "Text2gestures: A transformer-based network for generating emotive body gestures for virtual agents," 2021.

[27] C. Saund, A. Bîrlădeanu, and S. Marsella, "CCFM: An architecture for realtime gesture generation by clustering gestures by communicative function and motion clustering by communicative function," in *Proc. 20th Int. Conf. Autonomous Agents and Multiagent Systems (AAMAS 2021)*, vol. 9, 2021.

[28] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh, "Openpose: Realtime multi-person 2d pose estimation using part affinity fields," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2019.

[29] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," arXiv preprint arXiv:2002.05709, 2020.

[30] G. Lee, Z. Deng, S. Ma, T. Shiratori, S. Srinivasa, and Y. Sheikh, "Talking with hands 16.2m: A large-scale dataset of synchronized body-finger motion and audio for conversational motion analysis and synthesis," in *Proc. 2019 IEEE/CVF Int. Conf. Computer Vision (ICCV)*, pp. 763–772, 2019, doi:10.1109/ICCV.2019.00085.

[31] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, pp. 5998–6008, 2017.

[32] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proc. Int. Conf. Learning Representations*, 2018.

[33] W. Falcon and T.P.L. team, "PyTorch Lightning," 2019. [Online]. Available: https://www.pytorchlightning.ai, doi:10.5281/zenodo.3828935.

[34] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," in *Proc. 2019 Conf. Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, 2019. [Online]. Available: https://arxiv.org/abs/1908.10084.

[35] Y. Shiban, I. Schelhorn, V. Jobst, A. Hörnlein, F. Puppe, P. Pauli, and A. Mühlberger, "The appearance effect: Influences of virtual agent features on performance and motivation," *Computers in Human Behavior*, vol. 49, pp. 5–11, 2015, doi:10.1016/j.chb.2015.01.077.

[36] S. ter Stal, M. Tabak, H. op den Akker, T. Beinema, and H. Hermens, "Who do you prefer? The effect of age, gender and role on users' first impressions of embodied conversational agents in ehealth,"

Int. J. Human–Comput. Interaction, vol. 36, no. 9, pp. 881–892, 2020, doi:10.1080/10447318.2019.1699744.
[37] "The most popular talks of all time." [Online]. Available: https://www.ted.com/playlists/171/the_most_popular_talks_of_all. Accessed: Nov. 20, 2022.
[38] Y. Liu, G. Mohammadi, Y. Song, and W. Johal, "Speech-based gesture generation for robots and embodied agents: A scoping review," in *Proc. 9th Int. Conf. Human-Agent Interaction (HAI '21)*, pp. 31–38, New York, NY, USA, 2021, doi:10.1145/3472307.3484167.
[39] P. Wolfert, N. Robinson, and T. Belpaeme, "A review of evaluation practices of gesture generation in embodied conversational agents," *IEEE Trans. Human-Machine Syst.*, vol. 52, no. 3, pp. 379–389, 2022, doi:10.1109/THMS.2022.3149173.
[40] S. Olejnik and J. Algina, "Generalized eta and omega squared statistics: Measures of effect size for some common research designs," *Psychol. Methods*, vol. 8, pp. 434–447, 2003, doi:10.1037/1082-989X.8.4.434.
[41] M. Salem, S. Kopp, I. Wachsmuth, K. Rohlfing, and F. Joublin, "Generation and evaluation of communicative robot gesture," *Int. J. Social Robot.*, vol. 4, pp. 201–217, 2012.
[42] M. Salem, F. Eyssel, K. Rohlfing, S. Kopp, and F. Joublin, "To err is human (-like): Effects of robot gesture on perceived anthropomorphism and likability," *Int. J. Social Robot.*, vol. 5, pp. 313–323, 2013.
[43] C. Breazeal, K. Dautenhahn, and T. Kanda, "Social robotics," in *Springer Handbook of Robotics*, pp. 1935–1972, 2016.
[44] H. J. Smith and M. Neff, "Understanding the impact of animated gesture performance on personality perceptions," *ACM Trans. Graph. (TOG)*, vol. 36, no. 4, pp. 1–12, 2017.
[45] C. L. Sidner, C. Lee, L.-P. Morency, and C. Forlines, "The effect of head-nod recognition in human-robot conversation," in *Proc. 1st ACM SIGCHI/SIGART Conf. Human-Robot Interaction*, pp. 290–296, 2006.
[46] N. C. Krämer, N. Simons, and S. Kopp, "The effects of an embodied conversational agent's nonverbal behavior on user's evaluation and behavioral mimicry," in *Proc. 7th Int. Conf. Intelligent Virtual Agents (IVA 2007)*, pp. 238–251, 2007, Springer.
[47] J. Cassell, T. Bickmore, M. Billinghurst, L. Campbell, K. Chang, H. Vilhjálmsson, and H. Yan, "Embodiment in conversational interfaces: Rea," in *Proc. SIGCHI Conf. Human Factors Comput. Syst.*, pp. 520–527, 1999.
[48] K. Bergmann, S. Kopp, and F. Eyssel, "Individualized gesturing outperforms average gesturing–evaluating gesture production in virtual humans," in *Proc. 10th Int. Conf. Intelligent Virtual Agents (IVA 2010)*, pp. 104–117, 2010, Springer.
[49] C.-C. Chiu and S. Marsella, "Gesture generation with low-dimensional embeddings," in *Proc. 2014 Int. Conf. Autonomous Agents and Multi-agent Systems*, pp. 781–788, 2014.
[50] C. N. Gunawardena and F. J. Zittle, "Social presence as a predictor of satisfaction within a computer-mediated conferencing environment," *Am. J. Distance Educ.*, vol. 11, no. 3, pp. 8–26, 1997.
[51] J. N. Bailenson, K. Swinth, C. Hoyt, S. Persky, A. Dimov, and J. Blascovich, "The independent and interactive effects of embodied-agent appearance and behavior on self-report, cognitive, and behavioral markers of copresence in immersive virtual environments," *Presence*, vol. 14, no. 4, pp. 379–393, 2005.
[52] A. Kendon, *Gesture: Visible Action as Utterance*. Cambridge University Press, 2004.
[53] D. McNeill, *Gesture and Thought*. University of Chicago Press, 2005.
[54] N. Zielinski and E. M. Wakefield, "Language proficiency impacts the benefits of co-speech gesture for narrative understanding through a visual attention mechanism," in *Proc. Annu. Meet. Cogn. Sci. Soc.*, vol. 43, 2021.
[55] Z. Zhang, T. Ao, Y. Zhang, Q. Gao, C. Lin, B. Chen, and L. Liu, "Semantic Gesticulator: Semantics-Aware Co-Speech Gesture Synthesis," *ACM Trans. Graph.*, vol. 43, no. 4, pp. 1–17, 2024.
[56] H. Pang, T. Ding, L. He, M. Tao, L. Zhang, and Q. Gan, "LLM Gesticulator: Leveraging large language models for scalable and controllable co-speech gesture synthesis," *arXiv preprint arXiv:2410.10851*, 2024.
[57] N. Gao, Z. Zhao, Z. Zeng, S. Zhang, D. Weng, and Y. Bao, "GesGPT: Speech gesture synthesis with text parsing from ChatGPT," *IEEE Robot. Autom. Lett.*, vol. 9, no. 3, pp. 2718–2725, 2024, doi: 10.1109/LRA.2023.3241801.
[58] L. Zhu, X. Liu, X. Liu, R. Qian, Z. Liu, and L. Yu, "Taming diffusion models for audio-driven co-speech gesture generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Vancouver, BC, Canada, 2023, pp. 10544–10553.

GHAZANFAR ALI completed his PhD from UST in 2023. His research topic primarily focuses on infusing emotional Intelligence into virtual humans in AR/MR scenarios by leveraging deep learning of human behavior. He did his BE Computer Engineering from NUST–Pakistan. He is a former Faculty Member at a Pakistani University. In the Fall of 2017, he came to KIST as a UST student in the Integrative program and started working in the MR Lab. He is currently working as PostDoc at Intelligence and Interaction Research Center, KIST, South Korea.

WOOJOO KIM received his Ph.D. in Industrial and Systems Engineering from the Korea Advanced Institute of Science and Technology (KAIST) in 2022 and his B.S. in Human and Systems Engineering from the Ulsan National Institute of Science and Technology (UNIST) in 2015. He is currently an Assistant Professor in the Division of Liberal Studies, College of Global Future Convergence, Graduate School of Data Science at Kangwon National University, Chuncheon, Republic of Korea. Prior to joining Kangwon National University in March 2023, he was a Post-Doctoral Researcher at the Korea Institute of Science and Technology (KIST) and served as a Managing Consultant at LG CNS.

MUHAMMAD SHAHID ANWAR received his Ph.D. in Information and Communication Engineering from Beijing Institute of Technology in 2021 and his M.Sc. in Telecommunications Technology from Aston University in 2012. He is an Assistant Professor (Research) in the Department of AI and Software at Gachon University, South Korea. His research focuses on multimedia quality assessment, immersive media (VR/AR), and QoE evaluations using machine and deep learning techniques, with applications in healthcare and education.

JAE-IN HWANG is a professor at KIST School and a principal research scientist in Intelligence and Interaction Research Center at Korea Institute of Science and Technology. He received the BS, MS, and PhD degrees in computer science and engineering at Pohang University of Science and Technology,Korea, in 1998, 2000, and 2007, respectively.

AHYOUNG CHOI is an Associate Professor in the Department of AI and Software at Gachon University. She participated in healthcare projects at Samsung Electronics, developing algorithms for wearable devices to assess health and wellness. She also joined the Virtual Human project at the Institute of Creative Technology, USC, as a Postdoctoral Researcher under Jonathan Gratch. Dr. Choi completed her M.S. and Ph.D. at the Gwangju Institute of Science and Technology under Professor Woontack Woo in 2005 and 2011, respectively.