Shahrood University of Technology

Journal of Artificial Intelligence and Data Mining (JAIDM)
Journal homepage: http://jad.shahroodut.ac.ir

Research paper

Skeleton-Based Sign Language Generation Using a Transformer-based Generative Model

Rozhin Mohammadizand and Razieh Rastgoo*

Electrical and Computer Engineering Department, Semnan University, Semnan, 3513119111, Iran.

Article Info

Article History:
Received 07 June 2025
Revised 15 August 2025
Accepted 22 November 2025

DOI:10.22044/jadm.2025.16369.2759

Keywords:
Sign Language Generation, Skeleton, Transformer, Generative Adversarial Network (GAN), Sign Language Recognition.

**Corresponding author: rrastgoo@semnan.ac.ir (R. Rastgoo).*

Abstract

Sign language is a structured, non-vocal form of communication primarily used by individuals who are deaf or hard of hearing, who often face challenges interacting with non-signers. To address this, translation systems between sign and spoken language are essential, encompassing sign language recognition and production. In this work, we focus on sign language production and propose a deep learning framework for generating skeleton-based video representations of sign language at the word level. Our approach employs a conditional Generative Adversarial Network (cGAN) with transformer embeddings in both the generator and discriminator, augmented with bone-length and joint-angle constraints and a classifier-guided loss to ensure anatomically plausible and semantically consistent gestures. We further introduce a novel loss function to improve human keypoint generation for sign representation. Extensive experiments on three benchmark datasets demonstrate that our method outperforms state-of-the-art approaches according to statistical (MMD) and perceptual (FID) metrics, while qualitative analyses confirm that the generated gestures are temporally smooth, anatomically accurate, and semantically meaningful. These results highlight the effectiveness of our model in advancing word-level sign language synthesis.

1. Introduction

Sign language is a complex, structured form of communication used primarily by individuals who are deaf or hard of hearing, relying on both manual elements, such as hand gestures and movements, and non-manual elements, including facial expressions, body posture, and mouth shapes to convey meaning and grammatical structure [1,2]. Globally, over 5% of the population requires treatment for hearing impairments, with more than 60 million completely deaf individuals, and projections estimate that by 2050, over 2.5 billion people will experience some degree of hearing loss [3,4]. Given this widespread prevalence, developing systems for translating between sign language and spoken language is essential for effective communication [5,6]. The lack of a universal sign language and the existence of distinct regional sign languages introduce significant challenges for robust translation systems [7,8]. Moreover, many deaf individuals are not literate in written spoken languages, making sign language their primary mode of communication [9,10]. Unlike spoken languages, sign languages have unique grammatical rules and can express multiple meanings simultaneously within a single gesture, with sequential structuring that differs from spoken language [11,12]. While substantial progress has been made in sign language recognition (SLR), which translates signs into spoken language [13], sign language production (SLP), generating signs from spoken input, remains less explored [14]. Key challenges in SLP include accurately modeling hand orientation, movement, location, and shape, as well as addressing limited large-scale datasets and regional variations [15], all of which are critical for enhancing communication accessibility for the deaf and hard-of-hearing communities.

In the realm of sign language synthesis, various approaches leverage different input modalities [16,17]. While RGB-based models are effective in producing visually realistic outputs, our study adopts a skeleton-based approach due to its significant advantages in this domain. Skeleton-based methods inherently filter out extraneous visual information such as background noise, varying lighting conditions, and clothing artifacts, allowing the model to focus exclusively on the essential joint movements [18]. This significantly reduces data complexity, accelerates training convergence, and enhances the robustness of the model [19]. Furthermore, this approach naturally facilitates the enforcement of anatomical plausibility through direct manipulation of joint positions, leading to physically realistic movements. These methods also offer higher computational efficiency, making them particularly suitable for potential real-time applications, and provide greater interpretability by directly modeling the core kinematics of sign. These benefits underscore the practicality and reliability of leveraging skeletal representations for comprehensive sign language modeling.

To address these challenges and advance the field of sign language production, this research makes the following key contributions that the integration of all these contributions in a unified framework is the main novelty of this work:

• A deep neural network is designed to synthesize skeletal animations of sign language at the lexical level through a conditioned adversarial generative framework. The generator and discriminator incorporate transformer-based sequence embeddings, leveraging the attention mechanism to capture long-range and complex dependencies across sign sequences. This selection is motivated by the Transformer's ability to model holistic relationships between sequence elements, ensuring semantically coherent and anatomically plausible gestures. While LSTM and ST-GCN are effective for certain tasks, they tend to capture dependencies through recurrence or local graph operations, which can make representing global sequence context less direct; therefore, our choice is conceptually grounded in the Transformer's more direct attention-based mechanism for modeling holistic sequence relationships.

• A newly devised cost objective was formulated and implemented to refine the precision of bodily landmark creation for representing signed communication, while integrating bone-length and joint-angle constraints, as well as a classifier-guided loss to further improve semantic alignment and physical plausibility.

• The merit of our system was validated via assessments on a trio of standard data repositories, yielding considerable advancements in efficacy relative to contemporary methodologies and evaluated using both statistical (MMD) and perceptual (FID) metrics, complemented by qualitative analyses confirming that the generated gestures are temporally natural, anatomically plausible, and semantically meaningful.

2. Related works

Overall, we can categorize languages worldwide into two main types: those based on voice and hearing and those based on visual perception and movement [20-23]. The latter category includes sign languages, which are primarily used by deaf and hard-of-hearing individuals. To facilitate communication, translation systems play a crucial role. SLP generates sign language from another modality, while SLR interprets sign language into spoken or written form [8]. SLP can be categorized based on its input type: it may rely on spoken language input (such as text) or visual input (such as images and videos). Visual inputs can be represented in two ways: RGB frames, which contain high-resolution visual information but with higher complexity, and skeletal representations, which offer a lower-complexity alternative by focusing on key movement points rather than full visual details [2] (Figure 1).

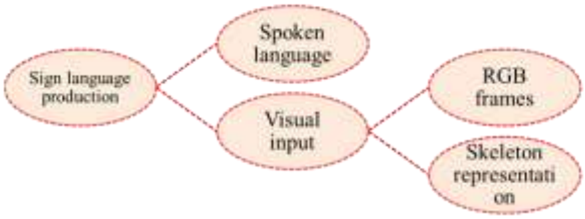

Figure 1. Illustration of categories in SLP.

With recent advancements in deep learning and computer vision, significant progress has been made in sign language recognition and production, as well as in related fields. Several researchers have explored different approaches to these tasks. For instance, the authors in [14] and [24] have worked on SLR from RGB videos using deep learning architectures. More specifically, [24] employed Convolutional Neural Networks (CNNs), while [14] leveraged Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRU), which are feedback-based learning models. Meanwhile, Natarajan et al. [4] proposed a hybrid approach for SLR in RGB videos, integrating CNNs with Bidirectional Long Short-Term Memory (CNN BiLSTM) networks. Additionally, Amorim et al. [17] focused on skeleton-based recognition using

Spatial-Temporal Graph Convolutional Networks (ST-GCNs). Similarly, Jiang et al. [11] worked with skeleton data for SLR.

Beyond SLR, recent studies have advanced SLP as well. For instance, Saunders et al. [7] introduced a model based on Progressive Transformers for end-to-end SLP, enabling direct translation of spoken language text into continuous 3D sign pose sequences. Qi et al. [20] developed a method leveraging latent diffusion models to generate sign language videos from text. Their work focuses on generating RGB videos. Walsh et al. [25] proposed a "sign stitching" approach, which constructs sign language sequences using dictionary examples and then refines them with a Generative Adversarial Network (GAN). Azevedo et al. [26] focused on generating non-manual gestures, particularly facial expressions, by integrating sentiment and semantic information using Spatio-Temporal Graph Convolutional Networks (ST-GCNs).

In the closely related field of human action generation, researchers have explored skeleton-based methods. Ivashechkin et al. [27] and Wang et al. [28] have both worked on skeleton-based human action generation but with different architectures. Ivashechkin et al. used a CNN-based approach, while Wang et al. employed a BiLSTM model for sequence generation. Yazdian et al. [29] investigated co-speech gesture generation using Vector Quantized Variational Autoencoders (VQ-VAE).

In the closely related fields of human motion estimation and action recognition, researchers have developed fundamental methods to analyze and understand human movement. Early works in pose estimation relied on statistical approaches such as Deformable Part Models (DPMs) [30], which partition the body into parts to estimate joint positions. With the advancement of deep learning, models like OpenPose [31] and High-Resolution Network (HRNet) [32] have demonstrated superior accuracy in predicting keypoint locations from visual data, providing reliable skeletal representations. Building on pose estimation, action recognition methods aim to classify activities based on body pose sequences. While earlier methods utilized handcrafted features with traditional machine learning models [33], recent deep learning methods such as Spatio-Temporal Graph Convolutional Networks (ST-GCNs) [34] and 3D CNNs [35] have shown state-of-the-art performance. A notable contribution in this area is the study by Rezaee et al. [36], which modeled abnormal walking of the elderly to predict fall risks using a Kalman filter and motion estimation approach. This work illustrates the application of motion estimation techniques for detecting abnormal movement patterns in real-world health-related scenarios.

In this work, we introduce a state-of-the-art model applicable to both sign language production and human action generation. We apply it to Persian Sign Language, focusing on manual gestures, and American Sign Language, incorporating facial expressions, body movements, and hand gestures. Our model demonstrates versatility and robustness across both domains.

3. Primitive concept

In this section, we briefly review Generative Adversarial Networks (GANs) [37], Conditional Generative Adversarial Networks (CGANs) [38], and Transformer Encoder [39].

3.1. GANs

GANs were introduced in 2014 [10] (Figure 2). As a deep learning-based generative model designed to learn the distribution of input data and generate new data that closely resemble it. A GAN consists of two main components: a generator and a discriminator. The generator takes a random noise vector as input and produces synthetic data, aiming to generate samples that appear similar to real data. The discriminator, on the other hand, is a binary classifier that distinguishes between real data, sampled from the true distribution Px, and generated data, sampled from the generator's distribution Pz. The generator and discriminator are trained in a competitive framework, where the generator continuously improves to fool the discriminator, while the discriminator learns to differentiate real from fake samples. This adversarial process is formulated as a minimax game:

$$\min_{G} \max_{D} V(D,G) E_{x \sim p_{data}(x)} \log[D(x)] + E_{Z \sim p_{Z}(z)} \log[1 - D(G(z))]$$

(1)

In (1), $V(D, G)$ represents the value function, which encapsulates the adversarial loss to be minimized (min) and maximized (max) by the generator (G) and discriminator(D), respectively. The first term, $E_{\mathrm{x} \sim P_{data(x)}}$ [logD(x)], represents the expected log-likelihood of the discriminator correctly identifying real data samples x drawn from the true data distribution $P_{data(x)}$, where $D(x)$ denotes the probability that the discriminator classifies x as real. The symbol E_x refers to the expectation, or average, over the real data samples x. This term drives the discriminator to correctly classify real data points. The second term, $E_{z \sim Pz(z)}[log(1 - D(G(z)))]$, corresponds to the expected log-likelihood of the discriminator correctly classifying fake samples generated by the generator G from a noise vector z, with E_z denoting the expectation over the noise vectors z, which are drawn from the prior distribution $P_{z(z)}$. Here, $G(z)$ represents the generator's output given z. The log refers to the logarithm. The generator aims to minimize this term by producing fake samples that closely resemble real data, while the discriminator maximizes it by distinguishing fake samples from the real ones.

3.2. CGANs

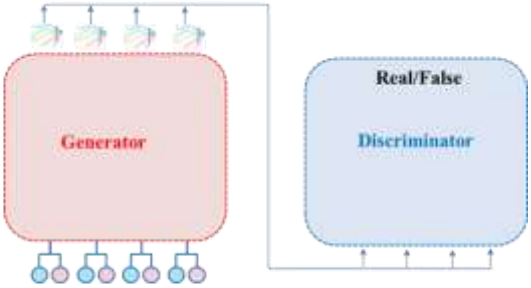

Figure 2. Schematic of the GAN Model.

CGANs were introduced [37] as an extension of GANs to address a key limitation: the lack of control over the generated data. While standard GANs are powerful generative models, they do not provide a mechanism to specify which type of data should be generated. This can lead to mode collapse, where the model produces samples from only a subset of the possible data distribution while ignoring others. In CGANs, additional control is introduced by conditioning both the generator and the discriminator on auxiliary information, such as class labels. This means that a label y is provided as input to both components, guiding the generator to produce samples corresponding to the specified class and helping the discriminator distinguish between real and generated samples within each class. The objective function of CGANs is formulated as:

$$\min_G \max_D V(D,G) E_{x \sim p_{data}(x)} \log[D(x,y)] + E_{z \sim p_z(z)} \log[1 - D(G(z,y),y)] \quad (2)$$

In (2), the formula elements are similar to standard GANs, as explained in part 3.1, with the key difference being the inclusion of the condition y, which represents a label of the data. The generator $G(z,y)$ takes a noise vector z and a condition y (the label), guiding it to produce samples that match the specified condition. The class labels. This means that a label y is provided as an input to both components, guiding the generator to produce samples corresponding to the specified class and helping the discriminator distinguish between real and generated samples within each class.

3.3. Transformer Encoder

The Transformer model [39] is a deep learning architecture designed for processing sequential data. Unlike traditional sequential models, the Transformer leverages parallel computing, significantly improving efficiency by processing entire sequences simultaneously. Additionally, it excels in capturing long-range dependencies within sequences. The model consists of six layers, with self-attention serving as its core mechanism, enabling it to weigh the importance of different input elements dynamically.

To summarize the layers of the Transformer model, the process begins with input embedding, which encodes the input sequence into a continuous vector representation. This is followed by positional encoding, which preserves the order of elements in the sequence by adding position-specific information to the embedding. Next, the multi-head attention mechanism enables the model to capture dependencies between different elements in the sequence. To achieve this, each input element generates three vectors: query(Q), key(K), and value(V). The attention mechanism computes the relationship between elements by taking the dot product of the query vector of one element with the transpose key vector (K^T) of all other elements. This result is then scaled by the square root of the key dimension ($\sqrt{d_k}$) and passed through a softmax function to obtain attention scores. The output is computed by performing a weighted sum of the value vectors based on these attention scores. The mathematical formulation of self-attention is given by:

$$Attention = \text{Softmax}\left(\frac{Q.K^T}{\sqrt{d_k}}\right)*V \quad (3)$$

In (3), Q (Query) represents the vector used to request information from other elements, K (Key) determines how much attention to pay to each element, and V (Value) contains the actual information used in the output. The key dimension (d_k) scales the dot product of queries and keys for stable training.

Following the multi-head attention layer, residual connections are applied, where the original input is added to the output of the attention mechanism. This sum is then normalized using layer normalization. The output is then passed through a position-wise feed-forward network (FFN), which applies a transformation independently to each token in the sequence. Another residual connection is applied, followed by layer normalization. In this manner, the input sequence undergoes a series of transformations, effectively encoding contextual information. Mathematically, this encoding process can be represented as:

$$F_{out} = Enc\left(F_{in}\right) \quad (4)$$

In (4), F_{in} represents the input, F_{out} represents the output of the encoding process, and Enc refers to the overall encoding process. This structured approach enables the Transformer to efficiently model long-range dependencies in sequential data while leveraging parallel computation for improved performance.

4. Proposed sign skeleton CGAN

A CGAN is proposed for generating skeleton-based isolated signs, where each frame consists of key points representing the human body and face. These key points are connected to reconstruct the human skeleton, capturing the essential body and facial movements involved in sign language gestures. This is a challenging task as it requires the generation of human-like, temporally consistent poses and movements, reflecting natural sign language motions. To accomplish this, we leverage the capabilities of the CGAN, which integrates a generator and a discriminator to work adversarially. The generator produces synthetic data, while the discriminator attempts to distinguish between real and generated data. Our CGAN model employs a dual-input mechanism for the generator, which takes both random noise and a conditional data label as inputs. The use of class labels as input was an intentional choice to allow the model to accurately learn the temporal and spatial structure of individual words. While textual input such as full sentences could be used, prioritizing word-level input ensures reliable and coherent word generation and keeps the system manageable, reflecting a deliberate focus on word-level performance rather than a limitation of the architecture. The random noise allows the generator to explore a diverse set of possible outputs, while the conditional label guides the generation process by providing specific information about the target gesture or pose. The discriminator, on the other hand, evaluates both real and generated data along with their corresponding labels. It differentiates between real data and generated data, providing feedback that helps refine the generator's ability to produce realistic and accurate sign language frames. In addition to the generator and discriminator, we introduce a classifier responsible for classifying both real and generated data with respect to their corresponding class labels (Figure 3).

In our proposed model, we aim to generate the skeleton human body parts, including the body, hands, and face, based on specific requirements. This generation process is highly complex because the produced data must not only resemble real human body structures but also represent meaningful gestures necessary for conveying sign language words. Due to this fundamental challenge, we employ a Transformer Encoder in our proposed adversarial generative model for both the generator and the discriminator. The Transformer architecture is particularly suitable for this task as it effectively captures dependencies within and across frames, ensuring coherent and realistic motion representation. Moreover, by leveraging self-attention mechanisms, the Transformer can model long-range relationships between key points, which is essential for generating fluid and natural movements. This capability is especially important in SLP, where subtle variations in motion carry significant linguistic meaning. By incorporating this approach, our model aims to enhance the accuracy and expressiveness of generated gestures, making them more interpretable and reliable for real-world applications.

The loss function is a crucial component of deep learning models, as it enables the model to improve its performance and learn essential features during training. In conditional GANs, there are two distinct loss functions: one for the discriminator (2) and one for the generator (5).

$$L_G = E_{z \sim P_z(z)}[\log\left(D\left(G\left(z,y\right),z\right)\right)] \quad (5)$$

In (5), L_G represents the generator loss, which the generator aims to minimize. The term $E_{z\sim Pz(z)}$denotes the expected value over the noise vector z. The generator G takes this noise z and the condition label y to produce a fake sample $G(z,y)$. The discriminator D then receives the generated sample and the label y to decide whether the sample is real or fake. By minimizing this loss, the generator learns to produce realistic samples that match the condition y and successfully fool the discriminator.

In our proposed model, we use the standard CGAN discriminator loss function (2), while the generator loss function incorporates the standard CGAN loss along with three additional components (6). The first term relates to the classifier, which categorizes both real and generated data into their corresponding classes, ensuring that the generated samples align with the target class labels. The second term is related to the skeletal structure of the human body, and the third concerns the orientation of joints. By incorporating these parameters, we ensure that the generated outputs maintain anatomical accuracy and realistic motion. These adjustments preserve the natural relationships between body parts, ensuring that the generated movements adhere to human biomechanics. Furthermore, we account for the natural limitations of body orientation—such as the fact that certain joint movements, like bending a knee, have specific constraints (e.g., the knee cannot bend in the opposite direction). This consideration helps ensure more realistic and anatomically plausible motion. Additionally, we make all confidence terms in the generator's loss function trainable, enabling the model to dynamically adjust its learning priorities. This adaptability enhances the generator's ability to refine its output, leading to higher-quality, more expressive, and contextually accurate motion representations.

$$L_{Gp} = reg_{gan} * L_G + reg_{class} * (real\_cl\_loss + fake\_cl\_loss) + reg\_bone * bone\_loss + reg\_angle * angle\_loss \quad (6)$$

The proposed generator loss function (6), L_{Gp}, includes several key components. L_G is the standard conditional GAN generator loss, weighted reg_GAN, a trainable confidence parameter. The term reg_class is the trainable confidence for the classification loss, which includes both $real\_cl\_loss$, the loss for real data classification, and $fake\_cl\_loss$, the loss for fake data classification. The component $bone\_loss$, weighted by reg_bone, a trainable parameter, is the bone length loss that calculates the difference in bone lengths between real and generated data. Similarly, $angle\_loss$, scaled by reg_angle, which is also trainable, computes the difference in joint angles between real and fake samples.

5. Results

In this section, details of the datasets, implementation, and experimental results are presented.

5.1. Datasets

To evaluate the performance of our proposed model, three diverse and widely recognized datasets are used. These datasets span different sign languages and human motion capture data, providing a comprehensive evaluation across various domains. One of the datasets is RKS-PERSIANSIGN [40], which is designed for Persian Sign Language. It contains video sequences of native signers with annotations for gestures and movements. The dataset includes data from 10 performers, spanning 100 classes, and consists of 10,000 samples [40]. Another dataset we used is ASLLVD (American Sign Language Lexicon Video Dataset), a comprehensive collection of American Sign Language videos with annotations for hand shapes, locations, and movements. This dataset serves as a benchmark for sign language recognition and generation, and it includes data from 6 performers, covering 3,300 classes, with a total of 9,800 samples [41]. Additionally, we used H3.6M (Human3.6M), a large-scale motion capture dataset widely used for human motion analysis and action recognition. It contains recordings of diverse actions with detailed 3D joint annotations. This dataset includes data from 11 performers, spanning 17 classes, and consists of 3,600,000 samples [42].

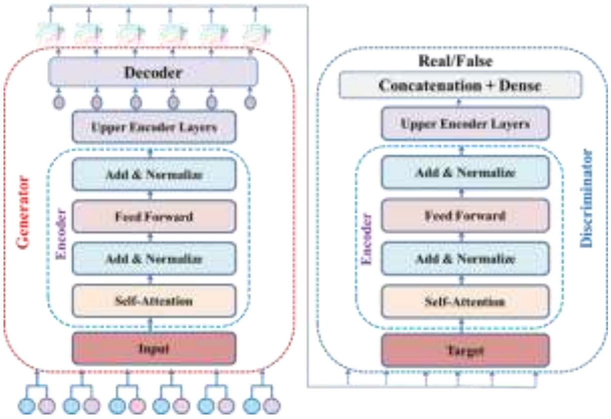

Figure 3. An overview of the proposed model.

5.2. Implementation details

Our model implementation is executed on Google Colab using Python and TensorFlow, providing an efficient environment for deep learning tasks. The model is trained for 20 epochs with a mini-batch size of 64, utilizing the Adam optimizer to optimize the parameters [40], with a decay rate of 0.5 applied every 5 epochs. The discriminator's learning rate is fixed at a smaller value of 0.00001 to ensure balanced training. To further stabilize the training process, momentum values of 0.5 and 0.9 are used for the first and second moments of the Adam optimizer, respectively. These configurations ensure stable, effective training and help the model achieve optimal performance over time.

5.3. Experimental results

Here, the performance of the proposed model is evaluated through experiments on three diverse datasets: RKS-PERSIANSIGN, ASLLVD, and H3.6M (Human3.6M). Our evaluation considers various aspects, such as structural configurations, model parameters, and loss functions, to understand their impacts on the model's output. The model's performance is assessed using the Maximum Mean Discrepancy (MMD) metric across all experiments. The results are presented in

(Table 1), illustrating the performance of different configurations. In Table 2, we compare our model with baseline models using the Human3.6M dataset, while Table 3 compares our base model with the improved model on the RKS-PERSIANSIGN and ASLLVD datasets. Additionally, the outputs are shown in Figures 4, 5, and 6, which display the output frames for these datasets. Besides MMD, we evaluate the generated gestures using the FID metric and conduct qualitative analyses to confirm that the gestures are temporally natural, anatomically plausible, and semantically meaningful. The resulting FID scores for the optimal setup demonstrate strong performance across datasets: RKS-PERSIANSIGN: 25.50, ASLLVD: 26.50, and H3.6M: 23.50.

6. Discussion

In this study, we proposed a novel approach for generating world-level skeleton sign language videos by integrating a CGAN with a Transformer Encoder. Our evaluation involved three diverse datasets: RKS-PERSIANSIGN, ASLLVD, and H3.6M, and we experimented with various configurations to optimize our model's performance. Initially, we evaluated the efficacy of BiLSTM networks for both the generator and discriminator. However, when we replaced the generator with a Transformer, we observed a notable improvement in performance. The best results were achieved when we employed Transformers in both the generator and discriminator, as evidenced by the lowest MMD scores, indicating superior performance in generating realistic sign language movements. This result suggests that the Transformer architecture is highly effective in capturing the complex temporal dependencies involved in sign language generation. Subsequently, we focused on optimizing key learning parameters, particularly the learning rates and the number of iterations for the generator. We found that a higher learning rate for the generator (0.0003) compared to the discriminator (0.00001) led to more stable convergence. Additionally, adjusting the generator's iterations significantly improved the quality of the generated sequences, striking a balance between efficient training and model performance. We also experimented the Transformer architecture configurations, specifically the number of attention heads and layers. Through our experiments, we determined that 13 attention heads and 6 layers produced the best performance, enabling the model to effectively capture multi-dimensional relationships while maintaining computational efficiency. Finally, we enhanced the model's loss function by introducing additional components related to bone length and joint orientation, alongside the standard CGAN and classification losses. The base loss function initially included CGAN and classification losses with non-trainable confidence parameters. In contrast, our second loss function extended this by incorporating two additional components, bone length and joint orientation, along with trainable confidence parameters for all four components. This modification enabled the model to better adapt and achieve more accurate results, outperforming the original loss function, which used non-trainable coefficients for these components. To evaluate the model's performance, the MMD metric has been used, which measures the similarity between generated and real data. Our model consistently showed lower MMD scores compared to baseline models, indicating that the generated sign language movements were more realistic and aligned with real-world data. These results demonstrate the potential of combining CGANs with Transformer encoders for generating high-quality sign language videos, providing a strong foundation for future advancements in this area. Besides MMD, we also assessed perceptual quality using FID and performed qualitative analyses, which confirmed that the generated gestures are temporally smooth, anatomically accurate, and semantically meaningful. It is worth mention that the proposed model focuses on generating individual sign language words, not arbitrary sequences. The Transformer, with its Attention mechanism, is specifically designed to learn the correct order and relationships between sequence elements, ensuring that the generated movements preserve the proper temporal and spatial structure of each word. Therefore, the model does not produce motions blindly, but generates coherent sequences that reflect the intended signs. The choice of Transformer was motivated precisely by this ability to capture dependencies across all elements in the sequence, which is crucial for meaningful sign generation.

Table 1. Results of structural configurations, model parameters, model configuration, loss function and the impact of the classifier module for the three datasets (RKS-PERSIANSIGN, ASLLVD, Human3.6M) based on the MMD metric (lower values indicate better performance).

Discriminator	**Generator**	**Generator training iterations**	**Learning rate**	**Number of heads**	**Number of layers**	**Loss function**	Classifier module	**RKS-PERSIANSIGN**	**ASLLVD**	**Human3.6**
BiLSTM	BiLSTM	2	D 0.00001 G 0.0003	13	6	Base model	Y	0.1440	0.1797	0.2104
BiLSTM	BiLSTM	2	D 0.00001 G 0.0003	13	6	Base model	N	0.2430	0.2624	0.2945
BiLSTM	Transformer	2	D 0.00001 G 0.0003	13	6	Base model	Y	0.1360	0.0876	0.1251
Transformer	Transformer	2	D 0.00001 G 0.0003	13	6	Base model	Y	0.1280	0.0829	0.1237
Transformer	Transformer	4	D 0.00001 G 0.0003	13	6	Base model	Y	0.1320	0.0835	0.1255
Transformer	Transformer	2	D 0.0002 G 0.0002	13	6	Base model	Y	0.1390	0.0877	0.1286
Transformer	Transformer	2	D 0.00001 G 0.0003	18	6	Base model	Y	0.1420	0.0831	0.1275
Transformer	Transformer	2	D 0.00001 G 0.0003	13	9	Base model	Y	0.1340	0.0872	0.1328
Transformer	Transformer	2	D 0.00001 G 0.0003	13	6	Proposed model	Y	0.0980	0.0821	0.1209

Table 2. Comparison of our model with baseline models using the Human3.6M dataset based on the MMD metric (lower values indicate better performance) [14].

Habibie et al.	**Cai et al.**	**Zhenyi Wang et al.**	**Our proposed model**
0.452	0.419	0.195	0.120

Table 3. Comparison between our base model and our proposed model on the RKS-PERSIANSIGN and ASLLVD datasets based on the MMD metric (lower values indicate better performance).

Dataset	**Basel model**	**Our proposed model**
RKS- PERSIANSIGN	0.1440	0.0980
ASLLVD	0.1797	0.0821

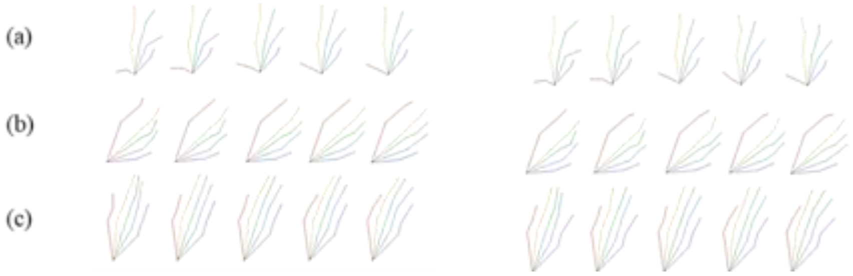

Figure 4. Visual results (synthetic data (right), Ground Truth (left)) on the RKS-PERSIANSIGN dataset: (a) Narahat, (b) Salam, (c) Tabrik.

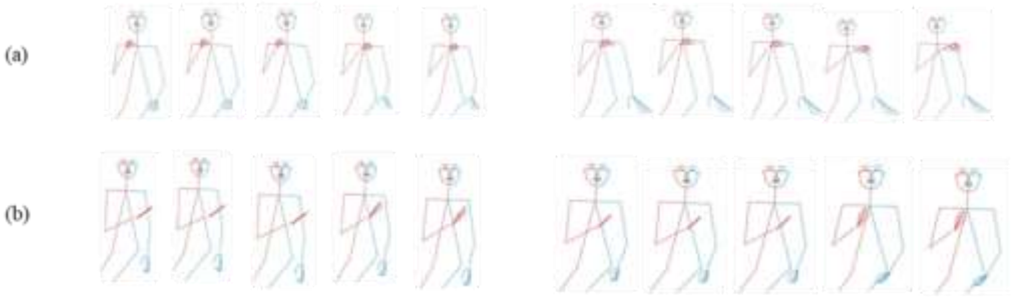

Figure 5. Visual results (synthetic data (right), Ground Truth (left)) on the ASLLVD dataset: (a) place, (b) poss.

.

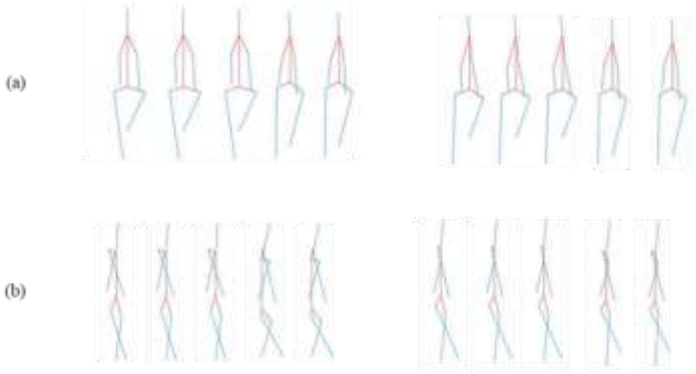

Figure 6. Visual results (synthetic data (right), Ground Truth (left)) on the Human3.6M dataset: (a) walking, (b) walking together.

7. Conclusion

In this paper, we proposed a model for generating skeleton-based sign language videos through the output of skeletal representations of sign language gestures. The model integrates a CGAN with a Transformer Encoder, along with an additional classifier to categorize the generated data into predefined classes. We rigorously evaluated the model on three datasets: RKS-PERSIANSIGN, ASLLVD, and H3.6M, demonstrating superior performance compared to baseline models. The results highlight the model’s ability to effectively capture complex body movements, making it particularly suitable for sign language production and related applications. By combining the power of a CGAN and a Transformer Encoder, the model was optimized to generate realistic body shapes and motions, showcasing its potential for a wide range of real-world applications in gesture recognition, motion synthesis, and human-computer interaction. Our evaluations, including both quantitative and qualitative assessments, confirm that the generated gestures are temporally coherent, anatomically plausible, and semantically meaningful, addressing potential concerns about motion realism. While our current model achieves robust word-level sign generation by focusing on primary skeletal keypoints, we acknowledge that critical non-manual components, such as facial expressions, gaze direction, and head posture, significantly contribute to the full expressive scope and grammatical nuances of sign language communication. Incorporating these elements represents an important direction for future work. Furthermore, our current research deliberately focuses on isolated word-level synthesis. Extending the model to handle more complex linguistic structures, such as sentence- or dialogue-level sign language generation, presents unique challenges given the distinct grammatical rules and simultaneous semantic expressions inherent in sign languages. Addressing these complexities is a natural progression for future research to enhance the model's real-world applicability.

References

[1] R. Rastgoo, K. Kiani, S. Escalera, "Sign Language Recognition: A Deep Survey," *Expert Systems with Applications*, vol. 164, 113794, 2020.

[2] R. Rastgoo, K. Kiani, S. Escalera, V. Athitsos, M. Sabokrou, A survey on recent advances in Sign Language Production, *Expert Systems with Applications* 243:122846, 2024.

[3] World Health Organization, https://www.who.int/. Access Date: May 28, 2025.

[4] B. Natarajan, E. Rajalakshmi, R. Elakkiya, Ketan Kotecha, Ajith Abraham, Lubna Abdelkareim Gabralla, " Development of an End-to-End Deep Learning Framework for Sign Language Recognition, Translation, and Video Generation," *IEEE Access*, vol. 10, pp. 104358-104374, 2022.

[5] R. Rastgoo, K. Kiani, S. Escalera, "Diffusion-Based Continuous Sign Language Generation with Cluster-Specific Fine-Tuning and Motion-Adapted Transformer," *CVPR*, pp. 4088-4097, 2025.

[6] R. Rastgoo, K. Kiani, S. Escalera, "A transformer model for boundary detection in continuous sign language," *Multimedia Tools and Applications*, vol. 83, pp. 89931–89948, 2024.

[7] B. Saunders, N. C. Camgoz, and R. Bowden, "Progressive Transformers for End-to-End Sign Language Production," in *CVPRW*, pp. 11070-11079, 2021.

[8] R. Rastgoo, K. Kiani, S. Escalera, M. Sabokrou, "Sign language production: A review," in *CVPRW*, pp. 3451-3461, 2021.

[9] R. Rastgoo, K. Kiani, S. Escalera, "A non-anatomical graph structure for boundary detection in continuous sign language," *Scientific Reports*, vol. 15, 25683, 2025.

[10] R. Rastgoo, K. Kiani, S. Escalera, "A deep generative Skeleton-based dynamic hand gesture production model," *Multimedia Tools and Applications*, vol. 84, pp. 48589–48608, 2025.

[11] S. Jiang, B. Sun, L. Wang, Y. Bai, K. Li, Y. Fu, "Skeleton Aware Multi-modal Sign Language Recognition," in *CVPRW*, pp. 3408-3418, 2021.

[12] R. Rastgoo, K. Kiani, S. Escalera, V. Athitsos, and M. Sabokrou, "All You Need in Sign Language Production," *arXiv:2201.01609v2*, 2022.

[13] R. Rastgoo, K. Kiani, S. Escalera, "A deep co-attentive hand-based video question answering framework using multi-view skeleton," *Multimedia Tools and Applications*, vol. 82, pp. 1401-1429, 2023.

[14] D. Kothadiya, C. Bhatt, K. Sapariya, K. Patel, A.B. Gil-González, J.M. Corchado, "Deepsign: Sign Language Detection and Recognition Using Deep Learning," *Electronics*, vol.11, 1780, 2022.

[15] R. Rastgoo, K. Kiani, S. Escalera, "ZS-GR: zero-shot gesture recognition from RGB-D videos," *Multimedia Tools and Applications*, vol. 82, pp. 43781-43796, 2023.

[16] R. Rastgoo, K. Kiani, S. Escalera, "Real-time isolated hand sign language recognition using deep networks and SVD," *Journal of Ambient Intelligence and Humanized Computing*, vol. 13, pp. 591-611, 2023.

[17] C.C. Amorim, D. Macêdo, C. Zanchettin, "Spatial-Temporal Graph Convolutional Networks for Sign Language Recognition," *arXiv:1901.11164v2*, 2020.

[18] R. Rastgoo, "A Multi-Stream Diffusion Graph Convolutional Model with Adaptive Motion-Aware Attention and Self-Supervised Pretraining for Continuous Sign Language Recognition," *Neurocomputing*, vol. 656, pp. 131567, 2025.

[19] R. Rastgoo, "A Persian Continuous Sign Language Dataset," *Journal of AI and Data Mining*, vol. 13, pp. 95-105, 2025.

[20] F. Qi, Y. Duan, H. Zhang, and C. Xu, "SignGen: End-to-End Sign Language Video Generation with Latent Diffusion," in *ECCV*, pp. 252-27, 2024

[21] R. Rastgoo, K. Kiani, S. Escalera, "Video-based isolated hand sign language recognition using a deep cascaded model," *Multimedia Tools and Applications*, vol. 79, pp. 22965-22987, 2020.

[22] R. Rastgoo, K. Kiani, S. Escalera, "Hand pose aware multimodal isolated sign language recognition," *Multimedia Tools and Applications*, vol. 80, pp. 127-163, 2021.

[23] R. Rastgoo, K. Kiani, S. Escalera, M. Sabokrou, "Multi-modal zero-shot dynamic hand gesture recognition," *Expert Systems with Applications*, vol. 247, 123349, 2024.

[24] L. Pigou, S. Dieleman, P.J. Kindermans, B. Schrauwen, "Sign Language Recognition Using Convolutional Neural Networks," in *ECCV*, pp. 572–578, 2015.

[25] H. Walsh, B. Saunders, and R. Bowden, "Sign Stitching: A Novel Approach to Sign Language Production," *arXiv:2405.07663v2*, 2024.

[26] R.V. Azevedo, T.M. Coutinho, J.P. Ferreira, T.L. Gomes, E.R. Nascimento, "Empowering Sign Language Communication: Integrating Sentiment and Semantics for Facial Expression Synthesis," *arXiv:2408.15159v1*, 2024.

[27] M. Ivashechkin, O. Mendez, and R. Bowden, "Improving 3D Pose Estimation for Sign Language," *arXiv:2308.09525v1*, 2023.

[28] Z. Wang et al., "Learning Diverse Stochastic Human-Action Generators by Learning Smooth Latent Transitions," *arXiv:1912.10150v1*, 2019.

[29] P. Jome Yazdian, M. Chen, and A. Lim, "Gesture2Vec: Clustering Gestures using Representation Learning Methods for Co-speech Gesture Generation," in *IROS*, pp. 5861-5868, 2022.

[30] P. F. Felzenszwalb, D. McAllester, and D. Ramanan, "A discriminatively trained, multiscale, deformable part model," in *CVPR*, pp. 119-12, 2009.

[31] Z. Cao, T. Simon, S. Wei, and Y. Sheikh, "Realtime multi-person 2D pose estimation using part affinity fields," in *CVPR*, pp. 7291-729, 2017.

[32] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *CVPR*, pp. 5693-570, 2019.

[33] H. Wang, A. Kläser, J. Schmid, and L. Van Gool, "Action recognition by dense trajectories," in *CVPR*, pp. 3169-317, 2011.

[34] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *AAAI*, pp. 7444-745, 2018.

[35] S. Ji, W. Xu, M. Yang, and K. Yu, "3D convolutional neural networks for human action recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 221-231, 2013.

[36] A. Rezaee, M. H. Razavi, M. H. Moradi, and M. R. Nazemzadeh, "Modelling abnormal walking of the elderly to predict fall risks using a Kalman filter and motion estimation approach," *J. Biomech*., vol. 49, no. 1, pp. 43-50, 2016.

[37] N. Esfandiari, K. Kiani, R. Rastgoo, “A conditional generative chatbot using transformer model,” *Journal of Modeling in Engineering*, vol. 23, pp. 99-113, 2025.

[38] N. Esfandiari, K. Kiani, R. Rastgoo, "A new transformer-based generative chatbot using CycleGAN approach," *Neural Computing and Applications*, vol. 37, no. 31, pp. 26125-26156.

[39] A.M. Ahmadi, K. Kiani, R. Rastgoo, "A Transformer-based model for abnormal activity recognition in video," *Journal of Modeling in Engineering*, vol. 22, no. 76, pp. 213-221, 2024.

[40] R. Rastgoo, K. Kiani, S. Escalera, "Hand sign language recognition using multi-view hand skeleton," *Expert Systems with Applications*, vol. 158, 113336, 2020.

[41] R. Rastgoo, K. Kiani, S. Escalera, "Word separation in continuous sign language using isolated signs and post-processing," *Expert Systems with Applications*, vol. 249, 12369, 2024.

[42] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu, "Human3.6M: Large scale datasets and predictive methods for 3D human sensing in natural environments," *IEEE Trans. Pattern Anal. Mach. Intell*., vol. 36, no. 7, pp. 1325–1339, Jul. 2014.

تولید زبان اشاره ناشنوایان مبتنی بر اسکلتون بدن با استفاده از یک مدل مولد مبتنی بر ترنسفورمر

روژین محمدی زند و راضیه راستگو*

دانشکده مهندسی برق و کامپیوتر، دانشگاه سمنان، سمنان، ایران.

ارسال ۲۰۲۵/۰۶/۰۷؛ بازنگری ۲۰۲۵/۰۸/۱۵؛ پذیرش ۲۰۲۵/۱۱/۲۲

چکیده:

زبان اشاره شکلی ساختارمند و غیرآوایی از ارتباط است که عمدتاً توسط افراد ناشنوا یا کم‌شنوا مورد استفاده قرار می‌گیرد؛ افرادی که اغلب در تعامل با افراد غیرآشنا به زبان اشاره با چالش‌هایی مواجه هستند. برای رفع این مشکل، سامانه‌های ترجمه میان زبان اشاره و زبان گفتاری، که شامل تشخیص و تولید زبان ا شاره می شوند، ضروری ه ستند. در این پژوهش، ما بر تولید زبان ا شاره تمرکز کرده و یک چارچوب یادگیری عمیق برای تولید نمایش‌های ویدیویی مبتنی بر اسکلت از زبان اشاره در سطح واژه پیشنهاد می‌کنیم. روش پیشنهادی از یک شبکه مولد تخاصمی شرطی مبتنی بر ترنسفورمر در هر دو بخش مولد و تفکیک‌کننده استفاده می‌کند و با اعمال محدودیت‌های طول استخوان و زاویه مفاصل، به‌همراه یک تابع زیان هدایت‌شده توسط دسته‌بند، حرکت‌هایی از نظر آناتومیکی قابل قبول و از نظر معنایی منسجم تولید می‌کند. علاوه بر این، یک تابع زیان جدید برای بهبود تولید نقاط کلیدی انسانی در نمایش زبان اشاره معرفی شده است. آزمایش‌های گسترده بر روی سه مجموعه‌داده معیار نشان می‌دهند که روش پیشنهادی نسبت به روش‌های پیشرفته موجود، از نظر معیارهای آماری (MMD) و ادراکی (FID) عملکرد برتری دارد. همچنین تحلیل‌های کیفی تأیید می‌کنند که ژ ست‌های تولید شده از نظر زمانی روان، از نظر آناتومیکی دقیق و از نظر معنایی معنادار ه ستند. این نتایج ن شان‌دهنده اثربخ شی مدل پی شنهادی در پیشبرد تولید زبان اشاره در سطح واژه است.

کلمات کلیدی: تولید زبان اشاره، اسکلت، ترنسفورمر، شبکه مولد تخاصمی، تشخیص زبان اشاره.