Received: 5 April 2023 | Revised: 21 June 2023 | Accepted: 18 July 2023

CAAI Transactions on Intelligence Technology

DOI: 10.1049/cit2.12321

The Institution of Engineering and Technology

WILEY

ORIGINAL RESEARCH

Improving diversity of speech-driven gesture generation with memory networks as dynamic dictionaries

Zeyu Zhao[1,2] | **Nan Gao**[1] | **Zhi Zeng**[3] | **Guixuan Zhang**[1,3] | **Jie Liu**[1,3] | **Shuwu Zhang**[3]

[1]Institute of Automation, Chinese Academy of Sciences, Beijing, China

[2]School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China

[3]Beijing University of Posts and Telecommunications, Beijing, China

Correspondence

Zhi Zeng.
Email: zhi.zeng@bupt.edu.cn

Funding information

National Key R&D Program of China, Grant/Award Number: 2022YFF0902202

Abstract

Generating co-speech gestures for interactive digital humans remains challenging because of the indeterministic nature of the problem. The authors observe that gestures generated from speech audio or text by existing neural methods often contain less movement shift than expected, which can be viewed as slow or dull. Thus, a new generative model coupled with memory networks as dynamic dictionaries for speech-driven gesture generation with improved diversity is proposed. More specifically, the dictionary network dynamically stores connections between text and pose features in a list of key-value pairs as the memory for the pose generation network to look up; the pose generation network then merges the matching pose features and input audio features for generating the final pose sequences. To make the improvements more accurately measurable, a new objective evaluation metric for gesture diversity that can remove the influence of low-quality motions is also proposed and tested. Quantitative and qualitative experiments demonstrate that the proposed architecture succeeds in generating gestures with improved diversity.

KEYWORDS

artificial intelligence, gesture

1 | INTRODUCTION

With the continuous development of technologies such as high-speed wireless communication, power-efficient mobile computation, holographic high-definition display, and data-driven artificial intelligence (AI), extended reality (XR) is widely considered to be the next-generation carrier of information that will replace multimodal video contents in the near future. A three-dimensional (3D) immersive environment will be provided by these synthetic realities that fascinate users with a sensory experience similar to the real world. Using two-dimensional (2D) user interfaces on mobile devices today in such realities may block users' direct and natural interaction with the media content, downgrading the experience to a non-immersive level. Thus, novel interaction methods that are more intuitive and human-friendly are needed to be perfected before the transition to the new era starts.

People exchange information mainly by communicating with other people. Using digital humans or virtual avatars as intermediaries when interacting in synthetic worlds can be regarded as intuitive and natural. More specifically, people mainly communicate by talking using language-encoded voices. The problem of generating voices from response text for digital humans in an interactive session has been well resolved by traditional and neural text-to-speech (TTS) technologies, as they can now effortlessly generate human-like voices with tones and emotions, which can be seen as relatively deterministic. However, simple observation in daily lives can show that people naturally talk while performing accompanied body gestures in most scenarios as well. Sometimes speeches without gestures can be viewed as unnatural when people communicate. Systems that can generate co-speech gestures given input speech text or audio can play important roles in future interactive sessions for making digital humans more natural.

Besides, co-speech gestures generated by such systems can also be helpful for those with some form of hearing loss [43] or understanding disability since co-speech gestures can provide auxiliary non-verbal information. However, the problem of generating gestures from speech voice audio or script text remains challenging for its indeterministic nature, where there exists a one-to-many mapping between the modalities. For the same speech, there exist considerably wider range of possibilities of acceptable co-speech gestures than voices.

Solutions to this kind of generation problem are being made increasingly feasible by the utilisation of neural networks in recent years. Given sets of synchronised speech and gesture data, neural models automatically learn the correlation between the modalities and gain the ability to give results from previously unseen speeches [2]. Nevertheless, it is observed that gestures generated by neural methods from in-the-wild speeches seem to be rather slow or dull, with less amount of movement shift than expected, which appears to be less diverse than those performed by a normal person. This may be explained by the total losses used in the learning process of the models that sum up the losses of all training samples. Guided by the losses, networks tend to perform best when taking all training samples into consideration, thus resulting in an averaged output.

To address this problem, we introduce dynamic dictionaries of pre-stored features to neural generation architectures, breaking the decoder's sole dependency on the averaged output by learnt parameters of the encoders. This grants the decoder of the generative network direct access to pose features dynamically stored in the dictionaries from the training datasets, making the generation process more flexible and comprehensive. Inspired by [3], we select memory networks to play the role of the dictionary [1], which are commonly used in numerous generation tasks with similar requirements in different fields of study. In experiments, the memory-network-based dictionaries succeeds in dynamically providing pose features to the generator based on generative adversarial networks (GAN) and improving the diversity of the generated gestures. Moreover, to ensure that the improvements we achieved are coming from the enhanced diversity instead of degraded quality such as lower rhythmicity, lower human-likeness, or higher randomness, we also propose a new relative objective metric that intuitively represents the level of diversity while eliminating those influences. With the new metric, reproducible experiments supporting our opinions are also conducted to show that said improvements are achievable using proposed architectures.

Our main contributions can be summarised as follows:

- To further improve the diversity of the speech-driven gesture generation results, we propose a generative model with an encoder-decoder architecture, multiple backbone choices, and memory networks introduced as dynamic dictionaries that explicitly store connections between speech and pose features.
- To accurately measure the diversity of the results with the effects of random, unsmooth, or falsely rhythmed motions eliminated, we propose a new metric for objective evaluation of gesture diversity, utilising velocity changes in pose sequences, with multiple threshold setups.
- To demonstrate those improvements are measurable and plausible, we design multiple quantitative and qualitative experiments including comparisons, ablation studies and user studies.

2 | RELATED WORKS

2.1 | Speech-driven gesture generation

Rule-based methods are used in early studies to blend pre-captured or pre-defined gestures stored in databases for primary interactive speech-driven gesture generation applications [23] before data-driven methods become mainstream. For example, by inferring the acoustic and semantic properties of the utterance for virtual characters, rules are defined in Ref. [24] to generate gestures and expressions. Rules to link a pre-defined set of unit gestures to keywords [25] are also used for real-world robots to perform actions when talking. Apart from utilising rules that need to be handcrafted, approaches based on probabilistic modelling are also used to solve this problem [23]. For example, the authors in Ref. [26] infer motion state distribution over a set of hidden states from the speech signal and design an optimal-policy controller based on conditional random fields to select the optimal. In Ref. [27], the problem is seen as a classification problem, and a neural model is proposed for selecting a proper gesture for a given speech context. Results given by these methods are more of a simple concatenation of pre-defined gestures than generating new gestures, drawing limited attention from researchers.

Data-driven methods such as neural networks and deep learning have made end-to-end speech-driven gesture generation applications possible and attention-drawing in recent years. These approaches view the problem as a regression problem instead of a classification problem. As a specialised generation problem, human gesture generation can leverage general deep generative models, including GANs [15] and variational autoencoders (VAE) [28]. Speech2Gestures [4] transform input audio spectrograms to generated gestures using convolutional encoders and decoders with the guidance of an adversarial loss. Seq2Seq [5] is designed to have an encoder-decoder architecture of attentional networks to generate sequences of poses from text inputs. JointEmbedding [6] creates human motion from description text by mapping text and motion to the same embedding space. TrimodalContext [2] adversarially trains multiple encoders and a decoder based on recurrent neural networks (RNN) that takes encoded multimodal data, such as text, audio, and speaker id, as input, to generate gestures. In more recent works, more complicated network architectures are designed to capture certain decoupled properties of gesture data. By splitting the latent code into shared and motion-specific codes for VAE, Audio2Gestures [29] explicitly models the audio-motion mapping for generation. The network in FreeMo [30] is divided into a rhythmic motion branch and a pose mode branch that uses conditional VAEs for improved performances. Fine-grained gestures are generated in HA2G [31] by extracting audio representations across semantic granularities so that the entire human pose can be gradually rendered in a hierarchical manner. The authors in Ref. [38] decouple gestures into rhythmic and semantic gestures and build correspondence between hierarchical embeddings of the speech and the motion. Also, new generative models such as flow-based models [39] and diffusion models, for example, DiffGesture, [40] are applied to the field. Good quality gestures can be generated controllably by these methods while they also come with limited diversity or increased architectural complexity compared to the proposed method. A comparison between these existing methods is shown in Table 1.

2.2 | Vision transformers

In the most recent years, vision transformers (ViT) [36] are becoming increasingly popular as a substitute for common models such as convolutional neural networks (CNN) in vision tasks, with the utilisation of the self-attention mechanism and parallelisable design. Originating in the field of natural language processing (NLP), transformer models that can process sequential input data all at once beat RNNs in performance and are used as backbones in many big models [8]. In theory, the multi-head self-attention mechanism with relative positional encoding can be viewed as a general case of convolution operations, which allows transformers to gain the ability to process images by directly viewing images divided into small sections as sequences. Despite the advantages, it requires much more effort to train transformers than CNN models from scratch for the complex structures inside the modules. Fortunately, we can instead fine-tune existing transformers pretrained from specific datasets, for example, ImageNet [37] and audio spectrograms [13], to utilise their abilities in speech-driven gesture generation tasks.

2.3 | Memory networks

Memory networks can work as external augmentation modules for common neural networks with long-term memory that stores explicit connections between domains or modalities. Early applications of memory networks can be found in algorithmic problem solving as neural Turing machines (NTM) [32], or problems in the field of NLP such as question answering (QA) tasks [33]. They are also introduced to problems, such as lifelong and one-shot learning, for their ability to memorise rare events [10]. Later, they are observed to work well on dealing with image features. For example, in Ref. [34], they are used for different tasks such as personalisation issues of image captioning, generating a descriptive sentence for a query image, and accounting for prior knowledge such as the user's active vocabularies in previous documents. In Ref. [35], memory networks are also used in alleviation of GAN training problems in image generation tasks. Based on the similarity of these applications, we believe that memory networks can also be utilised in speech-driven gesture generation to connect modalities of speech and pose features.

3 | METHOD

In this section, we formulate the speech-driven gesture generation problem and describe in detail how we introduce memory networks to the generative model as dynamic dictionaries of connections between speech and pose features, which breaks decoder's sole dependency on learnt parameters of the encoders and bring diversity improvements for the result of the speech-driven gesture generation.

3.1 | Problem formulation

The speech-driven gesture generation problem is the problem of generating non-verbal gesture sequences synchronised with input speech audio or text. A gesture sequence can be defined as a compound list of frames y containing 2D or 3D human joint position or rotation. The speech audio can be preprocessed and converted into spectrograms a so that they can work better with common well-performing feature extractors, for example, convolutional networks or transformers. The text is represented as a list of words aligned in time to the gesture frames, one word for each frame, which can be further embedded to vectors as network input t. For better non-

TABLE 1 Advantages and disadvantages of existing speech-driven gesture generation methods.

Category		Example methods	Advantages	Disadvantages
Traditional methods	Rule-based methods	[24, 25]	High-quality, smooth, and meaningful pre-defined gestures.	Concatenation of limited gestures without new gestures generated.
	Probabilistic modelling	[26, 27]		
Large-scale data-driven methods	Traditional generative models	[2, 4–6]	Automatic end-to-end learning with targeted design. New gestures generated controllably.	Requires huge amount of training data. Increased architectural complexity. Less-than-expected diversity.
	New architectural designs	[29–31, 38]		
	New generative models	[39, 40]		

sequential network compatibility, they are usually split into subsequences with fixed lengths, padded if necessary. The goal is to obtain a function $y = F(a,t)$ that maps audio or text input to the synchronising gesture sequence. Data-driven methods view this as a regression problem and design neural networks to fit such functions. In this work, we design a neural network which has a new architecture with memory networks that improves the diversity of the generated results.

3.2 | Overview

The proposed architecture is designed in a multimodal fashion, combining semantic and rhythmic connections between source *text* or *audio* and target *gesture* modalities without specifically designed NLP or other modules. The gestures are similarly defined as in Ref. [2] to be sequences of human poses. For each time frame in a sequence, 9 normalised directional vectors converted from 3D coordinates of 10 upper body joints (spine, head, nose, neck, left and right shoulders, left and right elbows, and left and right wrists), centred at the spine, are listed in a fixed order. Each directional vector points from a parent joint towards a child joint and has a length of 1. The introduction of directional vectors significantly stabilises the pose jittering between frames by eliminating the influence of root translation and bone length etc., since the locations of ground truth joints are inaccurately extracted from in-the-wild videos. More details of pose definition can be found in Ref. [1].

Before being fed into the generator or put into memory as input, all modalities of synchronised data are subdivided into groups of segments with equal temporal lengths l_{seq}, padded if necessary, and then processed by different feature extractors, to features with fixed lengths, avoiding the instability and inaccuracy of directly searching in raw data space. The text feature extractor which takes the embedded text representation as input is a sentence transformer fine-tuned from a pre-trained MiniLM [7], a compressed model distilled from bidirectional encoder representations from transformers (BERT) [8] models. This compressed design enables the extractor to provide text features that are distinctive enough with much less time consumption when processing. The feature extractor for audio that takes converted log-Mel spectrogram as input is the same as the audio encoder, which will be further introduced in the following sections. Also, the raw pose sequences are encoded by a CNN feature extractor trained in an unsupervised auto-encoder scheme from [2] as the pose features.

As illustrated in Figure 1, the whole model consists of two sub-networks: the dictionary network and the pose generation network. The dictionary network stores previously seen key-value pairs of text and pose features, providing a dynamic

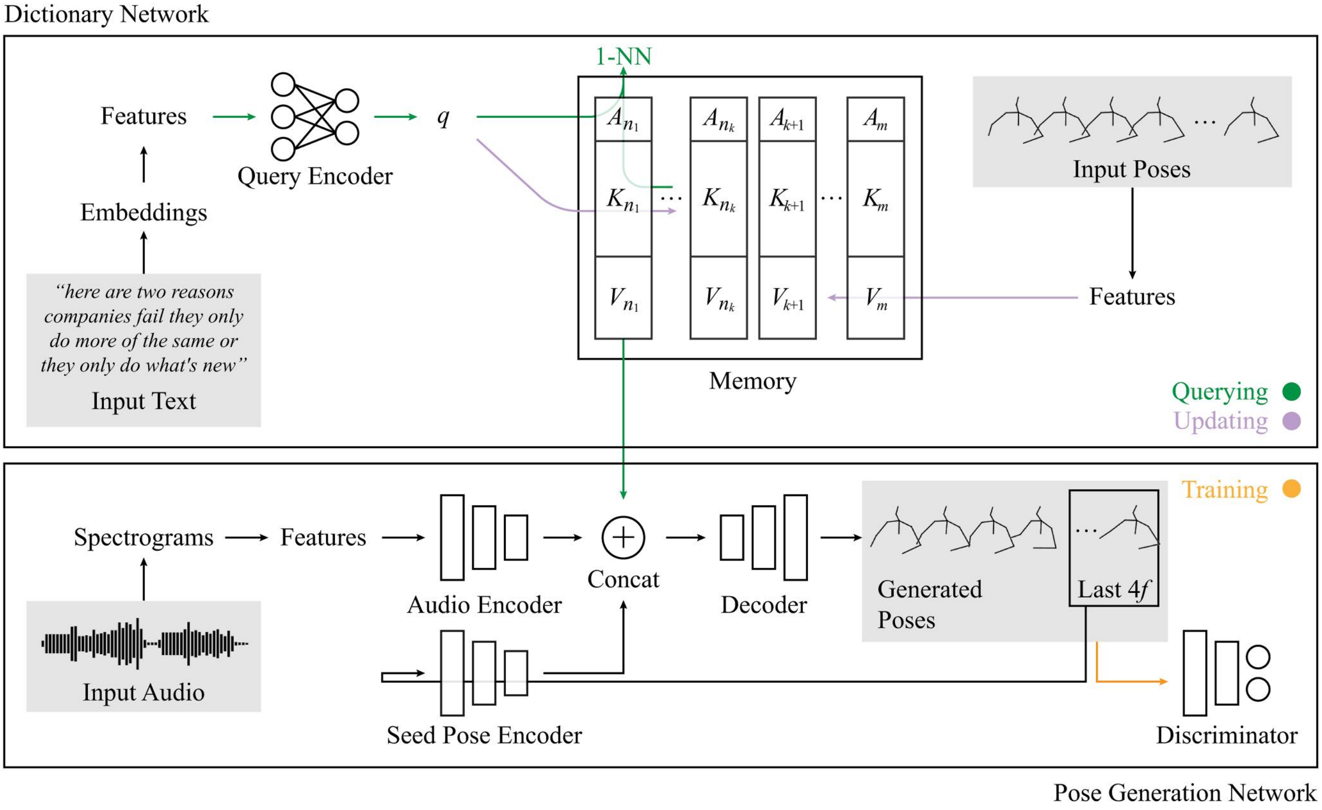

FIGURE 1 Overview of the proposed network architecture. The model is composed of two sub-networks. The dictionary network stores key-value pairs of text and pose features for the pose generation network to look up. During querying, the dictionary network searches for the nearest key and returns the corresponding value, where the keys and values came from the input text and pose features in the updating stage. The pose generation network then combines the audio, seed pose, and fetched pose features to generate the expected gestures. The adversarial training scheme is used with a discriminator for more realistic and human-like results.

dictionary with raw pose features that can be queried and fetched by the pose generation network. The decoder of the pose generation network then generates gestures combining those features and encoded input audio features by its encoder. In such a way, the decoder with access to the pre-stored raw pose features gains the ability to avoid the absolute dependency on the learnt parameters of the encoder, making generated gestures to be more diverse.

3.3 | Dictionary network

We choose memory networks as the backbone of the dictionary network. It works as a dynamic dictionary enabled by the ability of neural networks to learn from data. Inside the network, a list of previously stored key-value pairs of text and pose features, called the memory, is maintained for the generator to interact directly. Given a query as input, the dictionary network searches for the most approximate key in the memory and returns the corresponding value as output. For each key-value pair, an additional age value is attached for the updating process to determine the proper operation along with a threshold. This threshold is also used in the training process of the network. The structure of the dictionary network is shown in Figure 2, which is set up similarly to the one used in Ref. [3] for most mentioned attributes.

Formally, the memory M is a set of slots of key-value pairs with key $K_i, i = 1, 2, \ldots, m$, and value V_i, with an attached age $A_i \in \mathbb{N}$:

$$M = \{(K_1, V_1, A_1), (K_2, V_2, A_2), \ldots, (K_m, V_m, A_m)\}, \quad (1)$$

where all keys and values are randomly initialised and all ages are initialised at 0. To perform a query action, the input text feature X extracted from raw embedded text x is firstly processed by the query encoder $Q(X)$ and then normalised to construct the query q as in:

$$q = \frac{\mathrm{Q}(X)}{\|\mathrm{Q}(X)\|}, \quad (2)$$

where $\|q\| = 1$. In this work, a simplified linear query encoder with weights W and bias b

$$\mathrm{Q}(X) = WX + b, \quad (3)$$

is sufficient to fulfil its purpose, which is to make the search in memory keys more accurate by making the keys more distinctive, given the query, since all keys are coming from previously accepted queries. Then, the network calculates the cosine similarity between the query and keys in all slots, which are all guaranteed to be normalised, and the ones with top k similarity are selected as the k-nearest neighbours $\{(K_{n_1}, V_{n_1}, A_{n_1}), \ldots, (K_{n_k}, V_{n_k}, A_{n_k})\}$:

$$\{n_1, n_2, \ldots, n_k\} = \arg\max_i^{(k)} q \cdot K_i, \quad (4)$$

where $\|q\| = \|K_i\| = 1$. The pose feature value V_{n_1} of the 1-nearest neighbour $(K_{n_1}, V_{n_1}, A_{n_1})$ is then chosen to be the output of the dictionary network N_D:

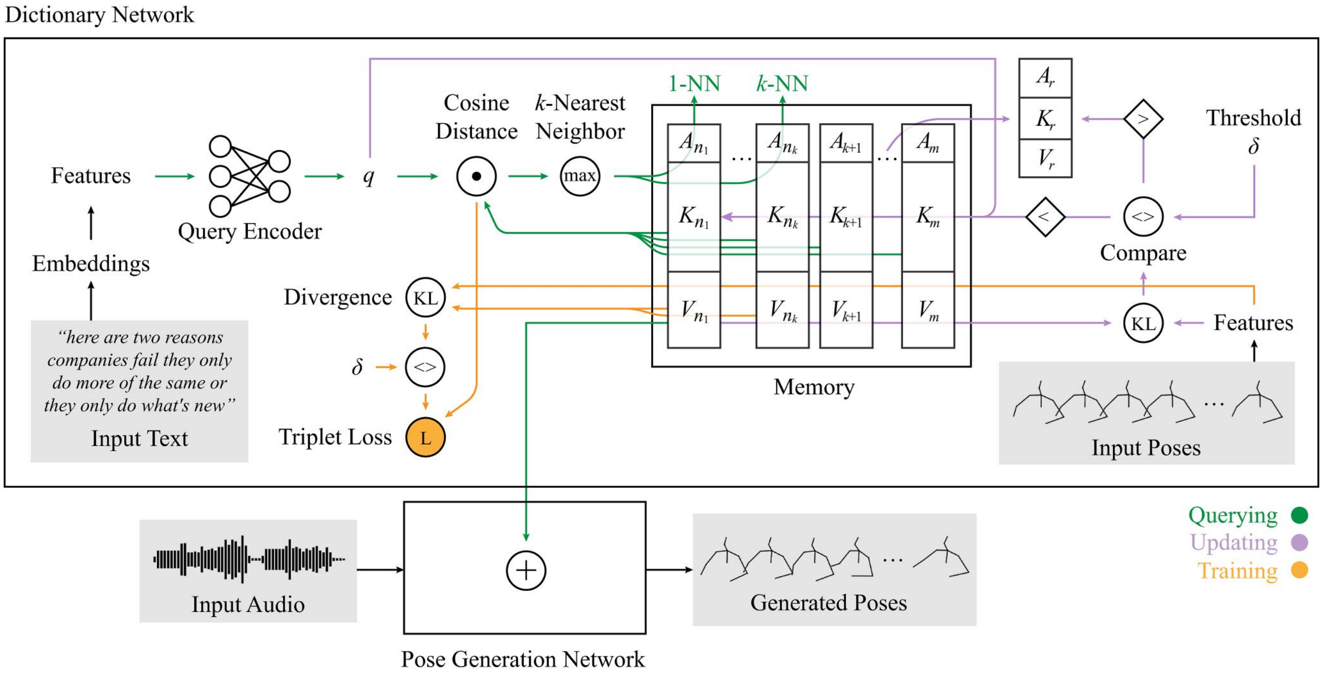

FIGURE 2 Detailed view of the dictionary network. When queried with input text features, the network calculates the cosine distances between the encoded query and the keys in memory and searches for k-NNs. The value of 1-NN is then returned as the result. The memory can then be updated if needed under different circumstances comparing the KL divergence between the value of 1-NN and the input pose features. The query encoder is also needed to be trained using a triplet loss constructed using the cosine distances according to different situations of the KL divergence between the value of k-NNs and the input pose features.

$$V_{n_1} = N_{\mathrm{D}}(X), \tag{5}$$

to complete the query process.

Each time a new input with the expected output or the ground truth pose features is provided, the memory can be updated under certain rules in the training stage of the model. Given input text feature X and target pose feature Y, as usual, the k-nearest neighbours of the constructed query are selected. Additionally, the ages of all slots are increased by one in this process. Then, we calculate the Kullback–Leibler (KL) divergence [9]:

$$\mathrm{KL}\ (v\|u) = \sum_{l} v_l \cdot \log \frac{v_l}{u_l}, \tag{6}$$

between the value of the neighbours $V_{n_j}, j = 1, 2, \ldots, k$, and the target output Y and compare that to a neighbouring threshold δ_n defined as a hyperparameter. Here, we define a positive neighbour if the calculated divergence from the neighbour is equal to or less than the threshold, which is

$$V_{n_j} \in \{v \mid \mathrm{KL}\ (v\|Y) \leq \delta_{\mathrm{n}}\}, \tag{7}$$

and a negative neighbour if the divergence is greater, which is

$$V_{n_j} \in \{v \mid \mathrm{KL}\ (v\|Y) > \delta_{\mathrm{n}}\}. \tag{8}$$

If the 1-nearest neighbour turns out to be positive, we consider that the memory slot containing the neighbour is reusable for this target since they share a similar value. The key is then updated by averaging the old key and the new query, followed by normalisation, and the age is reset to zero:

$$K_{n_1} \leftarrow \frac{q + K_{n_1}}{\|q + K_{n_1}\|}, A_{n_1} \leftarrow 0. \tag{9}$$

And if it is determined negative, it can indicate that there is no matched value in any memory slot that is similar enough to the target. In that case, one of the memory slots with the oldest age (K_r, V_r, A_r) is randomly selected and overwritten:

$$K_r \leftarrow q, V_r \leftarrow Y, A_r \leftarrow 0. \tag{10}$$

Note that only the 1-nearest neighbour is updated if necessary instead of all neighbours.

Furthermore, we also need to train the query encoder to generate queries and later keys with better distinguishability from one another. To do this, we use a triplet loss [10] with compatibility modifications to make it applicable to our situation. The original triplet loss is suitable for supervised situations since the positive or negative class labels of the samples are known to the model, which is not the case in this unsupervised setup. Therefore, similar criteria in the updating process with the threshold δ can be reused to determine the positive or negative class labels for the triplet loss. Given a query and the selected k nearest neighbours, we construct the triplet loss using the positive and negative neighbour with the smallest index $\left(K_{n_{\mathrm{p}}}, V_{n_{\mathrm{p}}}, A_{n_{\mathrm{p}}}\right), (K_{n_{\mathrm{n}}}, V_{n_{\mathrm{n}}}, A_{n_{\mathrm{n}}})$ in the ordered neighbour list:

$$L_{\mathrm{triplet}} = \max\left(q \cdot K_{n_{\mathrm{n}}} - q \cdot K_{n_{\mathrm{p}}} + \alpha, 0\right), \tag{11}$$

where α is the margin between positive and negative neighbours. The loss is truncated to be above zero to work with this margin. We can see that if the cosine similarity between the query and the positive neighbour is greater than that between the query and the negative neighbour added by a margin alpha, this loss will not be effective since in this case it is already distinctive enough for searching. The optimisation of this triplet loss

$$\min_{W,b} L_{\mathrm{triplet}}, \tag{12}$$

will maximise the similarity of the constructed query with the positive neighbours and at the same time minimise the similarity with the positive neighbours, allowing the keys to be more distinctive.

3.4 | Pose generation network

The pose generation network is an encoder-decoder network, with a similar architecture that can be commonly found in other generative setups, as seen in Figure 3. Since it is not the main focus of our network design, we reuse the basic structure from the translation model in Ref. [4] with additional modifications such as decoder input adjustments and backbone replacements. The network is composed of two encoders, which are the audio encoder E_a and the seed pose decoder E_s and a pose decoder D_p.

The audio encoder processes the log-Mel spectrogram M converted from the raw audio input and produces expected audio features U:

$$U = E_{\mathrm{a}}(M). \tag{13}$$

As shown in Figure 4, the backbone of the audio encoder can be chosen from a UNet [12], a fine-tuned Audio Spectrogram Transformer (AST) [13], or a mixed architecture of UNet and AST. A UNet audio encoder consists of ConvNorm modules connected residually which can process features hierarchically with a simplified UNet architecture. A single ConvNorm module here consists of a convolutional layer appended by a batch normalisation [41] layer and finally activated by a Leaky ReLU function [42]. In such architecture, the feature information extracted in the contracting path (left) at each layer can be reused by concatenating that with the spatial information at the corresponding layer in the expansive path (right). More spatial information is included in the features after UNet processing, making it easier for the decoder to produce gestures. An AST audio encoder is a fine-tuned

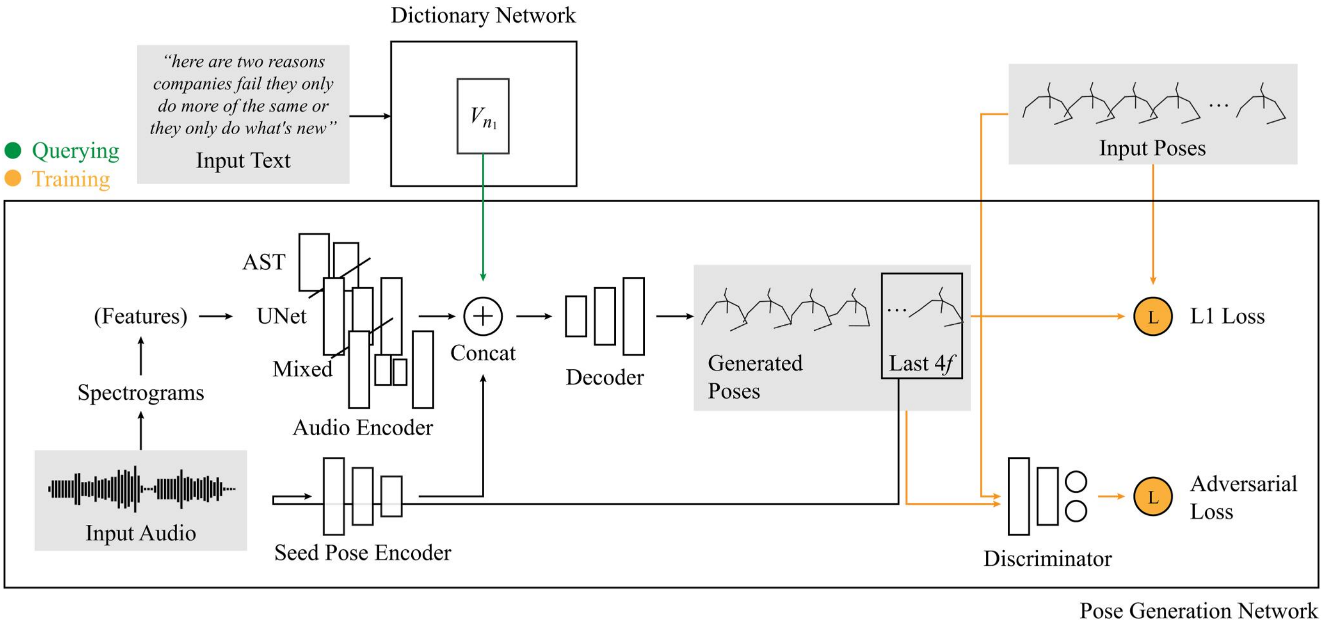

FIGURE 3 Detailed view of the pose generation network. The decoder receives concatenated audio, seed pose, and queried pose features to generate the expected gestures. The seed pose encoder utilises the last $l_{\text{seed}} = 4$ frames of the generated poses to initialise the next segment. The backbone of the audio encoder can be chosen from three different configurations. The network is supervised by the ground truth poses and trained in an adversarial scheme with a discriminator for more realistic and human-like results.

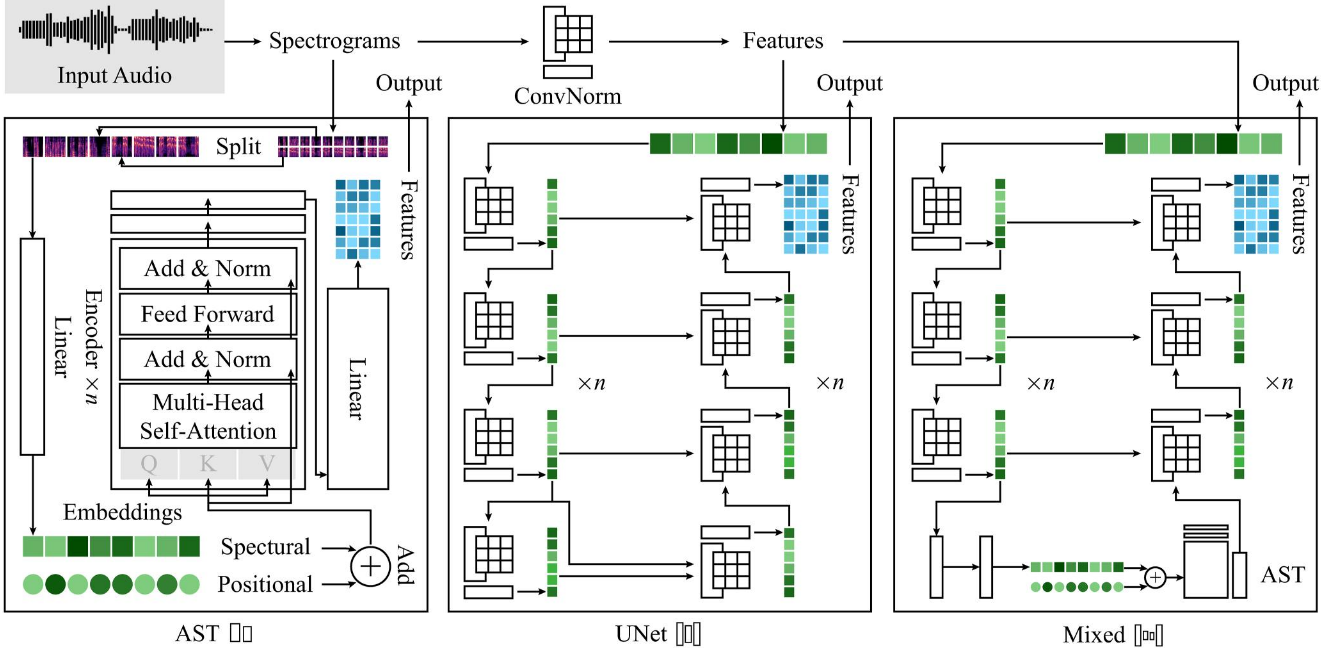

FIGURE 4 Three different backbones of the audio encoder, AST, UNet, and a mixed architecture of both. An AST audio encoder is a fine-tuned transformer model pre-trained for spectrograms with self-attention encoders. A UNet audio encoder uses residually connected convolutional layers with batch normalisation (ConvNorm) for hierarchical feature processing. A mixed audio encoder introduces AST as the bottom-layer feature processer to UNet for better output features.

transformer originally pre-trained for audio spectrograms. It has original transformer encoders based on multi-head self-attention with prepending and appending linear layers for processing input spectrograms and generating output audio features. The spectrograms are first split into patches and linearly projected as spectral embeddings, which is then added with trainable positional embeddings before being fed into the encoder. This AST configuration with a similar number of parameters (6 million) can be less time-consuming than UNet when processing data of the same scale due to the wider but shallower structure with more parameters per layer and less layer count. Compared to other image transformer models without pre-training at all or pre-trained on general image datasets, AST is chosen here for the reason that it is pre-trained on similar spectrogram data as ours, saving a huge amount of time to train a transformer from scratch. Also, according to [14], sometimes a mixed usage of UNet and transformer can improve the final results in practice. Such architecture combines the ability of both UNet and transformer for hierarchical and attentional feature extraction. Thus, we also design a mixed architecture to evaluate this insight. In this mixed architecture, an AST model is inserted between the last layer of the contracting path of a further simplified UNet and the first layer of the expansive path, with a linear layer prepended, increasing the number of parameters to more than 10 million. Note that before being fed into a UNet or a mixed audio encoder, the spectrogram is first processed by a fully convolutional (FC) [11] pre-extractor of multiple ConvNorm modules.

Then, the seed pose encoder with a linear design processes seed poses S to form the seed pose features p in a similar fashion:

$$P = E_{\mathrm{s}}(S). \tag{14}$$

Here, seed poses are used to shape the beginning frames of the generated pose segment and guarantee the continuity between segments. For the first segment, seed poses can be set to all zero or initialised by specified values by choice. For the second and subsequent segments with index $s > 1$, they are set to the last l_{seed} frames of the previous segment's generated poses:

$$S^{(s)} = \begin{cases} S_0, & s = 1, \\ S^{(s-1)}_{\left(l_{\mathrm{seg}} - l_{\mathrm{seed}} + 1, l_{\mathrm{seg}}\right)}, & s > 1, \end{cases} \tag{15}$$

where S_0 is the initialisation value of the seed poses, and the parenthesised superscript and subscript of S represent the index of the segment and the range of frame indices in the segment, respectively, and l_{seg} is the length of the segment. Guided by the losses, the network is required to generate the exact same poses at the beginning of the pose segment as the seed poses input. Thus, with extra post-generation smoothing, the last frames of the previous segment can be seamlessly merged with the beginning frames of the current segment.

Finally, the pose decoder generates the final pose segment output $\hat{y}$ by processing the encoded information combining the output of the dictionary network, the audio encoder, and the seed pose decoder:

$$\hat{y} = D_{\mathrm{p}}(V_{n_1} \oplus U \oplus P), \tag{16}$$

where $\oplus$ is the concatenation symbol.

For simplicity, we simultaneously optimise all encoders and decoders with the L1 loss between ground truth and the generated poses:

$$L_{\mathrm{L1}} = \mathbb{E}\|\hat{y} - y\|_1. \tag{17}$$

Instead of directly optimising this loss, we adopt the adversarial training scheme [15] for better realisticity and human-likeness of the generated results. We design a CNN binary discriminator D optimised to better distinguish between the ground truth and generated gestures, as "real" and "fake" gestures. Viewing it as a generator G, the forementioned part of the network should work to make the generated gestures as realistic as possible to fool the discriminator to be one that is unable to differentiate between generated and real gestures. This can be achieved by alternatively optimising the discriminator and the generator using the L1 loss and the adversarial losses with minimax optimisation. When the discriminator is being trained, the generator is fixed and the discriminator tries to better distinguish between real and fake gestures by maximising the discriminator loss:

$$\max_D L_{\mathrm{adv}_D} = \min_D - L_{\mathrm{adv}_D}, \tag{18}$$

where

$$L_{\mathrm{adv}_D} = \mathbb{E} \log D(y) + \mathbb{E} \log(1 - D\hat{y}). \tag{19}$$

When the generator is being trained, the discriminator is fixed and the generator tries to confuse the discriminator by minimising the generator loss:

$$\min_G L_G, \tag{20}$$

where

$$L_{\mathrm{G}} = \gamma_{\mathrm{L1}} L_{\mathrm{L1}} + \gamma_{\mathrm{adv}_G} L_{\mathrm{adv}_G}, \tag{21}$$

in which γ_{L1} and γ_{adv} are the weights for the respective L1 and adversarial losses, and

$$L_{\mathrm{adv}_G} = \mathbb{E} \log(1 - D\hat{y}). \tag{22}$$

This makes the generated gestures and the real ones to be more similarly distributed, which pushes the generated gestures to be more natural and human-like.

4 | EXPERIMENTS

In this section, we introduce the quantitative and qualitative evaluations and ablation studies we performed on the proposed model to demonstrate our achieved improvements are plausible and reproducible with the proposed objective metric for gesture diversity.

4.1 | Objective metric for gesture diversity

The simplest way to objectively evaluate a gesture generation model is to directly or indirectly [2] compare the generated sequences of pose landmarks, or in our case directional vectors, to the ground truth. Nevertheless, due to the indeterministic nature of the speech-driven gesture generation problem, such a comparison can be unsuitable in many cases. In theory, there exists an infinite number of possible gestures that can match the given input audio or text. In practice, different models trained on different datasets give totally different gestures which, on the contrary, can all be seen as valid for the same input. Over the years, major enhancements on objective evaluation metrics in related studies that are popular and widely used are still primarily made for this kind of comparison to evaluate how well the generated results match the ground truth. Some new high-quality metrics for better evaluating qualities such as rhythmic synchronisation are also proposed in some works. Compared to these metrics, the metrics for diversity are relatively neglected. Some of these works do provide their own basic metrics for diversity, but we find them weak in discernibility and inaccurate in our practice, rendering our improvements on diversity untraceable between different models.

During evaluating and experimenting with different metrics, we observe that the degree of rhythmic properties approximately matches the simple subjective perception of the gesture diversity level by human evaluators. It is easy to see that when a rhythmic shifting event occurs, sudden changes in frame-to-frame motion velocity appear in the motion sequences. Thus, we take advantage of this feature of velocity to form the new objective metric for diversity. By differentiating respective directional vectors between frames, we obtain the interframe velocity in the form of 3D vectors with magnitude and direction:

$$v_p = y_p - y_{p-1}, \tag{23}$$

where $p = 2, 3, \ldots, l_{\text{seg}}$. Some existing metrics utilising velocity vectors focus on the change of magnitude, which is hard to eliminate even small unrelated interferences such as different ranges of motion, interframe jittering, or random noises etc., in Euclidean space. We experimented on utilising the direction changes of the velocity vectors and found that the angle difference between frames is a good indicator of sudden changes in velocity at a rhythmic shifting event.

Specifically, we describe these changes by finding the angles between adjacent interframe velocities in degrees:

$$a_{\text{v},t} = \operatorname{degarccos} \frac{v_t \cdot v_{t-1}}{|v_t||v_{t-1}|}, \tag{24}$$

where $t = 3, 4, \ldots, l_{\text{seg}}$. Then, the angles are compared to an activation threshold $\delta_{a_{\text{v}}}$ which should be set empirically considering framerate etc. This threshold defines the minimum angle that should be viewed as the indicator of any sudden change in velocity direction. Finally, the ratio of angles above the threshold is calculated to be the absolute diversity metric:

$$e_{\text{absdiv}} = \frac{\sum_{a \in a_{\text{v}}} \mathbf{1}(a > \delta_{a_{\text{v}}})}{l_{\text{seg}} - 2}, \tag{25}$$

where $\mathbf{1}\ (\cdot)$ is the indicator function that returns 1 when the criterion behind is satisfied, or else it returns 0. Evidently, higher e_{absdiv} indicates more sudden changes in velocity directions which means better diversity in a pose sequence.

Nevertheless, this score can be fooled to be falsely high by very strong random noises or interframe jittering in the gesture sequences that are beyond our designed limits. This happens when properties such as rhythmicity or smoothness of the generated results are reduced or degraded. This makes the diversity quality unmeasurable using this metric and should be avoided by all means. Through experiments, we find that these interferences can be further eliminated by guaranteeing the rhythmic synchronisation between generated and ground truth gestures. As explained above, direct comparison in landmark coordinates is not suitable for this task. It strictly limits the generated poses to be the same in every aspect for higher similarity, which conflicts with the indeterministic nature of the problem. Hence, other ways to measure this synchronisation property are required to be found. Taking simplicity into consideration, we finally agree on what is suggested in Ref. [30], that two pose sequences with similar rhythm should have velocities with magnitudes that are temporally related. Stationary parts of gesture sequences share small magnitudes of velocities since in those frames the poses are similar to one another. If large movements in two matching sequences happen at the same point in time, regardless of the directions, there should be simultaneous increases in the magnitudes of the velocities. Therefore, we can form the synchronisation metric by directly comparing the magnitudes of the velocities:

$$e_{\text{syn}} = \mathbb{E}\, \| |\hat{v}| - |v| \|_1, \tag{26}$$

where $\hat{v}$ and v are velocities calculated from generated and ground truth gestures using the same method as above, respectively. For two sequences of gestures with good synchronisation, lower e_{syn} should be true since they are more similar in the velocity magnitudes. With the synchronisation score as the calibration factor, we can now form a simple relative diversity metric:

$$e_{\text{div}} = \frac{e_{\text{absdiv}}}{e_{\text{syn}}}. \tag{27}$$

4.2 | Dataset

For comparability and convenience, we train, test, and evaluate the proposed models on an extended version [2] of the TED gesture dataset [5], which is widely used and continuously improved in a number of studies. 1766 videos of TED talks containing people giving speeches on stages in the English language are collected and processed as the raw source of all data. Using OpenPose [16], 2D landmarks of human poses on each frame of the videos are detected and converted to 3D by a temporal convolutional estimator [17] trained on large-scale motion capture datasets. Transcribed English speech texts with onset timestamps of each word are extracted from the consisting audio using the Gentle forced aligner [18]. The extracted poses only contain the upper body part, with the segments where the full upper body cannot be clearly seen filtered out, resulting in 97 h' worth of synchronized sequences. The sequences are resampled to be 15 frames per second and subdivided into 34-frame segments ($l_{\text{seq}} = 34$) with a stride of 10. The length of the seed poses is set at $l_{\text{seed}} = 4$. When training, the initialisation value of the seed poses S_0 in the first segment of a sequence is set to be the same as ground truth value. Direction vectors introduced above are converted from the raw landmarks of the pose joints before being fed into the model. The dataset is divided the same way as the original authors [2], producing 199,384; 26,795; and 25,930 segments, respectively, for training, validation, and test sets.

4.3 | Quantitative evaluation

The model is trained, evaluated, and tested in different configurations. Most basic settings in the configurations are set to be identical, including the margin of the triplet loss $\alpha = 0.3$, the weight for the L1 loss $\gamma_{L1} = 100$, and the weight for the adversarial loss in the generator loss $\gamma_{\text{adv}} = 10$ etc. Also, some settings need to be adjusted dynamically for each different configuration, such as the total number of epochs, batch size, learning rate etc. These settings can affect the convergence speed of the model but are limited by the computing and memory performance of the running platform. We perform these adjustments with the goal of getting the final training losses to the same level with similar time consumption.

For the dictionary network, two memory configurations with different neighbouring thresholds are applied: $\delta_n = 0.8$ and $\delta_n = 0$. When set at 0.8, the threshold allows multiple similar pose features to share one memory slot with an averaged query key. We call this the dynamic memory configuration since the keys in the memory can be dynamically updated. With the zero threshold, however, the memory slots become static as every single pair of input text and pose features is a negative neighbour and saved to the memory as a new key-value pair. This makes the memory update static and hence can be called the static memory configuration. For the pose generation network, we tried three different backbone configurations, which are represented as UNet [12], AST [13], and Mixed [14], also mentioned in the Method section. Also, we provide two configurations of the activation threshold, $\delta_{a_v} = 90$ and $\delta_{a_v} = 120$, to see if there is any difference for the diversity metric when actions of smaller or larger magnitudes are regarded as sudden changes.

As baselines, 4 models are trained, evaluated, and tested in a similar manner using the same dataset, with different input modalities of speech audio, text, and mixed. Seq2Seq [5] is an attention-based model that generates gestures from text using RNN encoders and decoders with the attention mechanism. Speech2Gesture [4] generates gestures from speech audio with an architecture of UNet encoder-decoder trained in the adversarial fashion, which is also what our model is based on without our pipeline modifications. Thus, the comparison between this and our model can also be considered as an ablation study. JointEmbedding [6] uses a different representative approach that maps the text and motion to the same embedding space and creates motion from description text. Trimodal [2] combines trimodal context of text, audio, and speaker identity as joint input to learn co-speech gestures. All experiments are performed on the same machine with a 10-core (20-thread) Intel Xeon Silver 4210R @ 2.40 GHz CPU and two GeForce RTX 3090 GPUs.

Table 2 displays the comparison among the proposed model and 4 baselines in different memory configurations. We see that the proposed model with the static memory configuration has the highest score on absolute and calibrated diversity metric e_{absdiv}, e_{div} in all activation configurations, showing 30% and 35% improvements in gesture diversity over the best-performing baselines, respectively. Close synchronisation scores e_{syn} (less than 7% difference) also indicate that the rhythmic synchronisation between the generated and ground truth gestures is at similar levels, which means the better

TABLE 2 Comparison among 4 baselines and the proposed models in different memory configurations. We evaluate the models with the activation threshold δ_{a_v} of the diversity metric set at different levels and show the results separately (↓: Lower is better. ↑: Higher is better. **Bold**: Best in full comparison. Underline: Best in ablation study).

	Evaluation metric				
		$\delta_{a_v} = 90$		$\delta_{a_v} = 120$	
Method	e_{syn}↓	e_{absdiv}↑	e_{div}↑	e_{absdiv}↑	e_{div}↑
Seq2Seq [5]	0.01537	0.05164	3.35979	0.02905	1.89005
Speech2Gesture [4]	0.01599	0.17103	10.69606	0.09133	5.71169
JointEmbedding [6]	**0.01439**	0.01216	0.84503	0.00518	0.35997
Trimodal [2]	0.01450	0.10567	7.28759	0.05119	3.53034
Proposed (dynamic memory)	0.01694	0.21919	12.93920	0.11195	6.60862
Proposed (static memory)	0.01546	**0.22272**	**14.40621**	**0.12244**	**7.91979**

Note: The bold values indicates best in full comparison.

performance in diversity is not likely caused by degraded results with random noises or interframe jittering etc. In the ablation study, our proposed model with the static memory configuration performs 3%, 30%, and 35% better in three metrics, respectively, than Speech2Gesture, indicating that the pose feature dictionary based on memory networks is the key factor to achieve such improvements. As for the comparison in the activation threshold, we do not see order-changing differences between the two configurations. Consistent with the theory, we notice that with a bigger threshold $\delta_{a_v} = 120$, the metric focuses on greater changes in vector directions, resulting in more precise scores with less sensitivity, which is better matched with human evaluation. Thus, we make this the default activation threshold in later-introduced results.

In Table 3, we show the results of different backbone configurations of the audio encoder in the pose generation network. We see that despite our targeted AST design, the UNet-based audio encoder still results in better performance in diversity. Nevertheless, the introduction of AST does bring a considerable reduction in time (29%) consumption while maintaining roughly the same level of performance as the UNet-based model in dynamic memory configuration, which can outweigh the performance degradation in certain scenarios that require faster inference. As suggested in Ref. [14], the mixed configuration outperforms the other two with the best score in synchronisation and diversity with still less time consumption than UNet. If real-time generation is not required, the mixed configuration should be used to guarantee the best generation results.

4.4 | Qualitative results

We visualise the generated results for human evaluation by rendering videos of both stickman-like skeletons and retargeted animation of rigged 3D human models. Skeletons in 3D space can be rendered as 2D frames by re-converting directional vectors in gesture sequences back to joint landmarks and drawing lines between parent and child joints using off-the-shelf plotting tools such as Matplotlib [19]. Based on the conversion, joint rotations in the skeletons can be further calculated and retargeted to the skeleton of a rigged 3D human model using linear blend skinning (LBS), with the assistance of scriptable 3D animating software such as Blender [20]. For this, we develop tools that can read and parse the mesh, skeleton, and animation data from the compressed custom data files provided by the famous 3D open-world video game Grand Theft Auto V (GTAV) [21]. Released in the year 2013, the game contains a large number of rigged 3D human models with good design at that time to construct a huge world that allows players to interact freely with non-player characters (NPC) from different cultural backgrounds and in different personal styles. The popularity of the game spawned a huge modding community that provides all kinds of modding tools with continuous support, such as CodeWalker [22], which makes the development of our tools much easier. Since videos cannot be displayed on paper, we still use sequences of stick skeletons to demonstrate the generation results in static figures for clarity. Rendered videos of animated 3D human models are used in the user study that is later introduced.

TABLE 3 Comparison among the backbone configurations of the audio encoder. Time consumption for processing the entire test set is recorded for all configurations (↓: Lower is better. ↑: Higher is better. **Bold**: Best in comparison).

	Evaluation metric			
Method	e_{syn}↓	e_{absdiv}↑	e_{div}↑	**Time (s)**↓
UNet [12] (dynamic memory)	0.01694	0.11195	6.60862	53.2
UNet [12] (static memory)	0.01574	0.11971	7.60546	51.1
AST [13] (dynamic memory)	0.01721	0.09376	5.44800	**36.4**
AST [13] (static memory)	0.01615	0.10164	6.29350	36.8
Mixed [14] (dynamic memory)	0.01710	0.11160	6.52632	47.2
Mixed [14] (static memory)	**0.01546**	**0.12244**	**7.91979**	46.9

Note: The bold values indicates best in full comparison.

Abbreviation: AST, Audio Spectrogram Transformer.

To better demonstrate the generation results of pose sequences, we manually select frames with expected posing behaviours in different circumstances, for example, in a sentence or at rest, from the gesture segments generated by the proposed and baseline models from the same input and make static figures of stick skeleton sequences, as shown in Figures 5 and 6. In Figure 5, we show an example of generation results of the proposed model in two configurations. In Figure 6, we compare gestures generated by baselines and the proposed models. We can see that for the same speech in both modalities of text and audio, the gestures generated by the proposed model contain noticeably more drastic motions with improved diversity than the baselines. When there is a short silence in audio, the model produces gestures with more scene-transition-like gestures than the baselines. When a long silence in audio is encountered, however, the model produces normal gestures with a similar level of motion shifts as the baseline models instead of gestures that look random caused by the improved diversity. This further proves that our improvements in diversity are not coming from interframe jittering or random noises.

Also, six sets of rendered videos of retargeted animation on rigged 3D human models are shown to human evaluators as questionnaires in the user study of the generation results. As shown in Figure 7, for each set of videos, we transfer the movements of each upper-body gesture sequence generated by the six baseline or proposed models from the same speech to the upper part of a randomly chosen full-body rigged 3D human model from the video game GTAV. The visibility of the original gesture sequence in the form of a stickman-like skeleton can be toggled on and off according to our needs. The resulting six videos are then reorganised into 15 pairwise comparison questions as a single questionnaire. We find 12 human evaluators unrelated to the field and send three questionnaires to each. The results of this user study are shown in Figure 8. We see that for two sequences of gestures generated

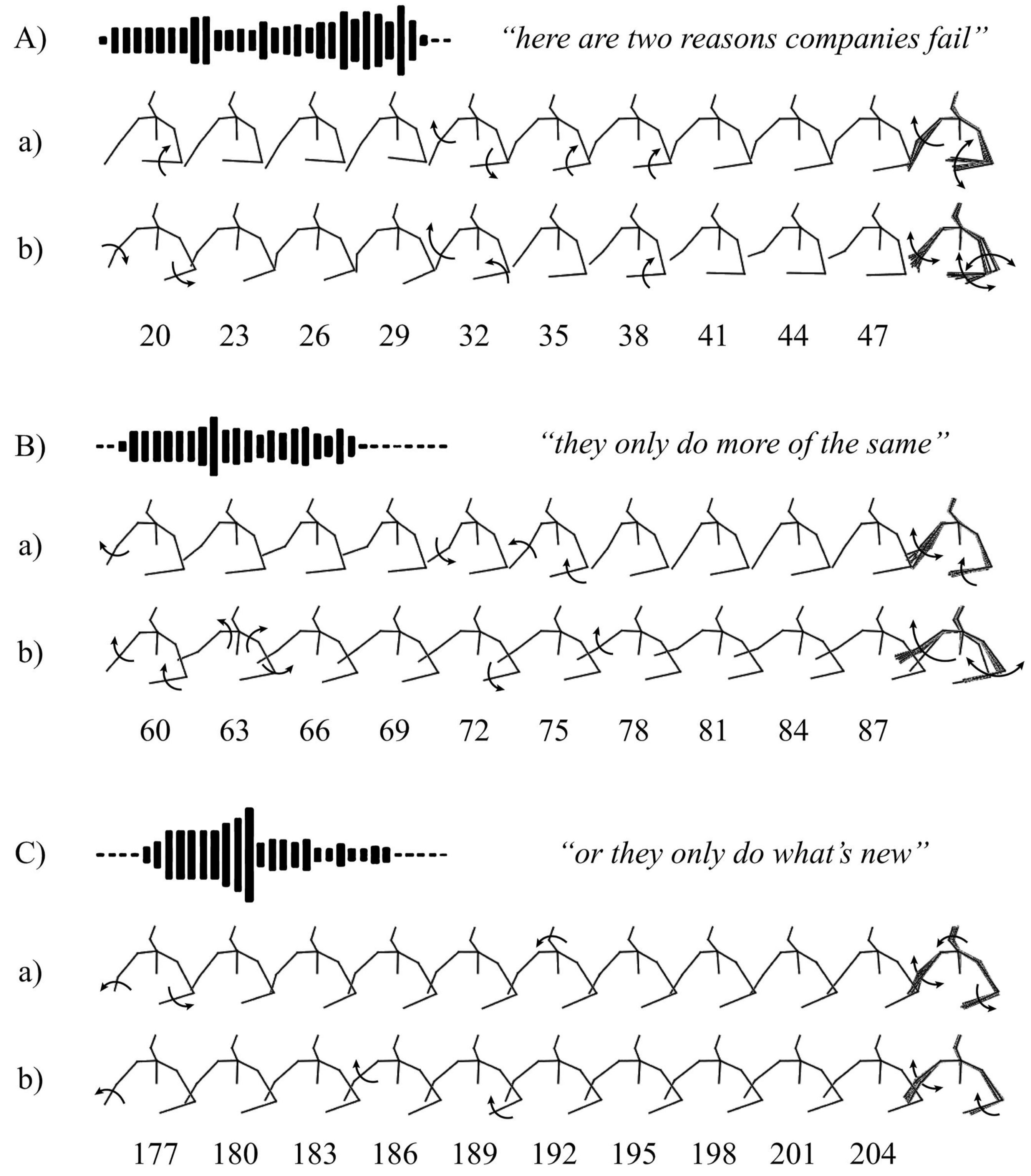

FIGURE 5 An example of generated gestures in stickman-like skeletons. Provided with a group of input audio waveform and text split (for demonstration only) into three segments (a), (b), and (c), the proposed model in (a) dynamic and (b) static memory configurations generates the gestures, respectively. Shorter segments of typical gestures are then selected in this figure. Note that each gesture segment has an appended stacked view of all poses. Frame numbers and arrows indicating directions of movements are shown below and on the skeletons.

by any two of the baseline or proposed models, the evaluators generally agree that the one with higher e_{div} score has better diversity than the other one with very few anomalies. All human evaluators agree that our model produces gestures with better diversity than the baselines without sacrifices in human likeness or smoothness.

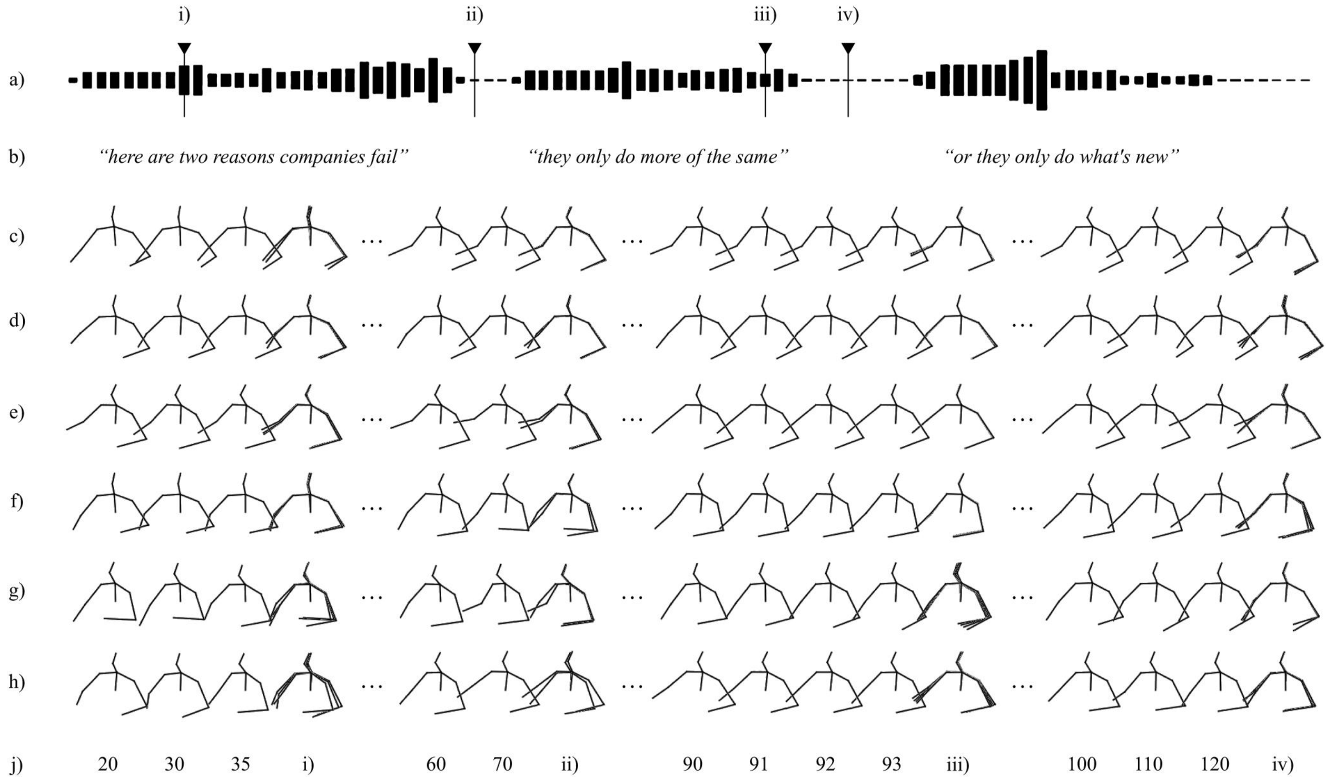
i)
ii)
iii)
iv)
a)
b)
"here are two reasons companies fail"
"they only do more of the same"
"or they only do what's new"
c)
d)
e)
f)
g)
h)
j)
20
30
35
i)
60
70
ii)
90
91
92
93
iii)
100
110
120
iv)

FIGURE 6 Qualitative comparison between results of different models in stickman-like skeletons. In the figure, we have (a) the testing audio waveform with time stamps indicating the approximate starting position of four selected ranges of frames for four different circumstances, which are (i) early in sentence, (ii) at short rest, (iii) late in sentence, and (iv) at long rest; and (b) corresponding text; and selected frames from gestures generated by (c) Seq2Seq, (d) Speech2Gesture, (e) JointEmbedding, (f) Trimodal, and the proposed model with (g) dynamic and (h) static memory configuration, with (j) frame numbers and corresponding labels of the four different circumstances with variable sampling intervals. Note that for each range of selected frames, the stacked view of each range is shown above the circumstance labels for a clearer comparison on the level of movement shifts.

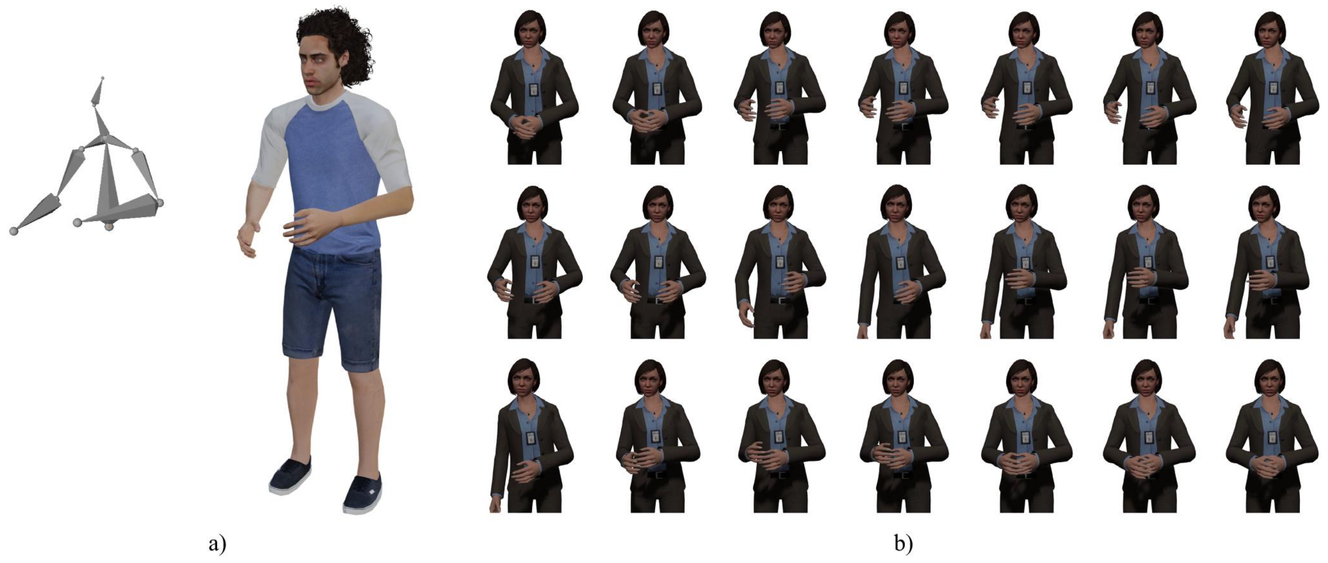

FIGURE 7 Examples of (a) retargeting animation of the skeletons to a rigged 3D human model and (b) some frames from the rendered videos.

5 | CONCLUSIONS

In this work, we introduce dynamic dictionaries to neural models for speech-driven gesture generation by coupling memory networks with encoder-decoder generative networks, allowing the pose generation network to utilise pre-stored pose features in the dictionary network for better diversity of the generated gestures. It solves the problem that gestures generated by neural methods are averaged and with an inadequate amount of movement shift due to their training schemes,

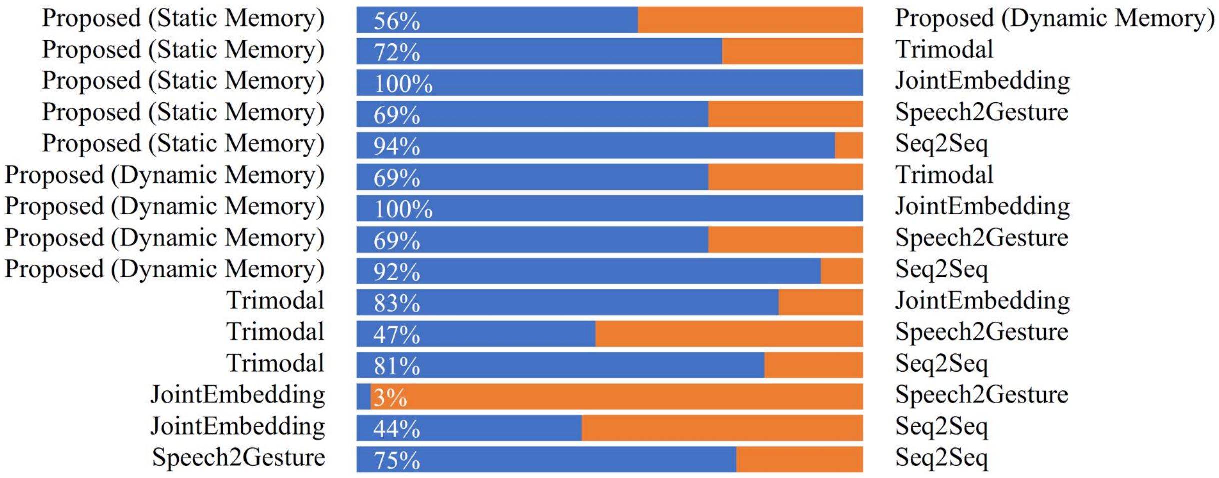

FIGURE 8 Results of the pairwise comparison on diversity between models by human evaluators. For example, the first line shows that 56% of the evaluators think the gestures generated by "Proposed (Static Memory)" have better diversity than those generated by "Proposed (Dynamic Memory)".

which can be described as slow or dull. By breaking decoder's sole dependency on the learnt encoder in the pose generation network, the model we designed successfully generates gestures with improved diversity. To demonstrate this, we propose a new objective metric for the evaluation of gesture diversity that utilises velocity direction changing angles to capture the occurrence of sudden changes in the gesture sequences, which is calibrated by a score of synchronisation to the ground truth. Quantitative experiments are conducted with the baseline and proposed models in different configurations. The results are compared in metrics with different settings showing that the proposed method can achieve best performances in diversity. User studies on the diversity improvements are also performed as qualitative evaluations to further guarantee those improvements are valid to potential human users.

As mentioned in Ref. [1], the model performs better in static memory configuration than in dynamic memory configuration, which is still the case in this work. However, we noticed that the diversity metric used in that work is far too sensitive, and further tests with different threshold settings are performed to make it better matched with human evaluation. In the future, we will conduct more experiments to see if the better performance is coming from the nature of the static memory configuration since it does provide richer and more specific keys than the dynamic one for different queries. Besides, this architecture of ours sees properties of the gesture data that can be decoupled into different types, such as rhythmic and sematic properties, as the same. It contains no structure specifically designed for these types that may improve the quality of generated gesture using hierarchical generation. Thus, we will also try new architectures of generation networks for better performance and look for greater uses of memory networks in other aspects of speech-driven gesture generation. Finally, this method cannot be directly used in some specific scenarios, such as real-world tasks, where training samples are lacking or extremely difficult to obtain. We will also look for ways to apply techniques like few-shot learning to our model for better generalisability and practicality.

ACKNOWLEDGEMENTS

This work was supported by the National Key R&D Programme of China (2022YFF0902202).

CONFLICT OF INTEREST STATEMENT

The author declares no conflict of interest.

DATA AVAILABILITY STATEMENT

The extended TED Gesture Dataset used in the experiments can be found in https://github.com/ai4r/Gesture-Generation-from-Trimodal-Context. Contact us for additional data if necessary.

ORCID

Zeyu Zhao https://orcid.org/0009-0002-6612-9731

REFERENCES

1. Zhao, Z., et al.: Generating diverse gestures from speech using memory networks as dynamic dictionaries. In: 2022 International Conference on Culture-Oriented Science and Technology (CoST), pp. 163–168 (2022)
2. Yoon, Y., et al.: Speech gesture generation from the trimodal context of text, audio, and speaker identity. ACM Trans. Graph. 39(6), 1–16 (2020). Available from. https://doi.org/10.1145/3414685.3417838
3. Yoo, S., et al.: Coloring with limited data: few-shot colorization via memory augmented networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, CA, USA, pp. 11283–11292 (2019)
4. Ginosar, S., et al.: Learning individual styles of conversational gesture. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, CA, USA, pp. 3497–3506 (2019)
5. Yoon, Y., et al.: Robots learn social skills: end-to-end learning of co-speech gesture generation for humanoid robots. In: 2019 International Conference on Robotics and Automation (ICRA), pp. 4303–4309. IEEE, Montreal (2019)
6. Ahuja, C., Morency, L.P.: Language2pose: natural language grounded pose forecasting. In: 2019 International Conference on 3D Vision (3DV), pp. 719–728. IEEE (2019)
7. Wang, W., et al.: MiniLM: deep self-attention distillation for task-agnostic compression of pre-trained transformers. In: Larochelle, H., et al. (eds.) Advances in Neural Information Processing Systems, vol. 33, pp. 5776–5788. Curran Associates, Inc. (2020). https://proceedings.neurips.cc/paper/2020/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

8. Devlin, J., et al.: BERT: pre-training of deep bidirectional transformers for language understanding. In: Burstein, J., Doran, C., Solorio, T. (eds.) Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 4171–4186. Association for Computational Linguistics, Minneapolis (2019). Available from. https://doi.org/10.18653/v1/n19-1423
9. Kullback, S., Leibler, R.A.: On information and sufficiency. Ann. Math. Stat. 22(1), 79–86 (1951). https://doi.org/10.1214/aoms/1177729694
10. Kaiser, L., et al.: Learning to remember rare events. In: 5th International Conference on Learning Representations, ICLR. Toulon, France, pp. 1–10. OpenReview.net (2017). https://openreview.net/forum?id=SJTQLdqlg
11. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3431–3440. IEEE Computer Society, Los Alamitos, CA, USA (2015). https://doi.ieeecomputersociety.org/10.1109/CVPR.2015.7298965
12. Ronneberger, O., Fischer, P., Brox, T.: U-net: convolutional networks for biomedical image segmentation. In: Navab, N., et al. (eds.) Medical Image Computing and Computer-Assisted Intervention - MICCAI 2015, pp. 234–241. Springer International Publishing, Munich (2015)
13. Gong, Y., Chung, Y.A., Glass, J.: AST: audio spectrogram transformer. Proc. Interspeech 2021, 571–575 (2021)
14. Chen, J., et al. TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation. arXiv preprint arXiv:210204306. 1–13 (2021)
15. Goodfellow, I., et al.: Generative adversarial nets. In: Ghahramani, Z., et al. (eds.) Advances in Neural Information Processing Systems, vol. 27, pp. 1–9. Curran Associates, Inc., Montreal (2014). https://proceedings.neurips.cc/paper/2014/file/5ca3e9b122f61f8f06494c97b1afccf3-Paper.pdf
16. Cao, Z., et al.: OpenPose: realtime multi-person 2D pose estimation using Part Affinity fields. IEEE Trans. Pattern Anal. Mach. Intell., 1–14 (2019)
17. Pavllo, D., et al.: 3D human pose estimation in video with temporal convolutions and semi-supervised training. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7753–7762. Long Beach, CA, USA (2019)
18. Ochshorn, R., Hawkins, M.: Gentle: A Forced Aligner [Internet] (2016). [cited 2022 May 22]. https://lowerquality.com/gentle/
19. Hunter, J.D.: Matplotlib: a 2D graphics environment. Comput. Sci. Eng. 9(3), 90–95 (2007). https://doi.org/10.1109/mcse.2007.55
20. Blender Online Community: Blender - a 3D Modeling and Rendering Package [Internet] (1994). [cited 2022 Sep 19]. https://www.blender.org/
21. Games, R.: Grand Theft Auto V [Internet] (2013). [cited 2022 Sep 19]. https://www.rockstargames.com/gta-v
22. Dexyfex. CodeWalker GTA V 3D Map + Editor [Internet]. (2017) [cited 2022 Sep 19]. https://www.gta5-mods.com/tools/codewalker-gtav-interactive-3d-map
23. Wagner, P., Malisz, Z., Kopp, S.: Gesture and speech in interaction: an overview. Speech Commun. 57, 209–232 (2014). Available from:. 10.1016/j.specom.2013.09.008 https://www.sciencedirect.com/science/article/pii/S0167639313001295
24. Marsella, S., et al.: Virtual character performance from speech. In: Proceedings of the 12th ACM SIGGRAPH/Eurographics Symposium on Computer Animation. SCA '13, pp. 25–35. Association for Computing Machinery, New York (2013). Available from. https://doi.org/10.1145/2485895.2485900
25. Softbank, R.: NAOqi API Documentation [Internet] (2018). [cited 2022 May 22]. http://doc.aldebaran.com/2-5/index/dev/guide.html
26. Levine, S., et al.: Gesture controllers. ACM Trans. Graph. 29(4), 1–11 (2010). Available from. https://doi.org/10.1145/1778765.1778861
27. Chiu, C.C., Morency, L.P., Marsella, S.: Predicting Co-verbal gestures: a deep and temporal modeling approach. In: Brinkman, W.P., Broekens, J., Heylen, D. (eds.) Intelligent Virtual Agents, pp. 152–166. Springer International Publishing, Delft (2015)
28. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. In: Bengio, Y., LeCun, Y. (eds.) 2nd International Conference on Learning Representations, ICLR. Banff, AB, pp. 1–14. Canada (2014). http://arxiv.org/abs/1312.6114
29. Li, J., et al.: Audio2Gestures: generating diverse gestures from speech audio with conditional variational autoencoders. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 11293–11302 (2021)
30. Xu, J., et al.: Freeform body motion generation from speech arXiv preprint arXiv:220302291. 1–10 (2022)
31. Liu, X., et al.: Learning hierarchical cross-modal association for Co-speech gesture generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10462–10472 (2022)
32. Graves, A., Wayne, G., Danihelka, I.: Neural Turing Machines. pp. 1–26. arXiv preprint arXiv:14105401 (2014)
33. Weston, J., Chopra, S., Bordes, A.: Memory networks. CoRR 3916, 1–15 (2015). abs/1410
34. Chunseong Park, C., Kim, B., Kim, G.: Attend to you: personalized image captioning with context sequence memory networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern recognition. Honolulu, HI, pp. 895–903 (2017)
35. Kim, Y., Kim, M., Kim, G.: Memorization precedes generation: learning unsupervised GANs with memory networks. In: 6th International Conference on Learning Representations, ICLR, pp. 1–15. OpenReview.net, Vancouver (2018). https://openreview.net/forum?id=rkO3uTkAZ
36. Dosovitskiy, A., et al.: An image is worth 16x16 words: transformers for image recognition at scale. CoRR (2020). abs/2010.11929. https://arxiv.org/abs/2010.11929
37. Touvron, H., et al.: Training data-efficient image transformers distillation through attention. In: International Conference on Machine Learning, pp. 10347–10357. PMLR (2021)
38. Ao, T., et al.: Rhythmic gesticulator: rhythm-aware co-speech gesture synthesis with hierarchical neural embeddings. ACM Trans. Graph. 41(6), 1–19 (2022). https://doi.org/10.1145/3550454.3555435
39. Ye, S., et al.: Audio-driven stylized gesture generation with flow-based model. In: Computer Vision-ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part V, pp. 712–728. Springer (2022)
40. Zhu, L., et al.: Taming Diffusion Models for Audio-Driven Co-speech Gesture Generation (2023). arXiv preprint arXiv:230309119
41. Ioffe, S., Szegedy, C.: Batch normalization: accelerating deep network training by reducing internal covariate shift. In: International Conference on Machine Learning, pp. 448–456. pmlr (2015)
42. Maas, A.L., et al.: Rectifier nonlinearities improve neural network acoustic models. In: Proc. Icml. Vol. 30. Atlanta, Georgia, USA, pp. 3 (2013)
43. Sheik-Ali, S., Sheik-Ali, S., Sheik-Ali, A.: Hearing Impairment and Introduction of Mandatory Face Masks. SAGE Publications Sage CA, Los Angeles (2023)

How to cite this article: Zhao, Z., et al.: Improving Diversity of Speech-Driven Gesture Generation with Memory Networks as Dynamic Dictionaries. CAAI Trans. Intell. Technol. 1–15 (2024). https://doi.org/10.1049/cit2.12321