# Self-Supervised Learning of Deviation in Latent Representation for Co-speech Gesture Video Generation

Huan Yang Jiahui Chen Chaofan Ding Runhua Shi Siyu Xiong Qingqi Hong
Xiaoqi Mo Xinhan Di
Giant Interactive Group Inc & AI Lab Xiamen University

## Abstract

*Gestures are pivotal in enhancing co-speech communication, while recent works have mostly focused on point-level motion transformation or fully supervised motion representations through data-driven approaches, we explore the representation of gestures in co-speech, with a focus on self-supervised representation and pixel-level motion deviation, utilizing a diffusion model which incorporates latent motion features. Our approach leverages self-supervised deviation in latent representation to facilitate hand gestures generation, which are crucial for generating realistic gesture videos. Results of our first experiment demonstrate that our method enhances the quality of generated videos, with an improvement from 2.7 to 4.5% for FGD, DIV and FVD, and 8.1% for PSNR, 2.5% for SSIM over the current state-of-the-art methods.*

## 1. Introduction

Co-speech gestures, a fundamental aspect of human communication [16], convey information in tandem with speech. These gestures effectively transmit social cues [10], including personality, emotion, and subtext. However, current conditional video generation methods [3, 7, 12, 17, 18] often produce motion videos that lack realism and fine-grained details. Additionally, these networks typically require substantial computational resources [1, 5, 19] and full supervision of motion annotation [4] with high labour cost.

To address these challenges, we propose a novel approach for generating co-speech gesture videos. First, a deviation module is proposed to generate latent representation of both foreground and background. This deviation module is consisted of latent deviation extractor, a warping calculator and a latent deviation decoder. Second, a corresponding self-supervised learning strategy is proposed to achieve the latent representation of deviation for high-quality video generation. In the evaluation of our first experiment, the results demonstrate high-quality co-speech video generation and outperform the state-of-the-art models in the evaluation of video quality.

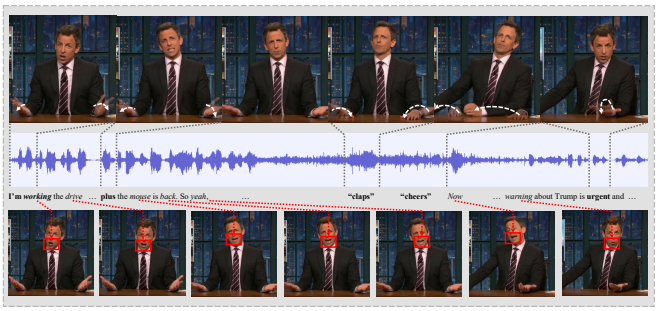


Figure 1. Examples of our generated gesture videos. White dashed arrows indicate gestures corresponding to bold words. The red dotted boxes indicate the mouth shapes corresponding to the italicized words.

## 2. Method

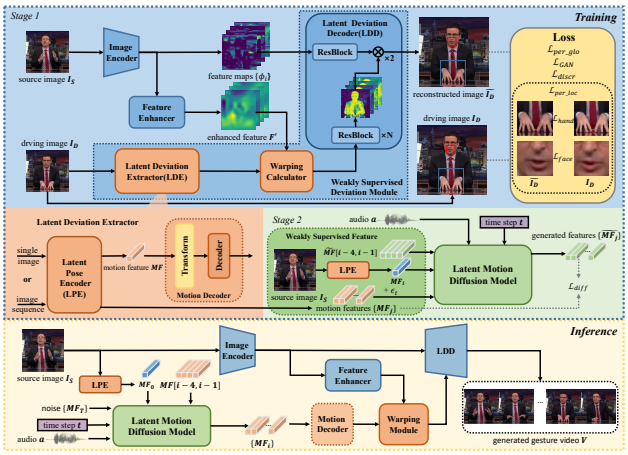


Figure 2. Co-speech gesture video generation pipeline of our proposed method consists of three main components: 1) the latent deviation extractor (orange) 2) the latent deviation decoder (blue) 3) the latent motion diffusion (green).

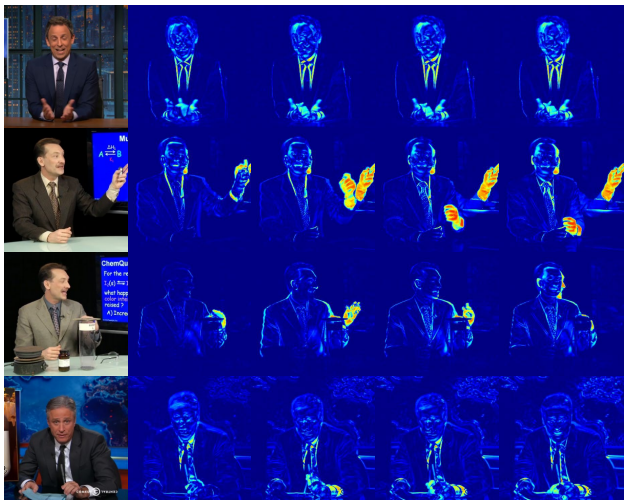


Figure 3. The deviation in latent representation.

We propose a novel method for generating co-speech gesture videos, utilizing a self-supervised full scene deviation, produces co-speech gesture video $V$ (i.e., image sequence) that exhibit natural poses and synchronized movements. The generation process takes as input the speaker's speech audio $a$ and a source image $I_S$. An overview of our model is shown in Figure 2.

We structured the training process into two stages. In the first stage, a driving image $I_D$ and a source image $I_S$ are used to train the base model. In one aspect, the proposed latent deviation module consisting of latent deviation extractor, warping calculator and latent deviation decoder is trained under self-supervision. In another aspect, other modules in the base model is trained under full supervision. In the second stage, the self-supervised motion features, consisting of $MF_i$, $\widetilde{MF}_{[}i-4, i-1]$, and the noise-added motion feature sequence $\{MF_j\}$, are used to train the latent motion diffusion model. In the following parts, we will introduce the two stages of training part and the inference part in detail.

## 2.1. Stage 1: Base Model Learning

### 2.1.1 Image encode.

First, we input the source image $I_S \in \mathbb{R}^{H\times W\times 3}$ into the image encoder $\mathcal{E}$ to obtain the feature $F$ and feature maps $\{\phi_{i\in N}\}$, as the layer-by-layer feature maps provide additional background details for subsequent steps.

$$F, \{\phi_{i\in N}\} = \mathcal{E}(I_S). \tag{1}$$

### 2.1.2 Feature Enhancer.

Directly inputting shallow spatial features $F$ into the warping calculator increases the training difficulty of LME and makes it more likely to decode blurred images with artifacts after applying optical flow transformations. Therefore, we use a feature enhancer to map the shallow spatial features to a higher-level space, achieving better results. The equation is as follows

$$F' = \frac{F-\bar{F}}{\sqrt{\sigma^2+\epsilon}}\gamma + \beta, \tag{2}$$

where $\bar{F}$ and $\sigma$ represent the mean and standard deviation of the features, while $\gamma$ is the scaling factor, and $\beta$ is the bias.

### 2.1.3 Self-Supervised Deviation Module.

We proposed a self-supervised deviation module which is consisted of three parts, latent deviation extractor, warping calculator and latent deviation decoder. The calculation is represented in the following.

**Latent Deviation Extractor.** To generate natural motion-transformed images, we first design latent pose estimation module (LPE) to extract a latent motion feature $MF \in \mathbb{R}^{1\times K}$. This latent motion feature then undergoes a non-linear pose transformation operation before being decoded. The compact representation of pose transformations within the latent space enables the generation of concise optical flow, which effectively drives the image.

$$\mathbf{V} = \psi(\mathcal{T}(LPE(I))). \tag{3}$$

**Warping Calculator.** To effectively integrate motion information into the source image, we first input the source image $I_S$ into the encoder, enhance the feature maps, and then warp $\mathcal{W}(\cdot)$ each enhanced feature using the rotation $\mathbf{R}$ and translation $\mathbf{T}$ matrices of optical flow $\mathbf{V}$ to obtain the deformed features $F'_{\mathcal{W}}$. This feature enhancement process amplifies key features while suppressing background noise, leading to a clearer representation of critical information in the image. The warping formula is as follows:

$$F'_{\mathcal{W}} = \mathcal{W}(\mathbf{R}, \mathbf{T}, F'). \tag{4}$$

**Latent Deviation Decoder.** Due to the occlusions and misalignments between $I_S$ and $I_D$, directly decoding after the warping operation often fails to achieve effective image reconstruction. Inspired by [4, 20], we add an full scene deviation $\delta_F$ into the decoder during the decoding process, which improves the accuracy of the reconstructed image $\widetilde{I}_D$. The full scene deviation formula is as follows:

$$\delta_F = L\frac{1}{1+e^{-(wF'_{\mathcal{W}}+b)}}, \tag{5}$$


Figure 4. Visual comparison with SOTAs. Our method generates gestures with more extensive accurate motions (dashed boxes), matching audio and semantics. Red boxes indicate unrealistic gestures generated by ANGIE [9], MM-Diffusion [12] and S2G-MDDiffusion [6].

which is a variant of the sigmoid function. Next, we interpolate and decode the deviation $\delta_F$ together with the feature $F$. This approach allows the motion features to be smoothly interpolated onto the source image features, resulting in more natural outcomes. The decoding process is as follows:

$$\mathbf{z} = \delta_F F + (1 - \delta_F) U(F), \tag{6}$$

where $U(\cdot)$ denotes image decoder. Moreover, we design a nonlinear activation function that can adjust the impact of negative input values. By tuning the parameter $c_\lambda$, we can control the contribution of negative input values to the output.

$$\mathbf{a} = max(0, \mathbf{z}) + c_\lambda min(0, \mathbf{z}). \tag{7}$$

**Training.** In the training process of co-speech gesture video generation, unlike vlogger [4] which applies a large-scale of annotation data for supervised dense representation of motion, we directly apply image reconstruction loss without dense/sparse supervised annotation of motion [2]. Following [2, 13, 14], we utilize a pre-trained VGG-19 [8] network to calculate the reconstruction global loss $\mathcal{L}_{per\_glo}$ between the reconstructed image $I_D$ and the generated image $\widetilde{I}_D$ across multiple resolutions:

$$\mathcal{L}_{per\_glo} = \sum_j \sum_i c_i |V_i(I_{Dj}) - V_i(\widetilde{I}_{Dj})|, \tag{8}$$

where $V_i$ is the $i$-th layer of the pre-trained VGG-19 network, and $j$ represents that the image is downsampled $j$ times. Additionally, we calculated the local loss $\mathcal{L}_{per\_loc}$ of the reconstructed image, which consists of hand loss $\mathcal{L}_{hand}$


Figure 5. Visualization results of fine-grained hand variations. The gesture videos we generate are clearer, more reasonable, more diverse and more natural in the same frame.

and face loss $\mathcal{L}_{face}$, based on the VGG-16 model:

$$\mathcal{L}_{per} = \lambda_{per\_glo}\mathcal{L}_{per\_glo} + \lambda_{per\_loc}\mathcal{L}_{per\_loc}. \tag{9}$$

To ensure realism, we employed a patch-based discriminator [11] and trained it using the GAN discriminator loss $\mathcal{L}_{discr}$ for adversarial training. Both the ground truth images and the generated fine images are converted into feature maps, where each element is classified as real or fake. Therefore, the final loss is the following:

$$\mathcal{L}_1 = \lambda_{per}\mathcal{L}_{per} + \lambda_{GAN}\mathcal{L}_{GAN} + \lambda_{discr}\mathcal{L}_{discr}. \tag{10}$$

## 2.2. Stage 2: Latent Motion Diffusion

In this stage, we adopted the latent motion diffusion model proposed by [6] and made modifications to it. Our diffusion model takes five inputs: the time step $t$, audio $a$, the noisy motion feature sequence $\{MF_{j\in M}\}$ , the motion feature $MF_i$ of the source image, and the predicted motion features $\widetilde{MF}_{[i-4,i-1]}$ from the previous four frames. The model then predicts a clean motion feature sequence $\{\widetilde{MF}_{i\in M}\}$ from noised $\{MF_{j\in M}\} + \epsilon_t$ and condition $c = (a, \widetilde{MF}_{[i-4,i-1]}, MF_i)$.

### 2.2.1 Feature Priors and Loss.

Both self-supervised deviation feature and fully-supervised feature are applied. We divide the loss into three components. The first is a motion feature loss calculated with MSE, which constrains the movements of the hands, lips, and head, promoting naturalness and coherence. Additionally, the loss includes implicit velocity and implicit acceleration losses [15] to prevent the overall motion from being too rapid. The final training loss is as follows:

$$\mathcal{L}_{diff} = \mathcal{L}_{MF} + \lambda_{im\_vel}\mathcal{L}_{imp\_vel} + \lambda_{imp\_acc}\mathcal{L}_{imp\_acc}. \tag{11}$$

| Name | Objective evaluation | | | Subjective evaluation | | | |
|---|---|---|---|---|---|---|---|
| | FGD ↓ | Div. ↑ | FVD ↓ | Realness ↑ | Diversity ↑ | Synchrony ↑ | Overall quality ↑ |
| Ground Truth (GT) | 8.976 | 5.911 | 1852.86 | 4.70± 0.07 | 4.45 ± 0.05 | 4.66 ± 0.08 | 4.70 ± 0.07 |
| ANGIE | 55.655 | 5.089 | 2965.29 | 2.01±0.07 | 2.38±0.07 | 2.11±0.07 | 1.98±0.08 |
| MM-Diffusion | 41.626 | 5.189 | 2656.06 | 1.63±0.07 | 1.93±0.08 | 1.57±0.09 | 1.44±0.08 |
| S2G-MDDiffusion | <u>18.131</u> | <u>5.632</u> | <u>2058.19</u> | <u>2.78±0.07</u> | <u>3.21±0.08</u> | <u>3.32±0.08</u> | <u>2.89±0.07</u> |
| Ours | **17.324** | **5.788** | **1997.96** | **3.82±0.07** | **3.94±0.07** | **4.11±0.06** | **3.93±0.05** |

Table 1. Quantitative results on test set. Bold indicates the best and underline indicates the second. For ANGIE [9] and MM-Diffusion [12], we cited the results in the S2G-MDDiffusion [6] paper. For S2G-MDDiffusion, we used the official open source code.

| Name | Hand gesture | | | Lip movement | | | Full image | | |
|---|---|---|---|---|---|---|---|---|---|
| | PSNR ↑ | SSIM ↑ | LPIPS ↓ | PSNR ↑ | SSIM ↑ | LPIPS ↓ | PSNR ↑ | SSIM ↑ | LPIPS ↓ |
| S2G-MDDiffusion | 22.37 | 0.625 | 0.106 | 23.84 | 0.689 | 0.092 | 29.39 | 0.952 | 0.034 |
| Ours | **23.91** | **0.756** | **0.054** | **29.10** | **0.882** | **0.038** | **31.79** | **0.976** | **0.018** |

Table 2. Comparison of gestures and mouth movements based on common image quality metrics.

| Name | Objective evaluation | | | Subjective evaluation | | | |
|---|---|---|---|---|---|---|---|
| | FGD ↓ | Div. ↑ | FVD ↓ | Realness ↑ | Diversity ↑ | Synchrony ↑ | Overall quality ↑ |
| w/o Dev. in LDD | **17.218** | 4.980 | <u>2112.81</u> | 3.03 ± 0.08 | 3.62 ± 0.08 | 3.62 ± 0.07 | 3.29 ± 0.08 |
| w/o $F'$ | <u>17.284</u> | <u>5.703</u> | 2252.88 | 3.62 ± 0.06 | 3.81 ± 0.09 | 3.80 ± 0.08 | 3.67 ± 0.08 |
| w/o Mo Dec. | 17.812 | 5.191 | 2201.66 | <u>3.66 ± 0.07</u> | <u>3.85 ± 0.09</u> | <u>3.86 ± 0.06</u> | <u>3.77 ± 0.06</u> |
| Ours | 17.324 | **5.788** | **1997.96** | **3.87±0.07** | **3.86±0.08** | **4.08±0.07** | **3.86±0.08** |

Table 3. Ablation study results. Bold indicates the best and underline indicates the second. 'w/o' is short for 'without'.

# 3. Experiments and Results

## 3.1. Dataset and Evaluation metrics.

Our experimental data comes from the PATS dataset [1],following the standard process from S2G-MDDiffusion [6], we conduct the first training stage to generate the gesture videos and compare with the state-of0the-art models. We employed 1) **Fréchet Gesture Distance (FGD)** 2) **Fréchet Video Distance (FVD)** and 3) **Diversity (Div.)** to assess quality of gesture videos [6]. Besides, we also apply PSNR, SSIM and LIPS [4] for the evaluation of generated hand gesture, lip movement and full scene.

## 3.2. Evaluation on Results(First Stage)

We compared our method with: 1) gesture video generation methods ANGIE [9] and S2G-MDDiffusion [6], and 2) MM-Diffusion [12].

In Figure 4, Our method(first stage) mitigate abnormal deformations in comparison with the SOTA models. In Figure 5 and Figure **??**, we also compare generation of hand gesture, facial expression and lip movements, our method also improves the quality of these parts. In detail, the proposed method mitigates jitter, blurred region(hands/faces/lips) and other abnormal deformations.

The quantitative results are shown in Table 1 and 2. Our method outperforms existing approaches in terms of FGD, Diversity, and FVD metrics, PSNR, SSIM and LIPS. Compared to S2G-MDDiffusion, our method achieved a 4.45% and 2.93% reduction in FGD and FVD, respectively, and a 2.77% increase in DIV. In terms of image quality, our method outperforms S2G-MDDiffusion across the metrics for hands, lips, and the entire image for PSNR(6.88%-22.06%) and for SSIM(2.52%-28.01%).

## 3.3. User Study and Ablation Study

We conducted a user study inspired by S2G-MDDiffusion [6]. Twenty participants were invited to provide Mean Opinion Scores (MOS) across four dimensions, including 1) **Realness**, 2) **Diversity** of gesture, 3) **Synchrony** between speech and gestures, and 4) **Overall quality**. Besides, We examined the effectiveness of the following components: 1) the deviation in the LDD, 2) the enhanced feature, and 3) the motion decoder in Table 3.

# 4. Discussion

We proposed the self-supervised learning of deviation in co-speech gesture video generation. The results(first stage) demonstrates improvement of the quality of gesture videos. We are conducting the second stage of the experiment.

## References


[1] Chaitanya Ahuja, Dong Won Lee, Yukiko I. Nakano, and Louis-Philippe Morency. Style transfer for co-speech gesture animation: A multi-speaker conditional-mixture approach. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XVIII*, volume 12363 of *Lecture Notes in Computer Science*, pages 248–265. Springer, 2020. 1, 4

[2] Goutam Bhat, Michaël Gharbi, Jiawen Chen, Luc Van Gool, and Zhihao Xia. Self-supervised burst super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10605–10614, 2023. 3

[3] Duygu Ceylan, Chun-Hao P Huang, and Niloy J Mitra. Pix2video: Video editing using image diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23206–23217, 2023. 1

[4] Enric Corona, Andrei Zanfir, Eduard Gabriel Bazavan, Nikos Kolotouros, Thiemo Alldieck, and Cristian Sminchisescu. Vlogger: Multimodal diffusion for embodied avatar synthesis. *arXiv preprint arXiv:2403.08764*, 2024. 1, 2, 3, 4

[5] Shiry Ginosar, Amir Bar, Gefen Kohavi, Caroline Chan, Andrew Owens, and Jitendra Malik. Learning individual styles of conversational gesture. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun 2019. 1

[6] Xu He, Qiaochu Huang, Zhensong Zhang, Zhiwei Lin, Zhiyong Wu, Sicheng Yang, Minglei Li, Zhiyi Chen, Songcen Xu, and Xiaofei Wu. Co-speech gesture video generation via motion-decoupled diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2263–2273, 2024. 3, 4

[7] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022. 1

[8] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 694–711. Springer, 2016. 3

[9] Xian Liu, Qianyi Wu, Hang Zhou, Yuanqi Du, Wayne Wu, Dahua Lin, and Ziwei Liu. Audio-driven co-speech gesture video generation. *Advances in Neural Information Processing Systems*, 2022. 3, 4

[10] Miriam Novack and Susan Goldin-Meadow. Learning from gesture: How our hands change our minds. *Educational psychology review*, 27:405–412, 09 2015. 1

[11] Weize Quan, Ruisong Zhang, Yong Zhang, Zhifeng Li, Jue Wang, and Dong-Ming Yan. Image inpainting with local and global refinement. *IEEE Transactions on Image Processing*, 31:2405–2420, 2022. 3

[12] Ludan Ruan, Yiyang Ma, Huan Yang, Huiguo He, Bei Liu, Jianlong Fu, Nicholas Jing Yuan, Qin Jin, and Baining Guo. Mm-diffusion: Learning multi-modal diffusion models for joint audio and video generation. In *CVPR*, 2023. 1, 3, 4

[13] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. First order motion model for image animation. In *Conference on Neural Information Processing Systems (NeurIPS)*, December 2019. 3

[14] Aliaksandr Siarohin, Oliver J. Woodford, Jian Ren, Menglei Chai, and Sergey Tulyakov. Motion representations for articulated animation. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun 2021. 3

[15] Li Siyao, Weijiang Yu, Tianpei Gu, Chunze Lin, Quan Wang, Chen Qian, Chen Change Loy, and Ziwei Liu. Bailando: 3d dance generation by actor-critic gpt with choreographic memory. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11040–11049, 2022. 3

[16] Barbara Tversky. Communicating with diagrams and gestures. *Research trends in science, technology, and mathematics education*, 2, 2007. 1

[17] Tan Wang, Linjie Li, Kevin Lin, Yuanhao Zhai, Chung-Ching Lin, Zhengyuan Yang, Hanwang Zhang, Zicheng Liu, and Lijuan Wang. Disco: Disentangled control for realistic human dance generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9326–9336, June 2024. 1

[18] Wilson Yan, Yunzhi Zhang, Pieter Abbeel, and Aravind Srinivas. Videogpt: Video generation using VQ-VAE and transformers. *CoRR*, abs/2104.10157, 2021. 1

[19] Youngwoo Yoon, Bok Cha, Joo-Haeng Lee, Minsu Jang, Jaeyeon Lee, Jaehong Kim, and Geehyuk Lee. Speech gesture generation from the trimodal context of text, audio, and speaker identity. *ACM Transactions on Graphics*, page 1–16, Dec 2020. 1

[20] Jian Zhao and Hui Zhang. Thin-plate spline motion model for image animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3657–3666, 2022. 2