Proceedings of CONF-SPML 2026 Symposium: The 2nd Neural Computing and Applications Workshop 2025
DOI: 10.54254/2755-2721/2026.TJ29059

Human Head Animation Generation Methods Driven by Multimodality

Peiyu Tsai

College of Information Science and Technology, Jinan University, Guangzhou, China
caipeiyu@stu2022.jnu.edu.cn

Abstract. With the rapid development of multimodal human-computer interaction, generating high-fidelity, emotionally rich, and naturally coordinated human head animation based on multi-source inputs such as language, images, and text has become a core issue in virtual human research. This article systematically reviews the representative methods in this field in the past five years, and conducts a classified analysis around Transformer-like sequence modeling, the gradual generation mechanism of diffusion models, the NeRF implicit three-dimensional modeling path, and other specialized architectures. Through vertical technological evolution and horizontal performance comparison, the advantages and bottlenecks of various methods in semantic understanding, emotion driving, perspective consistency, and style expression are revealed. It is further pointed out that the fine-grained control of the emotion-driven mechanism, the balance between 3D modeling efficiency and authenticity, and the high degree of customization of personalized appearance and behavioral style will determine the expressive boundaries of the virtual human image. Future research needs to continue to make breakthroughs in cross-modal feature fusion, semantic consistency modeling, and long sequence generation stability, so as to build a virtual human system with human-like interaction capabilities and provide theoretical support and technical reference for application scenarios such as digital human social interaction, education, and entertainment.

Keywords: multimodal human-computer interaction, human head animation generation, Transformer model, diffusion model, NeRF modeling

1. Introduction

Human head animation generation is a core component of virtual human technology, and its quality directly impacts the realism and expressiveness of interactions. This task relies on multimodal input signals such as speech and text, and requires the coordinated processing of key issues such as semantic understanding, emotion calculation, expression generation, and multi-viewpoint consistency.

With the development of deep learning, multimodal modeling technology has made significant progress. Researchers have begun to introduce the Transformer structure to enhance the temporal correlation between speech and face. For example, the EmoFace architecture proposed by Chang et al. builds a Chinese multi-emotion dataset based on the MetaHuman controller to achieve high-

DOI: 10.54254/2755-2721/2026.TJ29059

quality mapping from speech to expression [1]. Diffusion models are also gradually being applied to face generation tasks. For example, the HunyuanVideo-Avatar launched by Tencent Hunyuan integrates image injection and audio emotion mapping modules based on the Multimodal Diffusion Transformer (MM-DiT) architecture to achieve highly dynamic and emotionally consistent virtual human animation, supports multi-character audio driving, and performs well on multiple datasets [2]. In addition, Neural Radiance Fields (NeRF) technology has also been introduced into the field of human head modeling. The DreamWaltz framework proposed by Huang et al. combines the Skinned Multi-Person Linear model (SMPL) skeleton and the diffusion model posture before achieve text-driven 3D virtual human animation without mesh binding, showing high realism and coordination in multi-role interaction and scene expansion [3].In addition to the above mainstream paths, some studies have explored lightweight network architectures based on CNN or Gaussian Splatting.

This paper reviews multimodal driven human head animation generation methods, focusing on key technologies such as Transformer, diffusion model, and NeRF, comparing their performance differences in semantic understanding, expression generation, and three-dimensional consistency, summarizing challenges such as emotion coordination, cross-language adaptation, and data scarcity, and looking forward to future research directions to promote the development of human-like interactive virtual human systems.

2. Transformer category method

As a sequence modeling architecture based on a self-attention mechanism, the Transformer excels at capturing long-range dependencies and semantic associations in multimodal inputs. In the generation of human head animations, the Transformer often processes audio, text, or image features in an encoder-decoder format. Using a multi-head attention mechanism, it simultaneously focuses on lip movements, facial expressions, and head posture, achieving highly natural and emotionally consistent animation output.

Chang et al. proposed the EmoFace architecture, built a Chinese multi-emotion dataset for the MetaHuman controller device, and extracted per-frame controller parameters to train an audio-driven facial animation model [1]. The model supports flexible emotion control, generates natural and coordinated expressions, and enhances dynamic expressiveness by integrating blink and gaze controllers. In quantitative evaluations, EmoFace achieved a low mean absolute value (MAE) of 0.04024 in expression prediction tasks, a naturalness score of 5.514/7 in user studies, and a recognition accuracy of 0.805/1, close to the true label level. Its overall performance was better than FaceFormer and EmoTalk, demonstrating significant advantages in the synthesis of multi-emotional virtual human animations.

Drobyshev et al. proposed the EMO Portraits model, which achieved high-fidelity generation of strong expressions under low-domain data conditions by constructing a latent expression space and an innovative loss mechanism [4]. The model supports both speech and image-driven modes, enabling natural generation of lip movements, head rotations, and eye blinks. The accompanying FEED dataset covers extreme expressions and multi-viewpoint motion, enriching training and evaluation resources. In image-driven tasks, the model achieved an FID score of 59.6 and an expression transfer preference of 74.6%, significantly outperforming existing methods. In speech-driven mode, the video quality FID dropped to 28.5, with a lip sync offset of only 0.07, demonstrating excellent audio-visual coordination and dynamic reproduction.

Fei et al. introduced the EmpathyEar chatbot, which supports multimodal input to generate semantically and emotionally consistent virtual responses [5]. The system integrates a language

DOI: 10.54254/2755-2721/2026.TJ29059

model-driven semantic understanding module, a speech generator, and a talking digital portrait module to simultaneously generate emotionally rich speech and facial animation. From Chang's exploration of modeling refined 3D controllers driven by audio, to Drobyshev's implementation of cross-identity transfer and high-fidelity dynamic generation, to Fei's construction of a multimodal empathic response system integrating text, speech, and vision, the research focus of human head animation generation is gradually evolving from single voice input to multi-source information fusion and deep emotional expression. Related technologies continue to achieve breakthroughs in dynamic consistency, emotional controllability, and personalized synthesis, laying a solid foundation for the development of virtual human animation in terms of realism and expressiveness.

While the Transformer architecture excels in semantic modeling and emotional expression, such methods generally rely on high-quality annotated datasets, have limited adaptability to multiple languages, and face bottlenecks in dynamic consistency and real-time reasoning. Furthermore, the mapping accuracy of facial controller parameters is limited, making it difficult to capture subtle expressions and complex emotional changes, affecting the subtlety and naturalness of the generated animations.

Table 1 compares the differences between the three Transformer-type methods in terms of input modality, fusion strategies, and technical challenges. Table 2 compares the dataset constructed by EmoFace and the FEED dataset in terms of language, samples, emotions, and action dimensions.

Table 1. A horizontal data comparison table of the three representative transformer-like methods mentioned above

	EmoFace	EMOPortraits	EmpathyEar
Input Modal	Audio (Chinese)	Image + Audio	Text + Audio + Image
Fusion method	Content encoding +sentiment; encoding→ 512-dimensional vector	Expression latent vector z (dim=128 or 512)	Unified encoding of multimodal features → LLM reasoning
Challenges	Incomplete mapping, slow inference, small datasets, and weak details	No body generation, unstable expression, and distortion when turning the head	Non-end-to-end, cross-modal instability, and lack of evaluation standards

Table 2. Comparison between the new dataset built for EmoFace and the FEED

	Chinese Multi-Emotional Dataset	FEED
Language type	Chinese (Mandarin)	English(LibriSpeech pre-training + fine-tuning)
Sample size	Single actor material, limited quantity	23 subjects, 520 videos
Number of sentiment labels	7 categories (neutral, happy, sad, etc.)	7 standard emotions + a variety of non-traditional emotions
Expression Type	Basic Facial Expressions + Controller Parameters	Strong asymmetrical expressions, tongue-in-cheek movements, blinking, etc.
Action Dimension	Facial controller (174 dimensions)	Face + head movement + eye movement + gesture
Frame rate	60fps (interpolation alignment required)	Native multi-view video, frame rate unspecified

DOI: 10.54254/2755-2721/2026.TJ29059

3. Diffusion model methods

Diffusion models, with their powerful distributed modeling capabilities, are widely used in high-quality image and animation generation. Through a step-by-step denoising process, combined with audio features, emotion labels, and identity encoding, they accurately reproduce lip movements, blinks, head pose, and facial texture. Viewpoint embedding and 3D priors ensure seamless rendering from multiple viewpoints, while latent diffusion models optimize inference efficiency. Multimodal conditional injection enables flexible control of emotion and identity, while temporal regularization and contrastive learning enhance the naturalness and stability of long sequence generation.

Tencent Hunyuan launched HunyuanVideo-Avatar, which is based on the DiT architecture and integrates character image injection, audio emotion mapping, and facial perception adaptation modules to achieve highly dynamic and emotionally consistent virtual human animation [2]. This method supports multi-character audio driving, significantly improves the dynamics and consistency of the video, surpasses existing systems on multiple datasets, and provides a new path for immersive multi-character animation synthesis. Wang et al. proposed the EmotiveTalk method, which decouples audio and facial motion information to generate representations that accurately correspond to lip movements and facial expressions [6]. Combining a video diffusion framework with multi-source emotion control, this method achieves highly realistic and controllable speaker video generation. Experimental results show that this method significantly outperforms existing technologies in terms of emotion synchronization and dynamic representation.

From HunyuanVideo's emotion vector injection to EmotiveTalk's iterative emotion refinement and multi-source control, the diffusion model is moving from single-frame generation to a long-sequence, multimodal, and emotionally consistent, highly controllable synthesis path, significantly improving the realism and dynamics of virtual human expression.

While diffusion models offer advantages in detail generation and emotion control, their inference process is computationally expensive, making it difficult to meet the demands of real-time interaction. Furthermore, generating long sequences is prone to motion drift or emotional discontinuity. Especially in multi-character scenarios, semantic coordination and emotional consistency between characters remain technical challenges. Furthermore, the model's high reliance on the diversity and quality of training data limits its generalization capabilities in low-resource environments.

Table 3 compares the performance of HunyuanVideo and EmotiveTalk in multiple dimensions, including image quality, dynamic consistency, and voice synchronization. "↑" indicates that a larger value indicates better performance, while "↓" indicates that a smaller value indicates better performance.

EmotiveTalk's FID is lower than HunyuanVideo's, indicating that its generated images and videos are more realistic and natural. Its Sync-C score is higher, indicating better synchronization between speech and lip movements. In contrast, HunyuanVideo performs reliably in IQA and ASE, making it suitable for applications that prioritize image quality and expressive intensity. The overall distribution of these values reflects their different positioning in terms of generation accuracy and expressiveness.

DOI: 10.54254/2755-2721/2026.TJ29059

Table 3. Overall comparison of the two methods, where "↑" indicates that the larger the value, the better the performance, while "↓" indicates that the smaller the value, the better the performance

	HunyuanVideo	EmotiveTalk
FID↓	38.01/43.42	16.64/53.21
FVD↓	358.71/445.02	140.96/207.67
IQA↑	3.99/3.70	\
E-FID↓	\	0.54/0.57
Sync-C↑	5.30/4.92	8.24/6.82
Sync-D↓	\	7.09/7.43
ASE↑	2.54/2.52	\

4. Nerf category methods

NeRF (Neural Radiance Fields) is used to model implicit 3D structures, making it suitable for high-fidelity multi-view animation of human heads. This enables high-fidelity multi-view animation of human heads, rendering high-quality images from any angle and capturing previously invisible areas through perspective adjustment.

Huang et al. proposed DreamWaltz, a text-driven avatar generation framework that can create 3D avatars with complex appearance and can be animated [3]. This method combines the SMPL skeleton and diffusion model posture prior to achieve motion-driven animation without mesh binding. The system effectively avoids polyhedral artifacts and limb blur, supports arbitrary posture sequence driving, and demonstrates high realism and coordination in multi-role interaction and scene expansion. In the quantitative evaluation, the geometric structure and texture quality of the virtual image received 3.72 and 3.75 points (out of a maximum of 4), respectively. The rendering inferred a posed 3D image in each frame, maintaining a high degree of consistency. By learning pose-conditioned image priors, DreamWaltz can achieve animation effects of complex characters with significantly higher quality, providing an innovative path for the application of NeRF technology in virtual human generation.

HCI Survey proposed a multimodal human-computer interaction system framework, systematically sorted out the semantic drive and multimodal fusion paths in virtual human generation, and conducted a classification analysis of representative methods such as FaceFormer and AD-NeRF [7]. The review summarized common datasets and evaluation indicators, emphasized the importance of video generation efficiency in model performance evaluation, and provided theoretical support for the interactive design and application expansion of virtual human technology.

DreamWaltz integrates the SMPL skeleton and distillation strategy to support natural language-driven virtual human generation and animation rendering, addressing perspective and occlusion issues. The HCI Survey summarizes semantic-driven and multimodal interaction methods, revealing the evolution of NeRF technology toward semantic consistency and real-world integration, laying the foundation for immersive interaction.

Despite their impressive performance, NeRF-based approaches still face challenges such as low real-time rendering efficiency and unstable semantic control. Multi-role interaction and environmental adaptation also present scalability challenges, necessitating improvements in generation speed and semantic consistency to facilitate practical application.

DOI: 10.54254/2755-2721/2026.TJ29059

Table 4 provides a horizontal comparison of the aforementioned NeRF-based methods. DreamWaltz achieves high-precision 3D modeling by fusing NeRF with the SMPL skeleton, and introduces SDS distillation and density-weighted sampling mechanisms to improve consistency. However, it still faces problems such as high computing power consumption, limited resolution, and unstable semantic control.

Table 4. Is a horizontal data comparison table of the above NeRF methods

	3D modeling method	Consistency Mechanism	Challenges
Dream Waltz	NeRF + SMPL	SDS distillation+ density-weighted sampling	High computing power, risk of bias Limited resolution and efficiency Semantic control still needs optimization
HCI Survey	FaceFormer, AD-NeRF, etc.	Multimodal adaptation mechanism	The real-time performance of conversation video generation methods still needs to be improved.

5. Other method category

In addition to mainstream approaches, some studies have attempted to use specialized architectures to meet specific needs.

Arcelin et al. proposed Audio2Rig, a deep learning-based tool that can generate high-quality, stylized facial and lip-sync animations on Maya skeletons. This method combines audio-driven and emotional control, supports keyframe editing and emotional mixing, improves animation accuracy and artistic flexibility, and is suitable for studio-level automated production [8]. The system was trained on 9,471 frames of animation data provided by Stim Studio, taking approximately three hours. While the mouth and tongue animations performed well, the upper face still suffered from a lack of temporal coherence. Furthermore, the model's reliance on studio data limited its ability to generalize across projects, but it does have the potential to be extended to other creative platforms.

Peng et al. proposed SyncTalk++, which achieves high-fidelity, 101 frames per second speech-driven head video synthesis based on Gaussian Splatting, with high synchronization of identity, lip movement, expression, and head posture. The system introduces an expression generator and a torso restoration module to effectively improve adaptability to out-of-distribution speech, eliminate artifacts, and enhance visual quality [9]. Compared to existing methods, SyncTalk++ outperforms existing approaches in terms of synchronization, real-time performance, and deployment feasibility, making it suitable for interactive scenarios such as voice chat. Despite its excellent performance, the system still faces challenges such as complex attribute configuration, difficulty in identifying abnormal pixels, and the risk of deepfake abuse. Supporting detection mechanisms and ethical standards is urgently needed to ensure the safe application of this technology.

From Arcelin's emotion-driven skeletal animation to Peng's high-frame-rate 3D human head synthesis, voice animation technology is moving from structural control to a dual breakthrough in synchronization accuracy and visual realism, driving the evolution of virtual human generation towards high-quality, editable, and real-time interactive professional applications.

Specialized architectures demonstrate excellent performance in specific scenarios, but they still suffer from insufficient generalization capabilities when faced with diverse voice inputs, complex emotional expressions, or cross-domain applications.

DOI: 10.54254/2755-2721/2026.TJ29059

6. Conclusion

This article systematically reviews the research progress in speech-driven human head animation generation, focusing on four main approaches: the Transformer architecture excels in semantic understanding and emotional expression; the diffusion model excels in detail generation and multimodal control; NeRF technology, through implicit 3D modeling, enables multi-view rendering, driving the development of virtual humans towards immersive interaction; and specialized architectures such as SyncTalk++ and Audio2Rig offer innovative solutions in terms of real-time performance and artistic flexibility. Through a comparative analysis of representative approaches, this article reveals the evolving trends in semantic modeling, expression control, 3D consistency, and stylistic expression.

Overall, human head animation generation is evolving from single-modality driven to multi-source integration, from static synthesis to dynamic interaction, and gradually acquiring the capabilities of human-like expression and multi-role adaptation. Transformer methods continue to show potential in processing long sequences and semantic consistency, while diffusion models continue to achieve breakthroughs in generation quality and control accuracy. NeRF's 3D modeling capabilities provide more realistic spatial expression for virtual humans. Specialized architectures demonstrate efficiency and flexibility in specific application scenarios, driving this technology towards wider practical application.

Despite significant progress, current approaches still face challenges in real-time performance, cross-language adaptation, and emotional consistency. Future research could strengthen the stability of semantic modeling, enhance generalization capabilities in low-resource environments, and explore lightweight deployment and user-controllable mechanisms to promote the widespread application of virtual human systems in social, educational, and entertainment scenarios.

References

[1] Liu, C., Lin, Q., Zeng, Z.. (2024). Emoface: Audio-driven emotional 3D face animation. In 2024 IEEE Conference on Virtual Reality and 3D User Interfaces (VR) (pp. 387–397). IEEE.

[2] Chen, Y., Liang, S., Zhou, Z. (2025). HunyuanVideo-Avatar: High-fidelity audio-driven human animation for multiple characters. arXiv preprint arXiv: 2505.20156.

[3] Huang, Y., Wang, J., Zeng, A. (2023). Dreamwaltz: Make a scene with complex 3D animatable avatars. Advances in Neural Information Processing Systems, 36, 4566–4584.

[4] Drobyshev, N., Casademunt, A. B., Vougioukas, K. (2024). Emoportraits: Emotion-enhanced multimodal one-shot head avatars. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 8498–8507).

[5] Fei, H., Zhang, H., Wang, B. (2024). Empathyear: An open-source avatar multimodal empathetic chatbot. arXiv preprint arXiv: 2406.15177.

[6] Wang, H., Weng, Y., Li, Y. (2025). Emotivetalk: Expressive talking head generation through audio information decoupling and emotional video diffusion. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 26212–26221).

[7] Zhen, R., Song, W., He, Q. (2023). Human–computer interaction system: A survey of talking-head generation. Electronics, 12(1), 218.

[8] Arcelin, B., & Chaverou, N. (2024). Audio2Rig: Artist-oriented deep learning tool for facial and lip sync animation. In ACM SIGGRAPH 2024 Talks (pp. 1–2). ACM.

[9] Peng, Z., Hu, W., Ma, J. (2025). SyncTalk++: High-fidelity and efficient synchronized talking heads synthesis using Gaussian splatting. arXiv preprint arXiv: 2506.14742.