Theory and Practice of Science and Technology
2025, VOL. 6, NO. 6, 102-106
DOI: 10.47297/taposatWSP2633-456920.20250606

Toward Intelligent Virtual Anchor: Generative AI for Unified Multimodal Generation

Xinyu Zhong

Department of Electronics, University of Liverpool, Liverpool L69, United Kingdom

ABSTRACT
The rapid advancement of generative artificial intelligence has propelled a profound transformation in the field of virtual anchors, shifting them from traditional script-driven and pre-animated unimodal expression to an intelligent, multimodal collaborative generation paradigm. This paper systematically analyzes the technological evolution paths of virtual anchor systems across three key dimensions—speech synthesis, facial animation, and interaction mechanisms—and explores how generative AI significantly enhances the naturalness of expression, real-time interactivity, and emotional richness. Additionally, it points out existing challenges such as interaction cost, personalized expression, and ethical transparency, and proposes future directions for the intelligent design of virtual humans.

KEYWORDS
Virtual anchor; Generative artificial intelligence; Multimodal collaboration; End-to-end generation; Human-computer interaction

1 Introduction

1.1 Background

As a key application area of virtual human technology, virtual anchors have garnered growing attention in recent years across fields such as media dissemination, digital entertainment, and human-computer interaction. Traditional virtual anchors typically rely on pre-defined animations and concatenated speech technologies, with expression driven by scripted inputs. Such systems are limited in natural interactivity and emotional expressiveness.

With the advancement of generative artificial intelligence—particularly the increasing maturity of end-to-end models based on deep learning—virtual anchors are acquiring the capabilities of natural speech synthesis, facial animation, and real-time responsiveness. Their technical framework is evolving from modular composition toward unified modeling and intelligent multimodal coordination.

In recent years, research in the virtual anchor domain has made substantial progress under the impetus of generative AI. Studies have demonstrated that end-to-end speech synthesis models, employing semantically driven generative mechanisms, can effectively enhance the naturalness and emotional fluidity of synthesized speech [1]. In facial animation, models based on generative adversarial networks and neural rendering have significantly improved lip synchronization and emotional expressiveness [2]. Interactive systems have adopted cross-modal alignment mechanisms to integrate linguistic and visual information channels.

Moreover, multimodal modeling approaches are becoming mainstream, driving the evolution of virtual avatars from segmented control toward unified expressive frameworks [3]. Despite these breakthroughs at the technical level, current developments still focus predominantly on local optimizations. There is a lack of systematic investigation into collaborative generative mechanisms across the entire virtual anchor pipeline, and the inter-modal linkage among the three core components—speech, facial animation, and interaction—remains underanalyzed. These gaps form the motivation and entry point for this study.

1.2 Technological evolution

The development trajectory of virtual anchor technologies reflects a systematic transformation from traditional manual control mechanisms to intelligent, collaborative generative systems.

Traditional virtual anchors are largely constructed based on keyframe animation, concatenated speech synthesis, and templated interaction systems. In these systems, speech is typically synthesized from pre-recorded segments or rule-based generation models; facial expressions depend on animator-designed motion sequences; and interaction logic is often built on keyword matching and command trees. Although such systems are capable of basic humanoid representation, they are essentially assembled from discrete unimodal modules that lack a unified mechanism for semantic understanding and responsive generation. As a result, issues such as misalignment between speech and facial expressions, mismatch between tone and emotion, and disconnection between responses and context have long persisted [1-2].

The introduction of generative artificial intelligence has inaugurated a new paradigm of intelligent collaborative

modeling. End-to-end models such as VITS, StyleGAN-EX, and CLIP have redefined the generative paths of speech, vision, and language modalities. These models no longer rely on pre-defined templates or scripted commands. Instead, they employ deep neural networks to establish cross-modal semantic associations, enabling systems to autonomously learn contextual information from text, audio, or visual inputs and generate multimodal outputs in real time [3-4].

Unified modeling not only improves the naturalness of generated content but also empowers virtual anchors to express synchronized speech, facial expressions, and semantic understanding in complex scenarios such as conversation, singing, and lecturing.

From a structural perspective, this evolution signifies a fundamental shift in the logic of media generation: from rule-based concatenative generation toward semantically-driven collaborative generation; from static content playback toward dynamic emotional adaptation; and from functional output toward perceptually grounded expression. As Merleau-Ponty emphasized in his "Phenomenology of Perception," authentic expression must emerge from the integrated coordination of perception, action, and context [5]. Only by achieving endogenous coordination across modalities can virtual anchor systems transcend their nature as "simulacra" and evolve into emotionally aware and self-regulating digital beings.

2 Paradigm shift in technical architecture

The intelligent evolution of virtual anchor technologies centers on the progressive alignment of system architectures for "virtual humans." Virtual humans are typically constructed through the integration of technologies such as graphical rendering, speech synthesis, motion capture, and deep learning. These interactive virtual agents possess human-like appearances and behavioral patterns.

As sensing technologies advance, the concept of multimodal virtual humans has gradually taken shape. This refers to anthropomorphic systems capable of language understanding, emotional expression, perceptual interaction, and individual modeling. Their core modules span across multiple dimensions, including speech-driven control, facial animation, semantic generation, and modality fusion [6]. The current enhancement of virtual anchors in speech, vision, and interaction capabilities reflects a systemic transformation from unimodal control to multimodal perception and generation mechanisms.

2.1 Speech synthesis: from concatenation to end-to-end

Traditional speech synthesis systems—such as those based on Statistical Parametric Speech Synthesis (SPSS) or concatenative Text-to-Speech (TTS)—typically follow a three-stage process involving phonemes, prosody, and vocoders. The resulting speech often suffers from issues like disrupted prosody and rigid speaking rates [7]. These systems usually generate audio through concatenation of speech units or statistical modeling, making it difficult to maintain semantic continuity or adapt tone and emotion to context. As a result, virtual anchors tend to sound "like reading from a script" in live streaming or multi-turn dialogue scenarios, lacking naturalness.

Intermediate architectures such as Tacotron 2 have improved fluency to a certain extent but still rely on multi-stage modules and fail to fully resolve semantic discontinuity. In contrast, the VITS model introduces variational inference mechanisms and adversarial training frameworks to achieve end-to-end learning from text to audio [4]. This architecture bypasses the phoneme layer by encoding text into latent semantic vectors, directly generating mel-spectrograms, which are then reconstructed into audio waveforms via flow-based models. This significantly enhances both the naturalness of speech and its emotional dynamism [8].

Even more notably, VITS plays a central role in multimodal collaborative generation. Its output speech feature vectors can be directly used as synchronous inputs for facial animation modules, eliminating the need for separate mapping between voiceprints and facial expressions. This greatly reduces the likelihood of temporal misalignment. With this capability, virtual anchors can truly achieve coordinated expression where "what is said" aligns with "how it looks," establishing a consistent bridge between modalities [1].

The evolution path of virtual anchors clearly demonstrates this transition. For example, in early 2022, Neurosama used pre-recorded audio and a script-driven Live2D facial animation system, relying on template matching to respond to user comments—a typical case of technical concatenation. By mid-2023 to 2024, its developers integrated large language models (LLMs) and real-time TTS synthesis mechanisms, enabling the system to infer spoken responses directly from chat text. This approach allowed Neurosama to generate speech instantly during live interactions, seamlessly respond to MIDI comment interactions, and exhibit natural intonation and emotional variation. As a result, it evolved from a "tool-like character" to an intelligent virtual presence capable of real-time conversational responses.

2.2 Facial animation: from skeletons to neural rendering

In terms of facial modeling, traditional virtual anchors have largely adopted skeletal animation or BlendShape techniques, which drive facial control points to generate lip movements and eye expressions. A representative example is Apple's ARKit, which supports control over 52 facial parameters and has been widely used in Live2D systems and among early Hololive members [7]. While these methods enable basic facial switching, they lack dynamic perturbation mechanisms and are unable to generate micro-expressions based on speech rhythm or semantic context. This often results in perceptual dissonance, such as "mouth moving while the eyes remain expressionless."

Generative modeling fundamentally redefines the logic of facial animation. For instance, architectures like StyleGAN-EX decouple facial animation into two orthogonal spaces: identity vectors and expression perturbations. Facial expressions are controlled via input features such as phonemes or semantic representations [2]. Under this mechanism, expressions are no longer constrained by pre-defined parameters but are dynamically generated based on context. For example, when high-frequency segments in speech indicate emotional agitation, the system can produce larger eye openings, forehead movements, and mouth corner pulls—achieving a natural coupling between emotion and muscular motion.

A particularly important feature is that this model supports phoneme-level driving, meaning that each speech unit can be mapped to a continuous trajectory of micro-expressions, allowing speech rhythm and facial dynamics to synchronize at the millisecond level. When deployed jointly with the VITS model, the system can directly reuse intermediate tensors from the speech stream—such as mel-spectrograms, duration, and tension—to drive facial animation, thereby fully enabling a multimodal alignment pipeline from "speech to expression" [2][4].

This collaborative structure has been validated in the VASA-1 system developed by Microsoft Research Asia. VASA-1 takes a single facial image and any given audio input and processes them through neural networks to generate real-time lip movements and facial animations, synchronizing speech and visual output with high efficiency [9]. Experimental demonstrations show that the system can faithfully reproduce the speaker's speech rhythm and facial motion consistency in real time, demonstrating the technical feasibility of "end-to-end audio-driven facial modeling." Compared to traditional methods that rely on pre-set templates for facial micro-movements, neural rendering not only improves the expressiveness of facial animation but—more importantly—constructs a learnable generation logic for the facial responses of virtual anchors, thus enabling a data-driven path for emotional conveyance.

2.3 Interaction: from scripts to empathy

Traditional interaction modules in virtual anchor systems are typically based on keyword matching and fixed scripted replies. For instance, in early Live2D or Hololive platforms, anchors utilized LiveChat systems to read audience messages, where backend scripts would trigger pre-recorded voice lines or facial animation templates based on matched keywords. While this mechanism can achieve "quick responses" in certain scenarios, it often fails in cases involving free-form text, multi-turn dialogue, or emotional recognition.

Generative artificial intelligence has reshaped interaction logic through cross-modal perception mechanisms. Language modeling now employs architectures such as Transformer-XL, which offer long-sequence memory capabilities and can generate coherent semantic replies while retaining contextual history [10]. In the visual channel, the CLIP model is introduced to map user videos, facial expressions, or image-based comments into a shared vector space with language semantics, enabling semantic alignment between language and imagery [3].

This architecture endows virtual anchors with the ability for "non-instructional responses"—that is, they can infer users' underlying intentions without explicit emotional expression or direct requests. For example, during a live broadcast, if the system detects that a user is repeatedly posting negative comments or if the camera shows a downcast gaze or frowning, it can automatically switch to a softer vocal tone, adjust the lighting, or offer emotional comfort. This type of "soft perceptual capability" represents an intelligent response pattern that traditional script-based systems cannot achieve.

A notable example can be found in the 2024 livestreams of virtual anchor Kano. When a fan commented, "I've been feeling a lot of pressure lately," the system did not rely on keyword recognition. Instead, it generated a comforting vocal response with a caring tone through emotional recognition and dialogue history, while simultaneously switching to warm-toned lighting and slowing the background music. This intelligent reaction significantly enhanced audience immersion and emotional connection, marking a crucial step in the evolution of virtual anchors from "entertainment personas" to "digital companions."

3 Future technological trajectories for virtual anchors

3.1 Balancing interaction efficiency and generative cost

Although current end-to-end architectures have significantly simplified inter-module coupling processes, generative models still face challenges in practical deployment, such as high computational resource demands and response latency. For instance, inference with models like VITS and the StyleGAN family imposes substantial GPU workloads, making real-time operation on mobile or low-power devices difficult [4][11].

To address this issue, Neural Architecture Search (NAS) has become a prominent optimization strategy in recent years. NAS automatically explores optimal model architectures, reducing parameter count and inference latency while maintaining generation quality. It has already shown promising results in tasks such as speech generation and facial animation [8].

Additionally, knowledge distillation and other model compression strategies are being introduced into virtual anchor systems to retain the expressive power of large models while improving deployability [8]. When combined with hardware-friendly techniques like model pruning and quantization, these methods are expected to enable low-latency virtual anchor interactions on embedded platforms—extending their applications to emerging fields such as educational robots and mobile companions.

3.2 Personalization and ethical transparency

As virtual anchors increasingly approach human-like expressiveness, tensions between personalization and ethical control have become more pronounced. On one hand, users desire virtual anchors with stable and unique expressive styles that can adapt to individual audience preferences. On the other hand, the inherent opacity of generative systems introduces risks such as content hallucination, emotional manipulation, and identity deception [12].

One viable approach is to incorporate a "personality vector" control mechanism, which encodes stylistic features into adjustable parameters. These parameters can be modulated through user input or preference learning. This mechanism has been applied in affective dialogue systems and personalized speech synthesis [13][14].

At the same time, to improve ethical transparency, some researchers have proposed the use of visual watermarking and generation traceability indicators to clearly distinguish between human-created and AI-generated content, thereby mitigating users' cognitive risks [12]. At the regulatory level, the European Union's *Artificial Intelligence Act* draft stipulates that virtual agents must explicitly indicate their "non-human identity," thus providing a policy framework for embedding ethics into system design.

This trend suggests that future virtual anchor systems must be designed not only for functional performance but also for behavioral norms, personality boundaries, and expressive legitimacy—ultimately forming a design paradigm centered around "digital personhood."

3.3 Modeling strategies for cross-cultural expression

Current emotion modeling in virtual anchors is predominantly based on English-language corpora and Western facial behavior standards, such as the Facial Action Coding System (FACS) proposed by Ekman [7]. However, in real-world applications, different cultures interpret facial expressions, prosodic styles, and interaction rhythms in significantly different ways. For instance, East Asian users often express joy through a subtle "smiling with narrowed eyes," while Western users rely more on exaggerated mouth shapes and elevated intonation [7]. This cultural divergence has led to adaptation issues and emotion misclassification in existing generative models during cultural transfer.

To enhance cultural adaptability, systems must build finer-grained emotion expression models and implement cross-cultural emotion mapping mechanisms. Some studies have attempted to train multi-source style transfer models using large-scale multilingual data to reconstruct emotional expression across cultures within a unified semantic space [13]. Furthermore, the field of micro-expression synthesis has begun incorporating deformable space modeling to capture subtle expression variations in different cultural contexts [11].

In the future, embedding cultural parameters into the semantic generation pipeline may enable virtual anchors to become truly "globally communicative anthropomorphic entities."

4 Conclusion

This paper systematically articulates the internal logic and practical trajectory behind the transformation of virtual anchor technologies from traditional manual control systems to intelligent, collaboratively generative frameworks. The study demonstrates that generative artificial intelligence, by reconfiguring the three core channels—speech synthesis,

facial animation, and interaction response—has successfully enabled virtual anchors to achieve multimodal coherence and naturalistic expression.

In particular, technological innovations represented by the VITS speech model, the StyleGAN-EX facial animation model, and the real-time facial driving system VASA-1 have provided robust support for virtual anchors to evolve from functional agents into expressively lifelike digital beings.

At the same time, three key development directions are identified as focal points for future research and implementation: (1) improving interaction efficiency while reducing generative cost; (2) balancing personalized expression with ethical transparency; and (3) enhancing cross-cultural adaptability in emotion modeling.

Future research should aim to further refine end-to-end collaborative generation frameworks, improve technical performance, and ensure that ethical guidelines and cultural diversity are fully integrated into system design. This would enable virtual humans to interact with real users in ways that are more natural, human-centered, and emotionally intelligent—ultimately fostering richer and more meaningful digital companionship.

About the Author

Xinyu Zhong, sgxzhon2@liverpool.ac.uk.

Reference

[1] J. Kim, J. Kong, et al., "Conditional Variational Autoencoder with Adversarial Learning for End-to-End Text-to-Speech," in *Proc. 38th Int. Conf. on Machine Learning (ICML)*, PMLR 139, pp. 5530–5540, 2021.

[2] Y. Shirahata, R. Yamamoto, et al., "Period VITS: Variational Inference with Explicit Pitch Modeling for End-to-End Emotional Speech Synthesis," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2023.

[3] A. Sikarwar, A. Pawar, et al., "CLIP Model Image Search: A Deep Learning Approach for Finding Relevant Images," *arXiv preprint arXiv: 2111.14544*, 2021. [Online]. Available: https://arxiv.org/abs/2111.14544.

[4] J. Kim, J. Kong, et al., "End-to-End Speech Synthesis with VITS," *IEEE/ACM Trans. Audio, Speech, and Language Processing*, vol. 30, pp. 1849–1863, 2022.

[5] M. Merleau-Ponty, D. Landes, et al., *Phenomenology of Perception*, New York, NY, USA: Taylor & Francis, 2013.

[6] Y. Lingyan and Z. Yi, "Generative artificial intelligence empowers a new paradigm for the creation of virtual digital humans," Innovation and Application of Science and Technology, no. 16, pp. 1–4, 2025.

[7] P. Laukka, H. A. Elfenbein, et al., "Cross-Cultural Emotion Recognition and In-Group Advantage in Vocal Expression: A Meta-Analysis," *Emotion Review*, vol. 13, no. 1, pp. 3–11, Jan. 2021.

[8] C. Liu, B. Zoph, et al., "Progressive Neural Architecture Search," in *Proc. Eur. Conf. on Computer Vision (ECCV)*, 2018, pp. 19–34.

[9] S. Xu, G. Chen, et al., "VASA-1: Lifelike Audio-Driven Talking Faces Generated in Real Time," in Advances in Neural Information Processing Systems (NeurIPS), vol. 37, 2024.

[10] Z. Dai, Z. Yang, et al., "Transformer-XL: Attentive Language Models beyond a Fixed-Length Context," ACL, 2019.

[11] Y. Wang, Z. Liu, et al., "StyleGAN-EX: StyleGAN-based Expression Editing and Generation," CVPR, 2022.

[12] A. Möslein, "AI as a Challenge for Legal Regulation – The Scope of Application of the Artificial Intelligence Act Proposal," *Law, Innovation and Technology*, vol. 23, pp. 361–376, 2023.

[13] L. Devlin, M.-W. Chang, et al., "Multimodal Learning Frameworks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 3, pp. 963–976, Mar. 2022.

[14] L. He, Z. Wang, et al., "Multimodal Mutual Attention-Based Sentiment Analysis Framework Adapted to Complicated Contexts," *IEEE Trans. Circuits Syst. Video Technol.*, published online May 15, 2023.