



International Journal of Data Mining and Bioinformatics

ISSN online: 1748-5681 - ISSN print: 1748-5673

<https://www.inderscience.com/ijdmb>

Interactive AI virtual human synthesis technology based on multimedia visual images

Huichao Guo, Yongjian Hua, Ming Lei, Runhua Huang

DOI: [10.1504/IJDMB.2026.10079312](https://doi.org/10.1504/IJDMB.2026.10079312)

Article History:

Received:	30 December 2025
Last revised:	09 March 2026
Accepted:	13 April 2026
Published online:	13 July 2026

Interactive AI virtual human synthesis technology based on multimedia visual images

Huichao Guo*

School of Safety Science and Emergency Management,
Wuhan University of Technology,
Wuhan, 430000, Hubei, China
and
Zhejiang Business College,
Hangzhou, 310000, Zhejiang, China
Email: 334998@whut.edu.cn
*Corresponding author

Yongjian Hua and Ming Lei

Zhejiang Business College,
Hangzhou, 310000, Zhejiang, China
Email: yongjianh_zjbc@126.com
Email: work.leiming@foxmail.com

Runhua Huang

The Chinese University of Hong Kong,
Shenzhen, Shenzhen, 518100, Guangdong, China
and
Zhejiang Business College,
Hangzhou, 310000, Zhejiang, China
Email: runhuahuang@link.cuhk.edu.cn

Abstract: In the specific implementation, this paper first developed a multimodal feature-extraction process to extract semantic and rhythmic information from speech. It combined the emotional tags in the text with facial keypoint information from diverse visual inputs (e.g., static photos, dynamic video sequences, and 3D face models) to form a unified feature vector. Secondly, this paper designed MAM to dynamically adjust the weights of audio, text, and visual modalities and to synchronise expression changes with speech rhythm through real-time speech stream analysis and frame-level feature alignment. The dynamic weight adjustment in MAM operates along both the temporal dimension, aligning frame-level features across modalities, and the channel dimension, reweighting feature maps to emphasise modality-specific contributions to expression synthesis. This synthesis technology introduces a DM based on the U-Net structure, uses fused features as conditional inputs, and drives the progressive generation of virtual human image sequences.

Keywords: multimedia visual images; diffusion model; DM; multimodal attention mechanism; MAM; virtual human synthesis; emotion perception.

Reference to this paper should be made as follows: Guo, H, Hua, Y., Lei, M. and Huang, R. (2026) ‘Interactive AI virtual human synthesis technology based on multimedia visual images’, *Int. J. Data Mining and Bioinformatics*, Vol. 30, No. 7, pp.26–50.

Biographical notes: Huichao Guo received his Bachelor’s degree from Nankai University. Currently, he is chasing the PhD degree in School of Safety Science and Emergency Management, Wuhan University of Technology. His research interests include big data analysis, digital management, emergency management, and intelligent management.

Yongjian Hua is the Dean of the School of Cold Chain Technology at Zhejiang Business College and a leading academic in vocational education. Specialising in logistics platforms and supply chain optimisation, he has published over 50 academic papers, authored two monographs, and edited three textbooks, including a National 12th Five-Year Plan textbook. A strong advocate for industry-education integration, he has directed six provincial research projects and guided industrial development plans for 12 counties in Zhejiang. He is widely recognised for his outstanding contributions to China’s modern commercial education and logistics sectors.

Ming Lei is a dedicated Lecturer at Zhejiang Business College, holding a PhD in their specialised field. With a strong foundation in rigorous academic research and a passion for modern vocational education, he focuses on bridging theoretical frameworks with practical industry applications. As an active contributor to the college’s academic community, he is deeply engaged in advancing applied research, developing innovative curricula, and mentoring students. By emphasising practical skills and industry integration, he plays a vital role in cultivating highly qualified professionals equipped to meet the evolving demands of the modern commercial sector.

Runhua Huang received his BE degree from The Chinese University of Hong Kong, Hong Kong in 2014, MS degree from Columbia University, New York, NY, USA in 2019, and PhD degree in Data Science from The Chinese University of Hong Kong, Shenzhen. His research interests include smart transportation networks, population migration analysis, and urban mobility systems.

1 Introduction

As an important research direction in the field of human-computer interaction, virtual humans are gradually being integrated into scenarios such as intelligent customer service, virtual anchors, and educational counselling (Wang et al., 2025; Walker et al., 2023). With the growing strength of new media and immersive experiences, users have higher expectations for the realism of virtual human interaction, especially in terms of emotional expression, voice lip-syncing coordination, and video visual stability (Demirel et al., 2022; Golf-Papez et al., 2022). In real-world applications, most AI virtual human systems are limited to responding to single-modal inputs (Sorosrungruang et al., 2024; Zhu and Li, 2021). They are unable to effectively integrate information such as language, speech, and vision, resulting in a lack of naturalness and coherence in the generated results, particularly in visual expression and emotional response. They do not build a closed-loop

of perception-generation-feedback in real-time interactive scenarios, resulting in response delays and disconnection from emotional expression (Li et al., 2023; Cao et al., 2023). In the proposed architecture, the feedback loop is realised by integrating the emotion classification module, whose output is continuously injected as a conditional prompt into the diffusion model (DM) at each reverse step, thereby enabling the generated visual information to modulate subsequent expression intensity and style based on real-time analysis of the input speech's emotional state. In particular, when faced with speech inputs with changing language rhythm, fluctuating intonation, or varying emotional intensity, existing models show obvious deficiencies in lip alignment, expression conversion, and movement stability (Alnagrat et al., 2022; Abich et al., 2021).

Most current mainstream methods rely on generative adversarial network (GAN)-like generative structures. Although they possess certain generative capabilities, they remain limited in controlling the dynamic consistency across consecutive frames and in expressing emotional levels (Chai and Park, 2022; Ahuja et al., 2021). GAN methods have also often produced visible flickering of facial texture or occasional drift of facial landmarks across multiple frames when rendering very rapid speech or sudden emotional changes, resulting in artificial visual artefacts that degrade the user experience. For Interactive AI virtual humans to create a realistic visual feedback path, higher-dimensional inputs must be able to understand input mode features and align them (Lin and Yu, 2025; Di and Zheng, 2022). To solve these issues, comprehensive synthesis technologies must be developed to understand the semantics of the speech being performed, accurately determine the emotion being expressed, and present that information in a stable visual format for the interaction history. Research into Virtual Generation continues to develop, moving from early static techniques, such as rule mapping (Ding et al., 2023; Zhao et al., 2022), to dynamic generation models using deep learning. Wav2Lip synthesises lip movements from audio input that correlates phonemes with their respective lip motions (Chen et al., 2025; Lee et al., 2025). The StyleGAN series of models has shown considerable benefits for image clarity and structural integrity and is frequently used in the face generation process (Melnik et al., 2024; Kang and Jo, 2024). CycleGAN has also been applied to virtual human tasks to enable conversion between different styles. Although this multimodal, non-real-time system supports multispectral/3D vision and multiple inputs, the delay exceeds 200ms and cannot meet interactive requirements (Liu et al., 2023, 2025). However, these methods lack a mechanism for co-processing semantics, emotions, and vision. Audio-driven models (Niu et al., 2024; Cui et al., 2025) can achieve basic lip matching but are slow to respond to emotional changes, while image generation models (Jia et al., 2024; Jiang et al., 2022) achieve high clarity but struggle to maintain consistency between speech rhythm changes and visual dynamics. Some studies have introduced emotion recognition modules to improve the naturalness of expressions, but the models remain insensitive to changes in emotional intensity (Chen and Chen, 2023; Gupta et al., 2023). In addition, under continuous speech input or fast speech speed, the generated results often exhibit picture jumps and inconsistent frame-to-frame behaviour, which affects the final interactive experience. Current research faces challenges such as a weak fusion mechanism, poor temporal stability, and limited depth in emotional expression.

Researchers have sought to build cross-modal alignment networks and emotional regulation mechanisms in virtual humans to enhance their interactive realism (Slovak et al., 2023; She et al., 2025; Chamola et al., 2024). Some approaches enhance sentiment analysis of text input by incorporating image-based components and using the

Transformer architecture to improve the accuracy of semantic/image mapping, resulting in moderate improvements in output accuracy (Zhou et al., 2023). To achieve synchronisation of multiple streams of information through modal fusion, researchers employ multi-modal Transformers and cross-attention mechanisms to ensure consistent, accurate outputs throughout the entire creation process (Shaikh et al., 2024; Zhang et al., 2023). Furthermore, to achieve greater realism in image rendering, several researchers replaced GAN-type generators with DMs as their rendering source, capitalising on the gradual reconstruction of their content to achieve smooth video transitions and frame coherence (Liang et al., 2024; Alexanderson et al., 2023). Although these methods each have their own advantages, they still have limitations in terms of integration capabilities. Specifically, the combining of audio, text, and visual characteristics has limitations caused by a lack of a common reference system (including dimensionality), non-synchronisation of timestamps, and a lack of depth of emotion in the combined modal outputs producing images that do not accurately depict the emotional rhythm of speech or produce abrupt changes in expression between key frames, thus negatively impacting immersion. This paper solves the problem of modal separation in existing technologies by dynamically fusing speech, text, and visual features through multimodal attention mechanism (MAM). At the same time, DM is introduced to replace traditional GAN, and a progressive generation strategy is used to optimise temporal coherence, improve the naturalness of expressions, and speed up emotional response.

This study focuses on the generation of interactive AI virtual humans driven by multimodal visual images. By fusing speech, text, and visual signals, a controllable, high-fidelity image synthesis path is constructed to enhance the realism and emotional responsiveness of virtual human interaction (Chang Chien, 2024). The main goal of this method is to achieve synchronous alignment of speech rhythm and expression generation, improve the expression efficiency of cross-modal feature fusion, and enhance the emotional consistency between expression and speech content. To this end, this paper introduces a conditional generation model based on the U-Net diffusion structure, using the uniformly encoded multimodal features as a generation guide to achieve progressive reconstruction of sequence images. In view of the alignment deviation of multi-source data input in the time dimension, MAM is designed to dynamically weight audio semantics, intonation curves, text emotion vectors, and facial key point features to ensure accurate mapping of key emotional nodes in visual expression, and to achieve a real-time interactive closed loop with fast interactive feedback. By integrating the emotion classification module, the style changes and the expression of emotional intensity in the generated images are further regulated, enabling the virtual person to maintain consistency in expression and voice across emotional contexts such as anger, sadness, and surprise. Compared with existing methods, the model in this paper shows better performance in voice-driven response speed, image continuity stability, and emotion perception. The research results provide a feasible path to building a virtual human system with high interactivity and emotional expression capabilities, with strong application prospects and significant promotional value (Luo et al., 2024).

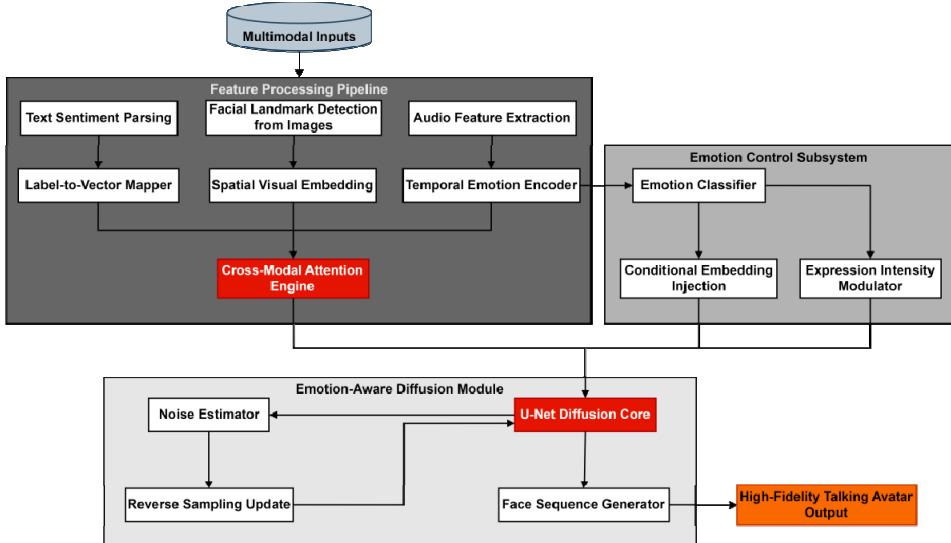
2 Methods

Figure 1 shows an interactive AI virtual human generation architecture that integrates audio, text, and images. The multimodal input to unified fusion model: this figure

provides an overview of how the different modalities contribute to a virtual human's expression and are integrated into its overall expression.

- **Input modes:** Starting with the primary input mode, the model collects speech rhythms, semantics, text emotion labels, and key facial locations from all three modalities into a single feature-processing pipeline.
- **Model components:** Both components will work together through a cross-modal attention module to couple the temporal and semantic relationships across the different inputs, producing a unified fused representation.
- **MSG module:** This module will also take the audio signal and extract the emotional state. Then it will provide real-time conditional control information (CC) to configure the original message to the DM processor, which will adjust both the expression's intensity and style when generating facial expressions.
- **Diffusion module:** The diffusion module uses a U-Net architecture to iteratively restore images at each time step, using conditional injection to converge to the highest-fidelity expressions under the control of multimodal drives.
- **Final output:** The entire system's output is a series of images of virtual humans with consistent emotional states and synchronised movements. Thus, there exists fine-grained synergy between cross-modal integration of information and emotion regulation in the generated outputs, providing dynamic consistency in spatial semantics, matched emotions, and the visual expression of the generated image.

Figure 1 Interactive AI virtual human generation architecture (see online version for colours)



2.1 Multimodal input tool extraction

Emotion tags are extracted from textual content, and semantic and rhythmic features are extracted as evidence of an emotional state. The face will serve as an indicator for

extracting key points from the pre-constructed image-based face model. These features will be combined into three input types for subsequent fusion operations: audio, text, and visual.

2.1.1 Semantic and rhythmic feature extraction of speech signals

An end-to-end (E2E) neural network model is used as a processing mechanism for E2E decoding, capturing semantic information and rhythmic elements from the original audio signal. The E2E neural network splits contiguous audio waveforms into multiple segments using frame-level processing (e.g., by chopping the waveform into short-term segments). Then it extracts the time-frequency characteristic data from the audio waveform signal using four convolutional layers with batch normalisation and ReLU activation (using a custom shallow CNN) to represent the temporal characteristics of a spectral change and the character of energy fluctuations over the duration of time (in the case of the audio frequency domain). The BiLSTM in the stacked time series learns an overall representation of semantic context and trends in intonational fluctuation data over time, using temporal information to provide additional information about meaning, emotion, or affect. The prosodic character of an audio waveform is quantified through analyses of the F0 trajectory data from the fundamental frequency (F0) of a particular speaker or source audio through the use of the intensity of changing sounds and the duration of events to help develop a basis for understanding emotions or affective content. To extract prosodic features from source audio files, a short-time Fourier transform is applied, and a normalisation strategy is used to enable the model to be more flexible across various pronunciations and speaking intensities. The self-attention mechanism dynamically adjusts and weights the contributions of different types of information to the synthesised expressions, resulting from the fusion of the semantic vector and prosodic features. The weights for the feature fusion are calculated using the following equation:

$$\alpha_i = \frac{\exp(e_i)}{\sum_{j=1}^n \exp(e_j)}, \quad e_i = \text{score}(q, k_i) \quad (1)$$

q represents the current decoding state vector, k_i is the i^{th} feature key in the input sequence, the score function measures the feature correlation, and α_i is the normalised weight. This mechanism ensures a dynamic balance between speech content and rhythmic information in expression, thereby improving the accuracy and delicacy of emotional transmission. The deep feature output of the speech module enhances the perception of emotional changes while accounting for the temporal consistency of the pronunciation rhythm, laying the foundation for multimodal information fusion.

2.1.2 Text emotion label analysis and facial key point visual feature extraction

The text information processing stage uses a pre-trained BERT-based sentiment analysis model, fine-tuned on the IEMOCAP corpus, to accurately assign emotion labels. The text sequence is first processed by word segmentation and part-of-speech tagging. Then, the deep bidirectional encoder representation technology is introduced to aggregate the emotion keywords in the sentence and construct the emotion vector. During training, the model uses a large-scale annotated dataset and designs a loss function for multi-category

emotion labels, thereby effectively improving its ability to discriminate fine-grained emotions. After the text emotion vector is normalised, it is used as one of the conditions for multimodal fusion to ensure a high correspondence between semantic emotions and visual expressions.

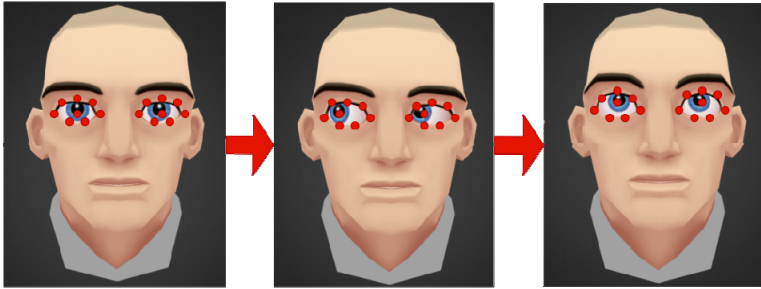
The facial visual feature collection uses a posture estimation network to extract 2D keypoint coordinates via a convolutional neural network. The network introduces a multi-layer supervision mechanism during training to optimise the spatial accuracy and temporal stability of key point positioning. The main facial attributes chosen as keypoint sets (such as eyes, eyebrows, and lips) are intended to represent the key regions of the face for accurate capture of facial expression detail. Temporal smoothing of the coordinate set for each keypoint reduces jitter, thereby allowing a better representation of how facial expression changes over time. The 3D video representation of the method described will be standardised before synchronising the 3D facial feature vector data with the audio and text data. After all modalities are synchronised using a timestamp mechanism, all three types of modality data will be organised on the time axis. The process is summarised:

$$\mathbf{F}_t = \frac{\mathbf{K}_t - \mu_{\mathbf{K}}}{\sigma_{\mathbf{K}}} \quad (2)$$

$\mathbf{K}_t$ represents the facial key point vector of frame t , $\mu_{\mathbf{K}}$ and $\sigma_{\mathbf{K}}$ are the mean and standard deviation of the sequence, respectively, and $\mathbf{F}_t$ is the normalised output. Normalisation stabilises feature values, which is convenient for subsequent network learning. The visual feature expression of this step reflects both static facial structure and dynamic expression information, providing a solid spatial foundation for expression synthesis.

Deep extraction of audio semantics and rhythmical features, precise analysis of text emotion vectors, and stable collection of visual key points make up the core of what supports multimodal inputs. The integration of the different modalities is efficiently achieved through unified feature encoding and time-alignment methods, thereby laying a strong foundation for the generation of natural expressions by interactive virtual humans.

Figure 2 Key point control of eyeball changes (see online version for colours)



Eyeball movement is an important part of synthesising interactive AI virtual humans. Figure 2 shows how eyeballs change from initial to intermediate to final. At the starting point, the eyeball position is set and marks the basic location of the eye, as shown at the key points. The first input signal is a voice command that tells the model to ‘look right first’ and then ‘look up’. With this command, the 3D model keypoints begin moving as the input signal changes, causing the model to adjust the position of its eyeball in

real-time. In the end position, the model has moved the eyeballs to the final target locations and has further optimised the distribution of keypoints so that the expression on the face is in sync with the rhythm of the speech. Fine-tuned control of keypoints has improved the eye's natural movement and sense of coherence, thereby enhancing interactive realism and immersion in virtual humans.

2.2 MAM design

This study builds a cross-modal attention model to map the rhythmic nature of speech and its associated emotional state to time-series data of facial expression changes. By dynamically adjusting the strength of the correlation between modalities via an attention mechanism, the actions created will be synced with the speech and enhance its emotional content.

2.2.1 Dynamic mapping and regulation of audio and expression sequences

The information contained in the speech rhythm is comprised of two parts: the temporal fluctuations of the speech waveform over time, and the rhythmic representation of the speech waveform. To create a facial action sequence that corresponds to the speech rhythm, an attention-mapping structure captures the dynamics of the speech waveform. This attention-mapping structure adds temporal position encoding to the frame-level features of the original speech, positioning them along the time axis and generating context-based representations of the speech. The temporal position encodings use the absolute sinusoidal position encoding format, which is then added element-wise to the speech frame embedding before being passed through the attention mechanism; thus preserving the sequential ordering of the speech signal. At each time step of the mapping network, the correlation weight between the speech frame and the current expression state (i.e., smile or grimace) is calculated to determine how strongly an expression will be triggered to change. To achieve this, a learnable query vector is used to focus on audio peaks throughout the audio sequence, thereby avoiding delayed or subtle expression (smile/grimace) changes that may have resulted from intonation fluctuations.

The weight calculation mechanism uses a scaled dot product function to define the similarity between modalities:

$$w_{i,j} = \frac{(q_i \cdot k_j)}{\sqrt{d}} \quad (3)$$

q is the facial action prediction query vector, k_j is the j frame speech embedding, and d is the embedding dimension. The calculated weight matrix is then normalised using softmax to determine the importance of each speech frame for the current expression action generation. The introduction of weight bias can further enhance the model's responsiveness in high-intensity intonation areas, giving the expression transformation a clear sense of temporal rhythm.

Using a multi-head attention mechanism enables the system to attend to multiple dimensions of speech types simultaneously and to consider the importance of different speech dimensions together (e.g., speech speed, pitch modulation, and rhythm). The multiple dimensions of the attention heads are converted into facial action control vectors

at the current time frame to enable detailed adjustments and maintain consistency in performing continuous speech-related actions.

2.2.2 *Synchronous fusion mechanism of emotional semantics and visual behaviour*

Emotion-related information can be derived from textual and auditory sources; however, there will be substantial inconsistencies in the degree of emphasis applied, depending on the context of each piece of emotion-based data. Therefore, to ensure that a Virtual Human can express emotions correctly ‘during’ an emotional display, an emotion-based attention-mapping path will be created, providing a common representation of text-based emotional labels, audio-based prosody parameters/information, and facial marker(s) spatial displacement patterns. This module uses a residual connection to transmit the visual state vector and applies an emotion perception mask to control the intensity of emotional impact.

During the mask generation process, the emotion vector is first converted into an attention offset factor via a fully connected network, and then superimposed with the visual modality output to produce the emotion regulation result for the current frame. This operation does not change the dimension of the original visual feature space, but only enhances or suppresses the expressiveness of specific actions in prominent positions. To further clarify the contributions of different modalities to emotion control, a gated attention mechanism is introduced to recalibrate the importance of each channel. The calculation method is as follows:

$$g_m = \sigma(W_m \cdot z_m + b_m) \quad (4)$$

z_m is the embedding vector of the m^{th} mode, W_m and b_m are learning parameters, and σ is the sigmoid function. The output g_m is used as a gating factor to adjust the control weight of the mode on the generated result. This mechanism effectively addresses imbalances in expression fluctuations caused by changes in the mode’s strength and maintains the consistency of the emotional level across the generated image.

During the entire fusion process, the alignment of the emotional dimension and the continuity of the visual action on the time axis are modelled simultaneously, avoiding sudden changes in expression or emotional interruption. The resulting multimodal collaborative features are input into the image generation model, driving the facial expressions to evolve naturally with the speech semantics and emotional intensity, achieving highly consistent audio-visual interactive performance.

2.3 *DM-driven image generation*

The U-Net architecture constructs the DM and controls the time step to produce images progressively. For input, the face’s first picture and the analysis of the fused multimodal features are used. Each step of back diffusion uses the noise prediction results from the current time step, thereby updating the detail level of the generated virtual human images in the series.

2.3.1 Noise sampling and initial state construction strategy

In the image generation task, to establish a reconstruction path from pure noise to a clear virtual human face image, a high-dimensional noise perturbation must be applied to each image frame. The perturbation process introduces noise incrementally at each time step, according to the specified time step, to construct multiple degraded image states. In the initialisation stage, a standard normal distribution is used to generate a random noise image, which is combined with a structural guidance map of the face area – obtained via a pre-trained facial landmark detector that produces a heatmap encoding keypoint positions – as a conditioning term and embedded into the latent space. The initial state contains the noise signal and is also attached to the encoded speech rhythm tensor, the emotion label tensor, and the facial key point set. The three are input into the potential representation channel after cross-modal fusion.

The perturbation process uses a fixed β scheduling strategy, gradually increasing a Gaussian perturbation to the state at each time step. The noise superposition process is modelled as follows:

$$x_t = \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1 - \bar{\alpha}_t} \varepsilon \quad (5)$$

x_0 represents the original image state, ε represents the sampling noise, and $\bar{\alpha}_t$ is the cumulative retention factor. This mechanism constructs a series of image paths from clarity to chaos, providing estimated gradients for subsequent reverse generation. Noise perturbation does not directly destroy the image structure. Still, it retains weak texture traces in salient areas, making it easier for the model in the reverse path to capture the mapping relationship between semantic information and visual details.

Reverse creation uses the created structural instruction guide as early as possible to construct a perceptual coordinate grid in high-dimensional space, allowing the gradual elimination of minor error details with each update relative to prior semantics. This structural framework ensures that the contour of the produced virtual human remains intact within each frame and that the emotional features of all frames are not deformed, distorted, or fragmented as they are produced over time.

2.3.2 Conditional guidance and time-series inversion update mechanism

During the image reverse generation process, the model removes noise at a fixed time step and embeds the multimodal fusion feature vector of the current moment at each step. This vector contains the speech-spectrum convolutional representation, text emotion encoding, and a face-point-posture tensor, which are input to the conditional channel of DM to guide the reverse denoising direction. The structural design uses a U-Net as the underlying feature transmission network, retains spatial relationships at intermediate scales, extracts coarse image contour features via the downsampling path, fuses detail residuals captured by the forward jump connection in the upsampling stage, and finally restores a realistic image frame.

The reverse process uses the method of predicting residual noise to iteratively perform noise correction operations:

$$\hat{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}} \left(x_t - \frac{1 - \alpha_t}{\sqrt{1 - \bar{\alpha}_t}} \cdot \varepsilon_0(x_t, c_t) \right) \quad (6)$$

$\varepsilon_\theta(x_t, c_t)$ is the residual noise predicted by the current model, and c_t is the conditional feature combination of the corresponding time step. This formula completes the state restoration from x_t to x_{t-1} at each step. The prediction network aims to perform residual regression, thereby learning to restore semantic clarity and facial structure stability from the local fuzzy state. The conditional features used to generate model outputs better emphasise parts of the face driven by emotion (eyebrows, eyes, and the corners of the mouth). This allows for a more natural progression of emotion and change in expression in relation to emotional tension.

The structure also incorporates a dynamic method for adjusting convolutional kernels and adaptive updating of network weights based on conditional features. Through this process, different strategies for generating detailed textures are applied depending on the current emotional state. For example, higher resolution textures at the corners of the mouth and eyelids would be produced when the emotion is high. Conversely, a greater emphasis is placed on the overall smoothness of the entire facial surface and lines when emotions are stable and/or low speed. As a result, the model output sequence will maintain the same contour across frames while simultaneously reinforcing the progressive logic of the emotions associated with the generated expressions.

Additionally, the generated output at each time step is self-consistent and maintains the same contour across all frames, while reinforcing the evolving logic of the emotional expression. Furthermore, the final generated video clips demonstrate a smooth transition of facial expressions and a precise fit between the voice emotion and the generated image, significantly enhancing the realism and interaction of the virtual human image.

2.4 *Emotion module integration*

An emotion classifier is included in the generation process to identify the emotion category of the input voice (in real-time) and input it into the DM as a conditional prompt to output a consistent expression by adjusting how intense the output will be and the Style used for that output.

2.4.1 *Emotion perception coding strategy and conditional embedding mechanism*

Rich acoustic signals exhibit rhythmic, pitch, and speech rate changes that are closely associated with the speaker's emotional state. Frame-level short-term speech segments are used to extract low-level features as units for emotion identification, including energy envelopes, cepstral coefficient sequences, pitch curves, and frequency band distribution curves; residual connections between them help stabilise variations in speed and noise during extraction. The feature tensor is processed in the time domain with a gated unit design, enabling the capture of emotional deformation trajectories over time. The bilinear mapping maps each output to a finite category space, thereby forming a static emotion-label prediction matrix. The bilinear mapping computes the outer product of the time-averaged feature vector and a learnt weight matrix, effectively capturing pair-wise relationships between temporal dynamics and spectral feature dimensions, creating a compact representation of emotion that captures time and features. The probability distributions over each category in the matrix indicate the dominant emotion in the current paragraph.

Basically, the predicted result does not mean you have finished classifying it; rather, it will be fed into the decision-making module at the current time step as a conditional prompt for real-time decisions. The DM may also modify its structural response to those predictive outputs by using the activation map at the end of each use case. The conditional prompt is inserted as a position-preserving vector into the mid-layer of the DM. It is structured using the following specific format:

$$z' = z + W_e \cdot y \quad (7)$$

z is the original feature tensor of the middle layer of the diffusion network, y is the emotion label mapping vector, and W_e is the emotion embedding transformation matrix. To evaluate the emotional aspect's spatial relation to the face, this formula reconstructs the original feature channel. Thus, allowing for more precise emotional correspondence between facial expression and the specified area related to its emotion's amplitude, e.g., the tension expression's eye, lip, and eyebrow arches, or the representation of physiological aspects (such as muscle tension and contour deviation) with respect to emotional state and time.

The time-stepping at each time step must keep in sync with the current size of the anti-trace disturbance; therefore, it enables real-time image generation and emotional flow synchronisation. Although the model uses discrete classification result signals as a primary method for controlling the signal to each of the four-dimensional magnitudes and identifying the four- to two-dimensional coordinates of each of the distinctive tonality states (statement, question, emphasis), it is used as a more real-time, dynamic instruction to provide guidance for the diffusion path in the emotional space and to allow for reconstructing with different style and amplitude. Furthermore, this mechanism enables the identification of tone-state differences (e.g., statement vs. question vs. emphasis) across various semantic segments. It supports the counterpoint coupling of speech emotion and facial emotion expression.

2.4.2 Expression amplitude regulation and structural response reconstruction mechanism

The emotional intensity level usually affects the realism of virtual facial expressions. A lack of regulation of emotional intensity may cause an additional problem: an exaggerated distortion or a stiff expression. To address that concern, the system will apply an amplitude regulation factor, allowing amplitude details to be dynamically adjusted and rhythmically altered in accordance with emotion-inferred research output expression actions. This adjustment procedure merges speech intensity features (i.e., the energy fluctuation rate with the feature flow mean gradient at the middle layer of the model) to create an amplitude modulator, and locally scales each U-Net sub-channel (via an upsampling weight) using its respective amplitude modulator. The feature flow mean gradient is computed from the face key regions (lips, eyes, and eyebrows) using a spatial mask based on facial landmark heatmaps, so that amplitude modulation is performed primarily in areas that are significant for emotional expressions.

The adjustment factor embedding form is as follows:

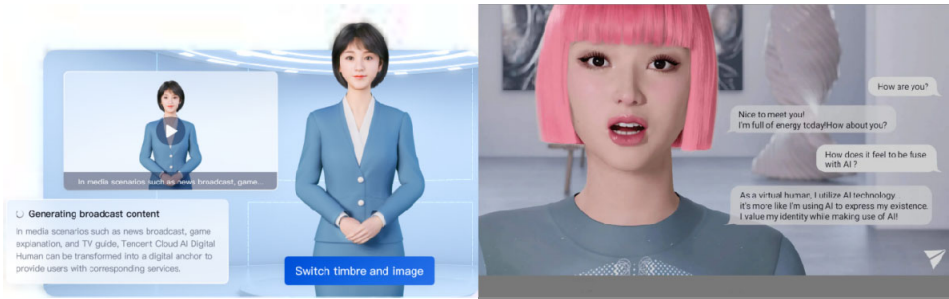
$$\tilde{F}_l = \gamma_l \cdot F_l + \beta_l \quad (8)$$

F_l is the l^{th} layer feature map, γ_l and β_l are the amplitude control coefficient and bias term, respectively. This control enables fine-grained intervention on the expression response dimension of the feature channel, so that the same emotion label yields style-adaptive expression patterns across different speech speeds or intonation conditions, thereby improving the diversity and naturalness of the virtual human face generation process.

The system maintains expression consistency by tracking the location of the last activation of the key expression area across time steps. Based on this information, the prediction area (i.e., current key area) is compared with the previously tracked key area. In addition, if mutations occur in the key area during continuous prediction of the same key expression area, then the model can reduce the intensity of the current key expression area, prioritising to keep the local structure shaped by the active key area in the prior time step, and avoid abrupt changes to the structure of the expression from sudden emotional changes resulting in stiffness.

In the final sequence generated, the expression’s rhythm aligns with speech, and face features transition gradually between styles and are logically connected across all emotional contexts. Thus, this avoids the frame-skipping challenge of abrupt changes in emotion. The integrated structure makes the model sensitive to the details of emotion in speech and ultimately provides the virtual human with greater emotional credibility. Experimentation on the application of virtual humans continues across different scenarios, such as intelligent customer service and virtual anchors, as illustrated in Figure 3.

Figure 3: Virtual human application scenario (see online version for colours)



3 Method effect evaluation

3.1 Experimental data and settings

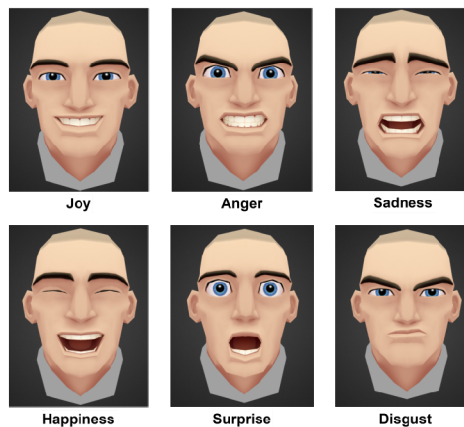
This study examines the performance of synthetic virtual humans in multi-channel input contexts – including audio (voice), written text, and visual information – for validating the models used to create them. Speech data was collected from the VoxCeleb2 Dataset, which contains many different speech types, characteristics, and acoustic environments (speed, accent, and background noise), as well as emotion labels from the IEMOCAP database for all respective written and/or spoken outputs. Facial keypoint detection was performed on all SCT video sequences from the 300-VW dataset to assess the quantitative accuracy of expression detail across all timeframes used to create these

virtual humans. The speech signal is processed frame-by-frame, and time-frequency domain features are extracted using a convolutional layer. A bidirectional long short-term memory network (BiLSTM) is then used to encode semantic and rhythmic information, ensuring synchronisation between speech rhythm and expression changes. The text input is fed into the pre-trained sentiment analysis model to generate a normalised sentiment vector that regulates the generated style. Visual guidance uses high-precision facial keypoint detection across key areas such as the eyes and lips to enhance precise control over expression details.

The model is trained E2E, DM is built on a U-Net architecture, and the noise scheduling uses a fixed β strategy that gradually introduces Gaussian perturbations to construct a degradation path. In the reverse generation stage, multimodal fusion features serve as conditional inputs to optimise expression details via a dynamic convolutional kernel adjustment mechanism. The emotion classification module analyses the intensity of speech emotions in real-time and is embedded in the middle layer of DM to enhance the consistency of expression and emotion.

The main evaluation indicators cover expression naturalness, lip synchronisation rate, visual coherence, and emotional consistency. The naturalness of expression is quantified by the Kullback-Leibler (KL) divergence of the distribution of action units (AUs) synthesised on six types of emotional expressions and the real data. The model's ability to generate expression synthesis is demonstrated in the results depicted in Figure 4. The lip-sync rate is the percentage of audio frames that correspond to lip movements, as calculated by the SyncNet model. Visual coherence is determined by measuring stability between frames using the structural similarity index (SSIM) and the peak signal-to-noise ratio (PSNR). Emotional consistency is measured as the percentage of matched emotions between expressions and manually rated emotions. The testing phase covered a variety of challenging scenarios, including standard speech rate, multiple dialects, and background noise, to assess the model's robustness.

Figure 4 Expression synthesis effect display (see online version for colours)



The core modules of MAMs include the audio-driven expression mapping and the emotion and visual fusion models, as detailed in Table 1. Collectively, these parameters achieve an accurate representation of rhythm and intonation variances by setting the embedding dimension and the number of attention heads. The numerical stability of the

model’s attention distribution is controlled by the scaling factor and the softmax temperature. The gating factor and residual weight of the fusion portion of the model will improve information flow and the strength of fusion during cross-modal feature integration. The parameters exhibited high convergence and generalisability during the training phase; consequently, they are well-suited for generating virtual human imagery from high-dynamic speech input.

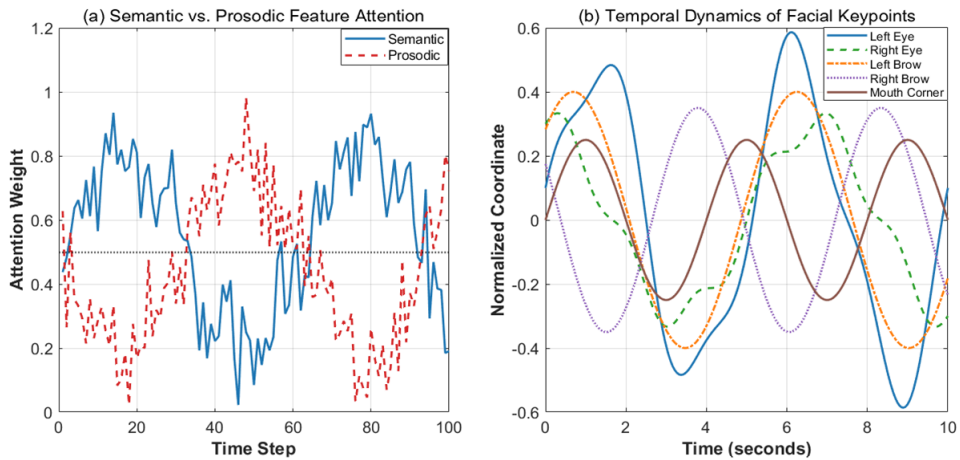
Table 1 MAM module core parameters

<i>Module</i>	<i>Parameter name</i>	<i>Value/range</i>	<i>Description</i>
Audio-to-expression attention	Embedding dimension	128	Ensures balance between expressiveness and efficiency
	Number of attention heads	8	Enhances the diversity of feature representation
	Scaling factor	11.31	Stabilises dot-product operations
	Softmax temperature	1	Controls the smoothness of weight distribution
	Weight bias range	± 0.05	Enhances sensitivity to prosodic variation
Emotion-visual fusion	Emotion vector dimension	64	Compacts emotional features for efficient fusion
	Fully connected neurons	128	Strengthens nonlinear mapping ability
	Gating factor range	[0.1, 0.9]	Constrains the modal weight fluctuation
	Residual connection coefficient	0.85	Preserves core information flow
	Batch size	32	Balances computational load and gradient stability
	Learning rate	2.00E-04	Ensures stable convergence during training

3.2 *Dynamic representation of multimodal features in virtual human generation*

Multimodal features can be sampled with their corresponding weight dynamic calculated over time. Specifically, a speech signal can be decomposed into different time frames, and the semantic and rhythmic characteristics of those frames can be extracted using a pre-trained model. The self-attention mechanism is then used to compute the weight distribution of these features at discrete time points, illustrating how the model’s focus changes across the spoken content throughout the process. At the same time, high-precision face recognition tools are used to extract facial key points, enabling us to observe fine movements in specific facial regions over time and further evaluate dynamic changes in visual features. All of the aforementioned data is then plotted in MATLAB as three-dimensional time-series trajectories, with weights indicating the correlation between modality inputs and their impact on visual performance throughout the virtual human generation process, to ensure correct synchronisation between facial expressions and the rhythm of speech.

Figure 5 Dynamic performance of multimodal features in virtual human generation (see online version for colours)



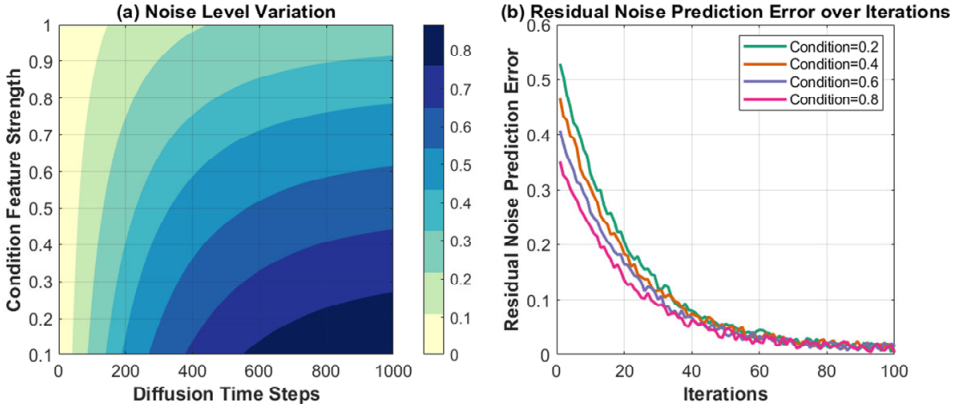
The dynamic changes in attention weights across all modal features (i.e., the various modalities used to convey messages) and facial region responses during generation are shown in Figure 5. This is evident in Figure 5(a), where the attention weights for semantic and prosodic features alternate between strengthening and weakening; this reflects the different attention strategies the model employs when attending to speech content. When semantic content is highly valued, the model tends to increase the weight assigned to semantic features, thereby improving accuracy in conveying the content's meaning. When there are large fluctuations in intonation or a rapid increase in the rhythm, the model assigns a higher guiding weight to rhythmic features, thus providing a stronger visual representation of the emotional rhythm. The range of the weighted fluctuations for both features is between zero and one. This cross-regulatory mechanism indicates that the model can adaptively allocate attentional resources to complex speech signals to generate more natural lip motion and facially synchronised motion. The graphical data in Figure 5b shows the trajectory (trajectory = direction of movement) over time of the coordinate changes of a standard human face and the different areas of the human face during the generation process (face generation). These trajectories demonstrate how dynamic (i.e., capable of changing motion) each facial area is due to the drive of speech events (speech-driven). Each facial area has a distinct movement pattern, yet all exhibit strong continuity of movement. The rhythmic differences in movement among the eye, eyebrows, and mouth indicate that the model has effectively captured the coupling and decoupling of the facial areas. The coordinates of the key points in the facial areas range from -0.6 to 0.6 in both the x- and y-directions. This type of multidimensional dynamic pattern enhances the subtlety of emotional expression and verifies that the visual guidance features are fully responsive to the temporal dimension. Thus, the results of applying the model demonstrate its strong ability to capture the key elements of cross-modal input and map them to stable, coherent visual sequences.

3.3 Error convergence analysis of DM noise dynamics and conditional constraints

The process of multiple experimental sampling of DMs across various time phases and strength conditions can be performed. Noise levels are determined by simulating the forward diffusion phase of the diffusion modelling process with non-conditional features. Then, the noise distributions (amplitude) at each time phase can be measured to allow comparative analysis of performance, using the comparative strength of the feature as a multimodal condition variable. The residual noise prediction error is computed as the difference between the model output from the reverse diffusion phase and the true image (real-world image), indicating how well the model has converged after training iterations. After many trials of varying the strengths of conditional parameters, I confirmed the trend of classification of suppression of noise (lowering occurrence of noise) and decreasing residual error (lowering the occurrence of error) to show whether the application and guidance of multimodal conditional information can effectively support and improve the prediction of noise and enhance the generation of detail and stability in DM-driven images. This provides quantitative evidence of the performance of DM-driven images.

Figure 6 shows the error convergence under DM noise dynamics and conditional constraints.

Figure 6 DM noise dynamics and error convergence under conditional constraints (see online version for colours)



The diffusion-time-step and feature-intensity dimensions' structures are represented in the two-dimensional heat map of noise levels in Figure 6(a). The noise level accumulates over time as the diffusion time step increases, thereby increasing its impact on image detail. For diffusion time-steps larger than 600, the increase in the amount of conditional feature strength leads to a decrease in the noise-growth-rate of the features. It thus indicates that the conditional information plays an important role in controlling the diffusion process and in reducing the effect of noise on the final image quality. This suggests that the fusion of multiple modal features can provide the model with enhanced guidance information, thereby improving the path and level of noise propagation. In Figure 6(b), the attenuation curves of residual noise prediction errors reflect DM's convergence performance under different conditional constraints. As such, the curves exhibit an exponential decay pattern when displayed graphically. Therefore, when trained

iteratively, the noise prediction error decreases gradually while the model's noise estimation accuracy improves. As the conditional parameter moved from 0.2 to 0.8 during training, the strength of the conditional signal increased, resulting in lower initial error, faster convergence, better detail, and greater stability and precision in the back diffusion process. In summary, there is an approximate correspondence between the amount of conditional information provided and improvements in the detail restoration and natural appearance of generated images. The changing trend of the data reveals the essential mechanism of the synergy between multimodal features and conditional guidance in DM, effectively improving the quality and expressiveness of virtual human image generation, and providing a scientific basis for dynamic adjustment in the generation process.

3.4 Expression naturalness

The probability distribution of AUs corresponding to six basic emotions (joy, anger, sadness, happiness, surprise, and disgust) was extracted from the real human face emotion dataset as a benchmark reference. The trained DM-MAM, Wav2Lip, StyleGAN3, and CycleGAN models were used to generate virtual human face image sequences corresponding to different emotions, and the distribution of AUs in each frame was extracted. For each type of emotion, by comparing the KL divergence between the model-generated distribution and the benchmark distribution, the difference in muscle movement patterns between the generated expressions and the real expressions is quantified. Finally, the divergence results of each model under the six emotions are visualised as a heat map.

Figure 7 KL divergence between synthetic and real AU distributions (see online version for colours)

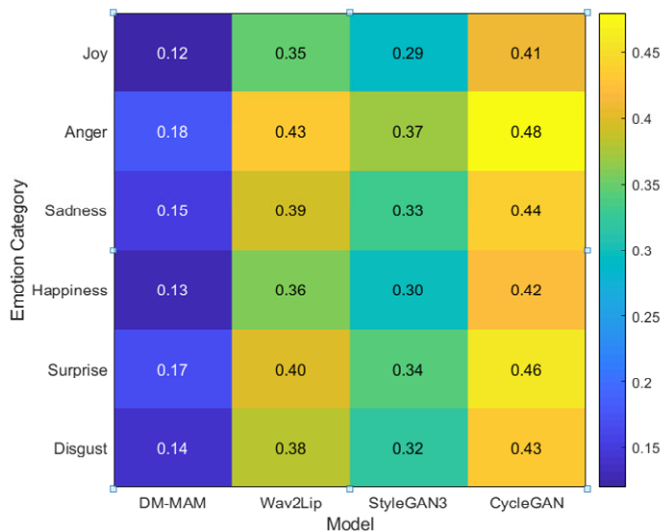


Figure 7 intuitively shows the differences in KL divergence between the distributions of facial AUs generated by different models and those of real human data across six basic emotions. Overall, the DM-MAM model shows low KL divergence across emotion

categories, ranging from 0.12 to 0.18, indicating that its synthetic expressions are closer to real expressions at the muscle movement level and exhibit strong naturalness and detail-restoration ability. In contrast, the divergence values of Wav2Lip and StyleGAN3 are generally higher, reflecting that they have certain limitations in capturing subtle facial muscle changes and emotional details, especially when expressing strong emotions such as anger and surprise, the deviation is more obvious. The divergence shown by CycleGAN is generally large, indicating that the generated expressions differ greatly from the real data in terms of AU activation patterns. This may be due to the generation mechanism's insufficient modelling of temporal dynamic features, resulting in the expression being incoherent and unnatural. The divergence between emotion categories also reveals differences in the model's adaptability in expressing them. The muscle movement patterns of emotions such as joy and happiness are relatively stable, and the model fitting effect is relatively good; the change in divergence reflects the different impacts of each model on the extraction and fusion mechanism of emotional features when processing multimodal input. The lower KL divergence ensures the realism of the virtual human's expression and promotes the coherence and fluency of its movement, further enhancing the user's immersive experience. The data results reveal the key role of improving the accuracy of AUs in enhancing the interactive expressiveness of virtual humans.

3.5 *Lip shape and speech synchronisation rate*

The synchronisation rate of the lip movements in each frame of a generated video and the audio of that frame is calculated by extracting a feature set with the pre-trained SyncNet model, aligning the lips to the audio using similarity measures, and counting the total number of synchronised frames. The resulting synchronisation rate is the number of successfully matched frames divided by the total number of frames. Three conditions will be used to test this methodology: normal Mandarin speech (rather than accented Mandarin), accented Mandarin, and Mandarin with background noise; these conditions represent a wide range of speech environments. Testing of groups of samples will be repeated under each condition to collect data on synchronisation rates, which will then be quantified and averaged, and the standard deviation, maximum, and minimum values will be calculated. As such, the degree of agreement (aligning) between lip movements and the content of speech will be quantitatively assessed to demonstrate the robustness and accuracy of speech synchronising across various models in complex environments and reduce subjectivity in evaluating the process, thus increasing the scientific rigor in which the evaluation will take place.

Table 2 summarises the comparative performance of lip-shape/phrase synchronisation across four models under three complex-speech input conditions (i.e., standard, Mandarin-speaking rate, dialectal accent, and background noise). Each model's synchronisation rate was calculated as the proportion of lip movements matched to speech (i.e., number of lip movements/speech-aligned video frames) using the pre-trained SyncNet. The average, median, standard deviation, minimum, and maximum values were calculated and used to evaluate how well each distribution metric represented the data's central tendency and variability. Compared to Wav2Lip, StyleGAN3, and CycleGAN, DM-MAM produced the highest average (94.2%, 92.1%, 90.6%) and median synchronisation rates for each condition, and had smaller standard deviations; thus, the generated lip shapes were more consistently synchronised to the input speech and less

impacted by variations in speech rates, dialectal influences, and background noise conditions. Whereas Wav2Lip produces relatively consistent synchronisations, its performance declines noticeably in non-standard conditions. In contrast, StyleGAN3 and CycleGAN showed inconsistent lip-to-speech synchronisation accuracy, primarily in background-noise conditions; the variance in the highest and lowest synchronisation rates showed that both performed poorly under disruptive audio conditions compared to DM-MAM and Wav2Lip. The data reflect differences in the synchronisation response capabilities of each model across a variable speech environment, providing a quantitative basis for analysing the model's robustness and adaptability.

Table 2 Comparison of lip shape and speech synchronisation performance

<i>Input condition</i>	<i>Model</i>	<i>Mean synchronisation rate (%)</i>	<i>Median (%)</i>	<i>Standard deviation (%)</i>	<i>Minimum (%)</i>	<i>Maximum (%)</i>
Standard mandarin	DM-MAM	94.2	94.5	1.3	91.8	96
	Wav2Lip	91.5	91.8	1.7	88.4	93.7
	StyleGAN3	87.8	88.1	2.1	83.5	90.2
	CycleGAN	83.6	83.3	2.8	78.9	86.5
Dialectal accent	DM-MAM	92.1	92	1.6	89.1	94.8
	Wav2Lip	89.3	89.5	2	85.4	91.9
	StyleGAN3	85.4	85.1	2.6	80.2	88.7
	CycleGAN	80.7	80.5	3	75.1	84.9
Background noise	DM-MAM	90.6	90.9	1.8	87	93
	Wav2Lip	86.7	87.1	2.3	82.9	89.2
	StyleGAN3	82.3	82.6	2.9	77.5	85.9
	CycleGAN	78.1	77.8	3.5	72	81.6

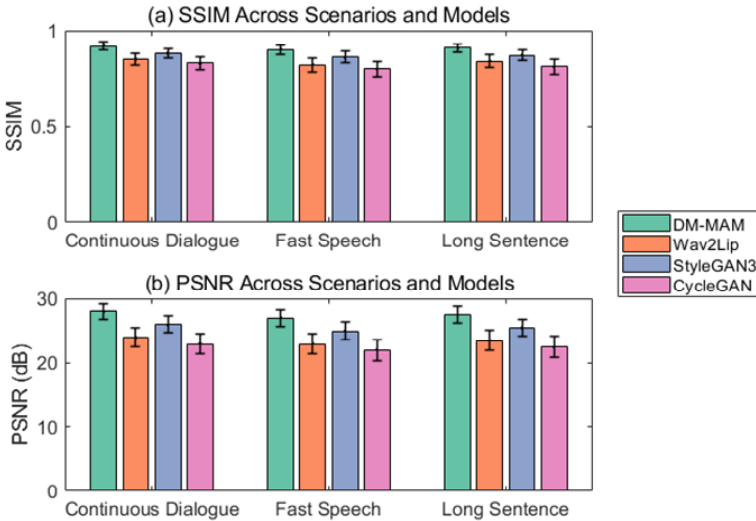
3.6 Visual coherence

For purpose of illustrating the performance of the Farneback optical flow method across three different types of video samples: conversationally, fast speech and input of long sentences (comparing the extracted motion vector field using two consecutive frames) and measuring the pixel-level change in each video frame through use of the SSIM and PSNR indices calculated on pairs of frames; to create a common evaluation metric based on the means and standard deviations of each sample group as defined by the testing objects that were constituted of four distinct models (DM-MAM and three others – Wav2Lip, StyleGAN3 and CycleGAN) using the same audio and text as inputs for a level playing field comparison; using the standard face synthesis benchmark set for each of the test matrices to prevent any data biasness from effecting the outcome of the experiments.

In Figure 8, the performance of various models was compared across three scenes based on SSIM and PSNR scores. SSIM measures the degree of preservation of visual detail between two video frames, and PNRS measures the quality and noise of an entire image. DM-MAM performed well with respect to both SSIM and PNRS in each of the three different scenes studied – continuous dialogues, fast speech rate, and long sentence input. The average SSIMs were 0.92, 0.90, and 0.91, and the average PNRS was about 28, 27, and 27.5 dB, respectively, indicating that DM-MAM produced more detailed,

coherent, high-quality images than the other models tested. In the continuous dialogue scene, all models performed similarly because the speech rate was relatively uniform throughout the video, resulting in smooth frame-to-frame transitions and minimal risk of visible jumps. Dynamic modes of action (DM-MAM), used in fast-speech cases, introduce tighter time constraints and increased demands on frame synchronisation and motion estimation, compromising SSIM and PSNR for many systems, indicating sensitivity to rapidly changing dynamic activity at high speed. When input to the system consists of long sentences, the complexity of the input's semantics increases the demand placed on facial systems to detect subtle differences in facial expressions. Because the DM-MAM method uses a multimodal fusion mechanism to coordinate the details of expression generation, the system produces facial images with high visual coherence. There is a measurable standard deviation across different video samples, indicating that the system's generated facial expressions exhibit some variability; thus, the system can produce facial expressions of varying quality due to the complexity of the input data and its robustness. The results of this study demonstrate the value of multimodal fusion techniques and dynamic adjustment mechanisms for enhancing overall visual stability in generated media, thus providing a basis for algorithm optimisation.

Figure 8 Comparison of SSIM and PSNR (see online version for colours)



3.7 Consistency of emotional expression

Evaluation of the consistency of emotional expression requires comparing the model's predictions against human-specified annotations, and six basic emotions serve as the basis for the test objects. The speech input generated for the video has been systematically assigned to emotions and compared with human annotations frame by frame. There were multiple rounds of speech input to account for variability in the sample data, thereby achieving statistical significance. The matching rate can be interpreted as a measure of the model's ability to recognise and respond appropriately to different types of emotion, relative to human recognition. Overall, this confirms the

fidelity and reliability of the generation systems in accurately producing emotion. Data collection and statistical analysis are strictly controlled to minimise variability by preventing external influences from affecting results and to ensure the overall evaluation process has a scientific basis for substantiation.

Table 3 Emotional expression consistency

<i>Emotion category</i>	<i>DM-MAM match rate (%)</i>	<i>Wav2Lip match rate (%)</i>	<i>StyleGAN3 match rate (%)</i>	<i>CycleGAN match rate (%)</i>
Happiness	92.3	88.1	83.7	79.5
Anger	90.7	85.4	81.2	77.8
Sadness	89.5	84.3	80.1	76.4
Joy	91.8	87.6	82.9	78.3
Surprise	88.9	83.7	79.6	75.1
Disgust	87.2	82.1	78.4	74

Table 3 presents a quantitative comparison of the consistency of emotional expression across the four generative models for the six basic emotions (joy, anger, sadness, happiness, surprise, and disgust). Model accuracy is assessed by comparing manually annotated emotion categories against predicted model outputs (matching rates). DM-MAM has demonstrated very strong performance in this area, with exceptionally high rates of match (joy – 92.3% / disgust – 87.2%), indicating the broad scope of emotions that can be modelled effectively with DM-MAM and the ability to accurately predict different types of emotion over time. Wav2Lip had the second-highest overall performance but still produced an average match rate slightly lower (82%–70%) than those of the other two models. Both StyleGAN3 and CycleGAN showed notable disparities in match rates, ranging from 80% to 75%, suggesting that these models do not accurately restore emotion-related expressions, particularly for more complex emotions such as disgust and surprise. Overall, the outcomes indicate that the flexibility DM-MAM provides in offering a variety of emotional inputs will produce greater consistency between a virtual human’s facial expressions and the emotions in their voice during interpersonal exchanges, resulting in more natural interactions between individuals.

4 Conclusions

This study aims to address the challenges of insufficient multimodal information fusion and unnatural expression generation, and constructs an interactive virtual human synthesis framework that combines multimodal attention and DM. Speech, semantic, and prosodic feature decoding, text emotion label extraction, and facial keypoint capture are combined into a unified feature vector to achieve deep coupling among the three modalities. The cross-modal attention network fine-tunes the weights of audio, text, and visual features, effectively coordinating the synchronisation of expressions and speech rhythm. DM uses the U-Net architecture to perform gradient-based self-consistent image reconstruction, and the noise scheduling and conditional prompts account for both global contours and local details. The final generated virtual human image sequence has significantly improved inter-frame coherence and emotional expression. After the emotion classifier is integrated into the generation process, the expression intensity and

style respond dynamically to the emotional state of the input voice, and the emotion switching process is natural and smooth. In the synthesis of six categories of emotions, the KL divergence between the distribution of model AUs and real AUs is between 0.12–0.18, and the matching rate of emotion categories is between 87.2% and 92.3%; the average synchronisation rate of lip shape and speech and the mean SSIM of each frame pair performed well, improving the multimodal information fusion and the naturalness of expression generation. Compared with existing audio-driven and image-generated solutions, this method is more robust in detail restoration and emotional interaction. In the future, it can be expanded to multi-person interaction scenarios and high-dimensional emotional spaces – specifically, continuous emotional dimensions such as valence and arousal – to enable finer-grained control over emotional expression and further improve micro-expression capture and continuous dialogue logic control. The research results provide a new approach to achieving immersive applications and intelligent human-computer interaction, with broad prospects for engineering deployment and industrialisation (Zeng et al., 2025).

Funding

- 1 Soft Science Research Program of Zhejiang Province, Project Number: 2025C35083, Project Name: Research on the Construction Path of AI Virtual Human-Based Science and Technology Popularisation in Zhejiang Province.
- 2 Research Project of Zhejiang Federation of Humanities and Social Sciences, Project Number: 2026B107, Project Name: Research on the Empowerment Mechanism of AI Virtual Humans in Live-Streaming E-Commerce Marketing in Zhejiang Province.
- 3 Research Findings of the Zhejiang Business College of Commerce Research Funding Sponsored Project, Project Number: SZYYB202402, Project Name: Research on AI Virtual Humans Empowering Live-Streaming E-commerce Marketing in Zhejiang.

Declarations

The authors declare that there are no conflicts of interest regarding the publication of this article.

References

- Abich IV, J., Parker, J., Murphy, J.S. et al. (2021) ‘A review of the evidence for training effectiveness with virtual reality technology’, *Virtual Reality*, Vol. 25, No. 4, pp.919–933.
- Ahuja, K., Ofek, E., Gonzalez-Franco, M. et al. (2021) ‘Coolmoves: user motion accentuation in virtual reality’, *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, Vol. 5, No. 2, pp.1–23.
- Alexanderson, S., Nagy, R., Beskow, J. et al. (2023) ‘Listen, denoise, action! Audio-driven motion synthesis with diffusion models’, *ACM Transactions on Graphics (TOG)*, Vol. 42, No. 4, pp.1–20.

- Alnagrat, A., Ismail, R.C., Idrus, S.Z.S. et al. (2022) 'A review of extended reality (XR) technologies in the future of human education: current trend and future opportunity', *Journal of Human Centered Technology*, Vol. 1, No. 2, pp.81–96.
- Cao, Q., Yu, H., Charisse, P. et al. (2023) 'Is high-fidelity important for human-like virtual avatars in human-computer interactions?', *International Journal of Network Dynamics and Intelligence*, Vol. 2, No. 1, pp.15–23.
- Chai, D.S. and Park, S. (2022) 'The increased use of virtual teams during the Covid-19 pandemic: implications for psychological well-being', *Human Resource Development International*, Vol. 25, No. 2, pp.199–218.
- Chamola, V., Bansal, G., Das, T.K. et al. (2024) 'Beyond reality: the pivotal role of generative AI in the metaverse', *IEEE Internet of Things Magazine*, Vol. 7, No. 4, pp.126–135.
- Chang Chien, C.Y. (2024) 'Design and research on the use of AI-based image recognition technology for collaborative campus safety management', *Journal of Information and Computing*, Vol. 2, No. 4, pp.47–58.
- Chen, X. and Chen, H. (2023) 'Emotion recognition using facial expressions in an immersive virtual reality application', *Virtual Reality*, Vol. 27, No. 3, pp.1717–1732.
- Chen, Y., Yao, Y., Li, Z. et al. (2025) 'Hyperlips: hyper control lips with a high-resolution decoder for talking face generation', *Applied Intelligence*, Vol. 55, No. 2, pp.1–16.
- Cui, J., Wan, Q. and Shin, S. (2025) 'Human-AI integration and sound-vibration technology-driven enterprise digital transformation: The mediating role of technological innovation', *Sound & Vibration*, Vol. 59, No. 1, pp.1733–1733.
- Demirel, H.O., Ahmed, S. and Duffy, V.G. (2022) 'Digital human modeling: a review and reappraisal of origins, present, and expected future methods for representing humans computationally', *International Journal of Human-Computer Interaction*, Vol. 38, No. 10, pp.897–937.
- Di, X. and Zheng, X. (2022) 'A meta-analysis of the impact of virtual technologies on students' spatial ability', *Educational Technology Research and Development*, Vol. 70, No. 1, pp.73–98.
- Ding, W., Xu, C., Arief, M. et al. (2023) 'A survey on safety-critical driving scenario generation – a methodological perspective', *IEEE Transactions on Intelligent Transportation Systems*, Vol. 24, No. 7, pp.6971–6988.
- Golf-Papez, M., Heller, J., Hilken, T. et al. (2022) 'Embracing falsity through the metaverse: the case of synthetic customer experiences', *Business Horizons*, Vol. 65, No. 6, pp.739–749.
- Gupta, S., Kumar, P. and Tekchandani, R.K. (2023) 'Facial emotion recognition-based real-time learner engagement detection system in online learning context using deep learning models', *Multimedia Tools and Applications*, Vol. 82, No. 8, pp.11365–11394.
- Jia, Z., Zhang, Z., Wang, L. et al. (2024) 'Human image generation: a comprehensive survey', *ACM Computing Surveys*, Vol. 56, No. 11, pp.1–39.
- Jiang, Y., Yang, S., Qiu, H. et al. (2022) 'Text2human: text-driven controllable human image generation', *ACM Transactions on Graphics (TOG)*, Vol. 41, No. 4, pp.1–11.
- Kang, B. and Jo, Y. (2024) 'StyleGAN-based advanced semantic segment encoder for generative AI', *IT Professional*, Vol. 26, No. 2, pp.17–23.
- Lee, H.J., Lee, S., Kim, H.G. et al. (2025) 'Analyzing talking head generation under real-world audio conditions', *TECHART: Journal of Arts and Imaging Science*, Vol. 12, No. 2, pp.18–24.
- Li, S.W., Kemp, M.W., Logan, S.J.S. et al. (2023) 'ChatGPT outscored human candidates in a virtual, objective, structured clinical examination in obstetrics and gynecology', *American Journal of Obstetrics and Gynecology*, Vol. 229, No. 2, pp.172.e1–172.e12.
- Liang, H., Zhang, W., Li, W. et al. (2024) 'Intergen: diffusion-based multi-human motion generation under complex interactions', *International Journal of Computer Vision*, Vol. 132, No. 9, pp.3463–3483.

- Lin, Y. and Yu, Z. (2025) 'Learner perceptions of artificial intelligence-generated pedagogical agents in language learning videos: embodiment effects on technology acceptance', *International Journal of Human-Computer Interaction*, Vol. 41, No. 2, pp.1606–1627.
- Liu, D., Chen, Y. and Wu, Z. (2023) 'Digital twin (dt)-cyclegan: enabling zero-shot sim-to-real transfer of visual grasping models', *IEEE Robotics and Automation Letters*, Vol. 8, No. 5, pp.2421–2428.
- Liu, D., Li, Z., Wu, Z. et al. (2025) 'Digital twin/MARS-CycleGAN: enhancing sim-to-real crop/row detection for MARS phenotyping robot using synthetic images', *Journal of Field Robotics*, Vol. 42, No. 3, pp.625–640.
- Luo, Z., Shao, X. and Ma, X. (2024) 'Enhancing learners' performance in contest through knowledge mapping algorithm: the roles of artificial intelligence and blockchain in scoring and data integrity', *Journal of Organizational and End User Computing*, Vol. 36, No. 1, pp.1–21.
- Melnik, A., Miasayedzenkau, M., Makaravets, D. et al. (2024) 'Face generation and editing with StyleGAN: a survey', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 46, No. 5, pp.3557–3576.
- Niu, L., Xie, W., Wang, D. et al. (2024) 'Audio2AB: audio-driven collaborative generation of virtual character animation', *Virtual Reality & Intelligent Hardware*, Vol. 6, No. 1, pp.56–70.
- Shaikh, M.B., Chai, D., Islam, S.M.S. et al. (2024) 'From CNNs to transformers in multimodal human action recognition: a survey', *ACM Transactions on Multimedia Computing, Communications and Applications*, Vol. 20, No. 8, pp.1–24.
- She, Y., Liu, F., Yang, B. et al. (2025) 'Converging real and virtual: embodied intelligence-driven immersive VR biofeedback for brain health modulation', *IEEE Transactions on Computational Social Systems*, Vol. 12, No. 3, pp.938–946.
- Slovak, P., Antle, A., Theofanopoulou, N. et al. (2023) 'Designing for emotion regulation interventions: an agenda for HCI theory and research', *ACM Transactions on Computer-Human Interaction*, Vol. 30, No. 1, pp.1–51.
- Sorosrungruang, T., Ameen, N. and Hackley, C. (2024) 'How real is real enough? Unveiling the diverse power of generative AI-enabled virtual influencers and the dynamics of human responses', *Psychology & Marketing*, Vol. 41, No. 12, pp.3124–3143.
- Walker, M., Phung, T., Chakraborti, T. et al. (2023) 'Virtual, augmented, and mixed reality for human-robot interaction: a survey and virtual design element taxonomy', *ACM Transactions on Human-Robot Interaction*, Vol. 12, No. 4, pp.1–39.
- Wang, Y., Wang, L. and Siau, K.L. (2025) 'Human-centered interaction in virtual worlds: a new era of generative artificial intelligence and metaverse', *International Journal of Human-Computer Interaction*, Vol. 41, No. 2, pp.1459–1501.
- Zeng, C., Wu, W., Yao, J., Xia, J., Li, X., Wang, J. and Chen, S. (2025) 'AI judge assistant: a new upgrade of the enabling judicial system', *Journal of Intelligence Technology and Innovation*, Vol. 3, No. 2, pp.74–92.
- Zhang, J., Xie, Y., Ding, W. et al. (2023) 'Cross on cross attention: deep fusion transformer for image captioning', *IEEE Transactions on Circuits and Systems for Video Technology*, Vol. 33, No. 8, pp.4257–4268.
- Zhao, L., Bi, Z., Hawbani, A. et al. (2022) 'ELITE: an intelligent digital twin-based hierarchical routing scheme for software-defined vehicular networks', *IEEE Transactions on Mobile Computing*, Vol. 22, No. 9, pp.5231–5247.
- Zhou, Y., Huang, H., Yuan, S. et al. (2023) 'MetaFi++: Wifi-enabled transformer-based human pose estimation for metaverse avatar simulation', *IEEE Internet of Things Journal*, Vol. 10, No. 16, pp.14128–14136.
- Zhu, Y. and Li, N. (2021) 'Virtual and augmented reality technologies for emergency management in the built environments: a state-of-the-art review', *Journal of Safety Science and Resilience*, Vol. 2, No. 1, pp.1–10.