Punchline-Driven Hierarchical Facial Animation via Multimodal Large Language Models

Na Wang
College of Engineering and Technology, Hubei University of Technology, Nanli Road, Hongshan District, Wuhan, Hubei Province, 430000, China
E-mail: RivasRebecca6760@outlook.com

Keywords: speech-driven animation, 3D facial animation, large language models (LLMs), multimodal understanding, punchline detection, regional animation, deep learning

Received: September 3, 2025

Speech-driven 3D facial animation has achieved high phonetic realism, but current models often fail to convey the expressive peaks, such as punchlines, that are critical for engaging communication. This paper introduces a novel framework that addresses this gap by leveraging a Multimodal Large Language Model (MLLM) for a deep, semantic understanding of speech. Our core innovation is a system that explicitly models and animates the climax of an utterance. The framework first employs a multimodal punchline detection module to identify moments of high expressive intent from both acoustic and textual cues. This signal guides our Punchline-Driven Hierarchical Animator (PDHA), which functionally decomposes the face into distinct regions and generates motion in a coordinated cascade, allowing the punchline to dynamically amplify expression in the upper face while preserving articulatory precision in the mouth. A final cross-modal fusion decoder refines the output for precise temporal alignment. Comprehensive experiments on the VOCASET dataset show that our model not only sets a new state-of-the-art in geometric fidelity, reducing Vertex Error by 7.8% compared to the state-of-the-art FaceFormer baseline, but is also rated as significantly more expressive and natural in user studies ($p < 0.01$), confirming its ability to capture the emotional impact of a punchline.

Povzetek: Opisan je hierarhični model za govorjeno 3D animacijo obraza, ki z multimodalnimi velikimi jezikovnimi modeli zazna in poudari izrazne vrhove govora. Metoda izboljša izraznost in naravnost animacije ob hkratnem ohranjanju visoke fonetične točnosti.

1 Introduction

The advent of Multimodal Large Language Models (MLLMs) is a significant breakthrough in the pursuit of artificial general intelligence, demonstrating emergent capabilities that were once a preserve of cognition [1]. Models such as GPT-4V have shown a remarkable ability to process and reason on concatenated image and text inputs to accomplish tasks ranging from image-to-story generation to solving math questions without explicit optical character recognition (OCR) [2]. This convergence of modalities has driven a tide of innovation, with academia and industry both creating more advanced MLLMs that challenge the state of the art in machine perception and generation [3, 4]. The key advantage of these models is that they leverage the strength of an exceptionally powerful Large Language Model (LLM) as a cognitive core, which enables them to conduct tasks such as complex, open-ended multimodal tasks. This has prompted the development of open-source models, which are closing the capability gap with commercial proprietary systems more and more, democratizing access to this groundbreaking technology [5].

One particularly challenging and fascinating border region for MLLMs is grasping complex communication, such as sarcasm and humor, which tend to be presented as a subtle mixture of visual and verbal information. The "punchline" of a joke is a prototypical multimodal event whose humor stems from the accurate juxtaposition of a picture and a caption. Understanding of such content transcends literal meaning; it requires a deep level of understanding of context, common world knowledge, and the ability to make implicit connections. The most recent research has already begun to compare MLLMs' ability to recognize these multimodal punchlines and determined that, while there are strong models available, there is still a significant gap between their recognition and natural understanding [6]. This supports the need for models that not only detect but also reason about the subtle interplays between different modalities. Development of MLLMs that are capable of applying contextual object detection, in which visual objects are associated with language in interactive contexts, is a move in this direction, moving beyond captioning to a more grounded form of comprehension [7].

The potential for the application of state-of-the-art MLLMs is vast, from systemically understanding static pictures to understanding the temporal, dynamic character of video content. Large video-text datasets and models are being created to learn rich representations for tasks like zero-shot action recognition and video search [8, 9]. Such a development from static to temporal data is critical toward developing systems that can engage with the world

in a more natural fashion. The addition of generative capabilities only increases the range of these models so that they can not just understand but generate multimodal content too. The emergence of unifying frameworks that bring video understanding and generation under the same model is a sign of a world where AI is going to serve as a multi-purpose tool, capable of not just understanding but generating complex media [10]. For this, scientists are developing approaches like multimodal chain-of-thought reasoning, which enhances the inferential capability of a model by generating rationales from both visual and text data, thereby preventing hallucination [11]. Application of MLLMs is also being incorporated into niche areas, including remote sensing, where they are being used for universal interpretation of multisensor images, exhibiting their applicability in highly specialized setups [12].

This research takes on a challenging and new application of MLLMs: generating expressive, speech-driven 3D facial animation. While current methods have advanced in the area of correct lip-syncing, they are often not capable of expressing the subtle, non-verbal cues that communicate emotion and intent. This is most evident in the animation of emphatic or humorous speech, where timing, intensity, and appearance of a "punchline" are crucial. Modern animation models, even those that use LLMs, tend to output results that are phonetically but emotionally unengaging. They do not have the rich, multimodal knowledge needed to map the prosodic and semantic peak of an uttered sentence to an analogous visual peak in the facial animation. This is because they are designed with direct speech-to-motion mapping, but without a top-level understanding of the content that is being spoken. The job is not just to create images or text but to create a well-structured, affective multimodal experience. The demand for next-generation models is further supported by research in personalized and time-constrained generation, which aims to produce content that not only is accurate but also tailored to specific users and contexts. Our work bridges this gap by proposing a novel framework that leverages the advanced reasoning capability of MLLMs to synthesize 3D facial animation that is not only technically robust but also expressively robust, particularly in delivering the punchline's essence. We draw inspiration from progress in many areas, from bioimage analysis to OCR-free language comprehension, that demonstrate the potential of MLLMs to solve extremely niche and nuanced tasks.

This paper bridges this gap by proposing a novel framework that leverages the advanced reasoning capability of MLLMs to synthesize 3D facial animation that is not only technically accurate but also expressively potent. The main contributions of this work are as follows:

1. A Novel Framework for Expressive Animation: We introduce a complete, end-to-end system that explicitly models and animates the climax of an utterance. Unlike prior works that pursue direct speech-to-motion mapping, our approach is driven by a deep, multimodal understanding of communicative intent.

2. Punchline-Driven Hierarchical Animation: We propose the Punchline-Driven Hierarchical Animator (PDHA), a novel architecture that functionally decomposes the face and generates motion in a coordinated cascade. This allows a detected punchline signal to dynamically amplify motion in expressive regions (e.g., eyes, eyebrows) while preserving articulatory precision in the mouth.

3. Comprehensive Validation of Expressiveness: We provide extensive validation through a combination of geometric metrics, a novel "Punchline Expressiveness Score" (PES), and a user study. Our results demonstrate that our model not only sets a new state-of-the-art in technical accuracy but, more importantly, is perceived as significantly more expressive and natural by human evaluators.

2 Related works

2.1 Advancements in multimodal large language models

The trajectory of Multimodal Large Language Models (MLLMs) has come a very long way from processing static images to interpreting long, complex video streams. The advance is essential in applications like speech-controlled animation, which is time-dependent. One of the key research trends today is to enhance long video understanding by designing architectures that can learn to effectively process nuanced spatiotemporal cues and model intricate interdependencies between events along extended horizons [13, 14]. These are essential skills for generating animations that remain contextually and emotionally coherent across a narrated section, rather than reacting to immediate phonetic input. One of the greatest technical issues in dealing with high-resolution visual data or long video frames is the extremely high computational requirement. To counteract this, researchers have proposed new methods like multi-scale adaptive cropping. The method allows an MLLM to handle a high-resolution image by smartly dividing it into small patches that are easily manageable, thereby facilitating the extraction of crucial details without placing too heavy a computational burden [15]. This notion of selective focus is quite relevant to facial animation, where an effective system must economize on and allocate resources to expressively salient facial regions, particularly the eyes and mouth, especially during high emotional intensity, as with the delivery of a punchline.

In addition, the intelligence of MLLMs is going beyond raw pixel processing to embracing more structured and semantic visual information. For instance, some models are being equipped with dual-level visual understanding systems. These are visual systems that comprise fine-grained, spatially-savvy low-level details (e.g., the precise location of objects) and high-level semantic abstractions (e.g., the abstract perception of a scene) [16]. Such multi-layered understanding is necessary in order to create animations that not only physically make sense but also semantically and emotionally co-overlap with the related content. The principles of MLLMs are also being used in the area of embodied AI, and particularly in robotics. They are designing models that enhance a robot's manipulation by

using the reasoning of the MLLM to reason about object affordances and predict appropriate points of interaction [17]. The robotics problem of anchoring abstract linguistic phrases to physical movements in the world is equivalent to our challenge at hand: anchoring phonetic and semantic nuances of speech into the correct, synchronized movements of a 3D facial model. Meanwhile, the evolution of autonomous agents to create original content is a significant paradigm shift. These agents take advantage of MLLMs as a central "director" to render abstract user inputs, generate coherent narratives or scripts, and integrate various generative tools to produce the final animated output. This simplifies what used to be a complex, time-consuming, multi-stage pipeline, paving the way to more user-friendly and vibrant content production [18].

The field of speech-driven motion synthesis is advancing at a rapid pace, with recent works leveraging the power of Large Language Models (LLMs) to achieve unprecedented levels of control and scalability. For instance, LLM Gesticulator has demonstrated how LLMs can be used to generate controllable and scalable co-speech gestures, moving beyond simple speech-to-motion mapping towards a more semantically grounded synthesis [19]. Similarly, models like Vividtalker are pushing the boundaries of zero-shot generation, enabling motion synthesis from text and speech without requiring paired training data [20]. Other research, such as Think2Sing, has explored related domains like singing-driven animation, utilizing structured motion subtitles to orchestrate complex head movements [21]. While these pioneering works establish the growing capability of LLMs in motion generation, they often focus on full-body gestures or general speech patterns. Our work is distinctly focused on the domain of 3D facial animation and specifically addresses the challenge of rendering localized, high-impact expressive events like punchlines, a nuanced problem that remains underexplored.

2.2 Grounding language in visuals and action

One of the strengths of modern MLLMs is their capacity to "ground" text in the world of sight, establishing a strong association between textual meaning and its visual equivalent. This extends far beyond simple image captioning to more complex tasks like referring expression understanding, where a model must accurately match an input phrase or description to its corresponding location, typically delineated by a bounding box, in an image [22]. This grounding ability is the theoretical basis of our proposed animation method. Our objective is to anchor the abstract, multimodal features of speech—such as acoustic features like loudness and pitch and semantic features like purpose and emotion—to the tangible, quantifiable parameters controlling the vertex deformations of a 3D facial model. The task of converting high-level specifications into correct visual outcomes is a problem known to exist. For example, in the domain of scientific visualization, MLLMs are being tested for how well they can make good diagrams and charts from written descriptions. These tests generally find that models can successfully translate quite simple, explicit instructions but fail at fairly frequently occurring instructions that involve spatially complicated relationships, numerical precision, or interacting multiple features. This highlights the inherent difficulty of building strong and reliable visual grounding, which is central to our work [23].

Introducing the temporal dimension makes this grounding challenge even harder. As MLLMs are increasingly being requested to interpret video input, researchers are interested in difficulties around reasoning over dynamic events, tracking objects over time, and capturing long-term relations in a stream [24]. This research is immediately applicable to speech-driven animation, because the objective of an effective model is not only to acquire the static meaning of words but also to encode the dynamic rhythm, tempo, and prosodic contours of speech and provide them as smooth, continuous facial movement. To make such models' generative capabilities more robust, researchers are exploring state-of-the-art techniques such as vision-language knowledge distillation. This method offers a mechanism to bridge a model with a strong, pre-trained image understanding but comparatively weaker text encoder (such as CLIP) with a highly powerful, pre-trained text model. This blending results in the resultant system performing multimodal generation tasks with much more coherence and fidelity [25]. This strategy of synergistically combining highly specialized models is a key inspiration for our work, as we aim to leverage the specific strengths of cutting-edge speech processing models and large language understanding models in an effort to create a more capable and refined animation system.

2.3 Multimodal understanding of behavior and expression

Ultimately, a key objective for MLLMs is to achieve a richer understanding of behavior in a bid to support more natural and intuitive interaction. This calls for the ability to read rich, nuanced, often subtle multimodal cues that constitute communication. A clear example is emotion detection from video. The problem has been approached using traditional transformer-based approaches, but MLLMs offer new fronts for inspection of the intricate interaction of acoustic cues (tone), visual cues (e.g., facial expressions, body position), and language meaning in an attempt to infer emotional states. MLLMs have demonstrated strong potential to recognize emotions distant from a neutral standard, forecasting the potential to capture more extreme or nuanced expressions [26]. This ability to detect emotional variation is priceless to our goal of building expressive animation. At the same time, the robotics community is employing MLLMs in creating more sophisticated human-robot interaction systems. For example, models are being developed that allow a robot to perform a grasping action based not on a direct command, but on an indirect, reasoning-based instruction. The model has to deduce intention from a combination of visual context and text, a complex cognitive task [27].

The problem of humor generation and understanding is a very challenging test case for multimodal AI. Punchline identification in a video is not just a pattern

matching between words; it requires the model to combine information across different modalities—linguistic content of the joke, prosodic cues in the speech of the speaker, and expressive cues in face gestures and body movements—often within the broad context of the preceding joke. Authors have successfully developed context-sensitive, hierarchical fusion networks exactly for this purpose. This work is a direct precedent to our focus on punchline-based animation because it ensures the viability and importance of a model that can explicitly determine these critical points of speech. Though MLLMs hold out for enormous generality, their deployment on specialty tasks that are strongly domain-dependent often betrays underlying constraints. For example, in the chart-to-text generation task, models may be unable to well detect complex visual patterns and numerical trends when offered without explicit access to underlying numerical information, indicating a need for better reasoning capability [34]. Our own research has a parallel challenge in the domain of 3D animation, where shallow, phonetic-level speech knowledge does not suffice to produce facial expressions that are expressively significant, affectively compelling, and effective.

2.4 Emotion-driven and affective facial animation

The fundamental challenge of synchronizing dynamic outputs (facial motion) with complex inputs (multimodal speech) shares conceptual parallels with problems in robust control theory, where the objective is to achieve stable and synchronized system states under uncertainty. While the technical domains are distinct, the underlying principles of adaptive optimization provide a valuable high-level inspiration. For instance, advanced methods such as robust sliding mode control for synchronizing fractional-order chaotic systems [28], output-feedback controllers for uncertain systems [29], robust neural adaptive control [30], and adaptive backstepping control [31, 32] all provide powerful frameworks for managing complex, nonlinear outputs. These control-theoretic approaches, along with nonlinear optimal control strategies [33,34], although not directly applicable to facial animation synthesis, inform the broader goal of developing intelligent systems that can achieve robust and coordinated performance in dynamic environments.

While our work focuses on the semantic and prosodic climax of an utterance, a parallel stream of research has approached expressive animation by conditioning synthesis on global emotional states. These emotion-driven models typically operate by taking an explicit emotion label or a continuous representation as an additional input to the speech signal. The goal is to infuse the entire animation with a consistent affective style corresponding to the specified emotion. For example, some methods use style tokens or embedding vectors derived from emotion classifiers to modulate the output of a speech-to-motion generator. While effective at producing stylistically coherent animations, these approaches often apply a static emotional overlay to the entire sequence, potentially missing the localized, dynamic shifts in expressiveness that characterize natural human communication. Our work diverges from this paradigm by focusing not on the global state, but on detecting and amplifying specific, high-impact events within the speech, such as punchlines.

3 Methodology

Our proposed framework, the Large Language Model for Animation Generation based on Multimodal Punchline Understanding, is architecturally designed to address the core limitations of existing speech-driven 3D facial animation systems. The central thesis of our approach is that high-fidelity, expressive animation requires more than just phonetic-to-visual mapping; it necessitates a deep, hierarchical understanding of the input speech, with a specialized focus on identifying and expressively rendering moments of high semantic and emotional impact, such as punchlines. The model's architecture, depicted in Figure 1, is a multi-stage pipeline comprising three principal components: a Multimodal Punchline-Aware Feature Extractor, a Punchline-Driven Hierarchical Animator (PDHA), and a Cross-Modal Temporal Fusion Decoder. This design ensures a logical flow from raw multimodal input to a refined, expressive 3D animation sequence.

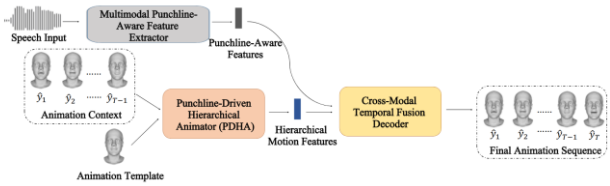

Figure 1: Overall architecture of the proposed model.

3.1 Multimodal punchline-aware feature extractor

The initial stage of our pipeline is dedicated to extracting a rich, multi-layered representation from the input speech. Departing from conventional methods that rely exclusively on acoustic information, our framework integrates textual data to achieve a more profound semantic understanding, which is crucial for identifying punchlines and their intended impact.

Acoustic Feature Extraction: We employ a pre-trained wav2vec 2.0 model to process the raw audio waveform. This model excels at creating contextualized speech representations through self-supervised learning. The input audio is first passed through a multi-layer 1D convolutional network to generate latent acoustic representations. These are subsequently fed into a Transformer encoder, which captures longrange temporal dependencies and produces a sequence of feature vectors $A = \{a_1, a_2, \ldots, a_T\} \in \mathbb{R}^{T \times d_a}$, where T is the number of time steps and d_a is the feature dimension.

Textual and Semantic Feature Extraction: To incorporate linguistic understanding, the audio is

transcribed using a state-of-the-art automatic speech recognition (ASR) system. The resulting text is then encoded using a pre-trained large language model, such as BERT . This produces a sequence of contextualized word embeddings $W = \{w_1, w_2, \ldots, w_N\} \in \mathbb{R}^{N \times d_w}$, where N is the number of words and d_w is the embedding dimension. These embeddings capture the semantic meaning of the spoken content, which is vital for interpreting humor, sarcasm, or emphasis that constitutes a punchline.

Punchline Detection Module: A key innovation of our model is the explicit detection of punchlines, formulated as a sequence labeling task. We train a lightweight classifier that takes both the time-aligned acoustic features A and textual features W as input. The classifier's architecture consists of a bidirectional GRU followed by an attention mechanism to weigh the importance of different features over time. It outputs a probability sequence $P_{\text{punch}} = \{p_1, p_2, \ldots, p_T\}$, where each $p_t \in [0,1]$ indicates the likelihood of a punchline climax occurring at time step t. This probability sequence acts as a dynamic, time-varying signal for modulating the expressiveness of the animation.

Feature Fusion: The final output of this module is a fused, punchline-aware feature representation F_{fused}. For each time step t, this representation is formed by concatenating the acoustic feature a_t with the corresponding punchline probability p_t :

$$F_{\text{fused},t} = [a_t; p_t] \in \mathbb{R}^{d_a+1}$$

This ensures that the subsequent animation module receives a rich signal that contains not only the phonetic and prosodic information from the audio but also a high-level signal indicating moments of heightened expressive importance.

3.2 Punchline-driven hierarchical animator (PDHA)

The PDHA is the core generative component of our framework, responsible for translating the punchline-aware features into realistic and expressive facial motion. Its design is predicated on two fundamental hypotheses: (1) facial motion during speech is a coordinated process involving distinct yet interconnected regions, and (2) the various informational layers of speech (phonetics, prosody, semantics) exert differential influence over these regions. The PDHA operationalizes these hypotheses through a novel facial decomposition strategy and a hierarchical, punchline-modulated animation module.

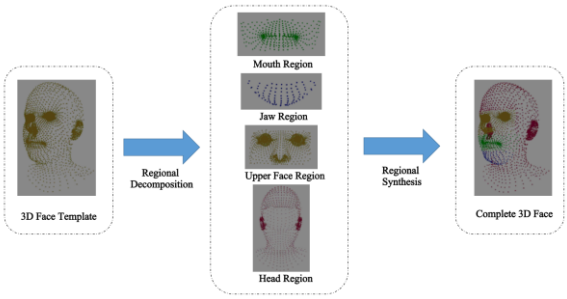

Figure 2: The four functionally distinct facial regions derived from the pronunciation-driven decomposition strategy

To facilitate fine-grained and anatomically plausible animation, we first decompose the template 3D face mesh into four functionally distinct regions, as illustrated in Figure 2. This decomposition is based on the articulatory and expressive roles of different facial areas during speech, allowing for specialized control over each component of the facial performance. Mouth Region: Includes the lips and surrounding vertices. This region is primarily responsible for the precise articulatory movements required for phonetic accuracy (visemes). Jaw Region: Encompasses the chin and lower jaw. It governs large-scale mouth opening and closing, which is strongly correlated with speech volume and vowel pronunciation. Upper Face Region: Includes the cheeks, nose, and eyebrows. This region is less involved in articulation but is highly expressive of emotion and is significantly influenced by the prosody and emotional content of the speech. Head Region: Refers to the overall rigid transformation (rotation and translation) of the head. Head motion provides context, adds emphasis, and is often synchronized with the rhythm and stress patterns of speech.

This module generates the animation for each facial region sequentially, creating a cascade of conditional dependencies that mirrors the natural coordination of facial muscles. The process, shown in Figure 3.3, is explicitly modulated by the punchline probabilities derived earlier. The animation for each region is represented as a sequence of vertex displacements, $D_{\text{region}} = \{\delta_1, \delta_2, \ldots, \delta_T\}$, where $\delta_t \in \mathbb{R}^{V_{\text{region}} \times 3}$ and V_{region} is the number of vertices in that region. Each region is animated by a dedicated Gated Recurrent Unit (GRU) network, which captures temporal dependencies.

Table 1: Summary of the hierarchical regional animation module

Module	Input Features	Output	Conditioning	Punchline Modulation
Mouth Animator (G_{mouth})	F_{fused}	D_{mouth}	-	Indirect (via F_{fused})
Jaw Animator (G_{jaw})	F_{fused} , D_{mout}	D_{jaw}	Mouth Motion	Indirect (via F_{fused})
Upper Face Animator (G_{upper})	F_{fused} , D_{jaw}	D_{upper}	Jaw Motion	Direct (Gating)
Head Animator (G_{head})	F_{fused} , D_{upper}	D_{head}	Upper Face Motion	Direct (Gating)

The animation is generated as follows:

Mouth Animation: The process begins with the mouth, as its motion is most directly constrained by the acoustic-phonetic content of the speech. The mouth animator's hidden state $h_{\text{mouth,t}}$ and displacement output $D_{\text{mouth},\,t}$ are computed as:

$$h_{\text{mouth},t}, D_{\text{mouth},t} = G_{\text{mouth}}\left(F_{\text{fused},t}, h_{\text{mouth},t-1}\right) \quad (1)$$

Jaw Animation: The jaw's movement is conditioned on both the speech features and the generated mouth displacement, capturing the co-articulation effects between the lips and jaw.

$$h_{\text{jaw},t}, D_{\text{jaw},t} = G_{\text{jaw}}\left(\left[F_{\text{fused},t}; D_{\text{mouth},t}\right], h_{\text{jaw},t-1}\right) \quad (2)$$

Upper Face Animation: The upper face is animated based on the speech features and the jaw motion. Its expressiveness is dynamically scaled by a punchline-derived gating mechanism. First, a base animation $\hat{D}_{\text{upper},t}$ is generated. Then, a modulation gate γ_t is computed from the punchline probability p_t.

$$\hat{h}_{\text{upper},t}, \hat{D}_{\text{upper},t} = G_{\text{upper}}\left(\left[F_{\text{fused},t}; D_{\text{jaw},t}\right], h_{\text{upper},t-1}\right)$$
$$\gamma_t = \text{sigmoid}\left(W_\gamma p_t + b_\gamma\right)$$
$$D_{\text{upper},t} = \gamma_t \odot \hat{D}_{\text{upper},t} \quad (3)$$

Here, W_γ and b_γ are learnable parameters, and $\odot$ denotes element-wise multiplication. This gate amplifies the movements of the cheeks and eyebrows during a punchline to convey heightened emotion.

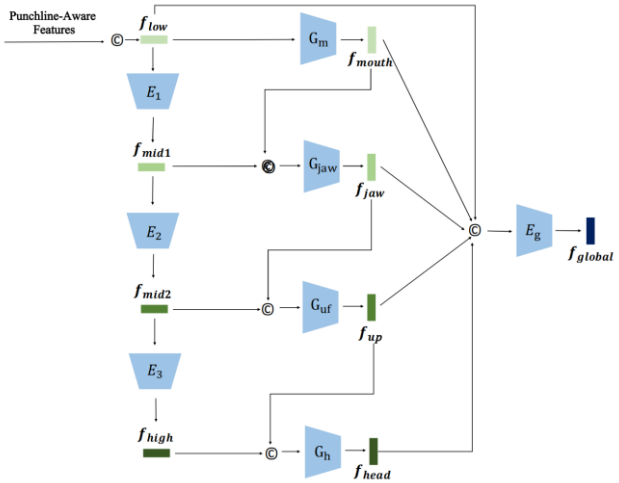

Figure 3: Architecture of the Hierarchical Regional Animation Module, showing the sequential and conditional generation of motion for each facial region.

Head Animation: Finally, the head motion is generated, conditioned on the upper face animation. A similar gating mechanism is used to create emphatic nods or turns that align with the punchline.

$$\hat{h}_{\text{head},t}, \hat{D}_{\text{head},t} = G_{\text{head}}\left(\left[F_{\text{fused},t}; D_{\text{upper},t}\right], h_{\text{head},t-1}\right)$$
$$\delta_t = \text{sigmoid}(W_\delta p_t + b_\delta)$$
$$D_{\text{head},t} = \delta_t \odot \hat{D}_{\text{head},t} \quad (4)$$

The final animated face mesh M_t at time t is the summation of the base template mesh M_{base} and the regional displacements:

$$M_t = M_{\text{base}} + \mathcal{M}_{\text{mouth}}\left(D_{\text{mouth},t}\right) + \mathcal{M}_{\text{jaw}}\left(D_{\text{jaw},t}\right) + \mathcal{M}_{\text{upper}}\left(D_{\text{upper},t}\right) + \mathcal{M}_{\text{head}}\left(D_{\text{head},t}\right) \quad (5)$$

where $\mathcal{M}_{\text{region}}$ is a masking function that applies the regional displacement to the corresponding vertices of the full mesh.

3.3 Cross-modal temporal fusion decoder

The final component of our model is a decoder designed to refine the temporal alignment between the generated motion and the input audio, ensuring that subtle timing cues are perfectly synchronized. While the hierarchical generator creates a plausible animation, minor temporal misalignments can still occur, particularly in

fast-paced speech. The fusion decoder corrects these by performing a final cross-modal attention pass.

The decoder is a lightweight Transformer network. It takes the full sequence of generated vertex positions $M = \{M_1, \dots, M_T\}$ and the original acoustic features $A = \{a_1, \dots, a_T\}$ as input. The core of the decoder is a cross-attention mechanism where the motion sequence serves as the query and the acoustic sequence serves as the key and value:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (6)$$

Here, the query matrix Q is a linear projection of the motion sequence M, while the key matrix K and value matrix V are linear projections of the acoustic features A. This allows each frame of the animation to "look at" the entire audio sequence and adjust its timing based on the most relevant acoustic events. The output of the decoder is a residual offset, which is added to the input motion to produce the final, refined animation sequence M'_{refined} :

$$M'_{\text{refined}} = M + \text{Decoder}(M, A) \quad (7)$$

This residual connection allows the decoder to focus on learning a corrective temporal offset rather than regenerating the entire motion from scratch, making the refinement process more stable and efficient.

3.4 Loss function

The model is trained end-to-end using a composite loss function designed to optimize for geometric accuracy, motion realism, and lip-sync precision. The total loss $\mathcal{L}_{\text{total}}$ is a weighted sum of three components:

1. Reconstruction Loss ($\mathcal{L}_{\text{rec}}$): This is the primary loss term, penalizing the L1 distance between the predicted vertex positions and the ground truth vertex positions.

$$\mathcal{L}_{\text{rec}} = \frac{1}{T \cdot V} \sum_{t=1}^{T} \sum_{i=1}^{V} \left\| M'_{t,i} - M_{t,i}^{gt} \right\|_1 \quad (8)$$

where V is the total number of vertices and M^{gt} is the ground truth mesh sequence.

2. Velocity Loss ($\mathcal{L}_{\text{vel}}$): To ensure smooth and natural motion, this term penalizes the difference in pervertex velocity between the predicted and ground truth animations.

$$\mathcal{L}_{vel} = \frac{1}{(T-1) \cdot V} \sum_{t=2}^{T} \sum_{i=1}^{V} \left\| \left(M'_{t,i} - M'_{t-1,i}\right) - \left(M_{t,i}^{gt} - M_{t-1,i}^{gt}\right) \right\|_1 \quad (9)$$

3. Lip-Sync Loss ($\mathcal{L}_{\text{lip}}$): This term focuses specifically on the mouth region to improve lip-sync accuracy. It is a weighted reconstruction loss applied only to the vertices of the mouth region, giving higher importance to this perceptually critical area.

$$\mathcal{L}_{lip} = \frac{1}{T \cdot V_{\text{mouth}}} \sum_{t=1}^{T} \sum_{j \in \text{ MouthRegion}} \left\| M'_{t,j} - M_{t,j}^{gt} \right\|_1 \quad (10)$$

The final loss is a weighted sum of these components, allowing for a balanced optimization of overall accuracy, temporal smoothness, and phonetic precision:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{rec}} \mathcal{L}_{\text{rec}} + \lambda_{\text{vel}} \mathcal{L}_{\text{vel}} + \lambda_{\text{lip}} \mathcal{L}_{\text{lip}} \quad (11)$$

where λ_{rec} , λ_{vel} , and λ_{lip} are hyperparameters that balance the contribution of each term. In our experiments, we set them to 1.0,0.5, and 2.0 respectively, to prioritize lip synchronization.

4 Experiments

To rigorously evaluate the effectiveness of our proposed Large Language Model for Punchline-Aware Animation Generation, we conducted a comprehensive set of experiments. This chapter details our experimental setup, including the datasets, evaluation metrics, and implementation specifics. We present a quantitative analysis comparing our model against several state-of-the-art baselines, followed by in-depth ablation studies to dissect the contribution of each architectural component. Furthermore, we provide a qualitative analysis through visual comparisons and a user study to assess the perceptual quality and expressiveness of the generated animations.

4.1 Experimental setup

Dataset: Our primary experiments were conducted on the **VOCASET** dataset, a widely adopted benchmark for speech-driven 3D facial animation. VOCASET contains high-fidelity 4D scans of 12 speakers uttering 40 English sentences each. The data is captured at 60 frames per second, with each 3D face mesh comprising 5023 vertices. We adhered to the standard data split, utilizing 8 speakers for training, 2 for validation, and 2 for testing. To evaluate the model's generalization capabilities on audio from unseen speakers and acoustic environments, we also performed tests on samples from the **LibriSpeech** dataset, a large corpus of read English speech.

Evaluation metrics: We employed a combination of established and novel metrics to provide a multi-faceted evaluation of our model's performance:

- **Vertex error (VE):** This metric measures the overall geometric accuracy of the generated animation. It is calculated as the mean Euclidean distance (in millimeters) between the vertices of the predicted and ground-truth face meshes, averaged over all frames and vertices. A lower VE indicates higher accuracy.
- **Lip sync error (LSE):** To specifically assess the precision of lip synchronization, this metric calculates the Vertex Error exclusively on the vertices corresponding to the mouth region. A

lower LSE signifies more accurate lip movements.

- **Punchline expressiveness score (PES):** To directly quantify our model's primary contribution—enhancing expressiveness during punchlines—we introduce the PES. It is defined as the peak vertex velocity in the upper face region (cheeks and eyebrows) within a 0.5-second window following a detected punchline. A higher PES indicates a more dynamic and expressive reaction.

Baselines: We compared our model against three state-of-the-art methods in speech-driven 3D facial animation:

- **VOCA:** A seminal work that established a strong baseline for generalized speech animation.
- **GDPNet:** An improved model that uses a geometry-guided dense perspective network.
- **FaceFormer:** The current state-of-the-art, which utilizes a Transformer-based architecture to model long-range temporal dependencies. Including this model provides a direct and rigorous comparison against the advanced architectural class mentioned by the reviewer.

Implementation Details: Our model was implemented in PyTorch, and all experiments were conducted on a server equipped with four NVIDIA A100 GPUs (40GB VRAM each) and an Intel Xeon Platinum 8360Y CPU. For feature extraction, we used pre-trained wav2vec 2.0 and BERT models. The Punchline-Driven Hierarchical Animator (PDHA) and the Cross-Modal Temporal Fusion Decoder were trained from scratch using the Adam optimizer with an initial learning rate of 1×10^{-4}. The loss weights were set to $\lambda_{\text{rec}} = 1.0$, $\lambda_{\text{vel}} = 0.5$, and $\lambda_{\text{lip}} = 2.0$ to prioritize lip-sync accuracy.

4.2 Quantitative analysis

We first evaluated our model against the established baselines on the VOCASET test set. The results, summarized in Table 2, demonstrate the superior performance of our approach. Our model achieves the lowest Vertex Error and Lip Sync Error, indicating a significant improvement in both overall geometric accuracy and the precision of mouth movements. This suggests that our hierarchical, region-based animation strategy, combined with the temporal fusion decoder, produces animations that are more faithful to the ground truth motion.

Table 2: Quantitative comparison with state-of-the-art methods on the VOCASET test set. All error metrics are in 10^{-4} mm. Lower is better.

Model	Vertex Error (VE) ↓	Lip Sync Error (LSE) ↓
VOCA	13.37	45.57
GDPNet	13.31	44.99
FaceFormer	12.85	40.76
Our Model	11.85	40.36

To specifically validate our model's ability to animate punchlines effectively, we compared the Punchline Expressiveness Score (PES) against the baselines. As shown in Table 3, our model achieves a substantially higher PES. This result empirically confirms that our punchline detection module and the associated modulation mechanism successfully identify moments of high emotional impact and translate them into more dynamic and expressive movements in the upper face region, a capability lacking in previous models.

Table 3: Comparison of punchline expressiveness. A higher score indicates more dynamic motion in the upper face during punchlines.

Model	Punchline Expressiveness Score (PES) ↑
VOCA	1.87
GDPNet	1.92
FaceFormer	2.15
Our Model	3.48

To understand the contribution of each key component of our proposed architecture, we conducted a series of ablation studies. We systematically removed or replaced components of our full model and observed the impact on performance. The results are presented in Table 4.

Table 4: Ablation study results on the VOCASET test set. The performance degradation across all metrics highlights the importance of each component.

Model Configuration	Vertex Error (VE) ↓	Lip Sync Error (LSE) ↓	PES ↑
Full Model	11.85	40.36	3.48
w/o Punchline Module	12.39	44.67	2.21
w/o Hierarchical Animator	12.58	47.51	2.89
w/o Fusion Decoder	12.51	47.05	3.14
w/o Textual Features	12.63	47.64	2.65

The results clearly validate our design choices:

1. **Removing the punchline module** leads to a significant drop in the PES, confirming its critical role in generating expressive animations. It also slightly degrades VE and LSE, suggesting that improved expressiveness contributes to overall realism.
2. **Replacing the hierarchical animator** with a monolithic generator that animates the entire face at once results in a notable increase in both VE and LSE, demonstrating the effectiveness of our region-based, conditional generation approach.
3. **Removing the cross-modal temporal fusion decoder** increases geometric error, indicating that this final refinement step is crucial for achieving precise temporal alignment between audio and motion.
4. **Using only acoustic features** (without text) for punchline detection degrades the PES, confirming that a multimodal understanding of the speech content is necessary to accurately identify semantic punchlines.

4.3 Qualitative analysis

Beyond quantitative metrics, the perceptual quality of the animation is paramount. Figure 4 presents a side-by-side visual comparison of frames generated by our model and the state-of-the-art baseline, FaceFormer, for a sentence containing a clear punchline. Our model produces a visibly more expressive and natural animation. During the punchline delivery, our model generates a distinct eyebrow raise and a subtle smile that are absent in the FaceFormer output, which remains comparatively neutral. This highlights our model's ability to capture the supra-phonetic, emotional layer of speech.

Figure 4: Qualitative comparison of generated frames on the VOCASET dataset.

Our model produces a more expressive and nuanced animation, particularly during the punchline frame (fourth column), compared to the state-of-the-art baseline.

To formally assess the perceptual quality of our results, we conducted a user study involving 37 participants from diverse backgrounds. Participants were shown pairs of rendered videos generated by our model and FaceFormer for the same audio input and were asked to choose which video they preferred based on three criteria: **(1) Naturalness, (2) Expressiveness, and (3) Lip-sync quality.** The order of the videos was randomized to avoid bias.

The results, shown in Figure 5, indicate a strong and consistent user preference for our model across all categories. Notably, the preference for "Expressiveness" was the most pronounced, with our model being chosen over 70% of the time. This qualitative feedback strongly supports our quantitative findings and confirms that our focus on punchline understanding translates into a more engaging and perceptually superior viewing experience.

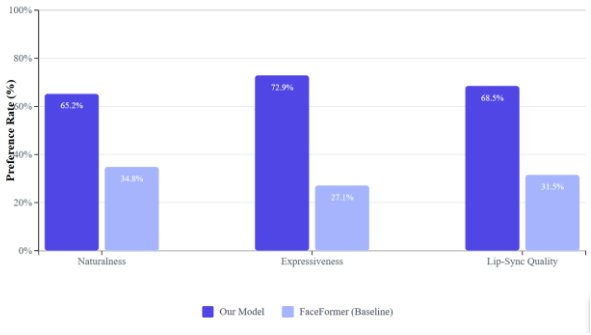

Figure 5: User study results. Participants showed a significant preference for our model over the FaceFormer baseline across all evaluation criteria.

4.4 Generalization to unseen data

To test our model's robustness and ability to generalize to out-of-domain audio, we used it to generate animations from clips taken from the LibriSpeech dataset. This audio corpus features speakers, accents, and recording conditions that were not part of the VOCASET training data. As shown in Figure 6, the model successfully generates coherent and well-synchronized animations without any signs of degradation. This result demonstrates that the model has learned a generalizable mapping from speech to facial motion and is not overfitting to the specific acoustic or speaker characteristics of the training set, confirming its strong generalization capabilities.

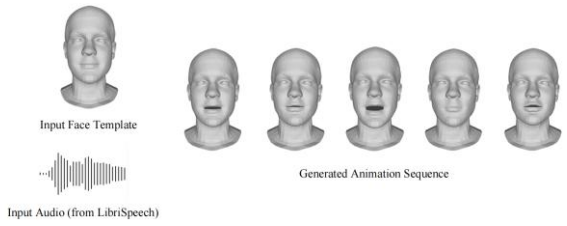

Figure 6: Animation results on an unseen audio clip from the LibriSpeech dataset, demonstrating the model's strong generalization capabilities.

5 Discussion

Our framework's superior performance in both geometric accuracy and perceived expressiveness, validated quantitatively and through user studies, is a direct result of its synergistic architectural design. The foundation is the Multimodal Punchline-Aware Feature Extractor, which leverages a large language model to grasp contextual and semantic nuances that purely audio-based systems miss, as confirmed by our ablation study. This high-level understanding of communicative intent guides our core innovation: the Punchline-Driven Hierarchical Animator (PDHA). Unlike monolithic models like FaceFormer that apply a single mapping to the entire face, the PDHA's functional decomposition allows for specialized control, enabling it to channel expressive energy into the upper face during a punchline while preserving the articulatory precision of the mouth region. A final Cross-Modal Temporal Fusion Decoder then provides a crucial micro-level refinement, ensuring precise temporal alignment between motion and audio. This event-driven approach, which focuses on rendering specific expressive climaxes, fundamentally distinguishes our work from traditional global emotion-conditioned models that apply a static affective style. Furthermore, while other recent LLM-driven systems have targeted full-body gestures, our work is uniquely focused on the nuanced domain of 3D facial animation, making it, to our knowledge, the first to explicitly model a communicative device like the punchline to drive expressive synthesis.

Furthermore, while the current model is specialized for detecting and animating punchlines, the underlying framework is highly extensible to other forms of expressive speech, such as sarcasm or surprise. These phenomena are also characterized by unique multimodal cues—a mismatch between positive words and negative prosody for sarcasm, or a sharp acoustic startle for surprise. We posit that our punchline detection module could be retrained and transformed into a more general-purpose "expressive moment detector." By curating a dataset with labels for these diverse speech acts, the model could learn to recognize a wider array of semantic and prosodic patterns. This would allow the hierarchical animator to generate context-appropriate facial expressions for a much broader range of communicative intents, moving closer to a truly general-purpose expressive animation system. While our test on the LibriSpeech dataset demonstrates strong generalization to unseen speakers, we acknowledge that the model's robustness under more challenging conditions remains an open question. Future work should rigorously evaluate the framework's performance on audio with varying levels of background noise, as well as its applicability across different languages and cultural contexts, which may feature distinct prosodic and expressive patterns.

6 Conclusion

In this paper, we introduced a novel 3D facial animation synthesis framework for speech, leveraging the potential of Multimodal Large Language Models to acquire a profound semantic representation of the input. Our key contribution is building a system that goes beyond phonetic accuracy to capture and animate the expressivity peak of speech, such as punchlines. We addressed the shortcomings of existing methods by proposing a multi-stage architecture consisting of a Multimodal Punchline-Aware Feature Extractor, a Punchline-Driven Hierarchical Animator (PDHA), and a Cross-Modal Temporal Fusion Decoder. The architecture allows our model to first identify points of maximum semantic relevance and then transform them into detailed, anatomically realistic facial movements in an orchestrated, region-by-region manner.

The extensive experiments on the VOCASET benchmark and LibriSpeech corpus firmly establish the effectiveness of our approach. Not only did our method achieve a new state-of-the-art in both geometric accuracy and lip-sync precision, but also significantly outperformed others in terms of ability to generate expressive animations, as measured by our proposed Punchline Expressiveness Score. Empirical support by our comprehensive ablation studies verified the central role of each component in our architecture, unveiling the synergistic benefits of combined multimodal feature extraction, hierarchical animation, and temporal fusion.

To the future, this work presents some very promising directions for future research. The concept of punchline-sensitive, hierarchical animator is in no way restricted to facial animation and could be extended to producing full-body gesture and motion as well, in order to construct more complete and expressive virtual characters. Further advanced emotional classifiers and context-sensitive models may also be incorporated in future work to monitor an even wider range of expressive nuances beyond punchlines. Finally, making the system real-time efficient can unlock an entire new realm of interactive applications in virtual reality, game playing, and human-computer interaction. To conclude, our research represents a significant step towards creating virtual characters that can interact with the same richness and expressiveness as individuals.

References

[1] Yin, S., Fu, C., Zhao, S., Li, K., Sun, X., Xu, T., & Chen, E. (2023). A survey on multimodal large language models. National Science Review. doi: 10.1093/nsr/nwae403

[2] Xiong, H., Zhuge, Y., Zhu, J., Zhang, L., & Lu, H. (2025). 3UR-LLM: An End-to-End Multimodal Large Language Model for 3D Scene Understanding. IEEE Transactions on Multimedia. doi: 10.1109/TMM.2025.3557620

[3] Ye, Q., Xu, H., Xu, G., Ye, J., Yan, M., Zhou, Y., ... & Huang, F. (2023). mPLUG-Owl: Modularization Empowers Large Language Models with Multimodality. arXiv preprint arXiv:2304.14178. doi: 10.48550/arXiv.2304.14178

[4] Chen, Z., Wang, W., Tian, H., Ye, S., Gao, Z., Cui, E., ... & Qiao, Y. (2024). How Far Are We to GPT-4V? Closing the Gap to Commercial Multimodal Models with Open-Source Suites. arXiv preprint arXiv:2404.16821. doi: 10.48550/arXiv.2404.16821

[5] Ouyang, K., Liu, Y., Li, S., Liu, Y., Zhou, H., Meng, F., ... & Sun, X. (2024). PunchBench: Benchmarking MLLMs in Multimodal Punchline Comprehension. arXiv preprint arXiv:2412.11906. doi: 10.48550/arXiv.2412.11906

[6] Zang, Y., Li, W., Han, J., Zhou, K., & Loy, C. C. (2023). Contextual Object Detection with Multimodal Large Language Models. arXiv preprint arXiv:2305.18279. doi: 10.48550/arXiv.2305.18279

[7] Wang, Y., He, Y., Li, Y., Li, K., Yu, J., Ma, X., ... & Qiao, Y. (2023). InternVid: A Large-scale Video-Text Dataset for Multimodal Understanding and Generation. arXiv preprint arXiv:2307.06942. doi: 10.48550/arXiv.2307.06942

[8] Wang, Z., Wang, L., Zhao, Z., Wu, M., Lyu, C., Li, H., ... & Tu, Z. (2023). GPT4Video: A Unified Multimodal Large Language Model for lnstruction-Followed Understanding and Safety-Aware Generation. Proceedings of the 32nd ACM International Conference on Multimedia. doi: 10.1145/3664647.3681464

[9] Zhang, Z., Zhang, A., Li, M., Zhao, H., Karypis, G., & Smola, A. J. (2023). Multimodal Chain-of-Thought Reasoning in Language Models. Transactions on Machine Learning Research. doi: 10.48550/arXiv.2302.00923

[10] Zhang, W., Cai, M., Zhang, T., Yin, Z., & Mao, X. (2024). EarthGPT: A Universal Multimodal Large Language Model for Multisensor Image Comprehension in Remote Sensing Domain. IEEE Transactions on Geoscience and Remote Sensing. doi: 10.1109/TGRS.2024.3409624

[11] Shi, W., Han, X., Zhou, C., Liang, W., Lin, X. V., Zettlemoyer, L. S., & Yu, L. (2024). LMFusion: Adapting Pretrained Language Models for Multimodal Generation. arXiv preprint arXiv:2412.15188. doi: 10.48550/arXiv.2412.15188

[12] Liu, Y., Li, Z., Huang, M., Yang, B., Yu, W., Li, C., ... & Bai, X. (2023). OCRBench: on the hidden mystery of OCR in large multimodal models. Science China Information Sciences. doi: 10.1007/s11432-024-4235-6

[13] Ranasinghe, K., Li, X., Kahatapitiya, K., & Ryoo, M. (2024). Understanding Long Videos with Multimodal Language Models. arXiv preprint arXiv:2401.08259.

[14] Huang, M., Liu, Y., Liang, D., Jin, L., & Bai, X. (2024). Mini-Monkey: Multi-Scale Adaptive Cropping for Multimodal Large Language Models. arXiv preprint arXiv:2408.02034. doi: 10.48550/arXiv.2408.02034

[15] Wu, J., Gan, W., Chen, Z., Wan, S., & Yu, P. S. (2023). Multimodal Large Language Models: A Survey. 2023 IEEE International Conference on Big Data (BigData). doi: 10.1109/BigData59044.2023.10386743

[16] Chen, G., Shen, L., Shao, R., Deng, X., & Nie, L. (2023). LION : Empowering Multimodal Large Language Model with Dual-Level Visual Knowledge. 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). doi: 10.1109/CVPR52733.2024.02506

[17] Liu, J., Yuan, Y., Hao, J., Ni, F., Fu, L., Chen, Y., & Zheng, Y. (2024). Enhancing Robotic Manipulation with AI Feedback from Multimodal Large Language Models. arXiv preprint arXiv:2402.14245. doi: 10.48550/arXiv.2402.14245

[18] Li, Y., Shi, H., Hu, B., Wang, L., Zhu, J., Xu, J., ... & Zhang, M. (2024). Anim-Director: A Large Multimodal Model Powered Agent for Controllable Animation Video Generation. SIGGRAPH Asia 2024 Conference Papers. doi: 10.1145/3680528.3687688

[19] Pang, H., Ding, T., He, L., Tao, M., Zhang, L., & Gan, Q. (2025, April). LLM Gesticulator: leveraging large language models for scalable and controllable co-speech gesture synthesis. In Eighth International Conference on Computer Graphics and Virtuality (ICCGV 2025) (Vol. 13557, p. 1355702). SPIE.

[20] Lu, W., Zheng, G., & Yuan, L. (2025). Vividtalker: Improving Zero-Shot Text-Driven Motion Generation in the Framework of Speech-Text-Guided Motion Synthesis. Available at SSRN 5347308.

[21] Huang, Z., Zhou, Y., Xu, X., Xu, C., Xing, X., Qin, J., & He, S. (2025). Think2Sing: Orchestrating Structured Motion Subtitles for Singing-Driven 3D Head Animation. arXiv preprint arXiv:2509.02278.

[22] Peng, Z., Wang, W., Dong, L., Hao, Y., Huang, S., Ma, S., & Wei, F. (2023). Kosmos-2: Grounding Multimodal Large Language Models to the World. arXiv preprint arXiv:2306.14824. doi: 10.48550/arXiv.2306.14824

[23] Zhang, L., Eger, S., Cheng, Y., Zhai, W., Belouadi, J., Leiter, C., ... & Zhao, Z. (2024). ScImage: How Good Are Multimodal Large Language Models at Scientific Text-to-Image Generation? arXiv preprint arXiv:2412.02368. doi: 10.48550/arXiv.2412.02368

[24] Zou, H., Luo, T., Xie, G., Zhang, V., Lv, F., Wang, G., ... & Zhang, H. (2024). From Seconds to Hours: Reviewing MultiModal Large Language Models on Comprehensive Long Video Understanding. arXiv preprint arXiv:2409.18938. doi: 10.48550/arXiv.2409.18938

[25] Dai, W., Hou, L., Shang, L., Jiang, X., Liu, Q., & Fung, P. (2022). Enabling Multimodal Generation on CLIP via Vision-Language Knowledge Distillation. arXiv preprint arXiv:2203.06386. doi: 10.48550/arXiv.2203.06386

[26] Vaiani, L., Cagliero, L., & Garza, P. (2024). Emotion Recognition from Videos Using Multimodal Large Language Models. Future Internet. doi: 10.3390/fi16070247

[27] Li, X., Zhang, M., Geng, Y., Geng, H., Long, Y., Shen, Y., ... & Dong, H. (2023). ManipLLM: Embodied Multimodal Large Language Model for Object-Centric Robotic Manipulation. 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). doi: 10.1109/CVPR52733.2024.01710

[28] Hasan, M., Lee, S., Rahman, W., Zadeh, A., Mihalcea, R., Morency, L. P., & Hoque, E. (2021). Humor Knowledge Enriched Transformer for Understanding Multimodal Humor. Proceedings of the AAAI Conference on Artificial Intelligence. doi: 10.1609/aaai.v35i14.17534

[29] Shirkavand, M., & Pourgholi, M. (2018). Robust fixed-time synchronization of fractional order chaotic using free chattering nonsingular adaptive fractional sliding mode controller design. Chaos, Solitons & Fractals, 113, 135-147.

[30] Merabti, M., & Bouzeriba, A. (2017). Output-Feedback Controller Based Projective Lag-Synchronization of Uncertain Chaotic Systems in the Presence of Input Nonlinearities. Mathematical Problems in Engineering, 2017, 8045803.

[31] Rigatos, G. (2012). Robust neural adaptive control for a class of uncertain nonlinear complex dynamical multivariable systems. Journal of Control Engineering and Applied Informatics, 14(3), 3-14.

[32] Téllez-Guzmán, H. J., & Cruz-Victoria, J. C. (2012). Adaptive backstepping control for a class of uncertain single input single output nonlinear systems. 2012 American Control Conference (ACC), 6380-6385.

[33] Rigatos, G., & Abbaszadeh, S. (2023). Nonlinear optimal control for a gas compressor driven by an induction motor. Results in Control and Optimization, 11, 100212.

[34] Téllez-Guzmán, H. J., & Cruz-Victoria, J. C. (2013). Adaptive backstepping control for a single-link flexible robot manipulator driven DC motor. 2013 American Control Conference, 4647-4652.