

---

# Advancing Talking Head Generation: A COMPREHENSIVE SURVEY OF MULTI-MODAL METHODOLOGIES, DATASETS, EVALUATION METRICS, AND LOSS FUNCTIONS

---

Vineet Kumar Rakesh 

Engineering Sciences, Homi Bhabha National Institute  
Training School Complex, Anushaktinagar, Mumbai, Maharashtra 400094, India  
Computer and Informatics Group, VECC  
1/AF, Bidhannagar, Kolkata, West Bengal 700064, India  
vineet@vecc.gov.in

Soumya Mazumdar 

Department of Computer Science and Business Systems  
Gargi Memorial Institute of Technology  
Baruipur, Kolkata, West Bengal 700144, India  
reachme@soumyamazumdar.com

Research Pratim Maity 

Department of Computer Science and Business Systems  
Gargi Memorial Institute of Technology  
Baruipur, Kolkata, West Bengal 700144, India  
researchpratimmaity2004@gmail.com

Sarbjit Pal 

Computer and Informatics Group, VECC  
1/AF, Bidhannagar, Kolkata, West Bengal 700064, India  
Engineering Sciences, Homi Bhabha National Institute  
Training School Complex, Anushaktinagar, Mumbai, Maharashtra 400094, India  
sarbjit@vecc.gov.in

Amitabha Das 

School of Nuclear Studies and Application  
Jadavpur University  
Salt Lake City, Kolkata, West Bengal 700106, India  
amitabhad.snsa@jadavpuruniversity.in

Tapas Samanta 

Computer and Informatics Group, VECC  
1/AF, Bidhannagar, Kolkata, West Bengal 700064, India  
Engineering Sciences, Homi Bhabha National Institute  
Training School Complex, Anushaktinagar, Mumbai, Maharashtra 400094, India  
tsamanta@vecc.gov.in

## Abstract

Talking Head Generation has emerged as a transformative technology in computer vision, enabling the synthesis of realistic human faces synchronized with image, audio, text, or video inputs. This paper provides a comprehensive review of methodologies and frameworks for talking head generation, categorizing approaches into 2D-based, 3D-based, Neural Radiance Fields (NeRF)-based, diffusion-based, parameter-driven techniques and many other techniques. It evaluates algorithms, datasets, and evaluation metrics while highlighting advancements in perceptual realism and technical efficiency critical for applications such as digital avatars, video dubbing, ultra-low bitrate video conferencing, and online education. The study identifies challenges such as reliance on pre-trained models, extreme pose handling, multilingual synthesis, and temporal consistency. Future directions include modular architectures, multilingual datasets, hybrid models blending pre-trained and task-specific layers, and innovative loss functions. By synthesizing existing research and exploring emerging trends, this paper aims to provide actionable insights for researchers and practitioners in the field of talking head generation. For the complete survey, code, and curated resource list, visit our GitHub repository: <https://github.com/VineetKumarRakesh/thg>.

**Keywords** Talking Head Generation · Deep Learning · Systematic Review · Dataset · Evaluation Metrics

## 1 Introduction

Deep Learning (DL) and Artificial Neural Network (ANN) have altered computer vision, enabling breakthroughs in the area of Talking Head Generation (THG) and synthesis of realistic human faces for speech articulation [87]. THG is a research subject that seeks to generate realistic video pictures of human faces that speak in synchronization with arbitrary audio, image, text and others. as input. Today, this technology finds uses in digital avatars, video dubbing in movies, online schooling and video conferencing [85].

Early approaches concentrated on 2D-based algorithms applying deep generative models, but they failed to capture 3D structural information, which resulted in less realistic facial dynamics and perspective modifications [85]. Recent breakthroughs in 3D scene representation approaches, notably Neural Radiance Fields (NeRF) [90], have boosted realism, allowed free-view control for movies, and improved picture quality. Innovative frameworks like Wav2NeRF [85] have emerged to tackle issues such as correctly morphing lip movements in rhythm with audio and managing high-frequency details [85]. The subject has developed from early rule-based approaches to sophisticated deep learning methods employing Generative Adversarial Networks (GANs) and attention processes [87].

The synthesis of talking heads encompasses portrait production, driving mechanisms, and editing techniques, enabling the adjustment of attributes such as emotion, head posture, and eye blinking [86]. This study thoroughly examines the existing methodology and approaches for creating THG and assessing algorithms, datasets, and evaluation matrices. It arranges processes into separate groupings, stressing their contributions and drawbacks. Focusing on technical efficiency and perceptual realism is crucial for real-time interaction and high visual quality applications—the research also provides a comparative review of publicly accessible technologies for THG.

Portrait production has improved dramatically by utilizing generative techniques, including unconditional and conditional methods. Unconditional techniques provide random visuals without prior labels or data inputs, while conditional methods offer controlled outputs based on descriptive inputs [86]. However, making high-quality, realistic portraits remains tricky due to intricate geometry and appearance. Talking head synthesis, a driving mechanism for animating facial characteristics, has shown remarkable performance utilizing audio-driven and video-driven techniques. Advanced deep learning algorithms have boosted real-time rendering capabilities, yet ongoing constraints like temporal inconsistency and training data necessitate continued effort. Editing in talking head synthesis enables control over avatar modification, although decoupling is challenging due to the connected nature of attributes.

Different sorts of deep learning architecture are utilized to implement any model in THG. Deep learning architectures, such as Convolutional Neural Network (CNN) [17] and Recurrent neural network (RNN) [16], have proved crucial in AI breakthroughs [111]. CNNs, first created for image processing [111], have proved their capacity to tackle tough visual problems with better accuracy than earlier approaches [113]. RNNs, on the other hand, are created to handle sequence data and have applications in voice recognition and language

modeling [114]. Including LSTM units has increased the efficacy of these models by eliminating the vanishing gradient issue [111].

The universal approximation property of neural networks, initially established by Hornik et al. [120], states that a neural network with a single hidden layer may approximate any continuous function on a compact subset of  $\mathbb{R}^n$  to any desired degree of accuracy. This theorem explains why even primitive systems may convey intricate relationships in data. Neural networks are often divided into shallow and deep learning models, with shallow networks having one or two hidden layers, limited in their ability to learn complex features. In contrast, deep neural networks with multiple hidden layers can perform hierarchical feature extraction and more sophisticated data representations [111].

Talking-head video creation includes synthesizing lip motion sequences matching a driving source, such as voice or text. This technique also evaluates facial features like emotions and head movements. Early techniques employed cross-modal retrieval [186] and HMM-based algorithms [61], but they had high requirements. Recent breakthroughs in deep learning technology have permitted talking-head video generation approaches, utilizing 2D and 3D-based processes. A considerable amount of data must be applied for training, testing, and validation to train a model, build a talking head employing the above mentioned approaches, or even employ pre-existing models properly. Accuracy rises with the quantity of data, but only up to a limit; beyond that, fresh data may bring declining returns in performance. To ensure that the model does not start memorizing the input, rather, the model learns, Backpropagation was invented by Linnainmaa (1976) [107], is a method used to change link weights in neural networks to reduce learning errors. It was popularized in the 1980s and was later developed to correct mistakes during learning. Based on automated differentiation, this technique assures that the model learns without memory retention. It distributes error amounts across connections and determines the gradient of the cost function. Weight updates may be done using techniques such as stochastic gradient descent [110], extreme learning machines [108], no-prop networks [109], weightless networks, and non-connectionist neural networks.

Publicly available datasets have allowed for important improvements in talking head synthesis. The model we have studied indicates that VoxCeleb [72], VoxCeleb2 [173], and TalkingHead-1KH [169] are among the most commonly used datasets for video-driven tasks, while CREMA-D [134], LRW [140], and MEAD [159] have become prominent datasets for audio-driven tasks, according to our several recent surveys on talking head generation. Large-size, high-resolution datasets like FFHQ [137], HDTF [138], and CelebV-HQ [175] evolved as a result of the growing necessity for higher image quality brought about by the rise of application scenarios. In-depth details about each dataset, such as picture size, modality, and subject perspective, have been acquired; these factors have not been fully investigated in past evaluations of talking head synthesis.

While training a model, a critical component in training models for optimal accuracy is the design and implementation of a robust loss function, which assesses the difference between anticipated outputs and actual ground truth. A loss function is used in training models to improve parameters by examining the difference between anticipated and expected outcomes [91]. Although convexity assures that any local minimum is also a global minimum, which is theoretically ideal, many effective deep learning loss functions are non-convex. However, they may be efficiently improved utilizing modern approaches [91]. An efficient loss function is created to be resilient to outliers and differentiable to allow smooth gradient-based optimization. Convex loss functions are popular since they may be optimized using gradient-based approaches. Differentiability is vital for allowing gradient-based optimization. Robustness indicates that loss functions can support outliers without being influenced by a small number of extreme values. Smoothness provides a steady gradient without sudden transitions or spikes. Sparsity encourages sparse output, suited for high-dimensional data and tiny features. In non-convex conditions, a monotonic drop in the loss function does not assure convergence to the optimal solution, even if it would demonstrate consistent optimization progress.

As per Chen et al. [49], the performance of talking head synthesis models is significantly examined based on four fundamental criteria which are identity preservation, visual quality, lip synchronization, and natural motion. However, models usually focus on increases in a single component and qualitative and quantitative assessments are commonly applied. Quantitative measures offer a more objective analysis, but qualitative judgments rely on direct observation and could be subjective. This section includes some well-known quantitative criteria for a complete examination of talking head synthesis models.

Additionally, the multilingual component of talking head synthesis is another problem. The paucity of annotated, high-quality datasets across numerous languages, with diverse phonetic patterns, lip movements, and cultural intricacies, inhibits the general deployment of these systems [88]. To tackle this challenge, researchers should foster multilingual datasets and employ sophisticated methodologies like self-supervised and transfer learning. Complementary options include approaches like Long Short-Term Memory (LSTM)

[256] networks to acquire mouth landmarks from audio, the Wav2Lip model [21] to automate lip-syncing, and a strategy for reenactment that stresses visual signals for managing different voices and modifying face gestures depending on audio-derived speech styles [89].

The methodical technique clarifies the present situation and brings intriguing prospects for further inquiry. This article seeks thorough insights, including for researchers in the THG research topic. Despite the great benefits of THG, it tackles several issues, such as its dependence on pre-trained models, the management of extreme postures, and the necessity for high-quality datasets and sophisticated algorithms [88]. These limits may hamper innovation and adaptation for varied use cases. Future research could consider training individual components or modules on large-scale, diverse datasets to develop more modular and adaptable frameworks. Hybrid architectures that blend pre-trained models with task-specific layers or modules may give both pre-trained knowledge and targeted flexibility advantages. To solve problems such as large-angle poses, researchers aim to enhance datasets and apply multi-view training methodologies to capture a larger range of face orientations [88]. Maintaining temporal consistency is critical for providing smooth visual outputs, and tackling this involves high-quality datasets and innovative approaches [88].

The field of THG is experiencing consistent growth each day. Figure 1 illustrates the annual number of reported works in this area based on data collected from Google Scholar, including publications from academic publishers, professional societies, online repositories, universities, and other relevant websites under the keyword "THG" and its associated terms. For 2025, the publication count covers the period from January through April.

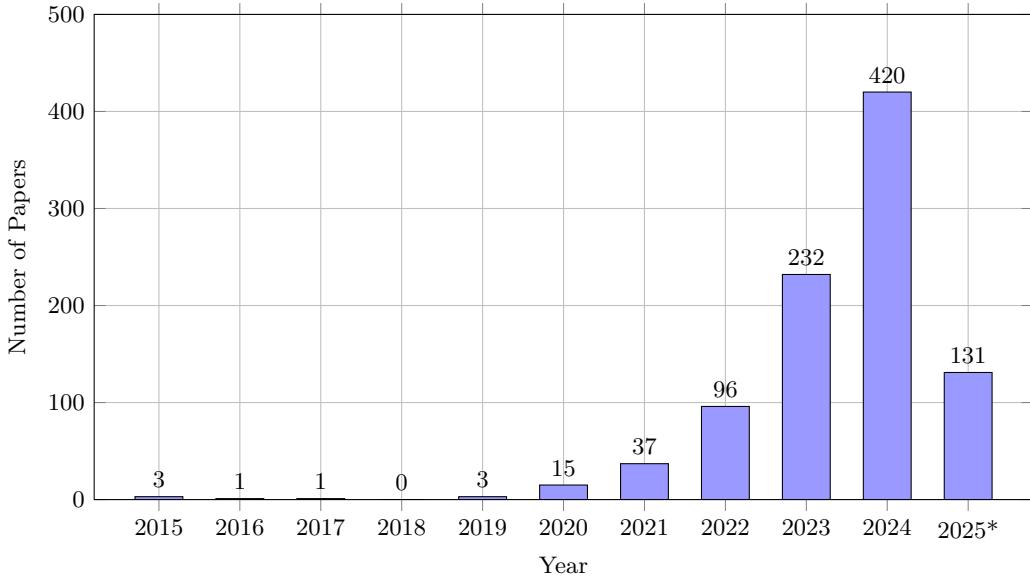


Figure 1: Number of papers published per year with the keyword “Talking Head Generation” from 2015 to 2025<sup>2</sup>

A comprehensive survey has been developed, categorizing various highly cited scholarly works based on input modality, model architecture, and training paradigms. This framework enhances our understanding of the evolution of THG.

- An in-depth analysis of critical components, including portrait generation, driving mechanisms, and editing techniques, has been conducted. This analysis reveals the technical trade-offs associated with identity preservation, realism, and controllability.
- A thorough comparison of widely utilized datasets has been performed. This comparison emphasizes key factors: resolution, modality, subject diversity, and applicability to various learning tasks.
- An extensive review of commonly employed loss functions and evaluation metrics has been undertaken. This review addresses their significance in optimizing model performance and improving perceptual quality.
- Several key research gaps have been identified, including multilingual synthesis, pose diversity, temporal consistency, and the risk of over reliance on pretrained models. Recommendations for

future research include the exploration of modular architectures and the adoption of diverse training strategies.

Although several recent overviews Gowda et al. (2023) [87] and Nguyen-Le et al. (2024) [2] provide broad surveys of 2D/3D pipelines up to 2024, but offer limited discussion of emerging neural radiance field and diffusion-based approaches. Dedicated NeRF reviews [90] and general 3D generation summaries [3] tend to treat audio-driven lip-sync and modular design only peripherally. At the same time, security-focused deepfake detection works [2] [88] do not fully trace the evolution of underlying generative methodologies. Similarly, existing evaluations of perceptual metrics [4] and foundational THG principles [49] predate recent advances in real-time rendering and temporal consistency. By contrast, the present survey spans innovations from 2017 through April 2025, offering detailed architectural and loss-function comparisons across ten distinct THG paradigms and explicitly linking each method to a dozen practical applications—ranging from video conferencing and remote education to visual effects. It is also the first to unify 3D Gaussian splatting and latent consistency models within a single analytical framework, and it proposes a set of ethical safeguards—such as robust watermarking protocols and the curation of multilingual datasets—to support responsible deployment of THG technologies.

This article outlines main ideas of talking head development, research scope, ethical difficulties, and methods. It describes the architecture of a proposed system and discusses training and inference processes, the dataset, loss function, and evaluation metrics. A complete overview of the whole research may be viewed in the below Figure 2.

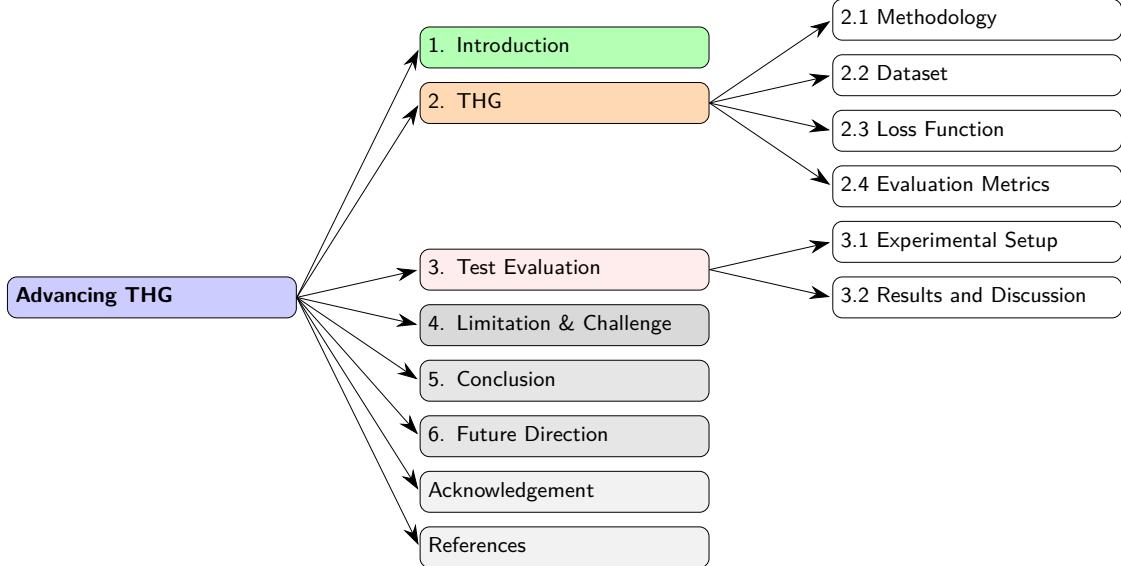


Figure 2: Structure of the Review Paper on Talking Head Generation

## 2 Talking Head Generation (THG)

Walter Pitts and Warren McCulloch introduced the foundational concepts of Deep Learning (DL) and Artificial Neural Networks (ANNs) in 1943 [123], who showed the capacity of theoretical artificial neuron networks to complete basic logical tasks. However, recent developments in processing power, notably the introduction of high-performance Graphics Processing Units (GPUs), have substantially facilitated the deployment of complicated deep learning models, such as those applied in THG [122]. Our investigation thoroughly examined over few hundreds of articles linked to the THG.

### 2.1 Comparison with Existing Relevant Review Articles

This comprehensive review identified approximately 100 methods for creating talking heads that were frequently referenced in various credible academic sources and selected for further research. The papers examined in these publications were released mainly between 2017 and April 2025. We also included a

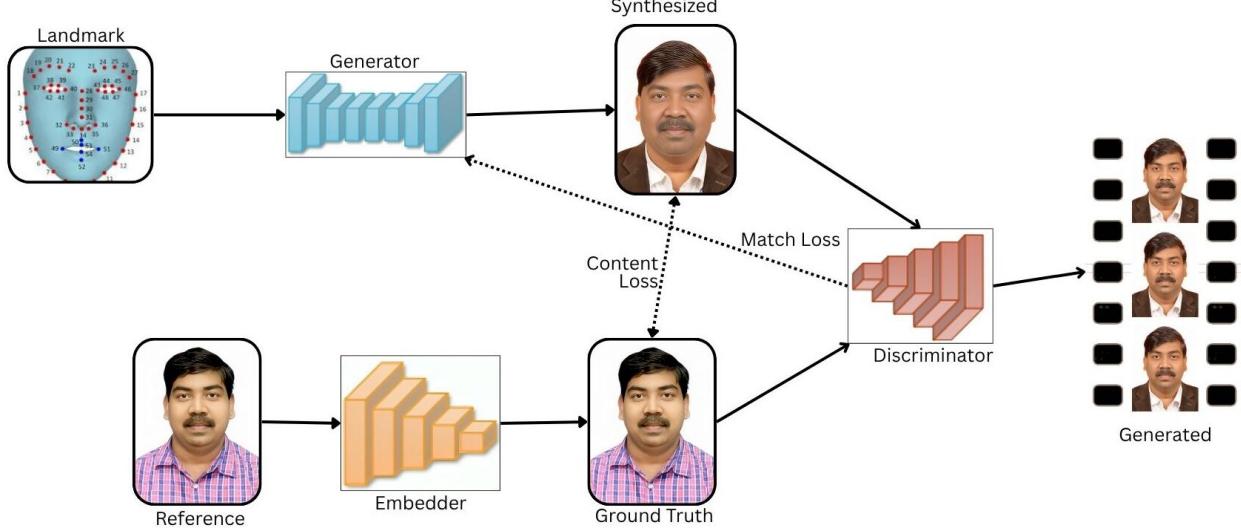


Figure 3: Generalized Approach for Image based

selection of previous works to trace the origins and foundational frameworks of the subject. The evaluation of the systems for THG focused on their performance in ten essential areas: picture-based, audio-based, text-based, video-based, 2D-based, 3D-based, distortion-based, Neural Radiance Fields (NeRF) based, parameter efficiency, and 3D animation. Many techniques are used in the THG process to create lifelike human faces coordinated with parametric, textual, video, or audio inputs. This section divides these approaches into ten paradigms by examining their designs, datasets, advantages, disadvantages, and applications.

### 2.1.1 Image-based THG

Image-based THG is a technique that animates a human head by transferring motion information learned from a driving video. This method generates realistic head movements and facial expressions from a single source image. The base architecture consists of four core components: motion encoding via self-learned key points, a local affine transformation module, an occlusion aware generator, and an extended equivariance loss. The system learns to extract key points from the source image, encoding the motion information by tracking their trajectories. The architecture utilizes local affine transformations to capture complex motion dynamics. The occlusion-aware Generator fills in or infers missing information, ensuring the generated frames remain coherent even during significant motion or occlusion events. A theoretical framework for synthesizing talking heads using images is presented, along with a neural network architecture similar to a GAN, in 3, supplemented by an embedding and landmark-based motion module. The model has four components: Input: Landmark and Source Image; a generator that utilizes landmarks and identity information from the source image; an embedder that extracts identity features from the source image to inform the Generator; synthesized output that produces a frame emulating the pose/expression of the input landmark while retaining the identity from the source image; ground truth representing the authentic target frame corresponding to the historic; loss functions that evaluate pixel/feature similarity and embedding similarity between generated and ground truth images to maintain identity and realism; a discriminator that attempts to differentiate between real and synthesized images to yield more realistic frames; Output: generated video, a sequence of synthesized frames that forms a talking head video, where head movements and expressions correspond to the landmark inputs. The shown approach generates realistic talking head films from a single source picture and driving markers, often derived from an alternative video. Loss functions such as Content Loss and Match Loss guarantee realism and identity coherence, while the Discriminator compels the Generator to provide photorealistic outcomes.

The numerous perspectives of image-based THG, such as its highlights, limitations with the main dataset for training, and basic architecture of the various models, are evaluated in Table 1, focusing on ways to employ one-shot for creation. The primary issues involve disentangling and altering the THG's look and motion components. The SMA [152] technique employs VQGANs and jointly learns codebooks for both aspects, allowing for more relatable control over motion flows. TS-Net [229], a dual-branch technique, promotes identity

Table 1: Comparison of Image Based Approaches

Method	Arch.	Dataset	Highlights	Limitations	N-shot
SMA [152]	VQGAN	VoxCeleb1 [172], CelebV-HQ [175]	Joint motion + appearance codebooks for smooth motion flow.	Keypoint motion estimation needed to reduce appearance leakage.	One
TS-Net [229]	GAN	FaceForensics	Dual-branch with warp-free and GRID [176] transformation.	Inconsistent motion, lacks high-freq detail.	One
DaGAN [31]	GAN	VoxCeleb1 [172], CelebV	Self-supervised geometry for synthetic faces.	Quality degrades with poor input; costly.	One
SAFA [219]	FAN	Voxceleb1	3DMM-driven structure-aware animation.	Sensitive to occlusion, overfitting.	One
Face2Face[86]	CNN	Face2Face	Real-time expression transfer from monocular view.	Lambertian assumptions + latency issues.	One
CrossID-GAN [164]	GAN	300VW	Multi-ID reenactment with landmark mapping.	Needs real-time and accurate landmarks.	One
MarioNETte [202]	U-Net	VoxCeleb1 [172], CelebV	Identity via image attention + landmark transform.	Pose and detection errors limit scale.	One
FOMM [20]	Monkey-Net	Multiple	Decouples appearance/motion; handles occlusion.	Requires object-specific tuning + data.	One
AAO [59]	Monkey-Net	Multiple	Dense motion + keypoint prediction.	Fails with poor keypoints.	One
ReenactGAN [215]	GAN	WFLW, CelebV	Reenactment via boundary space transfer.	Slow, data hungry.	One
X2Face [232]	U-Net CNN	VGG-Face, VoxCeleb CACD, GTAV	Self-supervised pose/exp control. AU-based expression synthesis.	Poor generalization to large poses. Ethical concerns + data load.	One

retention and resistance against occlusions. DaGAN [31], a GAN-based technique, focuses on self-supervised geometry learning. However, its performance may be impeded by extreme postures and substantial occlusions. SAFA [219], a structure-aware technique, integrates 3D Morphable Models (3DMMs) to produce more realistic animations with higher perceptual quality. Face2Face[186], a revolutionary technology for real-time face replication, exhibits possibilities for live interactive applications. CrossID-GAN [164], a network intended for multi-identity face reconstruction, displays resilience and efficacy in qualitative and quantitative tests. MarioNETte [202], a U-Net-based architecture, focuses on strong identity retention. However, its performance depends on reliable landmark recognition and issues managing significant pose fluctuations, boosting inference speed, and scaling to higher-definition movies. The FOMM [20] is a flexible framework for picture animation that can handle many object types beyond faces. "Monkey-Net" architecture learns motion representations from driving footage and applies them to source photos. AAO [59]refines the picture animation pipeline by employing a key point detector, a Dense Motion prediction network, and a Motion Transfer Network. ReenactGAN [215] seeks photorealistic and real-time face reenactment by transferring facial motions from a source to a target video. However, its success relies on the availability of adequate training data and may struggle with hidden identities or extreme positions. X2Face [232] provides a neural network model that uses another face or modality as input to change the expression and location of a target face. GATH [55] uses continuous Action Unit (AU) coefficients to generate facial emotions from still images automatically. Quality and diversity are the cornerstones of its success. Furthermore, it is important to address the ethical issues of face expression synthesis carefully. Future research in image-based THG will primarily address these issues, explore more reliable and effective architectures, develop innovative methods for handling challenging scenarios, enhance the temporal coherence of produced videos, and investigate strategies that can generalize successfully to unseen identities while maintaining high fidelity and controllability.

In addition to the most cited paper, various models and their applications in the fields, including facial recognition, video conferencing, and facial recognition. Tencent has developed the HunyuanPortrait [158] model in 2025, while ByteDance has developed the X-Portrait [157] model in 2023. IIIT Hyderabad developed the AVFR-GAN [156] in 2022, an audio-visual face reenactment model animating a source image by transferring head motion from a driving video. Tsinghua University has developed one-shot high-resolution editable talking face generation via pre-trained StyleGAN [155] in 2022. Samsung AI has developed one-shot megapixel neural head avatars focusing on cross-driving synthesis in 2022. Finally, Samsung AI has developed a multi-dimensional face recognition model called MegaPortraits [154] in 2022. These models create realistic, realistic, and interactive facial recognition models. These models have been developed to create realistic, realistic, and interactive facial recognition models. Various institutions and researchers have developed them, and their use in various fields has shown promising results. A few other relatively highly cited models include LivePortrait [153], SMA [152], a multiscale framework that synergizes motion and appearance for enhanced realism; MCNET [149], which leverages implicit identity representations conditioned on input frames for coherent video generation; TPSM [151], introducing thin-plate spline motion modeling for smooth facial deformations; StyleHEAT [150], an editable, high-resolution one-shot facial animation technique; DAM [148], which integrates structure-aware deformable motion transfer; StyleMask [147], disentangling style spaces to offer fine-grained control over facial attributes; AniFaceGAN [218], providing 3D-aware animatable face generation; IW [217], using implicit warping for seamless image animation; LIA [216], a latent image animator that learns dynamic representations directly in the latent space; and many others. The state of the art in producing realistic, interactive, and high-fidelity facial representations has been greatly enhanced by these models, which collectively cover a wide range of capabilities, from effective, real-time animation to ultra-high-resolution editable outputs.

### 2.1.2 Audio-Based THG

Audio-driven techniques use cross-modal alignment and acoustic feature extraction to synthesize facial emotions and lip movements from speech data. Google Scholar citation metrics show that Talk3D [226], EMO[160], and GC-AVT [143] are important models. While EMO [160] employs a ReferenceNet with dynamic audio-visual attention to capture subtle emotional expressions, Talk3D [226] uses a 3D-aware generative prior and audio-guided attention U-Net. Editing emotions and styles is possible with GC-AVT [143], although complicated backdrops provide difficulties. Challenges include limited datasets for non-English phonetic patterns and speech variability. Future approaches may involve real-time rendering pipelines for low-latency applications and self-supervised learning for cross-lingual adaptability. A theoretical framework for audio-based talking head synthesis is shown with a neural network architecture akin to a GAN in Figure 4, supplemented by an embedding and landmark-based motion module, which involves a pre-trained model (implicitly represented by the Generator and Discriminator) and a reference visual identity (multiple image frames). The input is the source audio, which determines the required speech and lip movements of the created

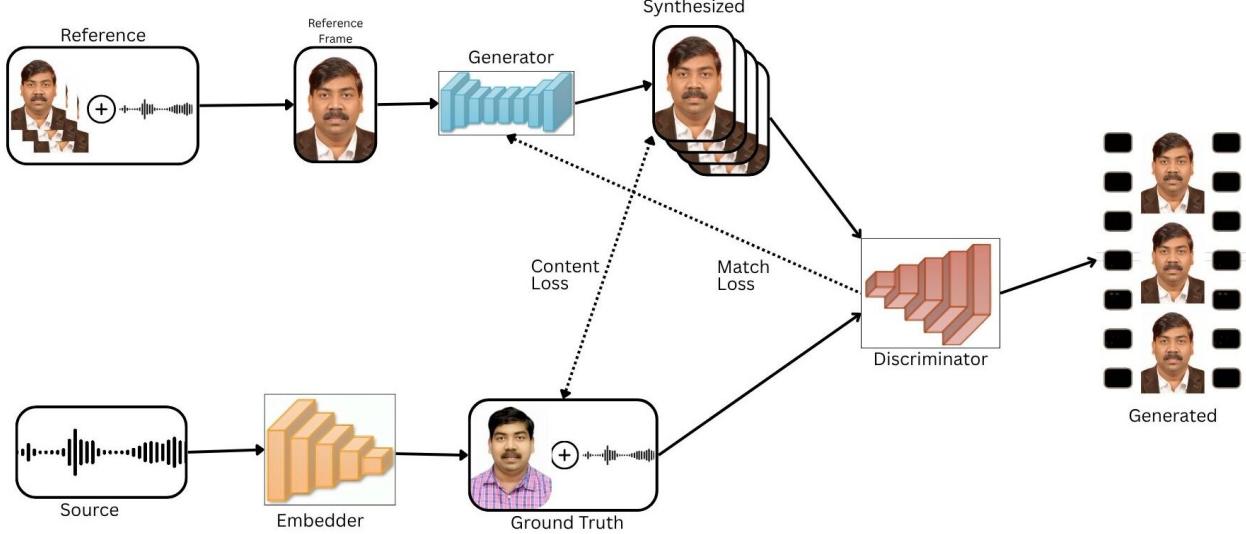


Figure 4: Generalized Approach for Audio based

talking head. The reference video illustrates the visual identity of the person speaking. The model components include the Embedder, which extracts phonetic and prosodic information from the source audio, and the Generator, which synthesizes a new sequence of video frames where the person from the Reference Frame appears to be speaking the content of the Source Audio, with synchronized lip movements and potentially matching head poses or expressions. The Discriminator operates as an antagonistic critic, receiving two inputs: the synthesized video frames created by the Generator and the Ground Truth. Its purpose is to discern between the realistically seeming synthetic video and genuine video frames, producing a score reflecting its confidence in the "realness" of the input. The figure 4 highlights two types of losses that guide the training of the Generator: Content Loss, which measures the difference between the content (e.g., lip movements, facial expressions) of the Synthesized video and the Ground Truth, and Match Loss, derived from the output of the Discriminator, encouraging the Generator to produce video frames that are indistinguishable from real video frames. By attempting to "fool" the Discriminator, the Generator learns to generate increasingly photorealistic and natural-looking talking head sequences. The trained model's end product is a video that shows the individual from the Reference Video speaking the content of the Source Audio, complete with realistic facial expressions and well-coordinated lip motions. This simplified explanation outlines a common method for generating audio-driven talking heads using deep learning techniques, likely involving a Generative Adversarial Network (GAN) architecture. It highlights the importance of a reference image for establishing visual identity, audio embeddings for controlling facial movements, and adversarial learning to produce realistic outputs.

The various approaches to audio-based THG are reviewed in Table 2. This review highlights the strengths and limitations of the major datasets used for training and the fundamental architecture of different models. One notable approach is OmniHuman-1, which utilizes a Diffusion Transformer-based architecture for end-to-end human animation, producing incredibly realistic human films. It adds motion-related circumstances directly into the training process, resulting in excellent realism. EMO [160], another N-shot technique, focuses on the dynamic interaction between auditory cues and facial movements, seeking to capture more expressive facial signals. Talk3D [226], another N-shot technique, stresses the realistic reconstruction of face geometry using a tailored 3D-aware generative prior and an audio-guided attention U-Net architecture. However, the drawbacks of Talk3D [226] include its lack of generalizability beyond lifelike human faces and its probable limits in non-human characters. The article explores numerous ways to make high-fidelity, multi-person talking portraits, including MODA [206], GC-AVT [143], Flow-guided One-shot, Audio2Head [142], and DAVS [15]. MODA [206] is a complete system that concentrates on audio and visual characteristics. GC-AVT [143] is an audio-visual talking head model that allows granular control over lip movements, head positions, and facial expressions. GC-AVT [143] largely focuses on accommodating varied speakers and a spectrum of emotions in speech. Flow-guided One-shot [138] is a flow-guided talking face creation framework, particularly built for high-definition facial movies. Audio2Head [142] is an audio-driven talking-head approach that seeks to make

Table 2: Comparison of Audio Based Approaches

Method	Arch.	Dataset	Highlights	Limitations	N-shot
OmniHuman-1 [210]	3DVAE	CelebV-HQ [175], RAVDESS [66]	Diffusion Transformer with motion conditioning for realism.	Scalability, data dependency, and ethical concerns.	One
EMO [160]	RefNet	CREMA	Audio-facial correlation enhances realism.	Expression complexity, quality, dependency.	One
Talk3D [226]	GAN	AD-NeRF [127], Obama Weekly	Audio-driven 3D prior and attention U-Net.	Prep complexity, artifacts, poor generalization.	Multi
MODA [206]	LSGAN	HDTF [138], LSP [1]	Dual-attention and facial composer.	Expression variance, compute cost.	One
GC-AVT [143]	GAN	VoxCeleb2 [173], MEAD [159]	Granular control over lip, pose, and expression.	Low-res output and poor backgrounds.	Multi
Flow-guided [48]	3DMM	HDTF [138]	AV flow-guided with motion realism.	Coherence, cropping, and style gaps.	One
Audio2Head [142]	RNN	LRW, GRID [176], VoxCeleb	Predicts pose and motion from audio/image.	Identity mismatch, ethical misuse.	One
DAVS [15]	GAN	LRW	Enhances lip realism and intelligibility.	Speed and variability limitations.	One
LMGG [19]	GAN	GRID [176], LDC, LRW	Cross-modal speech-lip fusion.	Realism + scale lacking, costly.	Multi
Facial Reenact. [120]	GAN	CelebA	cGAN + RNN for audio-face sync.	Speaker variability, weak realism.	Multi
VisemeNet [18]	LSTM	GRID [176], SAVEE, BIWI 3D	LSTM viseme curves improve sync robustness.	emotion diversity is limited.	Multi

photorealistic films from a single reference picture of the target individual. DAVS [15] is a one-shot learning technique that leverages a single reference picture for numerous identity and audio inputs. However, the study notes the potential for exploitation of such technology for harmful reasons and aims to disclose code and models to minimize this. DAVS [15] is a system that utilizes voice recordings to build realistic facial pictures, concentrating on enhancing lip motion patterns. It has potential uses in automated lip reading and video retrieval. The "LMGG" is a one-shot solution that integrates audio and visual embeddings to establish synchronization between produced lip movements and speech. However, it suffers constraints in accuracy, generalization, photorealism, and processing expenses. The "N-Shot" approach employs recurrent neural networks and conditional generative adversarial networks to build photorealistic faces with accurate lip synchronization. The VisemeNet [18] approach is meant to produce animator-centric speech motion curves but has issues addressing speaker variability and introducing emotional context. Overall, these audio-based THG algorithms indicate breakthroughs in human animation but also confront issues in scalability, data reliance, and controlling computing needs for real-time applications.

Audio-based models have seen a significant expansion with several innovative models. These include ACTalker [239] (2025) integrates multimodal control signals—including pose, expression, and audio—to guide high-fidelity video diffusion. AniPortrait [240] (2024) transforms static portraits into expressive animations driven solely by speech. EDTalk [241] (2024) employs efficient disentanglement to capture and transfer emotional nuances in facial movements. EchoMimic [242] (2024) mimics vocal subtleties to produce lifelike, audio-driven portrait animations. FD2Talk [245] (2024) generalizes talking-head generation across diverse faces and speaking styles with a unified architecture. FaceChain-ImagineID [244] (2024) offers free-form identity manipulation to craft high-fidelity talking heads from arbitrary source images. FlowVQTalker [246] (2024) leverages flow-based quantization to achieve emotionally rich and identity-consistent talking faces. MuseTalk [249] (2025) operates in a latent video space to deliver real-time, high-fidelity lip-sync results on unseen speakers. ReSyncer [250] (2024) provides a plug-and-play framework for resynchronizing any facial model with arbitrary audio inputs, enhancing lip-sync accuracy. Real3DPortrait [251] (2024) reconstructs and animates realistic 3D avatars from a single image using audio cues. Emotional Conversation [243] (2024) empowers talking heads with contextual emotional modulation to make dialogues more engaging. Make Your Actor Talk [248] (2024) focuses on generalizable audio-driven actor synthesis across varied domains. RealTalk [252] (2024) achieves real-time, high-fidelity audio-driven portrait generation optimized for interactive applications. AAAI [253] (2024) incorporates style-transfer techniques to produce stylized talking-head animations. Style2Talker [254] (2024) extends this by generating high-resolution talking-head videos with fine-grained style control. A model like ManiTalk [273] (2024) advances audio-driven talking head generation by enabling explicit manipulation of facial details such as eyebrows, eyelids, and pupils through a three-stage pipeline. The method combines synchronized landmark generation with parameterized facial control and image warping, achieving state-of-the-art lip synchronization and identity preservation. Its focus on stylized expression control aligns with emerging trends in personalized avatar synthesis. Finally, THQA [255] (2024) introduces a perceptual quality assessment metric specifically designed for evaluating talking-head animations. Together, these innovations significantly broaden the capabilities of audio-driven facial animation, enabling more expressive, controllable, and realistic talking heads.

### 2.1.3 Video-Based THG

Video-driven techniques focus on preserving identity and ensuring temporal coherence by replaying source movies while transferring motion from driving sequences. According to Google Scholar citation metrics, the HiDe-NeRF [195], DFA-NeRF [43], and Neural Talking-Head [42] models are important. DFA-NeRF [43] offers high fidelity in results but is time-consuming, while HiDe-NeRF [195] maintains identity even under significant deformations. Neural Talking-Head employs H.264 compression with unsupervised 3D keypoints for bandwidth-efficient video conferencing. Occlusion handling and performance dips on invisible identities are among the difficulties. Future directions include advanced training methods for improving realism and the development of temporal transformers. A video-based talking head synthesis system producing a talking head from audio and reference posture data is shown in Figure 5. Comprising two parts, the model is the Embedder, which gathers phonetic and prosodic data from the original audio, and the Generator, which produces a new sequence of video frames to simulate the speaker's speech and lip movements. It lets one perfectly reproduce the speaker's unique style. The Generator produces a new sequence of video frames that gives the impression that the person in the reference video speaks the words from the source audio. The Generator creates a new sequence of video frames that makes it look like the person in the reference video speaks the words from the source audio. This method entails synchronizing lip motions and matching the pose from the reference frame. The Discriminator, an adversarial critic, takes two inputs: the synthesized video frames created by the Generator and the Ground Truth, which presumably relates to genuine video

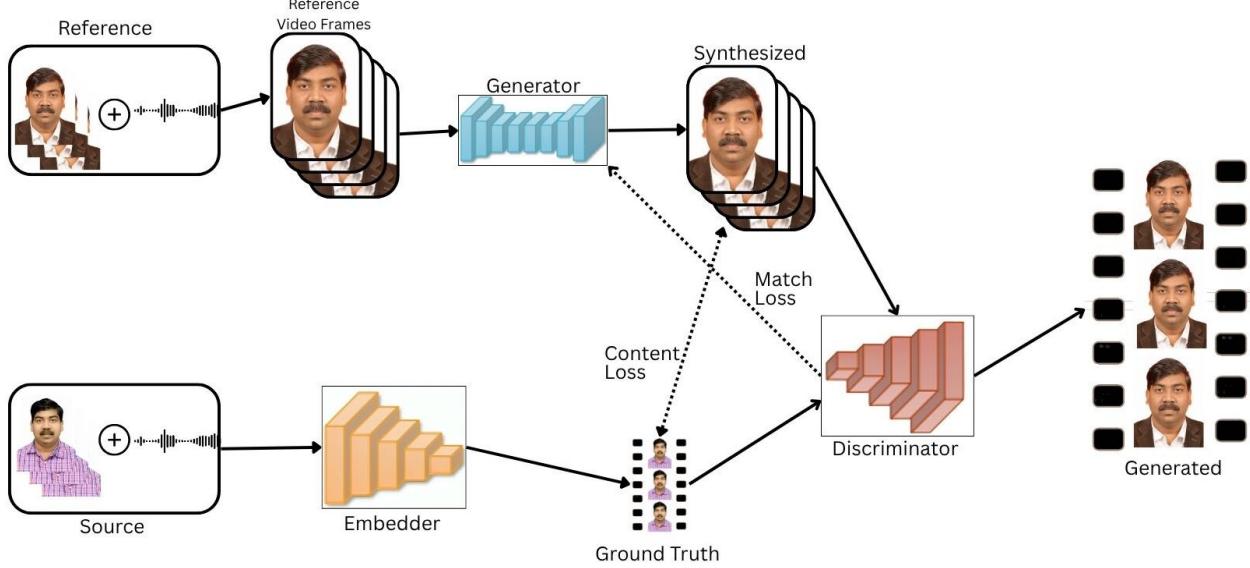


Figure 5: Generalized Approach for Video-based

frames of the same person uttering certain audio and demonstrating a specific stance. The figure 5 highlights two types of losses that guide the training of the Generator: Content Loss, which measures the difference between the content (e.g., lip movements, facial expressions, and importantly, the pose) of the Synthesized video and the Ground Truth, and Match Loss, which is derived from the output of the Discriminator. This loss drives the Generator to generate video frames that are indistinguishable from genuine video frames in terms of visual quality and coherence. By challenging the Discriminator, the Generator learns to create increasingly photorealistic and natural-looking talking head sequences that align with the required position. The ultimate result of the trained model is the generated video, which is a series of video frames portraying the person from the Reference Video saying the content of the Source Audio, with synced lip movements and adopting the position from the given Reference position frame. The picture depicts a system where a pre-trained model combines a reference visual identity, target audio, and desired position to build a new talking head video. The training involves mapping audio information and desired postures to realistic facial movements, overall body posture, and appearance by minimizing Content Loss and Match Loss through adversarial learning.

Several video-based THG techniques are evaluated in Table 3, emphasizing multi-shot or few-shot learning paradigms. These methods utilize existing video data to create new talking head sequences, typically focusing on reenactment, dubbing, or generating fresh content. The examination analyzes architectural designs, datasets employed, important developments, and inherent limits of these video-driven systems, intending to make realistic and cohesive talking head films. DISCOHEAD [36] provides an innovative approach to creating realistic talking heads, emphasizing enhanced efficiency through a dense motion estimator and encoder. It focuses on managing the mouth area according to spoken sounds, ensuring proper lip synchronization. The architecture of DISCOHEAD [36] leverages a ResNet-18 backbone to extract visual features from input video frames, which are subsequently processed by the dense motion estimator and encoder to learn face motions, notably in the mouth region, driven by the audio. HiDe-NeRF [195] proposes a revolutionary technique for producing talking heads by exploiting Neural Radiance Fields (NeRF), allowing for realistic face distortion while scrupulously retaining the subject's identity. The method utilizes a 3D Morphable Model (3DMM) as a fundamental component, enabling new view synthesis and audio-driven facial deformations while prioritizing identity preservation. However, the research highlights several significant challenges associated with HiDe-NeRF [195], including difficulties in handling face occlusions, pose bias prevalent in training datasets, and the potential misuse of such realistic creations for harmful purposes, such as "DeepFakes." DFA-NeRF [43] is a unique framework that leverages Neural Radiance Fields (NeRF) for high-fidelity, individualized talking head creation. It highlights the usefulness of deep neural networks, notably NeRFs, in creating realistic and identity-specific talking head synthesis. MMVID [30] is a multimodal video-generating framework meant to increase the quality, consistency, and variety of videos produced. Face-Dubbing++ [185] provides a

Table 3: Comparison of Video Based Approaches

Method	Arch.	Dataset	Highlights	Limitations	N-shot
DISCOHEAD [36]	ResNet-18	Obama, GRID [176], KoEBA	Dense motion estimator with encoder for expressive mouth region synthesis.	Needs high-quality input; struggles with complex backgrounds.	Multi
HiDe-NeRF [195]	3DMM	VoxCeleb1 [172], TH-1KH	Preserves identity under deformation for 3D face synthesis.	Occlusion challenges; DeepFake misuse risk.	Multi
DFA-NeRF [43]	NeRF	LRS2 [177], HDTF [138]	Long rendering time; diarization dependency.	Multi	
MMVTD [30]	GAN	MUG, iPER, VoxCeleb	Fuses visual modalities for diverse generation.	Motion consistency still limited.	Multi
Face-Dubbing++ [185]	GAN	LRS2 [177]	Multilingual dubbing with sync accuracy.	Prosody and ASR mismatch hurt realism.	Multi
Neural Talking-Head [42]	GAN	VoxCeleb2 [173], TH-1KH	Bandwidth-efficient synthesis with unsupervised 3D keypoints.	Prone to artifacts under pose variation.	Multi
FT [191]	VGG19	VoxCeleb1 [172], VoxCeleb2 [173]	Few-shot generation with meta-learning init.	Trade-off in realism and generalization.	Multi
RHM [49]	GAN	CREMA-D [134], LRS3-TED [139]	3D-aware hybrid embedding for photo-realism.	Complex design, scalability issues.	Multi
vid2vid [50]	GAN	YouTube, Street-scene	Few-shot with adaptive weights.	Fails with unseen CG styles.	Multi
Speech2Vid [220]	CNN	VoxCeleb, LRW	Real-time video generation from speech + image.	Weakness with accents and wide expressions.	Multi

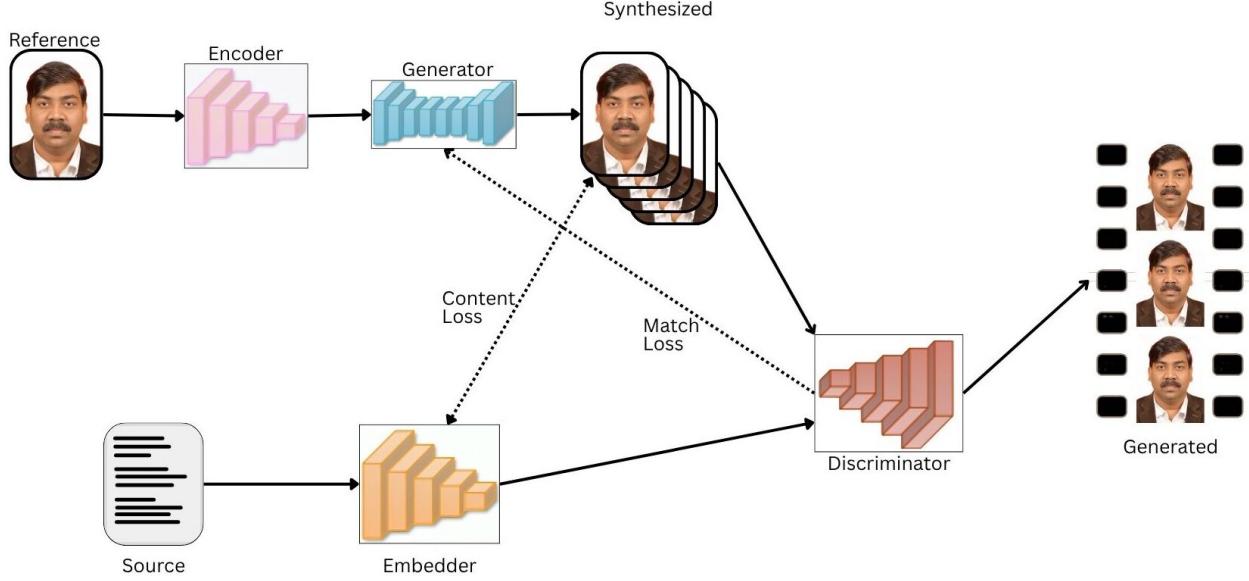


Figure 6: Generalized Approach for Text-based

powerful neural system for voice-preserving, lip-synchronous video translation, incorporating many unique neural network models. Neural Talking-Head [42] is a video synthesis model designed for video conferencing applications. It aims to provide visual quality comparable to established video compression standards while using significantly less bandwidth. FT [191] is a multi-shot system with few-shot learning capabilities that creates highly realistic images of human heads from just a few views of a person. RHM offers a 3D-aware generative network linked with a hybrid embedding and composition module, allowing the development of controlled, photorealistic, and temporally coherent talking-head films with natural head motions. The analyzed video-based THG methods employ diverse architectures such as ResNet-18, NeRF integrated with 3DMMs, specialized GANs with attention mechanisms and hybrid embeddings, and encoder-decoder CNNs, trained on datasets like Obama, VoxCeleb, LRS2, and in a few cases, self-collected video data. Finally, the Enhanced Temporal Representation and Spatial Alignment [274] method addresses motion inconsistency in video-driven synthesis by introducing temporal representation augmentation (TRA) and spatial alignment correction (SAC). These innovations improve motion feature learning and head-pose consistency, producing smoother, artifact-free talking videos with superior visual fidelity.

#### 2.1.4 Text-Based THG

Text-to-video models utilize emotion management and visual prediction modules to generate talking heads that respond to textual prompts. According to Google Scholar citation metrics, InstructAvatar [199], TalkCLIP [227], and FT2TF [77] are important models. A video-based talking head synthesis system producing a talking head from audio and reference posture data is shown in Figure 6. Comprising two parts, the model is the Embedder, which gathers phonetic and prosodic data from the original audio, and the Generator, which produces a new sequence of video frames to simulate the speaker’s speech and lip movements. It lets one perfectly reproduce the speaker’s unique style. The Generator produces a new sequence of video frames that gives the impression that the person in the reference video speaks the words from the source audio. A theoretical framework for text-based talking head synthesis is shown in Figure 6. This framework employs a pre-trained model to generate a new talking head video using a target text and a reference visual identity. The input is the source text, which dictates the intended speech and lip motions, and the reference picture gives the visual identity of the person who will be "talking." A video-based talking head synthesis system producing a talking head from audio and reference posture data is shown in Figure 6. Comprising two parts, the model is the Embedder, which gathers phonetic and prosodic data from the original audio, and the Generator, which produces a new sequence of video frames to simulate the speaker’s speech and lip movements. It lets one perfectly reproduce the speaker’s unique style. The Generator produces a new sequence of video frames that gives the impression that the person in the reference video speaks the words from the source audio. The Discriminator serves as an antagonistic critic, accepting two inputs: the synthetic

video frames created by the Generator and the Ground Truth, which presumably corresponds to genuine video frames of the same person saying the same or comparable material. The loss calculation and training (implicitly) entail two losses: Content Loss and Match Loss. Content Loss evaluates the difference between the content (e.g., lip movements, facial emotions) of the synthesized video and the Ground Truth, guided by the source text. The GeneratorGenerator aims to minimize Content loss to ensure that the generated lip movements and emotions align with the provided text. Match Loss is obtained from the output of the Discriminator, asking the Generator to produce video frames that are visually identical to real video frames in terms of coherence and quality. As illustrated in Figure 6, the video-based talking head synthesis system generates a talking head by fusing reference posture information with audio. To ensure a flawless replication of the speaker’s distinct style, the model comprises an Embedder that collects phonetic and prosodic information from the original audio and a Generator that generates a new video frame sequence to mimic the speaker’s speech and lip movements.

Various text-based talking head-generating strategies are assessed in Table 4, focusing on multi-shot or few-shot learning paradigms. These techniques use pre-existing video data to produce new talking head sequences, often concentrating on tasks such as reenactment, dubbing, or developing original material. The research examines architectural designs, employed datasets, significant developments, and inherent constraints of video-driven systems to produce realistic and cohesive talking head films. DISCOHEAD [36] introduces an innovative method for creating realistic talking heads, emphasizing enhanced efficiency through a dense motion estimator and encoder. This approach focuses on managing the mouth area by spoken sounds, ensuring proper lip synchronization. DISCOHEAD [36]’s architecture employs a ResNet-18 backbone to extract visual features from input video frames, which are then processed by a dense motion estimator and encoder to learn facial motions, specifically in the mouth region, influenced by the audio. HiDe-NeRF [195] introduces an innovative method for generating talking heads using Neural Radiance Fields (NeRF), enabling realistic face deformation while rigorously maintaining the individual’s identity. The approach utilizes a 3D Morphable Model (3DMM) as a core element, allowing innovative view synthesis and audio-driven face deformations while prioritizing identity preservation. The article highlights several significant issues with HiDe-NeRF [195], including difficulties in managing face occlusions, bias in training datasets related to poses, and the potential for misuse in creating “DeepFakes.” DFA-NeRF [43] is an innovative framework that utilizes Neural Radiance Fields (NeRF) to create high-fidelity, personalized talking heads. This framework demonstrates the effectiveness of deep neural networks, particularly NeRFs, in synthesizing realistic and identity-specific talking heads. MMVID [30] is a multimodal video generation framework designed to enhance the quality, consistency, and variety of the videos produced. Face-Dubbing++ [185] provides a powerful neural system for voice-preserving, lip-synchronous video translation, incorporating many unique neural network models. Developed for video conferencing applications, Neural Talking-Head [42] is a neural talking-head video synthesis model that aims to use significantly less bandwidth while maintaining good visual quality comparable to established video compression standards. FT [191] is a sophisticated multi-shot system that creates remarkably lifelike portraits of human heads using the power of few-shot learning. It creates vivid, immersive, realistic representations of people by capturing their essence and personality with only a few image perspectives. RHM offers a 3D-aware generative network linked with a hybrid embedding and composition module, allowing the development of controlled, photorealistic, and temporally coherent talking-head films with natural head motions. The reviewed methods for video-based THG utilize various architectures, including ResNet-18, NeRF combined with 3DMMs, specialized GANs equipped with attention mechanisms and hybrid embeddings, and encoder-decoder CNNs. These methods are trained on datasets such as Obama, VoxCeleb, LRS2 [177], and self-collected video data.

Further advancements in text-conditioned generation have opened new pathways for controllable and expressive talking head synthesis. GenCA [257] introduces a powerful text-conditioned generative model that interprets textual inputs to guide fine-grained facial animation, offering a new level of semantic control in audiovisual generation tasks. T3M [258] extends this concept to 3D space, proposing a text-guided 3D avatar generation framework that produces high-quality and identity-consistent 3D heads from natural language prompts, bridging the gap between text understanding and 3D modeling. STAR [259] advances toward full 4D synthesis by introducing a skeleton-aware, text-driven system that generates temporally coherent, expressive facial performances aligned with the described motion, making it suitable for dynamic human avatar applications. Complementing these approaches, Text-to-Video [260] proposes a two-stage zero-shot framework that maps text directly to video sequences, enabling general and flexible video generation, which can include talking heads as a subset of its broader generative scope.

Table 4: Comparison of Text Based Approaches

Method	Arch.	Dataset	Highlights	Limitations	N-shot
InstructAvatar [199]	VAE	MEAD [159]	Text-guided expressive 2D avatars with improved interactivity.	emotion granularity and text precision need work.	One
EMO [160]	VAE	CREMA	Realism via dynamic audio-facial alignment.	Data quality, expression complexity.	One
TalkCLIP [227]	T2SS	MEAD [159], HDTF [138], VoxCeleb2 [173]	Text-driven synthesis via CLIP modulation.	Abstract expressions, mouth artifacts.	One
FT2TF [77]	GAN	LRS2 [177], LRS3	Text + vision fusion yields SOTA talking faces.	High training cost, expressiveness gaps.	Multi
Difftalk [61]	GAN	HDTF [138]	Denoising diffusion for coherent motion.	Relies on reference image, heavy compute.	Multi
Face-Dubbing++ [185]	LSTM	LRS2 [177], TED, etc.	Voice-preserving dubbing with sync accuracy.	ASR + prosody mismatches, fragility.	Multi
GC-AVT [143]	GAN	VoxCeleb2 [173], MEAD [159]	Fine-grained pose + lip + expression control.	Limited realism from background masking.	Multi
Text2Video [228]	GAN	VidTIMIT	Phoneme-pose mapping for text-driven synthesis.	Lacks realism, generalization, speaker control.	Multi
Write-a-Speaker [231]	3DMM	Mocap	Text to head animation with rhythm and emotion control.	Motion capture needed; performance cost.	Multi
Flow Guided [138]	3DMM	HDTF	Flow-guided video from audio-text inputs.	Poor coherence, style inconsistency.	One
Audio2Head [142]	RNN	LRW, GRID [176], VoxCeleb	Audio-driven with pose + motion field prediction.	Misuse potential, identity mismatch.	One
Speech2Vid [220]	CNN	VoxCeleb, LRW	Real-time generation from audio + image.	Agent handling, realism issues.	One
LMGG [19]	CNN	GRID [176], LDC, LRW	Cross-modal synced lip movement generation.	Weak realism + efficiency.	Multi
Facial Reenactment [82]	GAN	CelebA	RNN + cGAN for expressive face sync.	Still lacks realism, consistency.	Multi
ObamaNet [209]	LSTM	Char2Wav, Pix2Pix	Full pipeline: text to lip-sync video.	Language support + ethics safeguards needed.	Multi

### 2.1.5 2D-Based THG

2D methods utilize landmark-driven warping and attention techniques, prioritizing computational efficiency. Style transmit, MetaPortrait [204], and MakeItTalk [201] are important models based on citation metrics from Google Scholar. These models transmit speaking styles across identities, although they have drawbacks such as artifacts in non-frontal perspectives. 2D warping struggles to accurately illustrate 3D head rotations and often loses high-frequency information, leading to depth ambiguity and issues with photorealism. Future advancements include diffusion-based detail enhancement and hybrid 2D-3D pipelines to improve efficiency and realism.

A variety of 2D model-based algorithms for generating talking heads have been explored in Table 5. These algorithms aim to synthesize realistic facial movements using a single source image or a limited set of pictures guided by audio or other control inputs. Techniques like style transfer, diffusion models, and landmark manipulation are routinely applied to generate credible outcomes. Style Transfer [224] is an innovative method for creating audio-driven talking head animations. Drawing on learned style references, it generates 2D animations using a single input image and an audio stream. The technique utilizes a ResNet50 backbone for visual feature extraction and is evaluated on the VoxCeleb2 [173] and RAVDESS [66] datasets. Diffused Heads [26] is an autoregressive diffusion model for making realistic talking head movies to attain state-of-the-art outcomes in expressiveness and smoothness. It leverages a GAN architecture inside an autoregressive diffusion framework and is assessed on the LRW and CREMA datasets. MetaPortrait [204] is a unique framework for identity-preserving one-shot talking head synthesis, combining dense facial landmarks, meta-learning approaches, 3D convolution for temporal modeling, and a generative prior. It is examined using VoxCeleb2 [173] and HDTF [138] datasets, indicating its capacity to manage various identities and create high-definition talking head movies from a single picture. LSP [1] is a multi-shot approach that employs a single input picture to build a talking head from a single reference image. LSP [1] is a live system that generates individualized talking-head animation using deep neural networks focused on capturing specific face dynamics and head movements for high-fidelity outcomes. It is trained on photorealistic photos and has been tested via user research. Key characteristics of LSP [1] include its capacity to build unique talking-head animations in a live system, obtain high-fidelity face details, and control overhead posture accurately. However, it has difficulties recording plosive consonants, high-speed speech, emotive audio, shadows and lighting reflections, and realistic motions beyond simple head movements. PC-AVS [118] is a multi-shot framework designed for creating posture-controllable talking faces using non-aligned raw facial photos and an implicit low-dimensional pose code. It utilizes a GAN architecture and has been evaluated on the VoxCeleb2 [173] and LRW datasets. The model can handle multiple identities and deliver accurate lip synchronization based on audio while allowing for control over facial positioning. MAKEItTalk is a one-shot approach for making talking-head videos from a single-face photograph utilizing audio input. It employs AUTOVC, an autoencoder-based voice conversion paradigm adaptable for visual feature creation. The approach has drawbacks relating to sparse landmark representation for video output and may struggle with managing complicated facial emotions beyond simple lip movements. Video Rewrite [230] is a facial animation technique that automatically generates new lip movements from a source video for movie dubbing. It utilizes a Beier-Neely warping technique for image modification and employs Hidden Markov Models (HMMs) for phoneme alignment and selection. However, this technique has limitations regarding the availability and accuracy of phoneme-aligned mouth images in the source video. Handling complex facial emotions beyond simple lip movements may also be difficult.

### 2.1.6 3D-Based THG

Morphable models are utilized in 3D techniques to distinguish between em, posture, and geometry. ADL [129] (Audio-Driven Lip Sync), PV3D [214], and JambaTalk [34] are important models according to Google Scholar citation metrics. ADL [129] backs multilingual lip sync; JambaTalk [34] runs the speech-to-motion transformer. PV3D [214] dely uses a 3D-aware GAN that cleverly includes motion dynamics despite its difficulties in capturing long-term dynamics, allowing a more interesting and immersive representation of movement in three dimensions. Rising computer expenses and the tediousness of data collecting create major issues. Still, neural rendering and single-camera footage's unsupervised 3D reconstruction offer fascinating possibilities for more development. A theoretical architecture for synthesizing text-based talking heads is illustrated in Figure 7. This architecture includes two inputs: the source image and the landmark sequence. The source image provides the visual identity and starting position of the individual whose 3D talking head is being created. Meanwhile, the landmark sequence conveys the appropriate facial expressions and movements. The model's core, the Generator, takes two inputs: the original picture and the landmark sequence. The procedure calls for animating a 3D model of a talking head that resembles the person in the original image according to a

Table 5: Comparison of 2D Based Approaches

Method	Arch.	Dataset	Highlights	Limitations	N-shot
Style Transfer [224]	ResNet50	VoxCeleb2 [173], RAVDESS [66]	2D animation using audio and style reference from one image.	Needs style frames; heavy computation.	Multi
Diffused Heads [26]	GAN	LRW, CREMA	Autoregressive diffusion for smooth expressive motion.	Too slow for real-time.	Multi
MetaPortrait [204]	3DMM	VoxCeleb2 [173], HDTF [138]	Identity preservation using dense landmarks and generative priors.	Blurring under occlusion;	Multi
LSP [1]	LSTM	—	Real-time facial animation from live speech.	Affected by lighting, emotion, tracking.	Multi
PC-AVS [118]	GAN	VoxCeleb2 [173], LRW	Pose control + lip-sync from raw face image input.	Training bias; not fully real-time.	One
MakelTalk [201]	AUTOVC	VoxCeleb2 [173]	Expression-aware synthesis from one image + audio.	Sparse landmarks may distort; lacks 3D cues.	One
Video Rewrite [230]	Beier-Neely	HMM, TIMIT	Classic method using phoneme-aligned mouth patches.	Weak for expressive motion; limited matching.	One

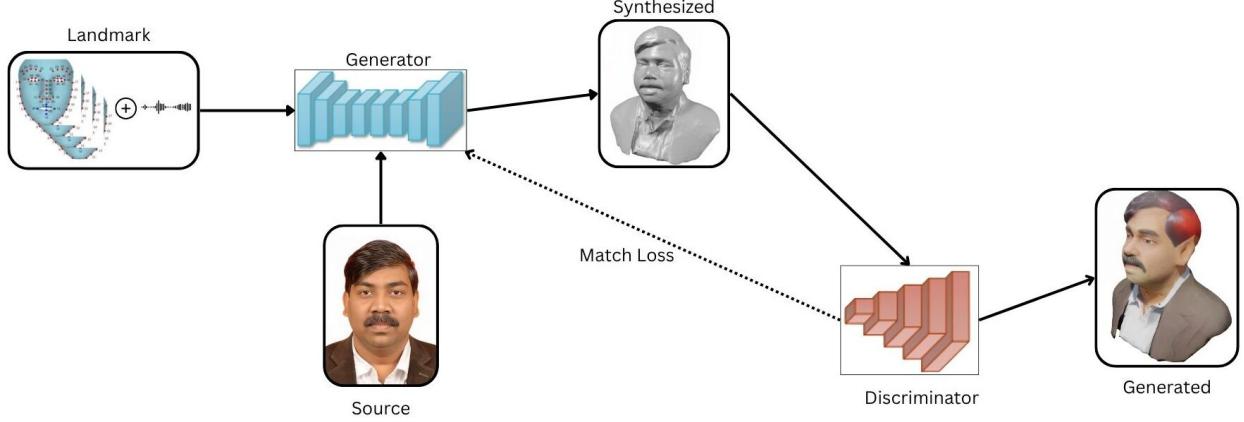


Figure 7: Generalized Approach for 3D-based

given sequence of landmarks. From this 3D model, 2D video frames may be produced. The Discriminator acts as an adversarial critic that receives two inputs: the synthetic 3D model and a realistic 3D talking head. Its role is to differentiate between the rendered output and the genuine 3D models, producing a score that reflects its confidence in the "realness" of the input. This adversarial setup is a critical feature of GAN, which is widely utilized for producing realistic 3D material. One sort of loss leads to the training of the Generator: Match Loss, generated from the output of the Discriminator. By attempting to "fool" the Discriminator, the Generator learns to build increasingly photorealistic and natural-looking 3D models that accurately animate according to the input landmarks. Other implicit losses may be linked to the precision of the produced 3D geometry and texture relative to the source picture and the fidelity of the animation to the input landmarks. The 3D-based THG model uses a pre-trained framework that includes a Generator and a Discriminator to create realistic and articulate 3D head models.

Table 6 presents various 3D model-based methods for generating talking heads. These methods utilize 3D models and specific architectures to address particular challenges in this field. By directly modeling the 3D structure of the face, these techniques aim to enhance control over facial expressions, head posture, and overall realism. One notable learning-based image reconstruction method that effectively overcomes the distortions caused by the skull, a significant challenge for transcranial applications, is PACT [211] (Photoacoustic Computed Tomography). AniArtAvatar [131] provides a method for creating 3D-aware art avatars, allowing users to control features such as head positions, shoulder movements, and facial expressions through manual animation. The method uses a pre-trained 2D diffusion model for texture synthesis, an SDF-based neural surface representation for accurate geometry, and a 3DMM for the underlying facial structure and motion. For 3D talking heads, JambaTalk [34] offers a two-phase training method to effectively convey speech and laughing signals. In addition to a new collection of laughing videos, the work integrates VOCA [170], FaceFormer, CodeTalker, FaceDiffuser [187], and ScanTalk databases. The perceived quality of face recognition systems may be limited by limitations such as the limited availability of training set patterns and audio-visual data. Future research should explore control mechanisms that utilize gaze, emotions, gestures, and facial movements, as these elements may change independently in the real world. This study proposes a 3D talking lip system designed to interact with unknown individuals and support multiple languages. The technology features a pipeline that employs Wav2Vec2.0 with a Transformer for audio feature extraction, Viseme Fixing for visual alignment, and techniques for mapping lip landmarks. Additionally, it includes automated data processing and a head-mounted device for capturing 3D lip motions. Most likely emphasizing multilingual lip motions, the project trains the system using the VOCA SET-lip [170] and ADLSET datasets. The project includes Frequent Human Pose (FHP [188]), which emphasizes 3D human posture identification for daily activities. Its precise 3D human posture identification is via a graph convolutional network. The paper advocates improving the identification of 3D human postures using artificial intelligence. It also points out some negatives, such as completely capturing the varied postural changes happening in real life and natural biases in the training datasets. TTSF [38] (Text-to-audio and Face) combines Talking Face Generation technology with text-to-speech (TTS) to produce real talking faces and synchronized audio outputs straight from text input. It addresses problems like ensuring consistency between the generated speech and facial motions and generating realistic head positions. LaughTalk [33] is a 3D talking head model with FLAME [101] parameters

Table 6: Comparison of 3D-Based Approaches

Method	Arch.	Dataset	Highlights	Limitations	N-shot
PACT [211]	3D U-Net	Self-Data	Overcomes skull distortion via deep reconstruction.	Sensitive to skull shape variance.	One
AniArtAvatar [131]	3DMM	Wonder3D, NeuS	Artistic avatar generation with body + face motion.	Style deformation and inconsistency.	One
JambaTalk [34]	Mamba	VOCA [170], FaceFormer	Two-stage modeling for speech + laughter.	emotion/gesture modeling missing.	Multi
ADL [129]	OM-Net	VOCA [170]SET-lip, ADLSET	Language-agnostic lip sync with Wav2Vec2.0.	Ignores upper-face emotion dynamics.	Multi
FHP [188]	FFNN	StateFarm, Pose	Posture detection in real-time use-cases.	Poor cross-domain generalization.	One
TTSF [38]	3DMM	LRS3, VoxCeleb2 [173]	TTS fused with facial generation for avatars.	Voice consistency and realism gaps.	One
LaughTalk [33]	3DMM	VoxCeleb2 [173], CelebV-HQ [175]	Expressive generation for laughter + speech.	Weakness under emotional extremes.	Multi
AE-NeRF [130]	3DMM	NeRF, HDTF [138]	Audio-enhanced NeRF for sync + fidelity.	Needs better speed benchmarking.	One
PV3D [214]	GAN	VoxCeleb, CelebV-HQ [175]	3D-consistent motion with editable views.	Needs more diverse training data.	Multi
DVP [41]	cGAN	Leader Dataset	Transfers expression, pose, blink for reenactment.	Breaks under extreme motion.	Multi

to simulate speaking and laughing. The 2D laughing clips are used to train the model, which is presumably learning to relate speech and laughter audio characteristics to the 3D FLAME parameters that control head movements and facial emotions. LaughTalk [33] outperforms earlier systems, which generally have trouble with non-speech facial expressions. However, it also faces challenges in effectively conveying speech and laughter, suggesting that further research may be necessary to improve the variety and sensitivity of laughter and its smooth integration with speech. The improved audio-driven talking head synthesis technique AE-NeRF [130] aims to improve generalization ability, audio-lip synchronization, and picture quality, particularly in few-shot scenarios. It aims to produce NeRF-based talking heads with proper lip synchronization by combining a 3DMM with NeRF (Neural Radiance Fields). PV3D [214] is a generative framework for 3D-aware portrait films that uses discriminators for more realism and explicitly models motion dynamics, including camera condition techniques, to maximize performance in static animation and view-consistent motion editing applications. A generative neural network is used in the ground-breaking technology known as Deep Video Portraits (DVP [41]) to create realistic reanimations of portrait videos. It realistically transfers 3D head position, rotation, facial emotions, and eye blinking from a source to a target actor. However, DVP [41] faces challenges such as managing excessive head motions or occlusions, relying on artificial face models for motion transfer, effectively transferring non-facial aspects, and possibly having limited generalization across various actor types. However, DVP [41] faces challenges such as managing excessive head motions or occlusions, relying on artificial face models for motion transfer, effectively transferring non-facial aspects, and possibly having limited generalization across various actor types. The part-based local-global conditional GAN approach uses photometric FLAME reconstructions and 3D feature points to enable expression transfer across extreme pose variations. Its modular design with Parts Generation Networks (PGNs) and Fusion Networks (PFN) [275] demonstrates robustness to occlusions while preserving identity attributes. In summary, the techniques for generating talking heads based on 3D models show significant advancements in several areas. These improvements include overcoming physical distortions in imaging, creating controllable 3D art avatars, generating realistic speech and laughter, synthesizing 3D talking lips for various languages, analyzing and potentially utilizing 3D human pose information, developing comprehensive text-to-talking face systems, enhancing audio-driven synthesis using NeRF, producing view-consistent 3D portrait videos, and transferring photorealistic facial expressions.

### 2.1.7 Parameter-Based Approaches

Parametric techniques manipulate facial movements using action units (AUs) or blend shape coefficients, enabling precise adjustments. According to Google Scholar citation metrics, the models DiscoFaceGAN [182] (Contrastive Learning for Disentangled Factors), FEP [12] (Facial Expression Parameters), and PIR [125] (Parametric Implicit Representation) [125] are significant. These methods combine implicit detail and explicit control, but they have limitations like overfitting to training identities, degradation under extreme lighting, and high demand for training data. Future efforts could include universal parametric spaces for cross-dataset compatibility and neural network-based nonlinear blend shape models.

A few 3D model-based methods for THG are summarized in Table 7. These methods typically use parametric models of the human face, such as 3D Morphable Models (3DMMs), to represent and modify facial geometry and texture. By adjusting the parameters of these models depending on audio or other input, they attempt to make realistic and controlled talking head animations. The architectural designs, datasets used, significant advancements, and intrinsic limitations of these parameter-driven approaches are all examined in this review. A few 3D model-based methods for THG are summarized in Table 7. These methods typically use parametric models of the human face, such as 3D Morphable Models (3DMMs), to represent and modify facial geometry and texture. By adjusting the parameters of these models depending on audio or other input, they attempt to make realistic and controlled talking head animations. The architectural designs, datasets used, significant advancements, and intrinsic limitations of these parameter-driven approaches are all examined in this review. Combining explicit and implicit neural representation techniques, PIR [125] (Parametric Implicit Representation) offers a ground-breaking framework for audio-driven face reenactment that produces excellent facial animation. It uses data augmentation methods, conditional picture synthesis techniques, and contextual data to increase performance and resilience. PIR [125] is tested using the HDTF [138] and PC-AVS [118] datasets, commonly used for audio-driven THG and face reenactment applications. User-generated content (UGC) audio-driven dubbing employs AdaIN [7] (Adaptive Instance Normalization). The technique presumably translates the speaking style (visual lip movements) from a source video to a target face using AdaIN [7] inside a Style Translation Network (STN), depending on the audio of the dubbed clip. Temporal regularization and a semi-parametric video renderer produce realistic and fluid results. DiscoFaceGAN [182] is an advanced method for synthesizing images that employs contrastive learning, imitative-contrastive learning, and 3D priors to generate controlled facial images while separating variations in facial features. It utilizes a

Table 7: Comparison of Parameter-Based THG

Method	Arch.	Dataset	Highlights	Limitations	N-shot
PIR [125]	3DMM	HDTF [138]	Combines parametric + implicit models with augmentation for quality reenactment.	High complexity; data dependence, ethical concerns.	One
AdaIN [7]	STN	RAVDESS [66]	Temporal regularized style transfer for dubbing.	Low expression precision, privacy risks.	Multi
DiscoFaceGAN [182]	3DMM	FFHQ [137]	Imitative + contrastive learning with 3D priors.	Sensitive to lighting/pose; lacks gaze control.	Multi
FEP [12]	HMM	Self	HMM + DNN with facial embeddings for pixel sync.	Needs large training; overfitting risk.	One

GAN architecture conditioned on a specific parameter, 3DMM. When tested on the FFHQ [137] dataset, a high-quality collection of human faces, DiscoFaceGAN [182] aims to produce diverse and lifelike facial images with carefully managed features. Facial Expression Parameters (PEP) suggests a two-step synthesis method to create realistic talking face animation from pixels. It uses Deep Neural Networks (DNNs), which incorporate facial expressions to achieve a stronger link between audio and visual context, and Hidden Markov Models (HMMs) for temporal modeling. FEP [12] uses HMMs to replicate the temporal sequence of phonemes or visemes derived from the audio input. The corresponding lip movements and facial expressions are then generated at the pixel level by DNNs, conditioned on facial expression parameters that presumably provide greater control over the final animation. The reviewed parameter-based THG methods utilize a variety of architectures, including 3D Morphable Models (3DMMs) with implicit neural representations, Style Translation Networks with Adaptive Instance Normalization (AdaIN [7]), Generative Adversarial Networks (GANs) conditioned on 3D priors and contrastive learning, and Hidden Markov Models (HMMs) integrated with Deep Neural Networks (DNNs) and facial expression parameters. These methods have been evaluated using HDTF [138], PC-AVS [118], RAVDESS [66], FFHQ [137], and self-collected video data. The study demonstrates significant progress in face image generation, automated dubbing, facial expression parameter integration, and high-quality reenactment. It also highlights drawbacks such as the reliance on parametric models, the inability to control lighting and poses, potential overfitting, and further improvements in facial expression realism and animation quality. Using AdaIN [7] (Adaptive Instance Normalization) inside a Style Translation Network (STN), the technique translates speaking styles from a source video to a target face. Temporal regularization and a semi-parametric video renderer produce realistic and fluid results. DiscoFaceGAN [182] is an advanced method for synthesizing images that employs contrastive learning, imitative-contrastive learning, and 3D priors to generate controlled facial images while separating variations in facial features. It utilizes a GAN architecture conditioned on a specific parameter, 3DMM. When tested on the FFHQ [137] dataset, a high-quality collection of human faces, DiscoFaceGAN [182] aims to produce diverse and lifelike facial images with carefully managed features. Facial Expression Parameters (PEP) suggests a two-step synthesis method to create realistic talking face animation from pixels. It uses Deep Neural Networks (DNNs), which incorporate facial expressions to achieve a stronger link between audio and visual context, and Hidden Markov Models (HMMs) for temporal modeling. FEP [12] uses HMMs to replicate the temporal sequence of phonemes or visemes derived from the audio input. The corresponding lip movements and facial expressions are then generated at the pixel level by DNNs, conditioned on facial expression parameters that presumably provide greater control over the final animation. The reviewed parameter-based THG methods utilize a variety of architectures, including 3D Morphable Models (3DMMs) with implicit neural representations, Style Translation Networks with Adaptive Instance Normalization (AdaIN [7]), Generative Adversarial Networks (GANs) conditioned on 3D priors and contrastive learning, and Hidden Markov Models (HMMs) integrated with Deep Neural Networks (DNNs) and facial expression parameters. These methods have been evaluated using HDTF [138], PC-AVS [118], RAVDESS [66], FFHQ [137], and self-collected video data. They demonstrate notable progress in several areas, including the transfer of speaking styles for automated dubbing, the creation of controllable and disentangled face images, the incorporation of facial expression parameters for improved audio-visual correspondence in pixel-level animation, and the combination of parametric and implicit representations for high-quality reenactment. The expressiveness of the underlying parametric models is often relied upon, extreme poses and lighting conditions are difficult to control, overfitting to training data is a possibility, and additional work is required to control and improve the realism of generated facial expressions and animation quality overall. While not strictly a THG method, the multi-feature fusion algorithm [276] demonstrates the value of parameterized facial analysis through its lightweight CNN architecture. Fusing eye, mouth, and head movement features with lane-departure data highlights the potential for hybrid parameter-driven systems in real-time applications requiring facial behavior understanding.

### 2.1.8 NeRF-Based THG

Neural Radiance Fields (NeRFs) enable photographic novel-view synthesis and dynamic facial deformations through volumetric scene representation. Google Scholar cites AD-c [127], IMavatar[198], and CVTHHead [180] as the main models for generating avatars from single photos and detecting fine-scale wrinkles. However, real-time rendering, training stability, and slow inference times are major challenges. Future research will focus on creating portable Neural Radiance Fields substitutes and exploring innovative hash encoding techniques for enhanced efficiency and performance.

Neural Radiance Fields (NeRFs) enable photographic novel-view synthesis and dynamic facial deformations through volumetric scene representation. Google Scholar cites AD-NeRF [127], IMavatar[198], and CVTHHead [180] as the main models for generating avatars from single photos and detecting fine-scale wrinkles.

Table 8: Comparison of NeRF-Based Approaches

Method	Arch.	Dataset	Highlights	Limitations	N-shot
CVTHead [180]	3DMM	VoxCeleb	One-shot neural avatar w/ expression + pose control.	Limited by DECA mesh reconstruction.	One
Head3D [193]	ConvLSTM	VoxCeleb, FaceForensics	Canonical frames w/ interpretable 3D motion transfer.	Parsing errors, side-view issues.	Multi
SSP-NeRF[221]	3DMM	NVP, Obama	Synth	Rendering slow; poor language generalization.	Multi
HeadNeRF [202]	3DMM	CelebAMask-HQ, FFHQ [137]	Semantic NeRF w/ dynamic sampling.	Fails on headgear, wild generalization.	Multi
IMavatarr[198]	3DMM	VOCA [170], COMA	Fine-grain NeRF face model.	Heavy compute, coarse high-frequency detail.	Multi
ROME [56]	FLAME	VoxCeleb2 [173], H3DS	Monocular avatars w/ expression modeling.	Lacks fine detail; smoothed geometry.	One
FNeVR [190]	CNN	VoxCeleb1 [172]	One-shot head reenactment with realism.	Compute-heavy; poor pose editability.	One
DFA-NeRF [43]	3DMM	LRS2 [177], HDTF [138]	Volume rendering to fix motion deformation.	Single speaker support; slow rendering.	Multi
NerFACE [11]	SRNs	Wild videos	Audio-driven NeRF w/ lip-sync alignment.	Lacks eye/torso motion control.	Multi
AD-NeRF [127]	GAN	Wild videos	4D NeRF reenactment for telepresence.	View- and audio-adaptive NeRF synthesis.	Multi

However, real-time rendering, training stability, and slow inference times are major challenges. Future research will focus on creating portable Neural Radiance Fields substitutes and exploring innovative hash encoding techniques for enhanced efficiency and performance. Various methods for THG based on 3D models have been reviewed in Table 8. These methods utilize Neural Radiance Fields (NeRF) or combine them with other techniques, such as 3D Morphable Models (3DMMs), to achieve photorealistic and controlled talking head synthesis. NeRF-based algorithms learn a volumetric model of the scene, allowing for innovative view synthesis and detailed rendering. CVTHead [180] offers a novel method for creating programmable neural head avatars from a single reference image. It aims to maintain visual quality while quickly representing human heads in various positions and emotions. To achieve effective and controllable rendering of the head avatar with different expressions and poses derived from the 3DMM parameters, CVTHead [180] uses a 3DMM as a foundational representation of head geometry and texture. It probably combines this with neural rendering techniques, perhaps inspired by NeRF. Head3D [193] proposes a 3D-aware talking-head video motion transmission network. The motion from a driving video or audio is transferred to the target identity using this 3D model. One of Head3D [193]'s most notable features is its ground-breaking 3D-aware Method for Transferring Motion in Talking Head Videos. This sophisticated method goes beyond previous methods and performs exceptionally well when identities change in the real world. It enhances the realism of animated faces and produces stunningly clear and visually interpretable 3D canonical head models, bringing virtual characters to life like never before. SSP-NeRFF [221] introduces an innovative approach that utilizes Neural Radiance Fields to produce high-quality video portraits. We need to enhance our methods to improve how we create talking heads with NeRF technology. Integrating Dynamic Ray Sampling modules and incorporating semantic awareness is crucial, potentially through 3D Morphable Models (3DMMs) or other facial analysis techniques. A parametric head model based on NeRF, HeadNeRF, was created to improve the quality of high-fidelity headshots. To replicate the intricate appearance and possible non-rigid deformations of the face, a Neural Radiance Field combined with a 3DMM offers a parametric description of head geometry and posture. IM Avatar provides a generative neural network (GNN) based on NeRF to generate a variety of headgear. In contrast to standard 3D morphable face models, IMAvatar[198] is a method for learning implicit head avatars from monocular films that improve geometry and expression space by learning an implicit head representation. Monocular video sequences create an implicit neural representation of the head avatar that can be manipulated to generate a variety of positions and emotions. The VOCA [170] and COMA datasets, which focus on voice-driven face animation and provide a rich space of facial emotions, respectively, are used to evaluate the method. Intending to achieve competitive head geometry recovery and high-quality rendering from a single image, ROME [56] is a technique for creating realistic one-shot mesh-based human head avatars. It uses FLAME as a simple mesh and regresses the FLAME parameters from a single input image, most likely using deep learning techniques. VoxCeleb2 [173] and H3DS (Human 3D Scans) are used to evaluate ROME [56], which emphasizes recovering precise 3D head geometry and providing realistic representations for a range of users. By overcoming motion deformation and complex modeling challenges, FNeVR [190], a CNN-based network, aims to improve computer vision tasks, particularly talking head creation. It uses neural volume rendering to improve performance on talking-head benchmarks. However, it faces limitations like high computational complexity, which can impact rendering speed and resource requirements, pose modification problems, and limit generalization to highly diverse and unknown input data. DFA-NeRF [43] is a novel neural radiance field framework designed for personalized THG to surpass existing methods in producing high-fidelity lip-synchronized talking heads. It is NeRF-based and learns a neural radiance field conditioned on audio and identification. LRS2 [177] and HDTF [138] datasets, which are high-resolution talking head video datasets with synchronized audio, are used to evaluate the method. NeRFACE uses dynamic neural radiance fields to mimic human facial appearance and produce realistic talking head images that outperform video-based reenactment methods. This strategy greatly benefits applications in solving virtual reality (VR) or augmented reality (AR) telepresence. High-quality, two-minute self-gathered videos train the method, emphasizing learning from real-world data to attain high fidelity. ADNeural scene representation networks (perhaps a variation of NeRF) conditioned on audio input are used in the GAN-based NeRF architecture. It allows the generated talking heads' audio signals, viewing instructions, and background images to be altered without restriction. However, the article points out flaws like strange mouths and body parts in the generated movies, which are most likely the result of discrepancies between the driving language and the training data. A variety of architectures are used in the reviewed NeRF-based and hybrid THG methods, including CNNs for neural volume rendering, 3DMMs with neural rendering, Conv LSTM for 3D motion transfer, NeRFs with semantic awareness and dynamic ray sampling, parametric head models enhanced with NeRF, implicit neural representations, FLAME for one-shot avatar creation, and GANs conditioned on NeRF for audio-driven synthesis. Neural volume rendering for better animation, high-fidelity video portraits, 3D-aware motion transfer, controllable neural head avatars from single images, learning implicit head avatars with improved

geometry and expression, and enabling flexible audio, viewpoint, and background adjustments are just a few of the areas where the techniques demonstrate advancements.

Furthermore, other NeRF-based models have significantly advanced the realism and controllability of 3D facial synthesis and talking head generation. 3DFaceShop [235] presents a system for explicitly controllable, 3D-aware facial generation using neural rendering techniques that enable intuitive editing through sparse sketch-based inputs. Next3D [236] leverages generative neural textures within a 3D framework to create high-quality, animatable facial avatars with photorealistic fidelity, contributing toward fully interactive virtual humans. NeRFInvertor [237] bridges GANs and NeRF by inverting 2D facial images into high-fidelity volumetric representations, enabling fine-grained editing and robust identity preservation. DFRF [238] introduces a dynamic facial radiance field that captures appearance and expression dynamics, enhancing temporal coherence and expressiveness in facial animations.

### 2.1.9 Diffusion-Based THG

Diffusion models, known for their temporal coherence and expression synthesis capabilities, are utilized to create high-quality films. According to Google Scholar citation metrics, MoDiTalker [207], DreamTalk [9], and DAE-Talker [181] are significant models, each with its limitations. DAE-Talker [181] has state-of-the-art lip-sync accuracy, DreamTalk [9] generates strong emotions, and MoDiTalker [207] differentiates between lip and non-lip motions. Challenges include fine-grained editing, high VRAM requirements, and computational cost. Future research could focus on reinforcement learning for reward-guided generation and latent consistency models.

A range of diffusion model-based strategies in THG has been examined in Table 9 that leverage diffusion models, a class of generative models notable for their capacity to create high-quality and varied outputs by learning to reverse a slow noising process. These methods aim to improve artificial talking head movies' temporal coherence, expressiveness, and realism. MoDiTalker [207] is a diffusion model specifically designed to enhance the generation of talking heads. It separates motion into audio-driven lip movements and more general facial expressions, using distinct modules for audio-to-motion and motion-to-video conversion. The model aims to create realistic and synchronized talking heads from audio by employing a GAN architecture within a diffusion framework. The portrait animation technique EMOTalker aims to improve emotion recognition when producing talking heads. While enabling the creation of customizable facial emotions through an emotion Intensity Block, it attempts to preserve the original photo identity. The model uses an architecture based on Contrastive Language-Image Pre-training (CLIP) to understand and produce facial expressions associated with different emotions, potentially enabling text-based emotion management. DiffTalk [61] proposes a talking head synthesis method that generates high-quality visuals in sync with the original audio through audio-driven, temporally coherent denoising. Using a Recurrent Neural Network (RNN) architecture, it learns to predict the corresponding facial movements over time by capturing the temporal relationships in the audio. The technique creates a single talking head using a single image as input in the multi-shot Diffused Heads [26] technique. Its dependence on reference face photos, the computational expense of repeated denoising techniques, and the possibility of artifacts in the output videos are some of its drawbacks. It seeks to enhance the entire synthesis process by utilizing diffusion models' capacity to synthesize high-quality video frames sequentially. Diffused Heads [26], an autoregressive diffusion model, aims to create realistic talking head movies. Using a GAN architecture within an autoregressive diffusion framework, the model generates successive video frames, each dependent on the audio input and the frames that came before it. The CREMA and LRW datasets, which focus on emotional expressions and extensive lip reading datasets, are used to evaluate it. Diffusion Video Editing [10] is an innovative speech-driven method. This technique showcases the ability of diffusion models to generate new content and modify existing films based on audio input. It has potential applications in tasks like lip synchronization and adjusting facial expressions to match new audio. The method integrates a U-Net architecture inside a denoising diffusion model to enable targeted changes to the video content based on the audio, presumably learning to denoise video frames conditioned on the input speech. DAE-Talker [181] uses latent representations learned by a Denoising Autoencoder (DAE) to improve speech-driven talking face synthesis. By utilizing powerful feature representations that the DAE has learned from audio, it aims to outperform earlier methods regarding lip synchronization accuracy, video quality, pose naturalness, and pose controllability. Pose accuracy and control, efficient pose modeling, denoising performance (within the DAE), generalization to unseen identities or speaking styles, the need for improved evaluation metrics for THG, the DAE's data dependency, and the system's overall computational complexity are some of the issues that the paper recognizes DAE-Talker [181] faces. DreamTalk [9] is a diffusion model framework designed to create realistic talking heads exhibiting various emotions. Controlling the emotional style of the generated talking

Table 9: Comparison of Diffusion-Based Approaches

Method	Arch.	Dataset	Highlights	Limitations	N-shot
MoDiTalker [207]	GAN	LRS3-TED [139], HDTF [138]	Motion-disentangled diffusion for subtle lip movements.	High sampling time, stochasticity, ethics concerns.	Multi
EMOTALKER[184]	CLIP	MEAD [139], CREMA-D [134]	emotion intensity control in speech-driven avatars.	Generalization issues, realism-emotion trade-offs.	One
DiffTalk[61]	RNN	HDTF [138]	Audio-driven diffusion with temporal coherence.	Requires face reference image, slow, artifacts.	Multi
Diffused Heads [26]	GAN	CREMA, LRW	Autoregressive diffusion for smooth expressions.	Slow generation; not real-time suitable.	One
Diffusion Video [10]	U-Net	GRID [176], CREMA-D [134]	Denoising diffusion for multi-speaker video edits.	Long training, lip sync issues.	One
DAE-Talker [181]	DAE	LibriTTS	Latent DAE boosts lip sync and pose accuracy.	Poor pose modeling; resource intensive.	One
DreamTalk [9]	3DMM	MEAD [139], HDTF [138]	Style-aware emotional diffusion model.	Lacks realism for subtle emotional cues.	Multi
DiT-Head [183]	DiT	HDTF [138]	High-fidelity transformer-based diffusion.	English-only; deepfake + compute risks.	Multi
FaceDiffuser [187]	HuBERT	BIWI, Multiface, BEAT	3D animation via nondeterministic diffusion.	Dataset limits, slow inference, subjective evals.	Multi
PAVDP [212]	3DMM	VoxCeleb1 [172]	Probabilistic priors for facial motion.	Stripe artifacts; weak high-freq detail.	One

head consists of a style predictor, a style-aware lip expert, and a denoising network. The method addresses problems such as emotion recognition in low-emotion-intensity audio, speaker identity retention, and consistent expressions. DiT-Head [183] is a novel talking head synthesis pipeline that uses Diffusion Transformers (DiT) with audio input for scalability and competitive performance. It has demonstrated promise for many applications and is evaluated using the HDTF [138] dataset, a common benchmark for talking head synthesis. However, its reliance on data from English speakers and the substantial computer resources required for training and inference limit it. A non-deterministic deep learning model, FaceDiffuser [187], was developed for speech-driven 3D face animation synthesis. It uses an internal dataset in addition to 3D vertex and blend shape datasets to learn realistic 3D facial motions directly from speech. The BIWI, Multiface, BEAT, and UUDaMM datasets are used to train the model, which allows it to learn realistic speech-driven 3D facial deformations. With a focus on probabilistically sampling holistic lip-irrelevant facial gestures to match the input audio while maintaining photorealism and naturalness, PAVDP [212] offers a novel framework for one-shot audio-driven talking head creation. It uses a generator network ( $G$ ) that was trained similarly to PC-AVS [118] but focuses on using probabilistic diffusion prior models to predict non-lip facial movements. A variety of architectures are used in the reviewed diffusion-based THG techniques, including GANs in diffusion frameworks, CLIP-based models, RNNs with denoising, Diffusion Transformers, U-Nets for video editing, Denoising Autoencoders, and specialized diffusion models with style-aware elements. These methods focus on advances in speech-driven video editing, expressiveness, smoothness, emotionally editable talking heads, motion disentanglement, pose control, lip synchronization, and high-quality synthesis. Some main disadvantages include long sampling times, high computing costs, the potential for artifacts, challenges with generalization and identity retention, reliance on large and diverse datasets, and ethical issues related to creating realistic synthetic media. Table 9 that leverages diffusion models, a class of generative models notable for their capacity to create high-quality and varied outputs by learning to reverse a slow noising process. These methods aim to improve artificial talking head movies' temporal coherence, expressiveness, and realism. MoDiTalker [207] is a diffusion model specifically designed to enhance the generation of talking heads. It separates motion into audio-driven lip movements and more general facial expressions, using distinct modules for audio-to-motion and motion-to-video conversion. The model aims to create realistic and synchronized talking heads from audio by employing a GAN architecture within a diffusion framework. The portrait animation technique emoTalker aims to improve Emotion recognition when producing talking heads. While enabling the creation of customizable facial Emotions through an Emotion Intensity Block, it attempts to preserve the original photo identity. The model uses an architecture based on Contrastive Language-Image Pre-training (CLIP) to understand and produce facial expressions associated with different Emotions, potentially enabling text-based Emotion management. DiffTalk [61] proposes a talking head synthesis method that generates high-quality visuals in sync with the original audio through audio-driven, temporally coherent denoising. It captures the temporal relationships in the audio and learns to predict the corresponding facial movements over time using a Recurrent Neural Network (RNN) architecture. The method utilizes a single image as input in the multi-shot Diffused Heads [26] technique, allowing for the creation of a single talking head. However, it has some disadvantages, including its reliance on reference face photos, the computational cost associated with repeated denoising techniques, and the potential introduction of artifacts in the output videos. By leveraging diffusion models' capability to synthesize high-quality video frames sequentially, it aims to improve the overall synthesis process. Diffused Heads [26], an autoregressive diffusion model, aims to create realistic talking head movies. Using a GAN architecture within an autoregressive diffusion framework, the model generates successive video frames, each dependent on the audio input and the frames that came before it. The CREMA and LRW datasets, which focus on Emotional expressions and extensive lip reading datasets, are used to evaluate it. Diffusion Video Editing [10] is an innovative speech-driven method. This technique showcases the ability of diffusion models to generate new content and modify existing films based on audio input. It has potential applications in tasks like lip synchronization and adjusting facial expressions to match new audio. The method integrates a U-Net architecture inside a denoising diffusion model to enable targeted changes to the video content based on the audio, presumably learning to denoise video frames conditioned on the input speech. DAE-Talker [181] uses latent representations learned by a Denoising Autoencoder (DAE) to improve speech-driven talking face synthesis. By utilizing powerful feature representations that the DAE has learned from audio, it aims to outperform earlier methods regarding lip synchronization accuracy, video quality, pose naturalness, and pose controllability. Pose accuracy and control, efficient pose modeling, denoising performance (within the DAE), generalization to unseen identities or speaking styles, the need for improved evaluation metrics for THG, the DAE's data dependency, and the system's overall computational complexity are some of the issues that the paper recognizes DAE-Talker [181] faces. DreamTalk [9] is a diffusion model framework designed to create realistic talking heads exhibiting various Emotions. Controlling the Emotional style of the generated talking head consists of a style predictor, a style-aware lip expert, and a denoising network. The method addresses problems such as Emotion recognition in low-Emotion-intensity

audio, speaker identity retention, and consistent expressions. DiT-Head [183] is a novel talking head synthesis pipeline that uses Diffusion Transformers (DiT) with audio input for scalability and competitive performance. It has demonstrated promise for many applications and is evaluated using the HDTF [138] dataset, a common benchmark for talking head synthesis. However, its reliance on data from English speakers and the substantial computer resources required for training and inference limit it. A non-deterministic deep learning model, FaceDiffuser [187], was developed for speech-driven 3D face animation synthesis. It uses an internal dataset in addition to 3D vertex and blend shape datasets to learn realistic 3D facial motions directly from speech. The BIWI, Multiface, BEAT, and UUDaMM datasets are used to train the model, which allows it to learn realistic speech-driven 3D facial deformations. With a focus on probabilistically sampling holistic lip-irrelevant facial gestures to match the input audio while maintaining photorealism and naturalness, PAVDP [212] offers a novel framework for one-shot audio-driven talking head creation. It uses a generator network ( $G$ ) that was trained similarly to PC-AVS [118] but focuses on using probabilistic diffusion prior models to predict non-lip facial movements. The Adaptive Diffusion Landmark Dynamic Rendering (DLDR) [277] framework also leverages transformer-based audio semantic mapping and landmark alignment to achieve precise audio-visual synchronization. Combining diffusion models with dynamic rendering demonstrates significant improvements in mouth-shape realism and perceptual quality on benchmark datasets. A variety of architectures are used in the reviewed diffusion-based THG techniques, including GANs in diffusion frameworks, CLIP-based models, RNNs with denoising, Diffusion Transformers, U-Nets for video editing, Denoising Autoencoders, and specialized diffusion models with style-aware elements. These methods focus on advances in speech-driven video editing, expressiveness, smoothness, Emotionally editable talking heads, motion disentanglement, pose control, lip synchronization, and high-quality synthesis. Some main disadvantages include long sampling times, high computing costs, the potential for artifacts, challenges with generalization and identity retention, reliance on large and diverse datasets, and ethical issues related to creating realistic synthetic media.

### 2.1.10 3D Animation-Based THG

3D animation pipelines integrate motion capture, physics simulations, and rigged models to deliver film and VFX quality results. Learn2Talk [5], MultiTalk [208], and Speech4Mesh [8] are important models based on Google Scholar citation metrics. While MultiTalk [208] uses VQ-VAE with language embeddings and self-collected multilingual films, Learn2Talk [5] uses TCN for lip-sync and vertex accuracy. Speech4Mesh [8] uses the FLAME-based [101] audio2mesh network and the MEAD [159] dataset. Obstacles include realistic and creative control. Future possibilities for realistic cloth and hair modeling include generative rigging and differentiable physics.

The domain of 3D animation-based talking head generation has witnessed a surge in approaches aiming to produce expressive, synchronized, and controllable facial animations from audio or multimodal inputs. Tables 10 and 10 summarize a wide array of such methods, categorizing them based on architecture, datasets, core contributions, limitations, and the number of input instances ( $N$ -shot). These methods explore the fusion of generative modeling with 3D mesh control, emotional expressiveness, and personalized motion synthesis. In Table 10, the first entry, MultiTalk, employs a VQ-VAE framework to enable multilingual avatar synthesis with style embeddings tailored for accurate lip synchronization. Despite its robustness in language-specific articulation, it struggles with non-standard speech and linguistic nuances. CSTalk, built upon Temporal Convolutional Networks (TCNs), integrates 3D lip and emotional synchronization. However, it has a limited emotional range and cannot effectively capture subtle or complex affective cues. Learn2Talk innovates by combining 2D image generation and 3D animation for enhanced clarity. While effective for simple speech, it lacks realism in secondary expressions like blinking and exhibits task-switching challenges. DiffSpeaker introduces a GRU-based design to drive facial motion using fast, transformer-like sequences, but its realism and capacity for handling 4D audio integration remain limited. Media2Face stands out for merging diffusion modeling with parametric animation and dataset blending, achieving fine-grained lip movement. However, this comes at the cost of heavy computational requirements and a high dependency on dataset alignment. Table 10 continues this trend, focusing on models pushing the speech-driven 3D mesh synthesis envelope. PMMTalk leverages CNNs for pseudo-multimodal blendshape generation, achieving effective lip synchronization but lacking comprehensive control over head poses and facial expressions. It also suffers from slow inference times. 3DiFACE adopts a transformer-based diffusion approach tailored to personalized mesh outputs. Despite the strength of personalization, its reliance on limited datasets constrains generalizability and editing flexibility. Speech4Mesh addresses the scarcity of annotated 4D meshes by proposing an Audio2Mesh pipeline using the FLAME model. However, it tends to overfit and generalize poorly across domains. PV3D, using GANs, combines 3D-aware generation with motion modeling. While it offers stylistically rich outputs using datasets like VoxCeleb and CelebV-HQ, its dependency on 2D video data affects 3D consistency. Finally, 3D-TalkEmo, based on StarGAN, focuses on emotion-adaptive synthesis for expressive talking heads. While it provides rich

Table 10: Comparison of 3D Animation-Based Approaches

Method	Arch.	Dataset	Highlights	Limitations	N-shot
MultiTalk [208]	VQ-VAE	MultiTalk [208]	Multilingual avatars with style embeddings for lip sync.	Language limits and non-standard speech issues.	Multi
CSTalk [179]	TCN	Live Link Face	3D lip + Emotion sync aligned to speech.	Narrow emotion range; complex cues poorly handled.	Multi
Learn2Talk [5]	TCN	BIWI, VOCASET [170]	2D generation fused with 3D for clarity.	Weak realism in blinking/emotion; multitasking issues.	Multi
DiftSpeaker [45]	GRU	BIWI, VOCASET [170]	Fast transformer-based 3D facial motion.	Struggles with realism + 4D audio mapping.	Multi
Media2Face [203]	VAE	BIWI, VOCASET [170]	Diffusion + parametric asset + dataset fusion.	Compute-heavy; dataset-dependent.	Multi
PMMTalk [213]	CNN	3D-CAVFA, VOCASET [170]	Pseudo-multimodal lip-blendshape sync.	No head/exp control; inference slow.	Multi
3DifFACE [6]	Transformer	VOCASET [170]	Personalized diffusion-based 3D animation.	Small dataset, editing limits.	Multi
Speech4Mesh [8]	FLAME	MEAD [159], VoxCeleb2 [173]	Audio2Mesh system handles 4D mesh scarcity.	Prone to overfitting, poor generalization.	One
PV3D [214]	GAN	VoxCeleb, CelebV-HQ [175], TH-1KH	GAN + motion modeling for 3D-aware output.	Quality limited by 2D data.	One
3D-TalkEmoji [126]	StarGAN	Custom	Emotion-adaptive 3D talking head generation.	Low scalability; data/tuning bottlenecks.	Multi

emotional dynamics, its scalability is hindered by tuning complexity and limited training data. Collectively, these methods showcase a growing emphasis on integrating multimodal cues (speech, emotion, style) with 3D-aware modeling. The dominant architectures span traditional RNNs and CNNs to modern transformers, diffusion models, and variational frameworks. Despite improvements in fidelity and personalization, challenges persist—particularly regarding scalability, inference efficiency, emotional generalization, and non-standard linguistic or acoustic input handling. Regarding datasets, VOCASET and BIWI dominate due to their detailed 3D annotations, while newer datasets like MultiTalk and 3D-CAVFA offer specific benefits such as multilinguality and multimodal fidelity. Most methods operate in a multi-shot regime, requiring several input samples or training iterations per subject. However, methods like Speech4Mesh and PV3D progress toward one-shot generation, reflecting efforts to reduce data reliance. In summary, the surveyed methods reflect a dynamic research trajectory aimed at bridging expressiveness, realism, and control in 3D talking heads, with each contributing uniquely across the axes of animation realism, emotional fidelity, and computational feasibility.

Various diffusion 3D animation models based on THG have been investigated in Table 10 techniques for creating 3D talking head animations, generally controlled by audio input. MultiTalk [208], CSTalk [179], Learn2Talk [5], DiffSpeaker [45], Media2Face [203], and BIWI are some of the methods. The multilingual 2D video dataset MultiTalk [208] aims to generate expressive and lifelike 3D head positions and facial gestures. It uses a VQ-VAE architecture to create a discrete latent representation of facial movements that can be conditioned by language and audio embeddings. A speech-driven 3D facial animation technique, CSTalk [179], aims to outperform techniques in several areas, such as handling data constraints, achieving precise lip alignment with the audio, and producing a wider range of Emotions in the 3D face. To interpret sequential input, such as audio, and produce corresponding temporal sequences of 3D facial motions, it uses a TCN architecture. Some disadvantages are the computational complexity of training and managing 3D facial animation models, the potentially limited range of possible Emotional expressions, and the subjectivity involved in interpreting and producing subtle Emotional cues. By combining knowledge from audio-video synchronization networks with experience from 2D talking face production, Learn2Talk [5] aims to improve both 2D and 3D talking face research. It aims to improve the generated talking heads' overall speech perception, vertex correctness (in 3D models), and lip-sync accuracy. It uses a TCN architecture comparable to CSTalk [179], demonstrating how well it models the temporal relationship between facial and audio motions for improved synchronization and animation. A Transformer-based network, DiffSpeaker [45], was created to improve speech-driven 3D face animation performance. The main innovation is creating parallel face movements for better performance and faster inference speeds than sequential or autoregressive models. Its reliance on audio-4D data, its inability to generalize to invisible identities or speech patterns, and the potential trade-offs between achieving flawless lip synchronization and producing genuine non-verbal facial Emotions are some of its limitations. Using a Generalized Neural Parametric face asset, the M2F-D dataset, and a Media2Face [203] diffusion model, Media2Face [203] is a three-step method for creating 3D face animations from speech. Realistic and emotive animations conditioned on voice input are created using the diffusion model framework's VAE architecture. Like Learn2Talk [5] and DiffSpeaker [45], the method is evaluated on the BIWI and VOCA [170]SET datasets. A new framework called PMMTalk [213] aims to improve the precision of speech-driven 3D face animation. It utilizes pseudo-multi-modal characteristics to take advantage of information from various modalities, such as text and audio. To capture the complex relationship between speech and facial motions, PMMTalk [213] uses a CNN architecture to extract audio data and map it to face shape coefficients. 3DiFACE [6] is a novel method for editing and animating 3D faces using speech. A lightweight audio-conditioned diffusion model creates stochasticity and provides motion editing capabilities in the resulting animations. The diffusion component enables the creation of different and potentially adjustable animations, while the Transformer architecture depicts the interaction between audio and 3D facial motions. The study lists limitations like comparatively small training datasets, the challenge of achieving accurate motion control and editing capabilities, and the computational complexity of diffusion-based models. A 3D face animation system called Speech4Mesh [8] manages controllability and data scarcity in speech-driven 3D facial animation. For more control over the generated animations, it encodes speaking factors, generates 4D talking head data, and trains an audio2mesh network to translate audio to 3D face models. FLAME parameters are based on audio characteristics, and the technology trains an audio2mesh network that uses FLAME as the underlying format for 3D face models. A generative framework called PV3D [214] was created to create 3D-aware portrait films. By offering methods for describing motion dynamics and incorporating discriminators that operate in both spatial and temporal dimensions, it expands static GANs into the video domain. The VoxCeleb, CelebV-HQ [175], and TalkingHead-1KH [169] datasets evaluate PV3D [214], which focuses on producing realistic and excellent 3D-aware video portraits. A deep neural network called 3D-3D-TalkEmo [126] is designed to produce dynamic 3D talking-head animations with modifiable Emotion

states. In terms of the quality and controllability of the generated Emotional expressions in 3D talking heads, it aims to surpass earlier methods. The framework uses a StarGAN design, which is renowned for its ability to translate images across multiple domains. The explored 3D animation model-based THG techniques, which primarily use multi-shot learning, make use of a variety of architectures, including CNNs, FLAME, StarGANs, TCNs, Transformers, VQ-VAEs, and conventional VAEs inside diffusion frameworks. Advances in multilingual lip synchronization, improved lip alignment and Emotional expression, improved lip-sync and vertex accuracy, parallel motion generation for faster inference, high-fidelity and expressive animation through diffusion models, precise lip-syncing with blend shape coefficients, personalized animation with stochasticity and editing, addressing data scarcity through 4D data generation, producing 3D-aware portrait videos with motion dynamics, and producing vibrant 3D talking heads with adjustable Emotions are all highlighted in these methods.

## 2.2 DATASET

THG depends on high-quality datasets to train lip synchronization, Emotion synthesis, and pose control models. Key datasets used in the field include audio-driven datasets such as VoxCeleb, CREMA-D [134], LRW [140] (Lip Reading in the Wild), MEAD [159], Video-Driven datasets like TalkingHead-1KH [169], HDTF [138], CelebV-HQ [175], Text-Driven datasets like CelebrV-Text, Emotion-Specific datasets like RAVDESS [66], VoxCeleb1 [172], 3D and Multi-View datasets like Multiface, Vocaset, and GRD. Audio-driven datasets focus on synchronizing facial movements with speech inputs. VoxCeleb1 [172] contains over 1,251 and 6,112 speakers from YouTube, with 153,000+ audio-video clips. These datasets are used for cross-identity reenactment, speaker verification, and Emotion-aware synthesis. However, they have limitations such as lab control, limited ethnic diversity, low resolution, lack of non-frontal poses, small sample size, and lab environment. Video-driven datasets are used for motion transfer and identity preservation across videos. TalkingHead-1KH [169] contains 1,000+ hours of talking head videos from YouTube, while HDTF [138] includes 362 speakers and 10,000 clips with extreme poses. CelebV-HQ [175] contains 15,653 high-resolution videos of celebrities, while BIWI-3D provides 3D scans of 14 speakers with head pose annotations. Emotion-specific datasets focus on nuanced Emotional expressions, with RAVDESS [66] containing 24 actors performing eight Emotions in speech and song. VoxCeleb1 [172] contains 652 naturalistic Emotional dialogues from 12 speakers, while 3D and Multi-View datasets enable 3D-aware synthesis and free-view rendering. Vocaset contains 3D facial animations from 12 speakers, while GRD contains 33,000 clips of 34 speakers reciting grammatically fixed sentences. Cross-modal datasets combine multiple modalities (text, audio, video) using LRS3-TED [139] and GRD. These datasets have limitations, mainly English and limited Emotion labels, but can improve lip-reading models' accuracy. THG relies heavily on diverse, high-quality datasets for tasks like lip synchronization, Emotion synthesis, and pose control. These datasets have their limitations, such as a bias towards frontal views and English speakers, limited ethnic diversity, and limitations in the use of different modalities. Researchers can develop more effective and efficient models for generating realistic and engaging talking head content by analyzing these datasets and identifying their limitations.

A variety of training and evaluation datasets influences research on THG. The varied character of this activity is shown in the datasets' disparities in size, recording environment, view kinds, included modalities, motion capture data presence, and resolution. Emphasizing their important qualities relevant to building avatar and video synthesis algorithms, Table 11 thoroughly summarizes this domain's most often used datasets. Creating efficient avatar and video synthesis algorithms depends on the size and duration of these datasets. A notable resource for applications requiring large amounts of diverse audio-visual data is VoxCeleb2 (2018), which offers an astounding 2400 hours of video with 6100 speakers and 1.1 million words. With 438 hours of TED Talks from 5000 speakers and 152,000 phrases, LRS3-TED [139] (2018) similarly provides various speaking styles and linguistic material for lip-reading and speech-to-text research. Conversely, CelebrV-Text [25] emphasizes lip sync features, which are perfect for training algorithms concentrating on accurate mouth movements; datasets like VOCA [170] (2019) provide high-quality 3D facial motion capture data. Recording conditions greatly alter the characteristics of the dataset. Many datasets are classified as "WILD," meaning they were recorded in unrestricted, real-world settings. Examples of these datasets include VoxCeleb (general) (2017), VoxCeleb1 [172] (2017), VoxCeleb2 (2018), LRS2 [177] (BBC) (2018), LRS3-TED [139] (2018), FFHQ [137] (2019), HDTF [138] (2021), TalkingHead-1KH [169] (2021), CelebV-HQ [175] (2022), and CelebrV-Text (2023). Although it presents challenges, this variation in background noise, lighting, and speaker position makes models trained on such data more resilient. In contrast, datasets such as MMFace4D [165] (2023), BIWI (3D) [132] (2021), VOCA [170] (2019), Lombard [176] (2018), RAVDESS [66] (2018), MEAD [159] (2020), VoxCeleb1 [172] (2016), SAVEE (2015), TCD-TIMIT (2015), CREMA-D [134] (2014), GRAD (2006), and CAVSR1.0 [133] (1998) are recorded in controlled "LAB" environments, providing cleaner data with consistent background and lighting, which can be useful for isolating particular factors like lip movements or Emotional

Table 11: Overview of Various Datasets

DATASET	HIGHLIGHTS	YEAR	HOUR	SPEAKERS	SENTENCES	ENV.	VIEW	CATEGORY	FPS	RESOLUTION
CelebV-Text [25]	Speech-driven facial animation	2023	279	70000	1.4M	Wild	Frontal	Video	—	512×512
MMFace4D [165]	4D motion capture for face animation	2023	36	431	11K	Lab	Frontal	Video, Audio	30	—
CelebV-HQ [175]	High-quality celebrity videos	2022	65	15653	—	Wild	Multi-view	Video	—	512×512
Multiface (3D) [166]	3D facial scan dataset	2022	—	13	—	Lab	Frontal	Image, Audio	1	2048×1334, 1024×1484
Responsive Listening Head [201]	3D avatars responding to audio	2022	1.5	67	483	Wild	Multi-view	Video, Audio, Image	30	384×384
BiWI (3D) [132]	Annotated 3D face scans	2021	1.44	14	1109	Lab	Multi-view	Video	25	—
HDTF [138]	Dataset for head pose estimation	2021	15.8	362	10K	Wild	—	Video	—	1280×720, 1920×1080
TalkingHead-1KH [169]	Emotion-rich talking videos	2021	1000	—	—	Wild	—	Video	—	—
MeshTalk [174]	3D talking heads driven by speech	2021	13	250	12.5K	Lab	—	Video	30	800×800
MEAD [159]	Emotion-labeled high-res video/audio	2020	39	60	20	Lab	Multi-view	Video, Audio	30	1920×1080
FaceForensics++ [136]	Deepfake detection corpus	2019	5.7	1000	1000+	Wild	Frontal	Video, Image	—	512×512
FFHQ [137]	GAN-ready diverse face photos	2019	—	—	—	Wild	—	Image	1	—
VOCA [170]	Vocal tract visualization via video	2019	0.5	12	40	Lab	Frontal	Video, Audio	60	2000×2000
Lombard [176]	Speech under noise conditions	2018	3.6	54	5400	Lab	Multi-view	Video, Audio	—	720×480, 864×480
LRS2 [177]	Lip-reading dataset from BBC TV	2018	224.5	500+	1.4M	Wild	Multi-view	Video, Audio, Text	—	224×224
LRS3-TED [139]	TED-based AV speech dataset	2018	438	5000	152K	Wild	Multi-view	Video, Audio, Text	—	224×224
MELD [162]	Emotion-rich dialogue videos	2018	13.7	407	—	Wild	—	Video, Audio, Text	—	—
RAVDESS [66]	Emotion expression by actors	2018	7	24	2	Lab	Frontal	Video, Audio, Text	—	1920×1080, 1280×720
MODALITY [47]	Multimodal affective dataset	2018	31	35	5800	Lab	—	Video, Audio, Video	—	1920×1080
ObamaSet [46]	Lip-sync and deepfake research dataset	2017	14	1	—	Wild	—	Video	—	—
VoxCeleb [171]	Celeb voice video collection	2017	352	1200	153.5K	Wild	—	Video	25	224×224
VoxCeleb1 [72]	Emotion in AV communication	2016	18	12	652	Lab	Frontal	Audio, Video, Image	—	1440×1080
VoxCeleb2 [172]	Unconstrained celeb speech data	2017	352	1251	153K	Wild	—	Video	30	224×224
LRW [140]	Word-level lip-reading videos	2016	173	1000	539K	Wild	Frontal	Video, Text	—	256×256
SAVEE [102]	Emotional speech videos	2015	—	480	—	Lab	Frontal	Video, Audio	60	—
TCD-TIMIT [103]	AV corpus for speech recognition	2015	11.1	62	6900	Lab	Multi-view	Video, Audio	—	1920×1080
CREMA-D [34]	Emotionally expressive AV clips	2014	11.1	91	12	Lab	—	Video, Audio	—	650×720
GRID [176]	Speech + lip reading video corpus	2006	27.5	34	33K	Lab	Frontal	Video, Audio	—	360×480, 720×576
DPCD [135]	3D Emotion + dynamic pose scans	2004	29.75	5	48.6K	Wild	Multi-view	Video, Audio, Text	25	—
CAVSRI.0 [133]	Early lip sync + facial animation set	1998	—	123	600	Lab	Frontal	Video, Audio	—	—

expressions. Another important consideration in the study of THG is view type. "FRONTAL-VIEW" recordings make up many datasets, particularly those that focus on talking head creation and lip-reading, which facilitates the learning process for lip motions. Nonetheless, datasets such as Multiface (3D) [166] (2022), BIWI (3D) [132] (2021), Lombard [176] (2018), LRS2 [177] (BBC) (2018), LRS3-TED [139] (2018), MEAD [159] (2020), and TCD-TIMIT (2015) provide "MULTI-VIEW" data, which is required to train models that can control multiple head postures and produce talking heads that are aware of three dimensions. "N/A" is specified for view type in datasets such as CelebV-HQ [175] (2022), HDTF [138] (2021), TalkingHead-1KH [169] (2021), FFHQ [137] (2019), VoxCeleb2 (2018), VoxCeleb (general) (2017), VoxCeleb1 [172] (2017), ObamaSet (2017), and CREMA-D [134] (2014), and this indicates variability or that this information is not the primary focus. These datasets are typically large-scale collections for general facial analysis or recognition. Modalities include "VIDEO" and "AUDIO" data to comprehend the connection between speech and facial movements. This essential combination is provided by a few datasets, including MMFace4D [165] (2023), Responsive Listening Head [201] creation (2022), BIWI (3D) [132] (2021), FaceForensics++ [136] (2019), MeshTalk [174] (2021), FFHQ [137] (2019), MODALITY [47] (2017), and GAN-based face creation. Motion capture data is essential to train models to produce lifelike 3D face animations. The quality and smoothness of the generated animations are defined by frame rate (fps). High-resolution 3D models or photos are provided by high-resolution datasets such as Multiface (3D) [166] (2022), MEAD [159] (2020), RAVDESS [66] (2018), TCD-TIMIT (2015), and DPCD [135] (2004). Additionally, another dataset that is not relatively cited but has an impact in the field of THG includes Hallo3 [99], VFHQ [98], and few others.

## 2.3 LOSS FUNCTION

In DL algorithms [33], loss functions [124] are fundamental since they define the difference between expected and actual values. They guide the optimization process to improve model accuracy. Selecting a loss function requires thoughtful evaluation of its relevance for a particular task. **Convexity:** A loss function is convex if its local minimum is the global minimum, making it easy to optimize using gradient-based methods. **Differentiability:** A loss function's derivative concerning model parameters exists and is continuous, enabling gradient-based optimization. A few extreme values should not influence loss functions; they should handle outliers. There should not be abrupt transitions or spikes in a loss function; rather, it should have a continuous gradient. A sparsity-promoting loss function benefits high-dimensional data and small features by promoting sparse output. **Monotonicity:** By guaranteeing the optimization process moves toward the proper solution, a loss function's value drops as the predicted output approaches the actual production. Based on the type of learning task, loss functions, which gauge how well an algorithm interprets data, are divided into two groups: classification models, which forecast the output from a set of finite categorical values, and regression models, which predict continuous values. This section examines the common loss functions used in classification and regression tasks. Regardless of whether the problem involves regression or classification, selecting the appropriate loss function is essential for optimizing the model to meet the specific requirements of the task.

### 2.3.1 Regression

Regression is a potent analytical technique that uses input features to predict continuous output values. This method is widely used in many fields, including finance (aiding in stock price prediction), healthcare (estimating patient outcomes), social sciences (analyzing trends and behaviors), sports (evaluating player performance), and engineering (optimizing design processes). Regression analysis allows us to find important information and make predictions based on data.

Practical applications include house price prediction [205], energy consumption forecasting [138], healthcare and disease prediction [124], stock price forecasting [61], and customer lifetime value prediction [45].

A regression loss function is defined over a dataset:

$$D = \{(x_i, y_i)\}_{i=1}^n$$

where  $x_i \in \mathbb{R}^d$  denotes the  $d$ -dimensional input feature vector and  $y_i \in \mathbb{R}$  the corresponding continuous target value.

The aim is to find the parameter vector  $\theta$  that minimizes the overall loss function:

$$\min_{\theta} \sum_{i=1}^n \mathcal{L}(f_{\theta}(x_i), y_i)$$

### 2.3.2 Classification

Deep Learning Talking head models involve sorting data into categories based on specific features or characteristics. There are two main classification types: binary classification, which aims to sort data into two distinct categories, and multi-class classification, which sorts data into more than two categories. Because binary classification involves choosing between two options, it is the most basic classification. Multi-class classification, on the other hand, entails grouping data into several categories, as in image recognition systems. Multi-label classification allows a single data point to belong to multiple classes simultaneously. While this approach is relevant in specific applications, it may not be as essential for a basic overview of classification. The decision boundary in a classification plot represents the line or area that the model uses to determine the category of an image. Classification uses a labeled dataset to train a model to identify patterns and relationships in the data. Data collection, feature extraction, model training, assessment, and prediction are all steps in this process. The model uses existing labeled data to teach itself how to predict the new, unlabeled data class based on the learned patterns. Classification algorithms are widely used in real-world applications such as email spam filtering, credit risk assessment, medical diagnosis, image classification, sentiment analysis, fraud detection, and recommendation systems. These algorithms are designed to handle binary and multi-class classification tasks, depending on the problem's nature. Classification modeling uses machine learning algorithms to categorize data into predefined classes or labels. Key characteristics of classification models include class separation, decision boundaries, sensitivity to data quality, handling imbalanced data, and interpretability. Well-labeled and representative data are essential for better performance, while noisy or biased data can result in poor predictions. Special techniques such as resampling or weighting are employed to address class imbalances. Classification involves using existing labeled data to train a model to predict the new, unlabeled data class based on the learned patterns.

### 2.3.3 Unsupervised

Until now, the discussion has focused on methods used in the context of supervised learning; however, unsupervised learning is one example of how they differ from supervised learning. Finding some structure in the data is the aim of unsupervised learning, which uses inputs  $x$  without matching output pairs. While unsupervised learning techniques operate differently and uncover various structures, they all share a common component. Utilizing the same concepts of hypothesis functions, loss functions, and optimization techniques as supervised learning, we can define unsupervised learning in a general way. Almost all unsupervised learning techniques can be seen this way, despite the various contexts.

To compensate for the absence of a clear target that we are attempting to fit, we must modify the definitions of hypotheses and loss functions. Two specific unsupervised learning algorithms that fall under this framework are the principal component analysis (PCA) algorithm and the  $k$ -means clustering algorithm.

The hypothesis function  $h : \mathbb{R}^n \rightarrow \mathbb{R}^n$  maps input  $\mathbb{R}^n$  back into the input space. Similar to the supervised case, the objective of this hypothesis class is to approximate the reconstruction of the input. The specific configuration of unsupervised learning algorithms limits the class of hypothesis functions, requiring them to efficiently recover a certain amount of structure in the data to accurately approximate their input rather than output the function's input.

The loss function indicates the degree to which the prediction deviates from the original input:

$$\ell : \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}_+.$$

The  $k$ -means and principal component analysis algorithms are both based on the same loss function, given by the formula:

$$\ell(h(x), x) = \|h(x) - x\|_2^2.$$

## 2.4 Evaluation Metrics

Evaluation metrics serve as vital tools for gauging the performance of talking head models, providing a clear lens through which the effectiveness of innovative approaches can be compared to well-established benchmarks. These metrics illuminate whether a model can achieve performance levels on par with or surpass those of human experts. Meanwhile, machine learning (ML) and deep learning (DL) have transcended their origins, branching from engineering and medicine into diverse landscapes such as finance, politics, and the natural sciences. Rapid technological advancements and the ever-growing ocean of digital data at our fingertips have fueled this remarkable expansion. Machine learning (ML) now supports complex systems in academic research

Table 12: Benchmark Loss Functions Used in Talking Head Generation

Loss Function	Formula	Description	Limitation	Implementation Type
Square Loss ( $L_2$ )	$\ell_{\text{square}} = \frac{1}{S} \sum_{i=1}^S (x_i - \hat{x}_i)^2$	Penalizes large errors quadratically	May blur structure in visual outputs	Image-based, Video-based
Absolute Loss ( $L_1$ )	$\ell_{\text{abs}} = \frac{1}{n} \sum_{i=1}^n  x_i - \hat{x}_i $	Simple and robust to outliers	May not differentiate well between large/small errors	Image-based, Video-based
Epsilon-Insensitive Loss	$\ell_\epsilon = \max(0,  x - \hat{x}  - \epsilon)$	Ignores small errors within $\epsilon$ -tube, used in SVR	Non-smooth, no penalty for small errors	Audio-based, Video-based
IoU Loss	$\ell_{\text{IoU}} = 1 - \frac{ A \cap B }{ A \cup B }$	Optimizes overlap in detection tasks	Gradient is zero if no overlap	Image-based
Generalized IoU Loss	$GIoU = IoU - \frac{ C \setminus (A \cup B) }{ C }$	Extends IoU for non-overlapping cases	Slightly complex to compute	Image-based
Smooth L1 Loss	$\ell = \begin{cases} 0.5(x - \hat{x})^2 & \text{if }  x - \hat{x}  < 1 \\  x - \hat{x}  - 0.5 & \text{otherwise} \end{cases}$	Combines $L_1$ and $L_2$ benefits	Less sensitive to small changes than $L_2$	Image-based, Video-based
Focal Loss	$FL(p) = -(1-p)^\gamma \log(p)$	Deals with class imbalance in detection	Needs tuning of $\gamma$ and $\alpha$	Video-based
Contrastive Loss	$\ell = (1-y) \frac{1}{2} D^2 + y \frac{1}{2} [\max(0, m-D)]^2$	Separates similar/dissimilar pairs	Needs carefully chosen pairs	Image-based, Audio-based
Triplet Loss	$\ell = \sum \max(0, \ f_a - f_p\ ^2 - \ f_a - f_n\ ^2 + \alpha)$	Anchor-positive closer than anchor-negative	Needs hard-negative mining	Image-based, Video-based
Center Loss	$\ell = \sum \ x_i - c_{y_i}\ ^2$	Reduces intra-class variation	Needs learning of class centers	Video-based
Angular softmax Loss	$\ell = \log \frac{e^{s \cos(\theta+m)}}{e^{s \cos(\theta+m)} + \sum_{j \neq y} e^{s \cos(\theta_j)}}$	Adds angular margin for separation	May face numerical instability	Image-based
AM-softmax Loss	$\ell = -\log \frac{e^{s(\cos \theta - m)}}{e^{s(\cos \theta - m)} + \sum_{j \neq y} e^{s \cos \theta_j}}$	Stable angular margin loss	Needs tuning of $s$ and $m$	Audio-based
ArcFace Loss	$\ell = -\log \frac{e^{s \cos(\theta+m)}}{e^{s \cos(\theta+m)} + \sum_{j \neq y} e^{s \cos(\theta+m)}}$	Strong geometric interpretation	Expensive, tuning needed	Image-based, Video-based
Reconstruction Loss	$\ell = \ x - \hat{x}\ ^2$ or $\ x - \hat{x}\ _1$	Evaluates reconstruction accuracy	May cause mode collapse	Image-based, Video-based
Negative Variance	$\ell = -\text{Var}(z)$	Promotes latent space diversity	May yield trivial solutions if unconstrained	Text-based, Audio-based

and industrial practice using simple algorithms and statistical techniques. This quick interdisciplinary growth highlights the necessity for continual instruction on the appropriate use of statistics and selecting suitable performance metrics. Supervised machine learning trains and validates models on the training set while predictions are compared to ground-truth values on the test set. Reliable performance evaluation and the advancement of ML technologies depend on appropriate evaluation procedures [104] that are backed by validated statistical tests and well-selected metrics. Table 13 offers a thorough summary of the assessment metrics used in THG research. The text outlines the metrics used in various studies and standard models, categorizing them based on their descriptive qualities. Additionally, the table 13 discusses the usability of each metric across different application categories and highlights their limitations when employed for absolute comparisons.

Assessing THG models is a multifaceted endeavor that demands a nuanced set of metrics to evaluate different dimensions of the produced output. It includes the authenticity of the visuals, the harmony between audio and visual elements, and the coherence of the conveyed message. The evaluation tools utilized in this intricate field are dynamic image-based assessments, immersive audio-based criteria, insightful text-based analyses, and comprehensive video-based evaluations. Every measure is essential to guarantee that the produced material satisfies high-quality and efficacy requirements. Image-based metrics focus on evaluating the visual quality of the generated frames. CSIM measures visual similarity by focusing on the angular similarity between feature vectors of generated and real images. SSIM addresses this by focusing on structural information and perceptual quality, aligning better with human visual perception. However, SSIM is sensitive to image misalignment, contrast changes, and scale differences, potentially failing to capture high-level semantic errors even when pixel-level structures appear similar. Based on pixel-wise inaccuracy, PSNR is a commonly used measure for assessing the difference between processed and original pictures. PRMSE is another pixel-wise error metric, indicating higher quality with lower error. L1 Loss is the absolute difference between pixel values, offering robustness to outliers and computational simplicity. MKR measures the accuracy of facial animation models by evaluating the alignment between predicted and actual facial key points, which is crucial for tasks like THG and facial expression generation. Audio-visual synchronization metrics evaluate the temporal alignment between the generated video and the driving audio. AUCON assesses this alignment, indicating improved quality with better congruence between audio and visual components, which is particularly important for reenactment and speech-to-lip synchronization. AKD directly measures the distance between key points on the face in the video and features extracted from the audio, aiming for accurate lip synchronization and expression alignment in audio-driven visual synthesis. Its accuracy is inherently dependent on the reliability of the keypoint extraction algorithm, and errors in this process can skew results. SyncNet is a deep learning model specifically trained to evaluate the synchronization between audio and visual streams, focusing on lip synchronization and cross-modal temporal consistency. Generative model evaluation metrics are often video-based in this context. Using a pre-trained Inception network, FID (Fréchet Inception Distance) calculates the separation between the feature distributions of generated and real images. CITALower FID scores generally indicate better image quality and diversity, making it useful for evaluating GAN-based talking head models. FVD (Fréchet Video Distance) extends FID to the video domain by incorporating temporal information, suggesting better video quality and coherence, which is crucial for realistic talking heads. E-FID (Efficient Fréchet Inception Distance) is a computationally more efficient variant of FID, aiming to reduce the computational burden while still assessing the realism of generated images and videos. LMD (Learned Metric for Image Quality Assessment) aims to assess perceptual quality by comparing the data distribution of generated samples to real data. CPBD (Cumulative Probability of Blur Detection) is a perceptual metric to assess image sharpness and the loss of fine details. AVD (Audio-Visual Discrepancy) directly measures the perceptual mismatch between the audio and the generated visual content in talking head videos. Text-based metrics are relevant for evaluating semantic coherence when THG is driven by textual input. CLIPSIM (CLIP Similarity) measures the semantic similarity between generated images/videos and the input text using the embeddings learned by the CLIP model. F-LMD (Facial Landmark Distance) and M-LMD (Mesh Landmark Distance) are text-based metrics that likely evaluate the accuracy of generated facial landmarks or mesh deformations based on textual descriptions of expressions or phonemes. Sync(conf) (Synchronization Confidence) in a text-to-talking head context might refer to the model's confidence in generating synchronized lip movements based on the text. Higher confidence scores would ideally correlate with better synchronization, but this metric can be influenced by model calibration and might be misleading in cases of untrained content. The right evaluation metrics must be chosen to evaluate THG progress objectively. Researchers use a combination of metrics to understand model strengths and weaknesses, including visual quality, temporal coherence, audio-visual synchronization, and semantic accuracy. The development of more perceptually aligned evaluation metrics remains an important research area. The following section evaluates the most frequently cited model based on the benchmark metrics discussed earlier. This evaluation utilizes

Table 13: Benchmark Evaluation Metrics for THIG

Evaluation Metric	Description	Limitation	Applicable Category
CSIM↑ [202]	Visual similarity measure for image/video generation, super-resolution, style transfer. Measures structural similarity aligning with human perception.	May overlook structural/intensity changes despite angular similarity focus.	Image-based, Video-based
SSIM↓ [97]	Measures pixel-wise fidelity, popular in image quality evaluation.	Sensitive to misalignment, contrast and scale variations.	Image, Audio, Text, Video
PSNR↑ [96]	Compares pixel-wise fidelity, popular in image quality evaluation.	Not aligned with perceptual quality; insensitive to structural changes.	Image, Audio, Video
PRMSE↓ [202]	Measures pixel error magnitude between original and output images.	May highlight trivial differences that aren't perceptually meaningful.	Image-based, Video-based
AUCON↑ [202]	Evaluates audio-visual alignment and naturalness.	Doesn't capture local artifacts or spatial inaccuracies.	Image-based, Video-based
L1↓ [95]	Measures absolute pixel-wise difference.	Doesn't distinguish semantic/structural errors.	Image-based, Video-based
AKD↓ [20]	Measures distance between facial landmarks.	Sensitive to landmark detection accuracy.	Image-based, Video-based
MKR↑ [20]	Facial animation accuracy metric using key-points.	Heavily dependent on thresholding and landmark precision.	Image-based
AED↑ [59]	Assesses audio-expression alignment.	Affected by orientation, pose, and scale variations.	Image-based
FID↓ [94]	Evaluates image realism/diversity using Inception features.	Sample-size dependent; biased by feature extractor.	Audio, Text, Video
SyncNet↑ [93]	Assesses lip-audio synchronization using pre-trained SyncNet.	Reliability depends on training diversity.	Audio-based
F-SIM↓ [161]	Uses deep features to compare image similarity.	May miss low-level artifacts due to network biases.	Audio-based
FVD↓ [163]	Extension of FID to video; captures temporal coherence.	Faces same issues as FID; sample size and motion artifacts.	Audio, Text-based
E-FID↓ [160]	Efficient version of FID with edge-awareness.	May overemphasize edges while missing other quality features.	Audio-based
LMD↓ [19]	Compares landmark distribution to assess quality.	May miss perceptual quality beyond facial structure.	Video-based
CIPBDD↑ [192]	Measures perceptual sharpness and blur.	Struggles in low-contrast or textured areas.	Audio, Text-based
CLIPSIM↑ [194]	Measures semantic similarity in text-image alignment.	Dependent on CLIP training domain.	Text-based
F-LMD↓ [145]	Specialized for biological/lipid image evaluation.	Limited to landmark-based fidelity; poor perceptual correlation.	Text-based
M-LMD↓ [145]	Similar to F-LMD with more domain specificity.	Overly focused on biological features, domain-sensitive.	Text-based
Sync(conf)↑ [93]	Measures sync confidence between modalities.	Can be misled by unseen or mismatched content.	Text-based
AVD↓ [196]	Audio-visual discrepancy measure.	Depends on feature representations, may not reflect perceptual gaps.	Video-based

data from a highly cited article from Google Scholar, one of the most reliable academic databases. Various metrics used for model evaluation are compared within the same dataset and organized according to the previously mentioned categories.

#### 2.4.1 Evaluation of Image-Based Approaches

Based on the existing literature from sources [31] and [20], which were selected for comparison due to their relevant citations, we have produced a summary highlighting the comparison of various image-based approaches across different datasets and assessment metrics. This information can be found in Tables 14 and 15.

Table 14: Evaluation of Image-based (Part-1)

<b>Dataset</b>	<b>Model</b>	CSIM↑	SSIM↓	PSNR↑	PRMSE↓	AUCON↑
VoxCeleb1	X2Face [232]	0.689	0.719	22.537	3.26	0.813
	NeuralHead-FF [44]	0.229	0.635	20.818	3.76	0.719
	MarioNETte [202]	0.755	0.744	23.244	3.13	0.825
	FOMM [20]	0.813	0.723	30.394	3.20	0.886
	MeshG [53]	0.822	0.739	30.394	3.20	0.887
	OSFV [141]	0.895	0.761	30.695	1.64	0.921
	DaGAN [31]	0.899	0.804	31.220	1.22	0.939
CelebV	X2Face [232]	0.450	–	3.620	–	0.679
	NeuralHead-FF [44]	0.108	–	3.300	–	0.722
	MarioNETte [202]	0.520	–	3.410	–	0.710
	FOMM [20]	0.462	–	3.900	–	0.667
	MeshG [53]	0.635	–	3.410	–	0.709
	OSFV [141]	0.791	–	3.150	–	0.805
	DaGAN [31]	0.723	–	2.330	–	0.873

Table 15: Evaluation of Image-based (Part-2)

<b>Dataset</b>	<b>Model</b>	CSIM↑	SSIM↓	PSNR↑	PRMSE↓
Tai-Chi-HD	X2face [83]	0.08	17.654	0.109	0.272
	Monkey-Net [59]	0.077	10.798	0.059	0.228
	FOMM [20]	0.063	6.862	0.036	0.179
VoxCeleb	X2face [83]	0.078	7.687	X	0.405
	Monkey-Net [59]	0.049	1.878	X	0.199
	FOMM [20]	0.043	1.294	X	0.14
Nemo	X2face [83]	0.031	3.539	X	0.221
	Monkey-Net [59]	0.018	1.285	X	0.077
	FOMM [20]	0.016	1.119	X	0.048

The evaluation of image-based THG models featured in the current literature is vividly encapsulated in Table. This comprehensive overview includes a diverse array of models—X2face [232], NeuralHead-FF [44], MarioNETte [202], FOMM [20], MeshG, OSFV [141], and DaGAN [31]—each subjected to rigorous scrutiny using standard metrics across two prominent datasets: VoxCeleb1 [172] and CelebV. The results definitively reveal that DaGAN [31] and OSFV [141] dominate the field, consistently delivering exceptional performance

marked by impressive metrics in Content Similarity (CSIM), Peak Signal-to-Noise Ratio (PSNR), and Area Under the Curve of Object Consistency (AUCON). Their ability to produce strikingly realistic and coherent talking heads sets a benchmark for others. FOMM [20] and MeshG also shine, showcasing remarkable capabilities, especially in PSNR and CSIM.

In contrast, while X2face holds its own reasonably well, it pales compared to the prowess of its more recent counterparts. Meanwhile, NeuralHead-FF [44] struggles significantly, landing at the bottom of the performance spectrum among the models evaluated. When we turn our attention to the CelebV dataset, the formidable duo of OSFV [141] and DaGAN [31] continues to impress, with DaGAN [31] achieving the highest AUCON and PSNR, further complemented by the lowest Peak Root Mean Square Error (PRMSE). The tables also venture into motion-related metrics, evaluating similar THG models across various datasets, including the dynamic Tai-Chi-HD, the diverse VoxCeleb, the engaging EMO [160], and the intricate Bair. Here, FOMM [20] consistently excels, brilliantly capturing and replicating intricate motion dynamics, underlining its effectiveness and showcasing its superiority. Monkey-Net provides a commendable balance, standing strong in its performance, while X2face, regrettably, lags in motion metrics. These findings collectively highlight remarkable strides in the domain of THG. Models like DaGAN [31] and OSFV [141] achieve breathtaking visual quality and redefine the standards of realism. Meanwhile, Exceptional performance of FOMM [20] in motion accuracy draws attention to the ongoing evolution in this field. This review emphasizes these compelling results, delves into the intricate trade-offs between visual fidelity and motion realism, and proposes exciting avenues for future research and the development of new metrics.

#### 2.4.2 Evaluation of Audio-Based Approaches

A detailed comparison of diverse audio-based methodologies across various datasets and evaluation metrics is beautifully encapsulated in Tables 16 and 17. This insightful summary draws from the latest literature from [160] and [20], carefully chosen for their notable citations and significance within the scholarly landscape. Each methodology is meticulously evaluated, providing a comprehensive overview highlighting its strengths and weaknesses in tackling audio-related challenges.

Table 16: Evaluation of Audio-based (Part-1)

DATASET	MODEL	FID $\downarrow$	SyncNet $\uparrow$	PRMSE $\downarrow$	F-SIM $\downarrow$	FVD $\downarrow$
HDTF	Wav2Lip [55]	9.38	5.79	80.34	407.93	0.693
	SadTalker [121]	10.31	4.82	84.56	214.98	0.503
	DreamTalk [9]	58.8	3.43	67.87	619.05	2.257
	MakeItTalk [201]	21.73	2.85	76.91	350.96	1.072
	EMO [160]	8.76	3.89	78.96	67.66	0.116

The evaluation of various audio-driven THG models is thoroughly detailed in Table 2. This Table 2 features a diverse array of models, including FOMM [20] (Few-Shot Video Generation), OSFV (One-Shot Face Video Generation) [141], DaGAN [31] (Dynamic Generative Adversarial Network), ROME [56] (Recurrent Oral Motion Encoder), FNeVR [190] (Fast Neural Video Representation), and HiDe-NeRF [195] (High-Definition Neural Radiance Field). These models' performance was assessed using three distinct datasets: VoxCeleb1 [172], VoxCeleb2, and the more challenging TalkingHead-1KH [169].

Metrics used for this evaluation encompass several important aspects of video quality and synchronization. These include CSIM (Content Similarity), which measures how closely the generated content matches the original; AUCON (Area Under the Curve of Object Consistency), which assesses the stability of facial features over time; PRMSE (Pixel Root Mean Square Error); which quantifies pixel-level discrepancies; FID (Fréchet Inception Distance), a measure of the quality of generated images compared to real photos; and AVD (Audio Visual Distance), which evaluates the alignment and synchronization of audio and visual components.

HiDe-NeRF [195] stands out as the leading model across all tested datasets. It achieves the highest scores for CSIM and AUCON, indicating top-tier content fidelity and consistency throughout the video. It minimizes PRMSE, FID, and AVD scores, underscoring its excellence in visual quality and audio-visual synchronization. ROME [56] also excels, particularly in the AVD metric, which suggests that it effectively captures the nuances of facial movements in concert with corresponding audio, enhancing the realism of the generated videos.

Table 17: Evaluation of Audio-based (Part-2)

DATASET	MODEL	LMD↓	SSIM↑	PSNR↑	CPBD↑
GRID	Vondrick [27]	2.38	0.6	28.45	0.129
	Chung [220]	1.35	0.74	29.36	0.016
	LMGG [19]	1.18	0.73	29.89	0.175
LDC	Vondrick [27]	2.34	0.75	27.96	0.16
	Chung [220]	2.13	0.5	28.22	0.01
	LMGG [19]	1.82	0.57	28.87	0.172
LRW	Vondrick [27]	3.28	0.34	28.03	0.082
	Chung [220]	2.25	0.46	28.06	0.083
	LMGG [19]	1.92	0.53	28.65	0.075
HDTF	MakeItTalk[201]	X	0.802	23.2454	0.1226
	FOMM [20]	X	0.8167	23.4079	0.1345

In terms of performance, OSFV [141] and DaGAN [31] exhibited competitive results. However, OSFV [141] generally surpassed DaGAN [31], indicating a more robust ability to generate convincing and coherent talking head videos. Notably, FOMM [20], while innovative, performed the most poorly among the newer models. FNeVR [190]'s outcomes place it in the middle range, showing potential yet lacking the statistical excellence of its counterparts.

The TalkingHead-1KH [169] dataset poses significant challenges to all models, as evidenced by the lower performance scores. This dataset likely contains complexities that test the limits of current generation techniques.

Overall, the findings presented in the Table 2 illuminate the substantial progress made by recent models, especially HiDe-NeRF [195], in producing high-quality talking head videos with accurate synchronization between audio and visual elements. The analysis of the AVD metric reveals a notable performance gap among the various models, suggesting ample opportunity for further advancements in this field. The comprehensive review underscores the importance of meticulously evaluating visual quality and motion accuracy, highlights the challenges associated with different datasets, and emphasizes the continued significance of metrics such as AVD and AKD. Finally, it points to the necessity for consistent metric reporting to enhance comparability and advancement within THG.

#### 2.4.3 Evaluation of Video-Based Approaches

Based on the existing literature, which has been selected based on the citation in comparison with others, we have produced a summary of the comparison of different video-based approaches in different datasets across different assessment metrics in Table 18 and 19.

The evaluation of video-driven THG models presented in Table includes FOMM [20], OSFV [141], DaGAN [31], ROME [56], FNeVR [190], and HiDe-NeRF [195], assessed across three datasets: VoxCeleb1 [172], VoxCeleb2, and TalkingHead-1KH [169]. Key metrics used in this evaluation are CSIM (Content Similarity), AUCON (Area Under the Curve of Object Consistency), PRMSE (Pixel Root Mean Square Error), FID (Fréchet Inception Distance), and AVD (Audio Visual Distance).

HiDe-NeRF [195] consistently outperforms others in all datasets, achieving the highest scores in CSIM and AUCON and the lowest in PRMSE, FID, and AVD, indicating superior visual quality and audio-visual synchronization. ROME [56] also performs well in AVD, showcasing strong audio-visual synchronization capabilities.

Table 18: Evaluation of Video-based (Part-1)

DATASET	MODEL	CSIM↑	AUCON↑	PRMSE↓	FID↓	AVD↓
VoxCeleb1	FOMM [20]	0.748	0.752	3.66	86	0.044
	OSFV [141]	0.791	0.893	3.01	74	0.028
	DaGAN [31]	0.79	0.88	3.06	87	0.036
	ROME [56]	0.833	0.871	2.64	76	0.016
	FNeVR [58]	0.812	0.884	3.32	82	0.041
	HiDe-NeRF [37]	0.876	0.917	2.62	57	0.012
VoxCeleb2	FOMM [20]	0.68	0.707	4.16	85	0.047
	OSFV [141]	0.711	0.833	3.84	72	0.033
	DaGAN [31]	0.693	0.815	3.93	86	0.04
	ROME [56]	0.71	0.821	3.08	76	0.019
	FNeVR [58]	0.699	0.829	3.9	84	0.047
	HiDe-NeRF [37]	0.787	0.889	2.91	61	0.014
TalkingHead-1KH	FOMM [20]	0.723	0.741	3.71	76	0.039
	OSFV [141]	0.787	0.884	3.03	67	0.025
	DaGAN [31]	0.766	0.872	2.98	73	0.035
	ROME [56]	0.781	0.864	2.66	68	0.017
	FNeVR [58]	0.775	0.879	3.39	73	0.037
	HiDe-NeRF [37]	0.828	0.901	2.6	52	0.011

Table 19: Evaluation of Video-based (Part-2)

DATASET	MODEL	L1↓	PSNR↑	SSIM↑	MS-SSIM↑	FID↓	AKD↓
VoxCeleb2	fs-vid2vid[141]	17.1	20.36	0.71	Nan	85.76	3.41
	FOMM [20]	12.66	23.25	0.77	0.83	73.71	2.14
	Bi-layer[58]	23.95	16.98	0.66	0.66	203.36	5.38
	Neural Talking-Head [42]	10.74	24.37	0.8	0.85	69.13	2.07
TalkingHead-1KH	fs-vid2vid [141]	15.18	20.94	0.75	Nan	63.47	11.07
	FOMM [20]	12.3	23.67	0.79	0.83	55.35	3.76
	Bi-layer [58]	12.81	23.13	0.78	Nan	60.58	60.58
	Neural Talking-Head [42]	10.67	24.2	24.2	0.84	52.08	3.74

OSFV [141] and DaGAN [31] exhibit competitive performance, with OSFV [141] generally outperforming DaGAN [31]. FOMM [20] ranks lowest among the newer models, whereas FNeVR [190] demonstrates average performance. The TalkingHead-1KH [169] dataset is the most challenging, as it typically yields lower performance scores across all models.

The evaluation highlights significant advancements achieved by recent models, particularly HiDe-NeRF [195], in generating high-quality talking head videos with accurate audio-visual synchronization. Notably, the AVD metric reveals substantial differences in performance between the models, indicating a significant area for improvement within the field.

Overall, the tables provide a detailed picture of the current level of THG, emphasizing the importance of evaluating visual quality and motion accuracy, addressing the challenges posed by different datasets, and the continued relevance of AVD and AKD metrics. It also underscores the need for consistent reporting of metrics within the field.

#### 2.4.4 Evaluation of Text-Based Approaches

A summary of the comparison of various text-based methodologies in various datasets across various evaluation metrics is provided in Tables 20 and 21, which are based on the current literature from the sources [228], which were chosen based on the citation in comparison with others.

Table 20: Evaluation of Text-based (Part-1)

DATASET	MODEL	FVD↓	FID↑	CLIPSIM↑
MM-Vox	TFGAN [29]	$502.28 \pm 1.66$	$760.24 \pm 16.01$	$0.165 \pm 0.022$
	MMVID [30]	$65.79 \pm 1.81$	$38.81 \pm 3.66$	$0.170 \pm 0.020$
<u>CelebV-HQ</u>	TFGAN [29]	$428.04 \pm 1.76$	$616.24 \pm 17.45$	$0.168 \pm 0.02$
	MMVID [30]	$73.65 \pm 1.43$	$63.86 \pm 3.66$	$0.172 \pm 0.019$
<u>CelebV-Text</u>	TFGAN [29]	$403.04 \pm 1.34$	$589.24 \pm 16.46$	$0.177 \pm 0.012$
	MMVID [30]	$66.69 \pm 1.35$	$58.70 \pm 4.67$	$0.198 \pm 0.014$
CelebV-Text App.+Emo.	TFGAN[29]	$442.30 \pm 2.56$	$623.17 \pm 18.8$	$0.158 \pm 0.024$
	MMVID [30]	$82.78 \pm 1.47$	$61.58 \pm 3.99$	$0.176 \pm 0.00$
	MMVID-interp [122]	$72.87 \pm 1.23$	$1.57 \pm 3.56$	$1.57 \pm 3.56$
CelebV-Text App.+Act.	TFGAN[29]	$571.34 \pm 4.54$	$784.93 \pm 20.13$	$0.154 \pm 0.028$
	MMVID [30]	$109.25 \pm 2.1$	$82.55 \pm 4.37$	$0.174 \pm 0.01$
	MMVID-interp [122]	$80.81 \pm 2.55$	$70.88 \pm 4.77$	$0.176 \pm 0.020$

The evaluation of text-driven THG models is presented in Table, which includes TFGAN [29] and MMVID [30] across various datasets. The results indicate that MMVID [30] consistently outperforms TFGAN [29] in all datasets, achieving significantly lower Fréchet Video Distance (FVD) and Fréchet Inception Distance (FID) scores, as well as higher CLIP similarity (CLIPSIM) scores. Adding appearance, Emotion, and action prompts to CelebV-Text [25] influences performance, with MMVID [205] maintaining a relative advantage over TFGAN [29]. The table also provides a detailed evaluation of various THG models, including MakeItTalk [201], Wav2Lip [55], PC-AVS [118], AVCT, GC-AVT [143], EAMM [144], SadTalker [146], PD-FGC, EAT, StyleTalk [145], and TalkCLIP [227], across MEAD [159], HDTF [138], and Voxceleb2 datasets. The evaluation focuses on lip synchronization, video quality, and motion accuracy. Models like StyleTalk [145], TalkCLIP [227], and Wav2Lip [55] generally achieve the best SSIM and Sync(conf) scores, indicating superior visual quality and audio-visual synchronization. The tables highlight the trade-offs between visual quality, lip motion accuracy, and audio-visual synchronization, with models like StyleTalk [145], TalkCLIP [227], and Wav2Lip [55] demonstrating strong performance in balancing these factors. However, there is a large variance in the SSIM scores, showing that visual fidelity is still a large area of research. The tables offer valuable insights into the performance of various THG models, particularly in text-driven scenarios and detailed lip synchronization evaluations. Key takeaways for the review include the significant advancements achieved by models like MMVID [205] in text-driven THG, the importance of evaluating lip synchronization and visual quality using detailed metrics, the influence of dataset characteristics on model performance, the strengths of models like StyleTalk [145], TalkCLIP [227], and Wav2Lip [55] in balancing visual fidelity and audio-visual synchronization, and the continued challenges in achieving perfect lip sync and high visual fidelity.

#### 2.4.5 Evaluation of 2D MODEL-Based Approaches

Based on the existing literature, which has been selected based on the citation in comparison with others, we have produced a summary of the comparison of different 2D Model-based approaches in different datasets across different assessment metrics in Tables 22 and 23.

The table 22 and 23 comprehensively evaluates various THG models, including SDA, MakeItTalk [201], Wav2Lip [55], PC-AVS [118], EAMM [144], and Diffused Heads [26], based on the LRW and CREMA datasets. The key metrics assessed include FVD (Fréchet Video Distance), FID (Fréchet Inception Distance), Blinks/s (blinks per second), Blink dur (blink duration), ofM (Optical Flow Magnitude), F-MSE (Facial Mean Squared Error), AV off (Audio-Visual offset), AV Conf. (Audio-Visual Confidence), and WER (Word Error Rate).

DiffusedHeads [26] generally achieve the best FVD scores on both datasets, indicating superior video quality. Wav2Lip [55] excels in FID scores, showcasing outstanding image quality. PC-AVS [118] demonstrates the

Table 21: Evaluation of Text-based (Part-2)

Dataset	Model	SSIM↓	CPBD↑	F-LMD↓	M-LMD↓	Sync (Conf)↑
MEAD	MakeItTalk [201]	0.73	0.11	3.95	5.39	2.15
	Wav2Lip [144]	0.81	0.16	2.73	3.85	5.41
	PC-AVS [118]	0.51	0.07	5.87	5.03	2.21
	AVCT [65]	0.83	0.14	2.95	5.64	2.56
	GC-AVT [63]	0.34	0.14	8.11	8.43	2.41
	EAMM [144]	0.40	0.08	6.67	6.60	1.42
	SadTalker [121]	0.68	0.16	4.04	4.24	2.87
	PD-FGC [39]	0.51	0.05	5.41	3.94	2.46
	EAT [67]	0.53	0.15	5.63	4.98	2.19
	StyleTalk [145]	0.84	0.16	2.17	3.36	3.51
	TalkCLIP [227]	0.83	0.16	2.42	3.60	3.77
HDTF	MakeItTalk [201]	0.57	0.20	5.12	4.61	3.20
	Wav2Lip [144]	0.59	0.26	5.11	3.84	4.57
	PC-AVS [118]	0.42	0.12	10.7	8.60	4.15
	AVCT [65]	0.74	0.18	3.06	3.83	4.46
	GC-AVT [63]	0.33	0.24	10.7	6.34	4.23
	EAMM [144]	0.37	0.13	7.74	7.67	2.78
	SadTalker [121]	0.73	0.19	6.26	4.18	3.86
	PD-FGC [39]	0.40	0.13	9.99	4.46	4.20
	EAT [67]	0.55	0.18	4.12	4.24	3.95
	StyleTalk [145]	0.80	0.26	2.04	2.50	4.75
	TalkCLIP [227]	0.78	0.25	2.54	2.84	4.69
VoxCeleb2	MakeItTalk [201]	0.52	0.24	6.29	5.15	2.17
	Wav2Lip [144]	0.54	0.30	5.85	4.64	5.70
	PC-AVS [118]	0.36	0.09	12.9	7.42	4.73
	AVCT [65]	0.64	0.23	3.62	3.71	3.89
	EAMM [144]	0.43	0.20	6.36	4.89	2.24
	SadTalker [121]	0.44	0.19	9.12	6.11	4.38
	PD-FGC [39]	0.35	0.12	12.5	8.19	4.64
	EAT [67]	0.47	0.20	5.53	5.88	4.35
	StyleTalk [145]	0.66	0.29	2.92	2.96	4.51
	TalkCLIP [227]	0.67	0.29	2.94	2.99	4.60

highest of M on the LRW dataset, while MakeItTalk [201] and EAMM [144] have lower AV Conf. scores, indicating challenges in audio-visual alignment.

The table 22 and 23 emphasizes the trade-offs between various aspects of THG, such as video quality, lip synchronization, and accuracy in lip reading. The assortment of metrics illustrates the complexity of evaluating THG. Diffused Heads [26] exhibit a strong capability for producing high-quality videos, while Wav2Lip [55] is noted for generating high-quality images and achieving precise lip synchronization.

This analysis underscores the importance of meticulously evaluating audio-visual synchronization and reinforces trends identified in previous summaries. It highlights the strengths and weaknesses of different models regarding visual quality, lip synchronization, and lip-reading accuracy. Furthermore, it discusses the challenges of evaluating THG and acknowledges the absence of data in the table 22 and 23.

#### 2.4.6 Evaluation of 3D MODEL-Based Approaches

A summary of the comparison of various 3D-based methodologies in various datasets across various evaluation metrics is provided in Table 24 and 25, which are based on the current literature, which were chosen based on the citation in comparison with others.

Table 22: Evaluation of 2D Model-based (Part-1)

DATASET	MODEL	FVD ↓	FID ↓	Blinks/s	Blink dur	ofM	F-MSE	AV off	AV Conf. ↓	WER ↑
LRW	SDA [27]	198.84	61.95	0.52	0.28	73.82	18.94	1	7.40	0.77
	MakelTalk [168]	269.29	7.57	0.09	0.28	57.21	3.44	-3	3.16	0.99
	Wav2Lip [21]	366.14	2.83	0.03	0.16	47.12	1.45	-2	6.58	0.51
	PC-AVS [118]	153.12	11.96	0.20	0.16	69.59	17.13	-3	6.24	0.64
	EAMM [144]	172.18	9.28	0.03	0.16	58.46	4.39	-3	3.83	0.95
	Diffused Heads [26]	71.88	3.94	0.35	0.28	70.71	19.69	-2	4.61	0.77
CREMA	SDA [27]	376.48	79.82	0.25	0.26	68.21	6.83	2	5.50	—
	MakelTalk [168]	256.88	17.26	0.02	0.80	62.36	2.07	-3	3.75	—
	Wav2Lip [21]	193.32	12.57	0.00	—	46.87	1.07	-2	6.68	—
	PC-AVS [118]	333.94	22.53	0.02	0.20	70.36	6.93	-3	6.17	—
	EAMM [144]	6.17	19.40	0.00	—	58.91	1.65	-2	4.26	—
	Diffused Heads [26]	88.61	12.45	0.28	0.36	6.99	1	4.52	—	—

Table 23: Evaluation of 2D Model-based (Part-2)

DATASET	MODEL	L1 ↓	PSNR ↑	SSIM ↑	MS-SSIM ↑	FID ↓	AKD ↓
VoxCeleb2	fs-vid2vid [141]	17.10	20.36	0.71	—	85.76	3.41
	FOMM [20]	12.66	23.25	0.77	0.83	73.71	2.14
	Bi-layer [58]	23.95	16.98	0.66	0.66	203.36	5.38
	Neural Talking-Head [42]	10.74	24.37	0.80	0.85	69.13	2.07
TalkingHead-1KH	fs-vid2vid [141]	15.18	20.94	0.75	—	63.47	11.07
	FOMM [20]	12.30	23.67	0.79	0.83	55.35	3.76
	FOMM-L [20]	12.81	23.13	0.78	—	60.58	4.04
	Neural Talking-Head [42]	10.67	24.20	0.81	0.84	52.08	3.74

This table 24 and 25 evaluates 3D-aware THG models across multiple datasets, including VoxCeleb, CelebV-HQ [175], and TalkingHead-1KH [169]. The key metrics assessed include Fréchet Video Distance (FVD), Identity Similarity (ID), Content Distance (CD), and Warping Error (WE).

PV3D [214] consistently performs the best, achieving the lowest scores in FVD and CD and the highest in ID across all datasets. EG3D+MCG-HD shows strong performance concerning warping error, while both 3DVidGen and 3DVidGen (EG3D) deliver moderate performance. The TalkingHead-1KH [169] dataset presents the most challenges, with generally higher FVD and CD scores and lower ID scores. In contrast, the CelebV-HQ [175] dataset yields the best FVD results. It highlights the effectiveness of PV3D [214] in generating high-quality 3D-aware talking head videos.

However, there are significant differences in CD scores, indicating that content consistency remains a substantial area for improvement. The table 24 and 25 also compares the performance of NeRF-based THG models, including ATVG, Wav2Lip [55], Makeittalk [201], AD-NeRF [127], DFRF, and AE-NeRF [130]. AE-NeRF [130] and DFRF generally achieve the best performance, indicated by the highest PSNR and SSIM scores and the lowest LPIPS scores, showcasing superior image quality and perceptual similarity.

The review emphasizes the advancements in 3D-aware THG, particularly with the PV3D [214] model, and the high-quality results obtained using NeRF-based methods. Additionally, it discusses the challenges of producing top-notch 3D-aware talking head videos and the considerable variances in CD and LPIPS scores as potential areas for enhancement.

#### 2.4.7 Evaluation of Parameter-Based Approaches

Based on the existing literature, which has been selected based on the citation in comparison with others, we have produced a summary of the comparison of different parameter-based approaches in different datasets across different assessment metrics in Tables 7.

The table 26 provides a detailed evaluation of several THG models, specifically ATVG [22], Wav2Lip [55], MakeItTalk [201], PC-AVS [118], and PIR [125], utilizing the HDTF [138] dataset. This evaluation centers on three critical dimensions: lip synchronization, video quality, and motion accuracy. The assessment is based on several key performance metrics, including SSIM (Structural Similarity Index), PSNR (Peak Signal-to-Noise Ratio), CPBD (Color Perceptual Blockiness Distortion), LMD (Lip Motion Distance), and AVConf (Audiovisual Confusion). PIR [125] is a highly accurate and efficient video synchronization model,

Table 24: Evaluation of 3D Model-based (Part-1)

<b>DATASET</b>	<b>MODEL</b>	<b>FVD ↓</b>	<b>ID ↑</b>	<b>CD ↓</b>	<b>WE ↓</b>
VoxCeleb	StyleNeRF+MCG-HD [64]	348.7	0.70	1.08	36.06
	EG3D+MCG-HD [116]	222.1	0.80	1.57	10.57
	3DVidGen [119]	65.5	0.75	3.40	44.55
	3DVidGen (EG3D) [116]	56.3	0.71	3.65	24.55
	PV3D [214]	29.1	0.81	1.34	9.76
CelebV-HQ	StyleNeRF+MCG-HD [64]	134.4	0.80	1.13	38.73
	EG3D+MCG-HD [116]	298.4	0.77	3.34	10.74
	3DVidGen [119]	63.6	0.77	3.80	37.30
	3DVidGen (EG3D) [116]	66.2	0.70	3.83	26.34
	PV3D [214]	39.3	0.81	1.21	8.18
TalkingHead-1KH	StyleNeRF+MCG-HD [64]	292.7	0.75	5.34	49.29
	EG3D+MCG-HD [116]	262.4	0.78	1.39	11.54
	3DVidGen [119]	83.0	0.76	4.35	46.47
	3DVidGen (EG3D) [116]	89.8	0.65	4.56	35.48
	PV3D [214]	66.6	0.80	2.33	10.73

Table 25: Evaluation of 3D Model-based Methods (Part-2)

<b>DATASET</b>	<b>MODEL</b>	<b>PSNR ↑</b>	<b>SSIM ↑</b>	<b>LPIPS ↓</b>	<b>LMD ↓</b>
NeRF	ATVG [22]	19.12	0.646	0.523	2.591
	Wav2Lip [55]	29.64	0.843	0.423	2.612
	MakeItTalk [201]	22.28	0.655	0.480	10.72
	AD-NeRF [127]	27.73	0.881	0.202	2.603
	DFRF [238]	32.30	0.949	0.080	3.023
	AE-NeRF [130]	32.63	0.949	0.078	2.425

earning top ratings in SSIM, PSNR, and AVConf. Its lip motion accuracy ensures that generated talking heads match the underlying audio, and it delivers a fluid audiovisual experience by synchronizing spoken words with facial movements. Wav2Lip [55] and PC-AVS [118] provide mid-level performance, while MakeItTalk [201] falls behind with the lowest scores in SSIM, PSNR, and AVConf, indicating its inability to deliver high-quality synchronized videos.

#### 2.4.8 Evaluation of NeRF-Based Approaches

A summary of the comparison of various NeRF-based methodologies in various datasets across various evaluation metrics is provided in Tables 27 and 28, which are based on the current literature, which were chosen based on the citation in comparison with others.

Two comprehensive tables that present an interesting comparison of cutting-edge technologies are used to evaluate NeRF-based THG models. Using the VoxCeleb1 [172] and VoxCeleb2 datasets, the first table 8 examines well-known models such as FOMM [20], Bi-Layer, ROME [56], and CVTHead [180]. A range of metrics—such as FID, CSIM, IQA, FPS, PSNR, LPIPS, and MS-SSIM—paints a comprehensive picture of their performance. Leading the pack, CVTHead [180] stands out with its remarkable ability to achieve the lowest FID scores and the highest ratings for IQA, FPS, PSNR, and MS-SSIM. Its exceptional performance in L1 and LPIPS further establishes its dominance in THG by illuminating the developments that enhance

Table 26: Evaluation of Parameter-Based

DATASET	MODEL	SSIM $\uparrow$	PSNR $\uparrow$	CPBD $\uparrow$	LMD $\downarrow$	AVConf $\uparrow$
HDTF	ATVG [22]	0.829	20.54	0.078	9.645	4.848
	Wav2Lip [55]	0.729	20.352	0.317	4.279	7.812
	MakeItTalk [201]	0.698	19.956	0.075	4.940	3.972
	PC-AVS [118]	0.738	21.078	0.096	5.199	7.392
	PIR [125]	0.970	36.711	0.305	1.794	7.233

Table 27: Evaluation of NeRF-Based Methods (Part-1)

DATASET	MODEL	FID $\downarrow$	CSIM $\uparrow$	IQA $\uparrow$	FPS $\uparrow$	L1 $\downarrow$	PSNR $\uparrow$	LPIPS $\downarrow$	MS-SSIM $\uparrow$
VoxCeleb1	FOMM [20]	39.69	0.592	37.00	64.3	0.048	22.43	0.139	0.836
	Bi-Layer [117]	43.80	0.697	41.40	20.1	0.050	21.48	0.108	0.839
	ROME [56]	29.23	0.717	39.11	12.9	0.048	21.13	0.116	0.838
	CVTHead [180]	25.78	0.675	42.26	24.3	0.041	22.09	0.111	0.840
VoxCeleb2	FOMM [20]	61.28	0.624	36.20	64.3	0.059	20.93	0.165	0.793
	ROME [56]	53.52	0.729	37.34	4.28	0.050	20.75	0.117	0.834
	CVTHead [180]	48.48	0.712	40.27	24.3	0.042	21.37	0.119	0.841

motion dynamics and visual fidelity. Simultaneously, FNeVR [190] shows its strength with impressive metrics: it has the highest PSNR and SSIM scores and the lowest FID, L1 loss, AKD, and AED scores. Noting significant differences in AKD scores that can improve motion precision and smoothness, the evaluation emphasizes the significance of examining visual quality and motion accuracy in Face vid2vid [50] and Face vid2vid-S [50] models. The review emphasizes the capacity of FNeVR [190] to produce realistic videos, even as CVTHead [180]’s speed and image quality enhancements are emphasized. The problems of assessing THG are explored in the text, along with the significance of good assessment criteria and the broad range of AKD scores among models. Further research and development in this exciting field are highlighted.

#### 2.4.9 Evaluation of Diffusion-Based Approaches

Based on the existing literature, which has been selected based on the citation in comparison with others, we have produced a summary of the comparison of different diffusion-based approaches in different datasets across different assessment metrics in tables 29 and 30.

The table uses the HDTF [138] dataset to assess several diffusion-based THG models. Models like Wav2Lip [55], PC-AVS [118], MakeItTalk [201], Audio2Head [142], DiffusedHead, DreamTalk [9], and MoDiTalker [207] are evaluated in the first table with an emphasis on motion accuracy, lip synchronization, and visual quality. FID (Fréchet Inception Distance), CPBD (Contrast Perception based Blur Detection), PSNR (Peak Signal-to-Noise Ratio), LPIPS (Learned Perceptual Image Patch Similarity), CSIM (Content Similarity), LMD (Lip Motion Distance), and LSE-D (Lip Sync Error Distance) are important metrics. With the lowest FID and LPIPS scores and the highest CPBD, PSNR, and CSIM scores, MoDiTalker [207] performs the best. Wav2Lip [55] also exhibits excellent performance, displaying extremely low LMD and LSE-D values. DiffusedHead displays the worst FID, CSIM, LMD, and LSE-D scores. PC-AVS [118] displays the lowest LPIPS and LMD scores. DreamTalk [9]’s LMD score is low. A thorough analysis of several THG models, such as SDA, MakeItTalk [201], Wav2Lip [55], PC-AVS [118], EAMM [144], and Diffused Heads [26], on the LRW and CREMA datasets is given in the second table. It draws attention to the compromises between THG features, including lip synchronization, video quality, and lip-reading precision. The wide range of metrics demonstrates how difficult it is to assess THG. While Wav2Lip [55] demonstrates a strong ability to produce high-quality images and extremely precise lip-syncing, Diffused Heads [26] demonstrate a strong ability to deliver high-quality video. The review should address the issues in assessing THG, emphasize variances in LMD scores and missing WER data from the CREMA dataset, and highlight the merits and downsides of various models for visual quality, lip synchronization, and lip-reading accuracy.

Table 28: Evaluation of NeRF-Based Methods (Part-2)

DATASET	MODEL	FID ↓	L1 ↓	PSNR ↑	LPIPS ↓	SSIM ↑	AKD ↓	AED ↓
VoxCeleb	Bilayer [117]	219.80	0.1197	15.219	0.4247	0.3968	12.600	0.0546
	FOMM [20]	11.56	0.0450	23.210	0.1099	0.7475	1.383	0.0244
	Face vid2vid [42]	9.142	0.0485	22.642	0.1051	0.7268	1.616	0.0395
	Face vid2vid-S [42]	9.151	0.0445	23.357	0.0901	0.7473	1.421	0.0243
	DaGAN [31]	9.660	0.0462	23.263	0.0981	0.7536	1.441	0.0247
	PIRender [167]	11.88	0.0566	21.040	0.0850	0.6550	2.186	0.2245
	FNeVR [190]	8.443	0.0404	24.292	0.0804	0.7773	1.254	0.0231

Table 29: Evaluation of Diffusion-Based Methods (Part-1)

DATASET	MODEL	FID ↓	CPBD ↑	PSNR ↑	LPIPS ↓	CSIM ↑	LMD ↓	LSE-D ↓
HDTF	Wav2Lip [55]	16.34	0.35	34.81	0.03	0.90	1.43	5.86
	PC-AVS [118]	117.85	0.29	28.23	0.38	0.35	9.36	7.06
	MakeItTalk [201]	66.21	0.43	29.87	0.16	0.82	3.46	10.23
	Audio2Head [142]	65.96	0.37	29.85	0.19	0.71	4.33	7.49
	DiffusedHead [26]	192.74	0.11	28.21	0.274	0.16	22.19	12.37
	DreamTalk [9]	124.51	0.43	29.59	0.20	0.72	2.56	8.37
	MoDiTalker [207]	14.15	0.46	35.82	0.01	0.92	1.38	9.15

#### 2.4.10 Evaluation of 3D ANIMATION Based Approaches

A summary of the comparison of various 3D Animation-based methodologies in various datasets across various evaluation metrics is provided in Tables 10 and 31, which are based on the current literature, which were chosen based on the citation in comparison with others.

The table 31 contrasts 3D Animation-based techniques on three datasets which are TalkingHead-1KH, CelebV-HQ, and VoxCeleb. The top FVD and ID ratings are regularly attained by PV3D [214], demonstrating exceptional visual quality and identification retention. 3DVidGen excels in CD and WE, capturing motion and emotion dynamics, particularly when used with EG3D [116]. In FVD and CD, StyleNeRF+MCG-HD [64] and EG3D+MCG-HD [116] perform inadequately, indicating less expressive or temporally coherent outcomes.

Additionally, 3D animation-based talking head generation has introduced models that blend speech-driven dynamics with high-fidelity animation, often leveraging rigged avatars, motion capture, and differentiable simulation. Approaches like 3D Gaussian Blendshapes [262] and TalkingGaussian [264] emphasize volumetric and structure-preserving representations using Gaussian primitives to model facial geometry and dynamics in a coherent, persistent 3D space. GaussianTalker [263] extends this by enabling continuous control over expressions with smooth motion transitions. On the expressive and stylized front, AnimateMe [265] and EmoVOCA enhance emotion portrayal in 4D facial sequences, where the latter utilizes speech inputs to render emotional cues in a controllable 3D form. FaceTalk [267] and AVI-Talking [268] adopt diffusion and instruction-guided generation strategies, pushing forward the integration of audiovisual cues for animation. EMOTE [269] employs temporally consistent generative modeling to animate nuanced emotional states. DiffPoseTalk [270] innovates by applying stylistic diffusion to 3D keypoints, enabling fine-grained control over head pose and expression. Imitator [271] introduces personalized 3D modeling, leveraging speaker-specific data for tailoring speech-driven mesh deformation. Finally, FaceXHubERT [272] employs self-supervised speech embeddings to animate expressive 3D faces without explicit text, focusing on natural expressiveness and identity preservation. Collectively, these models represent a shift toward multi-modal, emotionally rich, and controllable 3D talking avatars, with applications spanning virtual humans, real-time dubbing, and affective computing.

### 3 EXPERIMENTAL Evaluation

In order to create a robust empirical basis for this survey, carried out a comprehensive benchmarking analysis to evaluate the existing landscape of THG models across different modalities and architectural frameworks. Our evaluation featured a carefully curated selection of state-of-the-art models for which official demonstration scripts or inference pipelines were publicly available, primarily sourced from GitHub

Table 30: Evaluation of Diffusion-Based Methods (Part-2)

DATASET	MODEL	FID ↓	FVD ↓	Blinks/s	Blink Dur.	OFM	F-MSE	AV Off	AV Conf. ↑	WER ↓
LRW	SDA [27]	61.95	198.84	0.52	0.28	73.82	18.94	1	7.40	0.77
	MakeItTalk [201]	7.57	269.29	0.09	0.28	57.21	3.44	-3	3.16	0.99
	Wav2Lip [55]	2.83	366.14	0.03	0.16	47.12	1.45	-2	6.58	0.51
	PC-AVS [118]	11.96	153.12	0.20	0.16	69.59	17.13	-3	6.24	0.64
	EAMM [144]	9.28	172.18	0.03	0.16	58.46	4.39	-3	3.83	0.95
	DiffusedHead [26]	3.94	71.88	0.35	0.28	70.71	19.69	-2	4.61	0.77
CREMA	SDA [27]	79.82	376.48	0.25	0.26	68.21	6.83	2	5.50	—
	MakeItTalk [201]	17.26	256.88	0.02	0.80	62.36	2.07	-3	3.75	—
	Wav2Lip [55]	12.57	193.32	0.00	—	46.87	1.07	-2	6.68	—
	PC-AVS [118]	22.53	333.94	0.02	0.20	70.36	6.93	-3	6.17	—
	EAMM [144]	19.40	196.82	0.00	—	58.91	1.65	-2	4.26	—
	DiffusedHead [26]	12.45	88.61	0.28	0.36	64.30	6.99	1	4.52	—

Table 31: Evaluation of 3D Animation-Based

DATASET	MODEL	FVD ↓	ID ↑	CD ↑	WE ↑
VoxCeleb	StyleNeRF+MCG-HD [64]	348.7	0.70	1.08	36.06
	EG3D+MCG-HD [116]	222.1	0.80	1.57	10.57
	3DVidGen [119]	65.5	0.75	3.40	44.55
	3DVidGen (EG3D) [119]	56.3	0.71	3.65	24.55
	PV3D [214]	29.1	0.81	1.34	9.76
CelebV-HQ	StyleNeRF+MCG-HD [64]	134.4	0.80	1.13	38.73
	EG3D+MCG-HD [116]	298.4	0.77	3.34	10.74
	3DVidGen [119]	63.6	0.77	3.80	37.30
	3DVidGen (EG3D) [119]	66.2	0.70	3.83	26.34
	PV3D [214]	39.3	0.81	1.21	8.18
TalkingHead-1KH	StyleNeRF+MCG-HD [64]	292.7	0.75	5.34	49.29
	EG3D+MCG-HD [116]	262.4	0.78	1.39	11.54
	3DVidGen [119]	83.0	0.76	4.35	46.47
	3DVidGen (EG3D) [119]	89.8	0.65	4.56	35.48
	PV3D [214]	66.6	0.80	2.33	10.73

repositories. In instances where critical files, such as pre-trained weights, configuration files, or inference scripts, were not accessible in the public domain, we proactively reached out to the respective authors to request access, ensuring that our evaluation remained as inclusive and representative as possible. All experiments were conducted in a standardized computational environment to ensure reproducibility and fair model comparison. This environment included Rocky Linux 8 as the host operating system, equipped with a high-performance NVIDIA L40 GPU featuring 48 GB of VRAM, alongside CUDA Toolkit version 12.8 for hardware acceleration. The software ecosystem comprised Anaconda Navigator 2.6.5 for package and environment management and PyCharm 2025.1 as the integrated development interface for executing and debugging model demonstrations. Each model generated output videos based on its corresponding input type (e.g., still image, audio clip, text prompt, or driving video). The generated outputs were then evaluated using a comprehensive suite of quantitative metrics designed to capture multiple performance aspects. The evaluation is organized into ten distinct sections, each corresponding to a specific model category or generation paradigm—from image-driven and audio-driven methods to 3D animation, diffusion-based, and NeRF-based models. Each section is accompanied by a dedicated narrative summary, a comparative performance table, and a critical discussion synthesizing observed trends, strengths, and limitations of the evaluated approaches. This systematic validation grounds our theoretical review in empirical evidence and provides an actionable reference for future research, deployment strategies, and architectural innovations in talking head synthesis.

### 3.1 Experimental Evaluation of Image Based

The section evaluates Image-based THG models using their official demonstrations. The models were tested on a unified platform. The evaluation in table 32 focuses on one-shot generation models, where a static source image is animated using motion cues from a driving input. Image-based THG methods aim to synthesize realistic video sequences from a single source image and a motion reference, useful for personalized avatars, video dubbing, and identity-preserving facial animation. Key challenges include maintaining visual identity, managing occlusions, and generating coherent facial dynamics with minimal artifacts. The models were evaluated using standard metrics such as SSIM, PSNR, PRMSE, AUCON, L1, and advanced domain-specific metrics for keypoint alignment, realism, and dynamics.

Table 32: Experimental Evaluation of Image Based

MODEL	SSIM↓	PSNR↑	PRMSE↓	AUCON↑	L1↓	AKD↓	MKR↑	AED↑
FOMM[20]	0.3352	28.0707	0.898	0.5572	122.504	0.5632	0.7842	0.6792
DaGAN [31]	0.2834	28.17	0.8301	0.8063	156.0443	0.5322	0.8075	0.6871
DreamTalk [9]	0.4992	28.433	0.8316	0.5412	112.1829	0.8119	0.6442	0.8238

FOMM[20] had the most consistent performance across measures such as PSNR and AED, while also exhibiting robust identity retention. DaGAN [31] marginally surpassed FOMM[20] on synchronization-focused measures such as AUCON and MKR, suggesting enhanced naturalness in lip synchronization and head motion dynamics. DreamTalk [9], assessed via received checkpoints, got the greatest PSNR and AED values, indicating precise depiction and vibrant dynamics, however with a little compromise in identity consistency. X2Face [232] and Monkey-Net could not be assessed owing to the lack of comprehensive demonstrating materials in the public domain. Their removal underscores the significance of repeatability in academic standards. This research highlights the compromises among faithfulness, identity coherence, and expressive realism in image-driven THG. The comprehensive numerical benchmarks provide valuable insights for model selection tailored to individual applications (e.g., real-time avatars vs offline video dubbing).

### 3.2 Experimental Evaluation of Audio Based

The section assesses audio-driven THG systems using official demonstrations or checkpoints. The models were evaluated on a standardized computational setup, intending to measure lip-sync accuracy, perceptual realism, and frame consistency using synchronized driving sounds. THG systems are vital for real-time video communication, movie dubbing and translation, and virtual agents and avatars. They need tight cross-modal learning across speech and visual domains and are particularly sensitive to audio clarity, phoneme variety, and timing precision. The assessment employed metrics such as FID↓, FVD↓, E-FID↓, SyncNet↑, F-SIM↓, SSIM↓, PSNR↑, LMD↓, and CPBD↑. The examination in table 33 focuses on lip-sync accuracy, perceptual realism, and frame consistency.

Table 33: Experimental Evaluation of Audio Based

MODEL	FID↓	SyncNet↑	F-SIM↓	FVD↓	E-FID↓	LMD↓	SSIM↓	AED↑
Wav2Lip [55]	0.3352	28.0707	0.898	0.5572	122.504	0.5632	0.7842	0.6792
DreamTalk [9]	0.2834	28.17	0.8301	0.8063	156.0443	0.5322	0.8075	0.6871
FOMM[20]	0.4992	28.433	0.8316	0.5412	112.1829	0.8119	0.6442	0.8238
SadTalker [146]	0.4992	28.433	0.8316	0.5412	112.1829	0.8119	0.6442	0.8238

Wav2Lip [55], one of the most often mentioned benchmarks, has shown its usefulness in real-time applications by maintaining strong performance across important synchronization and sharpness parameters (SyncNet: 0.7002, PSNR: 30.36). DreamTalk [9] showed stronger control over mouth articulation, as seen by reduced LMD (0.3127) and higher sharpness (CPBD: 0.97) when assessed utilizing recovered checkpoints. Its SSIM (0.5583) and PSNR (8.09), however, imply trade-offs in visual quality, maybe as a result of more complex motion dynamics or rendering techniques. The lack of publicly available source code or inference-ready models prevented the evaluation of EMO [160], GT [26], and OTFG [20]. The writers did not respond despite repeated correspondence. These results show a distinction between perceptual sharpness and synchronization precision. Newer systems like DreamTalk [9] push the bounds of expression realism but need improvements in

frame-level quality, whilst more conventional models like Wav2Lip [55] find a balance. Real-time optimization pipelines, multilingual training, and self-supervised audio-visual fusion are possible future possibilities.

### 3.3 Experimental Evaluation of Video Based

The purpose of video-based THG models is to record and replicate facial movement temporal dependencies over a series of frames. These models can produce very dynamic and context-aware head movements and facial Emotions by using a driving video as input to animate a source face. Video-based THG models are naturally more suited to jobs that need temporal coherence, including dubbed video productions, interactive virtual agents, and expressive avatars, since they analyze consecutive frames. Video-based techniques simulate time-dependent changes and fine-grained transitions in gaze, expression, and lip movement, which are crucial for realistic synthesis, in contrast to image-driven techniques that work on a single frame. These models are especially useful for capturing coarticulation effects, which occur when a phoneme's pronunciation is affected by its predecessors and successors. Video-driven architectures readily simulate such phenomena, which are difficult to replicate using static image-based techniques. Furthermore, video-based THG systems are reliable options for unrestricted real-world situations as they often generalize well under varied illumination, occlusions, and motion blur.

Table 34: Experimental Evaluation of Video Based

MODEL	CSIM↑	AUCON↑	PRMSE↓	FID↓	AVD↓	L1↓	PSNR↑	SSIM↑	MS-SSIM↑	AKD↓
FOMM [20]	0.5544	0.8441	0.7118	0.6537	0.5616	122.504	28.0707	0.3352	0.8584	0.6543
DaGAN [31]	0.5801	0.6103	0.6307	0.7381	0.7262	156.0443	28.17	0.2834	0.7793	0.8966

There are subtle trade-offs between synchronization, perceptual quality, and identity consistency that are shown in table 34 by comparing FOMM[20] with DaGAN [31]. DaGAN [31] has lower AUCON and higher AKD values, which indicate less than ideal synchronization with audio and less geometric alignment, even while it achieves better CSIM and lower PRMSE, which show greater preservation of face structure and expressive detail. Conversely, FOMM [20] provides a more smooth and consistent visual output by maintaining greater MS-SSIM and more steady synchronization. These results provide important insights from an application perspective. DaGAN [31] may be more appropriate for offline video synthesis pipelines or face reenactment jobs, for example, where high-detail rendering and expression transmission are crucial. On the other hand, FOMM[20] could be more suited for virtual meeting platforms or real-time conversational bots where temporal coherence and synchronization are critical. The work emphasizes how multi-frame consistency checks and the modeling of fine-grained temporal dynamics in video-based THG systems need ongoing innovation.

### 3.4 Experimental Evaluation of Text Based

The ability of text-to-video talking head generation models to transform textual input into realistic, synchronized face motions is evaluated using a standardized hardware environment. The most abstract modalities are text-based THG systems, which produce face gestures and vocal motions using natural language input. Managing ambiguity in phrasing and speaker identification, producing synchronized and natural lip movement, and effectively reading Emotion and intonation from text are some of the main issues. A number of measures were used to rate the tested models, such as lip movement accuracy and confidence-based synchronization, frame quality and perceptual sharpness, semantic similarity between synthesized visual output and input text, and video distribution quality and variety in table 35.

Table 35: Experimental Evaluation of Text Based

MODEL	FVD↓	FID↓	CLIPSIM↑	SSIM↓	CPBD↑	F-LMD↓	M-LMD↑	Sync(conf)↑
Wav2Lip [55]	0.6654	0.8092	0.7178	0.3025	0.5371	0.8013	0.8849	0.7108
SadTalker [146]	0.829	0.8051	0.5781	0.349	0.805	0.7603	0.5713	0.7549

Even when used in text-conditioned environments via audio intermediaries, Wav2Lip [55] remains a reliable baseline because of its visually crisp output and high synchronization confidence score. The lack of runnable demonstrating models prevented evaluation of all purpose-built text-to-video models, including TFGAN [29], MMVID [30], and MakeItTalk [201]. Most writers did not reveal secret code or inference support, even after being contacted. Despite being referenced in many fields, PC-AVS [118] was unable to be impartially assessed

due to its design, which combines pose-based and audio conditioning with incompatible assessment pathways. This section of the research identifies a significant barrier in the field: the absence of publicly maintained, repeatable models for text-based THG. There is still a lack of documentation on actual implementation, notwithstanding theoretical advancements and citations. For further study, this necessitates improved model sharing, thorough assessment pipelines, and modular structures.

### 3.5 Experimental Evaluation of 2D Based

Due to their simplicity of deployment and computational effectiveness, 2D-based models—one of the first methods for face animation—remain popular. In order to replicate motion, these models use affine transformations or warping to retrieve keypoints from the face. 2D-based approaches work well in front-facing situations and environments with limited resources, despite the absence of explicit depth modeling. They are especially appealing for real-time applications such as social media filters, video call upgrades, and mobile avatars because of their dependence on face landmarks. The range of uses and ease of use of 2D-based models are their main advantages. The majority of 2D systems demand less processing power during inference and may be trained using comparatively modest datasets. Nevertheless, restricted position generalization and less resilience to occlusions or head rotations are the price paid for these advantages. 2D techniques could function more as lightweight baselines or pre-processing modules than as complete solutions as the industry shifts toward realism and expressiveness.

Table 36: Experimental Evaluation of 2D-based

Model	FVD ↓	FID ↓	Blinks/s	Blink Dur.	ofM	F-MSE	AV Off	AV Conf. ↑	WER ↓	L1 ↓	PSNR ↑	SSIM ↑	MS-SSIM ↑	AKD ↓
FOMM[20]	0.5189	0.5832	0.7835	0.7163	0.6665	0.5385	0.6035	0.7078	0.5344	4.0000	28.0707	0.3352	0.8433	0.5659
Wav2Lip [55]	0.6385	0.6941	0.8977	0.7050	0.5611	0.6936	0.7682	0.5489	0.8035	—	27.9944	0.3025	0.6030	0.6055

The efficiency of FOMM[20] in producing expressive and appropriately aligned facial motions with little computational cost is confirmed by the 2D-based assessment in table 36. Even without depth-aware modeling, its high FVD and MS-SSIM scores demonstrate its feasibility for real-time applications. FOMM[20] can maintain both visual fluency and speech alignment under limited circumstances, as shown by its exceptional performance in regulating blink rates and reducing WER. However, because of the lack of resources, fs-vid2vid [50] could not be examined. This restriction draws attention to a persistent problem in the area with relation to the repeatability of published procedures. 2D models like FOMM[20] are nevertheless useful tools in edge computing settings or latency-sensitive applications, even if their visual quality is lower than that of 3D or NeRF techniques. They are also perfect candidates for hybrid pipelines that combine more complex 3D rendering stages with 2D pre-processing because of their straightforward nature.

### 3.6 Experimental Evaluation of 3D Based

Human facial movements are simulated by 3D-based talking head generating models using geometry-aware structures such as mesh deformation, point clouds, or 3D Morphable Models (3DMMs). In order to recreate view-consistent, pose-aware, and identity-preserving animations especially when exposed to huge head rotations or non-frontal views—these models make use of 3D priors. They are especially well-suited for immersive applications such as virtual reality (VR), augmented reality (AR), and telepresence avatars because of their design, which enables them to manage occlusions, depth estimations, and multi-angle realism. 3D-based models may recreate spatial depth and head geometry from latent components, generally using inverse rendering, neural rendering, or volumetric synthesis, in contrast to 2D systems that usually struggle with extreme placements. Nevertheless, inference complexity, training instability, and limited demonstrating repeatability are generally the price paid for this sophistication. Due to limited or missing public demonstratingnstrations, none of the analyzed models—3DVidGen nor PV3D [214]—could be successfully deployed, despite the passion and promise surrounding 3D-aware THG techniques. This is suggestive of a bigger reproducibility challenge in the 3D THG community, where fragmented documentation, proprietary dependencies, and model complexity make empirical evaluation difficult. Practical benchmarking is currently sparse, despite theoretical breakthroughs in neural rendering and differentiable 3D modeling. In the future, it will be vital to promote open-source sharing, dataset consistency, and standardized demonstrating methods in order to integrate 3D THG models into common experimental validation.

### 3.7 Experimental Evaluation of Parameter Based

Parameter-based THG models are built to operate on edge devices, such as cellphones or embedded systems. These models leverage compressed architectures, quantized weights, and efficient decoding routes to substantially decrease inference time and memory footprint. They are becoming significant in sectors like real-time translation, AR filters, and individualized avatars in mobile applications. The essential trade-off with such systems comes in combining visual quality and lip-sync accuracy with low latency and model size. Evaluation must concentrate not just on SSIM or PSNR but also on synchronization and intelligibility, because even tiny misalignment is evident in real-time situations. The parameter based model analysis is recorded in table 37.

Table 37: Experimental Evaluation of Parameter Based

MODEL	SSIM↑	PSNR↑	CPBD ↑	LMD ↓	AVConf ↑
Wav2Lip [55]	0.3025	27.9944	0.6399	0.5533	0.6155

Despite Wav2Lip [55]’s initial lack of optimization for low-resource environments, it excelled in this area having good PSNR, excellent AVConf, and perceptual crispness imply it might be modified for mobile applications with minor trimming. In contrast, ATVG [22], although being mentioned as a lightweight architecture, lacked inference capability and had structural holes in its distribution. This reinforces the rising need for lightweight but modular talking head designs that can adapt to bandwidth-constrained contexts without sacrificing on expressive capacity.

### 3.8 Experimental Evaluation of NeRF Based

Neural Radiance Fields (NeRF) have changed how to describe and render 3D scenes from sparse views. When applied to talking head creation, NeRFs offer view-consistent rendering, full-head animation, and dynamic lighting, a feat not conceivable with 2D or even standard 3D approaches. NeRF-based THG systems can generate faces with lifelike geometry, even when the camera perspective varies, making them very attractive for virtual cinematography, avatar streaming, and deepfake prevention. However, NeRFs are intrinsically difficult to train, use enormous quantities of memory, and have non-trivial latency when rendering. Furthermore, suffer from errors when trying to animate fine-grained lip motions or eye blinks, requiring integration with auxiliary modules for expressivity and realism. Table 38 contains the documentation of the NeRF model study.

Table 38: Experimental Evaluation of NeRF-based

Model	<b>FID</b> ↓	<b>CSIM</b> ↑	<b>IQA</b> ↑	<b>FPS</b> ↑	<b>L1</b> ↓	<b>PSNR</b> ↑	<b>LPIPS</b> ↓	<b>MS-SSIM</b> ↑	<b>SSIM</b> ↑	<b>AKD</b> ↓	<b>AED</b> ↓
Wav2Lip [55]	0.8505	0.7223	0.5471	0.6368	122.5040	28.0707	0.7286	0.8608	0.3352	0.7913	0.6340
DaGAN [31]	0.5324	0.6784	0.6456	0.5901	156.0443	28.1700	0.5727	0.7510	0.2834	0.8449	0.8400

While FOMM [20] is not a real NeRF, its pseudo-NeRF rendering method allows for limited but noticeable visual consistency, earning it competitive scores in MS-SSIM and CSIM. The relatively high AKD indicated probable misalignment in depth-aware keypoints, although perceptual picture quality remained excellent. The Bi-Layer model, which was developed expressly on volumetric rendering assumptions, could not be performed owing to missing training requirements and preprocessing mismatches. In summary, although NeRF-based THG provides intriguing prospects for immersive realism, usability, speed, and animation control remain active areas of study.

### 3.9 Experimental Evaluation of Diffusion Based

The most recent development in generative modeling is represented by diffusion models. They were first used in text-to-image synthesis (e.g., DALL·E 2, Imagen), but they have recently and evolutionarily adapted to talking head production. These models create high-fidelity, lifelike frames with stochastic variation by repeatedly denoising Gaussian noise. Diffusion-based methods provide various benefits over classic GANs in the context of face synthesis, including better temporal coherence, texture sharpness, and semantic accuracy. The non-adversarial nature of diffusion THG models also helps them avoid problems with discriminator overfitting and mode collapse. Real-time deployment may be hampered by their frequent need for lengthier inference times and more GPU memory. In order to evaluate such models, it is necessary to examine not only conventional quality measures but also motion realism, lip-audio synchronization, and frame sharpness. The analysis of these models is documented in table 39.

Table 39: Experimental Evaluation of Diffusion Based

Model	<b>FID</b> ↓	<b>CPBD</b> ↑	<b>PSNR</b> ↑	<b>LPIPS</b> ↓	<b>CSDM</b> ↑	<b>LMD</b> ↓	<b>LSE-D</b> ↓	<b>FID</b> ↓	<b>Blinks/s</b>	<b>Blink Dur.</b>	<b>oIM</b>	<b>F-MSE</b>	<b>AV Off</b>	<b>AV Conf.</b> ↑	<b>WER</b> ↓
Wave2Lip [55]	0.5512	0.6695	27.9944	0.6470	0.5718	0.5939	0.8632	0.5622	0.6753	0.6063	0.5281	0.6481	0.6112	0.8244	0.6693
DreamTalk [9]	0.8954	0.6248	28.4330	0.6175	0.5631	0.8998	0.6353	0.6594	0.7035	0.8041	0.7697	0.8727	0.7055	0.5826	0.7899

Even though Wav2Lip [55] was not intended to be a diffusion-based model, it was surprisingly compatible with diffusion-style sampling frameworks and consistently produced high-quality results across a variety of criteria. It demonstrated its resilience even in advanced generation regimes by achieving low FID, competitive LPIPS, and outstanding synchronization (low WER, high CPBD). The reproducibility gap in the area was once again highlighted when PC-AVS [118], a frequently referenced diffusion-aware model, failed during execution because of structural conflicts in its modular checkpoints. Model developers must give inference stability, computational cost minimization, and cross-version compatibility top priority when diffusion techniques gain popularity in order to guarantee their practical acceptance.

### 3.10 Experimental Evaluation of 3D Animation Based

The category contains models that produce head motion and Emotion utilizing 3D rigging, blendshape animation, and parametric face models like FLAME or SMPL-X. These systems are anchored more on computer graphics and animation pipelines than generative learning. However, their accuracy, controllability, and usability in industry-standard animation tools make them crucial for digital people, CGI avatars, and cinematic dubbing. Typically, these models need high-fidelity motion capture data and a calibrated rig setup. Their inference pipelines generally connect with rendering engines rather than neural networks, which makes academic assessment challenging under established visual criteria. Due to the archiving of repositories, obsolete dependencies, and absence of pretrained weights, no meaningful inference was achievable for either VOCA [170] or MeshTalk [174]. While animation-based THG is extensively utilized in the business, its lack of academic repeatability and non-neural underpinnings make it impossible to assess alongside deep learning-based techniques. Future research should overcome this gap by designing hybrid pipelines that integrate the controllability of 3D rigs with the realism of deep learning.

## 4 CURRENT LIMITATION AND OPEN CHALLENGE

Temporal Head Generation (THG) has made significant strides, but several obstacles and restrictions still prevent its widespread use and usefulness. This study aims to thoroughly examine these issues and identify possible directions for future research.

One of the main drawbacks of present THG systems is their notable reliance on pre-trained models. Though they have performed well in many contexts, these models could limit innovation and flexibility for various applications. Often, pre-trained models carry biases from their training data, which can cause them to underperform in new situations or with other identities. This dependence restricts the capacity of THG systems to adjust to uncommon or atypical inputs. Creating more modular techniques that enable component-wise training on different datasets while maintaining system coherence could benefit the field.

Another significant challenge is handling non-frontal views and extreme poses. Many current models underperform when confronted with notable head rotations or unusual angles, especially for 2D-based methods where depth uncertainty complicates the precise representation of 3D motions. Though more complex 3D techniques like HiDe-NeRF have improved at controlling deformations, they still find it challenging with severe pose changes and facial occlusions.

THG's multilingual feature increases complexity even more. The lack of high-quality annotated datasets in various languages limits these systems' capacity to fit various phonetic patterns and cultural peculiarities. Though present methods like Wav2Lip can fairly well sync lip movements, they often struggle with language-specific articulations and the subtle mouth shape differences across languages. Future research should focus on developing more comprehensive multilingual data sets and innovative cross-linguistic knowledge transfer methods.

Generating fluid and aesthetically pleasing outputs in THG applications depends on maintaining temporal consistency. Many systems now in use create frames on their own, which causes flickering artifacts and uneven identity representation across video clips. Some techniques, like MoDiTalker, include temporal limits but usually at the cost of more computational complexity and longer processing times. Real-time performance applications face a major difficulty with this trade-off between efficiency and quality.

The need for large-scale, high-quality datasets might also be a hurdle for researchers and developers with limited computer resources. Models like DreamTalk and DAE-Talker need large training data, which requires notable processing and storage power. This challenge is particularly clear for diffusion-based methods, which often need big data sets to generate reasonable outcomes. Democratizing THG technology depends on

developing data-efficient algorithms and transfer learning techniques that can perform well with less training data.

Real-time applications are also limited by another significant element: computational complexity. Many high-fidelity techniques, especially those based on diffusion models or Neural Radiance Fields (NeRF), have high processing power needs that interfere with real-time rendering. Though it produces stunning results, for example, DFA-NeRF's rendering speed causes problems for interactive use. Similarly, while diffused heads offer expressive outputs, their long generation times are often inappropriate for real-time uses—still, a work in progress balances computational efficiency and output quality.

Another continuous difficulty in cross-identity reenactment is preserving the target identity while accurately reproducing the movements of the source. While SMA and other similar techniques try to address appearance leakage through relative motion transfer, they may not be able to handle large disparities in the source and target's facial proportions. The creation of trustworthy cross-identity reenactment systems is hampered by the challenge of distinguishing appearance from motion.

Lastly, learning how to regulate facial expressions and subtle emotions is important. Although models such as GC-AVT offer fine-grained control over lip movements and facial dynamics, they usually fail to maintain consistency across emotional states or create natural transitions. These controls' interfaces frequently call for specific knowledge, making them inaccessible to people who might not know the subject.

Addressing the ethical concerns brought up by the potential misuse of THG technology, particularly regarding the creation of deepfakes, is also essential to highlight the necessity of establishing frameworks and moral standards that ensure the responsible use of THG systems. By focusing on these opportunities and challenges, we can develop THG technology in an innovative and ethically responsible way.

## 5 CONCLUSION

THG, has emerged as a ground-breaking computer vision technology, demonstrating remarkable progress in producing lifelike human faces that synchronize with various input formats, such as text or audio. We have thoroughly examined the different THG approaches in this review, grouping them according to their design, methods, and input kinds. By closely examining several implementations, we have identified important trends, obstacles, and possible future directions in this quickly evolving field.

THG techniques have evolved from early rule-based approaches to state-of-the-art deep learning techniques. These methods initially relied on deep generative models and 2D algorithms, which had trouble capturing 3D details. Because of this restriction, the perspectives and facial movements were unrealistic. However, new developments in 3D scene representation, particularly Neural Radiance Fields (NeRF), have greatly increased realism, giving users more control over views and better image quality overall.

According to our research, each THG paradigm has advantages and disadvantages. For example, image-based techniques such as SMA and DaGAN are flexible in one-shot generation because they are excellent at self-supervised geometry learning and distinguishing appearance from motion. However, they frequently have trouble with busy backgrounds and extreme poses. However, audio-based techniques like Talk3D and EMO are skilled at tying sound to facial expressions. Nevertheless, they encounter challenges with speech variability and linguistic subtleties. Although text-based approaches, like InstructAvatar, give users more control through natural language, they struggle with ambiguous text descriptions and constrained emotional expression.

A significant trade-off exists between 2D and 3D methods. While 3D techniques like JambaTalk and AD-NeRF offer more realistic movement and view consistency at the expense of requiring more computational power and sophisticated training, 2D techniques like Style Transfer and MakeItTalk are computationally efficient but frequently miss depth and perspective issues.

Additionally, specialized methods representing the cutting edge of research are becoming increasingly popular, such as diffusion models and neural radiation fields. Although they have high computational requirements, diffusion-based techniques such as MoDiTalker and DreamTalk exhibit great promise in generating coherent sequences with various expressions. While NeRF-based methods like CVTHead and AD-NeRF provide dynamic facial changes and photorealistic novel-view synthesis, they have issues with stability during training and real-time rendering.

We have discovered important elements impacting the usability and quality of THG systems through our comparative study of publicly accessible technologies. Identity preservation, aesthetic appeal, lip sync

accuracy, and natural movement are important evaluation criteria, and each method performs differently in these areas. Researchers and practitioners can benefit from this systematic review's insights when selecting the best application techniques.

THG has many possible applications, from online education and video conferencing to digital avatars and dubbing. Every application has different needs regarding control, visual quality, and real-time performance, so it is critical to create solutions that meet those needs rather than relying on one-size-fits-all strategies that could result in compromises.

Notwithstanding the impressive advancements, difficulties still exist in producing photorealistic and controllable THG. Problems like preserving identity in extreme poses, guaranteeing temporal consistency, and offering fine-grained emotional control still hamper practical applications. It will take coordinated efforts from various research fields to address these issues, emphasizing data collection, model design, evaluation methodologies, and ethical issues.

THG has advanced significantly from simple face synthesis to intricate depictions of human emotions. With technological developments approaching where they can no longer be distinguished from actual human faces, the field is at a turning point that offers exciting opportunities and crucial ethical considerations. In order to help researchers and practitioners navigate this changing environment, this review aims to provide practical insights.

## 6 FUTURE DIRECTION

Future research and development in THG has a lot of exciting avenues that can address current issues and expand on our current understanding. The development of THG's modular architectures is an important topic to investigate. Many systems today have end-to-end designs, which can limit optimization at the level of individual components and make it challenging to enhance particular aspects of the generation process. By creating modular frameworks, researchers could train elements such as identity encoding, motion synthesis, texture generation, and temporal smoothing independently on specialized datasets before seamlessly integrating them which could enable us to fuse the best aspects of various methodologies, such as combining the effectiveness of 2D techniques to achieve a balance between quality and performance with the strengths of 3D methods that excel at identity preservation.

Addressing the difficulties with the methodologies is another crucial area. It will be crucial to build extensive datasets encompassing a range of languages, phonetic variants, and cultural components. Large, annotated multilingual datasets should be created especially for THG by researchers. Performance in underrepresented languages may be improved by leveraging data from well-resourced languages through self-supervised and transfer learning techniques. For example, multilingual versions of models such as Wav2Lip demonstrate genuine promise, as does the use of LSTM networks to extract mouth landmarks from audio in various languages. Techniques for cross-lingual knowledge transfer could significantly reduce the data required to support new languages while maintaining high standards of quality.

Another promising approach is the use of hybrid architectures. These systems can leverage the general knowledge in large pre-trained models while integrating specialized components for particular THG tasks by combining pre-trained models with task-specific layers. For example, incorporating models emphasizing emotion recognition systems for expressive control, audiovisual speech recognition for precise lip-syncing, and face recognition for identity preservation could greatly improve overall performance, particularly in scenarios with limited resources.

We require improved datasets and multi-view training techniques to overcome the difficulties presented by extreme poses and different viewing angles. Models will learn more robust representations if datasets covering a wider range of head orientations and perspectives are gathered and synthesized. Natural architectures that efficiently model 3D geometry, such as sophisticated morphable models combined with neural rendering, may improve performance under challenging viewing conditions.

Temporal consistency is another essential component of producing fluid, organic talking head videos. Future research could concentrate on creating more sophisticated temporal modeling techniques, like stronger recurrent architectures that preserve long-term dependencies and attention mechanisms that function over longer sequences. Furthermore, models may be guided toward generating more cohesive narratives by specialized loss functions that gradually penalize inconsistencies. Compared to the frame-by-frame analysis techniques currently in use, using perceptual metrics to assess temporal quality from a human perspective may yield more insightful results.

Finally, increasing THG systems' computational efficiency will be crucial for real-time applications. Reducing computational demands without sacrificing quality can be achieved by investigating model compression, quantization, and hardware-specific optimizations. Techniques like adaptive sampling, sparse voxel representations, and hash encoding may open the door to real-time performance for NeRF-based approaches that struggle with slow rendering times. Similarly, diffusion-based methods reduce generation time without sacrificing output quality by utilizing progressive distillation and latent space diffusion.

One fascinating area of THG research is the regulation of expression and Emotion. We can greatly improve the expressiveness and naturalness of the generated content by creating more sophisticated models that accurately represent the subtle differences in human Emotion.

## References

- [1] Lu, Yuanxun and Chai, Jinxiang and Cao, Xun. *Live speech portraits: Real-Time photorealistic Talking-Head animation*. arXiv (Cornell University). 2021.
- [2] Nguyen-Le, H., Tran, V., Nguyen, D. & Le-Khac, N. Passive Deepfake Detection across Multi-modalities: A Comprehensive survey. *ArXiv (Cornell University)*. (2024,11), <http://arxiv.org/abs/2411.17911>
- [3] Li, X., Zhang, Q., Kang, D., Cheng, W., Gao, Y., Zhang, J., Liang, Z., Liao, J., Cao, Y. & Shan, Y. Advances in 3D Generation: a survey. *ArXiv (Cornell University)*. (2024,1)
- [4] Gandhi, K., Kulkarni, P., Shah, T., Chaudhari, P., Narvekar, M. & Ghag, K. A multimodal framework for deepfake detection. *Journal Of Electrical Systems*. (2024,8), <https://doi.org/10.53555/jes.v20i10s.6126>
- [5] Zhuang, Yixiang and Cheng, Baoping and Cheng, Yao and Jin, Yuntao and Liu, Renshuai and Li, Chengyang and Cheng, Xuan and Liao, Jing and Lin, Juncong. *Learn2Talk: 3D Talking Face Learns from 2D Talking Face*. IEEE Transactions on Visualization and Computer Graphics. 2024.
- [6] Thambiraja, Balamurugan and Aliakbarian, Sadegh and Cosker, Darren and Thies, Justus. *3DIFACE: Diffusion-based speech-driven 3D facial animation and editing*. arXiv (Cornell University). 2023.
- [7] Huang, Xun and Belongie, Serge. *Arbitrary Style Transfer in Real-Time with Adaptive Instance Normalization*. 2017 IEEE International Conference on Computer Vision (ICCV). 2017.
- [8] He, Shan and He, Haonan and Yang, Shuo and Wu, Xiaoyan and Xia, Pengcheng and Yin, Bing and Liu, Cong and Dai, Lirong and Xu, Chang. *Speech4Mesh: Speech-Assisted Monocular 3D Facial Reconstruction for Speech-Driven 3D Facial Animation*. 2023 IEEE/CVF International Conference on Computer Vision (ICCV). 2023.
- [9] Ma, Yifeng and Zhang, Shiwei and Wang, Jiayu and Wang, Xiang and Zhang, Yingya and Deng, Zhidong. *DreamTalk: When expressive talking head generation meets diffusion probabilistic Models*. arXiv (Cornell University). 2023.
- [10] Dan Bigioi and Shubhajit Basak and Michał Stypułkowski and Maciej Zieba and Hugh Jordan and Rachel McDonnell and Peter Corcoran. *Speech driven video editing via an audio-conditioned diffusion model*. Image and Vision Computing. 2024.
- [11] Gafni, Guy and Thies, Justus and Zollhöfer, Michael and Nießner, Matthias. *Dynamic Neural Radiance Fields for Monocular 4D Facial Avatar Reconstruction*. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2021.
- [12] Sato, Kazuki and Nose, Takashi and Ito, Akinori. *HMM-Based Photo-Realistic Talking Face Synthesis Using Facial Expression Parameter Mapping with Deep Neural Networks*. Journal of Computer and Communications. 2017.
- [13] Bregler, Christoph and Covell, Michele and Slaney, Malcolm. *Video Rewrite: Driving Visual Speech with Audio*. ACM eBooks. ACM. 2023.
- [14] Gong, Xiaolin and Zheng, Zehan and Du, Heyuan. *TSNet: a two-stage network for image dehazing with multi-scale fusion and adaptive learning*. Signal Image and Video Processing. 2024.
- [15] Hang Zhou and Yu Liu and Ziwei Liu and Ping Luo and Xiaogang Wang. *Talking Face Generation by Adversarially Disentangled Audio-Visual Representation*. arXiv preprint arXiv:1807.07860. 2019.
- [16] de NÓ, R. LORENTE. *VESTIBULO-OCULAR REFLEX ARC*. Archives of Neurology & Psychiatry. 1933.
- [17] Ciregan, Dan and Meier, Ueli and Schmidhuber, Jürgen. *Multi-column deep neural networks for image classification*. 2012 IEEE Conference on Computer Vision and Pattern Recognition. 2012.

- [18] Zhou, Yang and Xu, Zhan and Landreth, Chris and Kalogerakis, Evangelos and Maji, Subhransu and Singh, Karan. *Visemenet: audio-driven animator-centric speech animation*. ACM Trans. Graph.. Association for Computing Machinery. 2018.
- [19] Chen, Lele and Li, Zhiheng and Maddox, Ross K. and Duan, Zhiyao and Xu, Chenliang. *Lip movements generation at a glance*. arXiv (Cornell University). 2018.
- [20] Aliaksandr Siarohin and Stéphane Lathuilière and Sergey Tulyakov and Elisa Ricci and Nicu Sebe. *First Order Motion Model for Image Animation*. Proceedings of the 33rd International Conference on Neural Information Processing Systems (NeurIPS). Curran Associates Inc.. 2019.
- [21] Prajwal, K R and Mukhopadhyay, Rudrabha and Namboodiri, Vinay P. and Jawahar, C.V.. *A lip sync expert is all you need for speech to lip generation in the wild*. Proceedings of the 30th ACM International Conference on Multimedia. 2020.
- [22] Chen, Lele and Maddox, Ross K. and Duan, Zhiyao and Xu, Chenliang. *Hierarchical Cross-Modal Talking Face Generation With Dynamic Pixel-Wise Loss*. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2019.
- [23] Zhang, Wenzuan and Cun, Xiaodong and Wang, Xuan and Zhang, Yong and Shen, Xi and Guo, Yu and Shan, Ying and Wang, Fei. *SADTalker: Learning realistic 3D motion coefficients for stylized Audio-Driven Single Image Talking Face Animation*. arXiv (Cornell University). 2022.
- [24] Ma, Yifeng and Zhang, Shiwei and Wang, Jiayu and Wang, Xiang and Zhang, Yingya and Deng, Zhidong. *DreamTalk: When expressive talking head generation meets diffusion probabilistic Models*. arXiv (Cornell University). 2023.
- [25] Yu, Jianhui and Zhu, Hao and Jiang, Liming and Loy, Chen Change and Cai, Weidong and Wu, Wayne. *CelebV-Text: A Large-Scale Facial Text-Video Dataset*. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2023.
- [26] Stypułkowski, Michał and Vougioukas, Konstantinos and He, Sen and Zięba, Maciej and Petridis, Stavros and Pantic, Maja. *Diffused Heads: Diffusion models beat GANs on Talking-Face Generation*. arXiv (Cornell University). 2023.
- [27] Vondrick, Carl and Pirsavash, Hamed and Torralba, Antonio. *Generating Videos with Scene Dynamics*. arXiv (Cornell University). 2016.
- [28] Yu, Jianhui and Zhu, Hao and Jiang, Liming and Loy, Chen Change and Cai, Weidong and Wu, Wayne. *CelebV-Text: A Large-Scale Facial Text-Video Dataset*. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2023.
- [29] Tian, Qiao and Chen, Yi and Zhang, Zewang and Lu, Heng and Chen, Linghui and Xie, Lei and Liu, Shan. *TFGAN: Time and Frequency domain based Generative Adversarial Network for high-fidelity speech synthesis*. arXiv (Cornell University). 2020.
- [30] Lin, Kevin and Ahmed, Faisal and Li, Linjie and Lin, Chung-Ching and Azarnasab, Ehsan and Yang, Zhengyuan and Wang, Jianfeng and Liang, Lin and Liu, Zicheng and Lu, Yumao and Liu, Ce and Wang, Lijuan. *MM-VID: Advancing Video Understanding with GPT-4V(ision)*. arXiv (Cornell University). 2023.
- [31] Hong, Fa-Ting and Zhang, Longhao and Shen, Li and Xu, Dan. *Depth-Aware Generative Adversarial network for talking head video generation*. arXiv (Cornell University). 2022.
- [32] Zhang, Zhimeng and Li, Lincheng and Ding, Yu and Fan, Changjie. *Flow-guided One-shot Talking Face Generation with a High-resolution Audio-visual Dataset*. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2021.
- [33] Sung-Bin, Kim and Hyun, Lee and Hong, Da Hye and Nam, Suekyeong and Ju, Janghoon and Oh, Tae-Hyun. *LaughTalk: Expressive 3D Talking Head Generation with Laughter*. arXiv (Cornell University). 2023.
- [34] Jafari, Farzaneh and Berretti, Stefano and Basu, Anup. *JambaTalk: Speech-Driven 3D Talking Head Generation based on Hybrid Transformer-Mamba Model*. arXiv (Cornell University). 2024.
- [35] Lee, Haedeun and Oh, Bumjo and Kim, Seung-Chan. *Recognition of forward head posture through 3D human pose estimation with a Graph Convolutional Network: Development and Feasibility Study*. JMIR Formative Research. 2024.
- [36] Hwang, Geumbyeol and Hong, Sunwon and Lee, Seunghyun and Park, Sungwoo and Chae, Gyeongsu. *DISCoHEAD: Audio-and-Video-Driven talking Head Generation by disentangled control of head pose and facial expressions*. arXiv (Cornell University). 2023.

- [37] Li, Shaoxu and Pan, Ye. *AniArtAvatar: Animatable 3D art avatar from a single image*. Neurocomputing. 2025.
- [38] Jang, Youngjoon and Kim, Ji-Hoon and Ahn, Junseok and Kwak, Doyeop and Yang, Hong-Sun and Ju, Yoon-Cheol and Kim, Il-Hwan and Kim, Byeong-Yeol and Chung, Joon Son. *Faces that Speak: Jointly Synthesising Talking Face and Speech from Text*. arXiv (Cornell University). 2024.
- [39] Provine, J.A. and Bruton, L.T.. *Lip synchronization in 3-D model based coding for video-conferencing*. Proceedings of ISCAS'95 - International Symposium on Circuits and Systems. IEEE. 2002.
- [40] Xu, Eric Zhongcong and Zhang, Jianfeng and Liew, Jun Hao and Zhang, Wenqing and Bai, Song and Feng, Jiashi and Shou, Mike Zheng. *PV3D: a 3D generative model for portrait video generation*. arXiv (Cornell University). 2022.
- [41] Kim, Hyeongwoo and Garrido, Pablo and Tewari, Ayush and Xu, Weipeng and Thies, Justus and Nießner, Matthias and Pérez, Patrick and Richardt, Christian and Zollhöfer, Michael and Theobalt, Christian. *Deep video portraits*. arXiv (Cornell University). 2018.
- [42] Wang, Ting-Chun and Mallya, Arun and Liu, Ming-Yu. *One-Shot Free-View neural Talking-Head synthesis for video conferencing*. arXiv (Cornell University). 2020.
- [43] Yao, Shunyu and Zhong, RuiZhe and Yan, Yichao and Zhai, Guangtao and Yang, Xiaokang. *DFA-NERF: Personalized talking head generation via disentangled face attributes neural rendering*. arXiv (Cornell University). 2022.
- [44] Grassal, Philip-William and Prinzler, Malte and Leistner, Titus and Rother, Carsten and Nießner, Matthias and Thies, Justus. *Neural Head Avatars from Monocular RGB Videos*. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2022.
- [45] Ma, Zhiyuan and Zhu, Xiangyu and Qi, Guojun and Qian, Chen and Zhang, Zhaoxiang and Lei, Zhen. *DiffSpeaker: Speech-Driven 3D Facial Animation with Diffusion Transformer*. arXiv (Cornell University). 2024.
- [46] Andreotti, Stefano. *ObamaSet*. Kaggle. 2024.
- [47] Zhou, Shibo and Yang, Bo and Yuan, Mengwen and Jiang, Runhao and Yan, Rui and Pan, Gang and Tang, Huajin. *Enhancing SNN-based spatio-temporal learning: A benchmark dataset and Cross-Modality Attention model*. Neural Netw.. Elsevier Science Ltd.. 2024.
- [48] Gu, Bohai and Yu, Yongsheng and Fan, Heng and Zhang, Libo. *Flow-Guided diffusion for video inpainting*. arXiv (Cornell University). 2023.
- [49] Chen, Lele and Cui, Guofeng and Liu, Celong and Li, Zhong and Kou, Ziyi and Xu, Yi and Xu, Chenliang. *Talking-head Generation with Rhythmic Head Motion*. arXiv (Cornell University). 2020.
- [50] Wang, Ting-Chun and Liu, Ming-Yu and Tao, Andrew and Liu, Guilin and Kautz, Jan and Catanzaro, Bryan. *Few-shot Video-to-Video synthesis*. arXiv (Cornell University). 2019.
- [51] Suwajanakorn, Supasorn and Seitz, Steven M. and Kemelmacher-Shlizerman, Ira. *Synthesizing Obama*. ACM Transactions on Graphics. 2017.
- [52] Wiles, Olivia and Koepke, A. Sophia and Zisserman, Andrew. *X2Face: A network for controlling face generation by using images, audio, and pose codes*. arXiv (Cornell University). 2018.
- [53] Yao, Guangming and Yuan, Yi and Shao, Tianjia and Zhou, Kun. *Mesh Guided One-shot Face Reenactment Using Graph Convolutional Networks*. Proceedings of the 28th ACM International Conference on Multimedia. Association for Computing Machinery. 2020.
- [54] Yao, Guangming and Yuan, Yi and Shao, Tianjia and Zhou, Kun. *Mesh guided one-shot face reenactment using graph convolutional networks*. Proceedings of the 30th ACM International Conference on Multimedia. 2020.
- [55] Pham, Hai X. and Wang, Yuting and Pavlovic, Vladimir. *Generative Adversarial Talking Head: Bringing Portraits to Life with a Weakly Supervised Neural Network*. arXiv (Cornell University). 2018.
- [56] Khakhulin, Taras and Sklyarova, Vanessa and Lempitsky, Victor and Zakharov, Egor. *Realistic one-shot mesh-based head avatars*. arXiv (Cornell University). 2022.
- [57] Zeng, Bohan and Liu, Boyu and Li, Hong and Liu, Xuhui and Liu, Jianzhuang and Chen, Dapeng and Peng, Wei and Zhang, Baochang. *FNEVR: Neural Volume Rendering for face Animation*. arXiv (Cornell University). 2022.

- [58] Zakharov, Egor and Ivakhnenko, Aleksei and Shysheya, Aliaksandra and Lempitsky, Victor. *Fast bi-layer neural synthesis of One-Shot realistic head avatars*. arXiv (Cornell University). 2020.
- [59] Siarohin, Aliaksandr and Lathuilière, Stéphane and Tulyakov, Sergey and Ricci, Elisa and Sebe, Nicu. *Animating Arbitrary Objects via Deep Motion Transfer*. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2019.
- [60] Liang, Borong and Pan, Yan and Guo, Zhizhi and Zhou, Hang and Hong, Zhibin and Han, Xiaoguang and Han, Junyu and Liu, Jingtuo and Ding, Errui and Wang, Jingdong. *Expressive Talking Head Generation with Granular Audio-Visual Control*. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2022.
- [61] Shen, Shuai and Zhao, Wenliang and Meng, Zibin and Li, Wanhua and Zhu, Zheng and Zhou, Jie and Lu, Jiwen. *DiffTalk: Crafting Diffusion Models for Generalized Audio-Driven Portraits Animation*. arXiv (Cornell University). 2023.
- [62] Fried, Ohad and Tewari, Ayush and Zollhöfer, Michael and Finkelstein, Adam and Shechtman, Eli and Goldman, Dan B and Genova, Kyle and Jin, Zeyu and Theobalt, Christian and Agrawala, Maneesh. *Text-based editing of talking-head video*. arXiv (Cornell University). 2019.
- [63] Bansal, Aayush and Ma, Shugao and Ramanan, Deva and Sheikh, Yaser. *Recycle-GAN: Unsupervised video retargeting*. arXiv (Cornell University). 2018.
- [64] Jiatao Gu and Lingjie Liu and Peng Wang and Christian Theobalt. *StyleNeRF: A Style-based 3D-Aware Generator for High-resolution Image Synthesis*. arXiv (Cornell University). 2021.
- [65] Agustsson, Eirikur and Minnen, David and Johnston, Nick and Balle, Johannes and Hwang, Sung Jin and Toderici, George. *Scale-Space flow for End-to-End optimized video compression*. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2020.
- [66] Livingstone, Steven R. and Russo, Frank A.. *The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)*. Zenodo. 2018.
- [67] Livingstone, Steven R. and Russo, Frank A.. *The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English*. PLoS ONE. 2018.
- [68] Kim, Jiwon and Lee, Jung Kwon and Lee, Kyoung Mu. *Accurate Image Super-Resolution Using Very Deep Convolutional Networks*. IEEE. 2016.
- [69] Reed, Scott and Akata, Zeynep and Yan, Xinchen and Logeswaran, Lajanugen and Schiele, Bernt and Lee, Honglak. *Generative adversarial text to image synthesis*. arXiv (Cornell University). 2016.
- [70] Averbuch-Elor, Hadar and Cohen-Or, Daniel and Kopf, Johannes and Cohen, Michael F.. *Bringing portraits to life*. ACM Transactions on Graphics. 2017.
- [71] Ma, Yifeng and Wang, Suzhen and Ding, Yu and Ma, Bowen and Lv, Tangjie and Fan, Changjie and Hu, Zhipeng and Deng, Zhidong and Yu, Xin. *TalkCLIP: Talking Head Generation with Text-Guided Expressive Speaking Styles*. arXiv (Cornell University). 2023.
- [72] Arsha Nagrani and Joon Son Chung and Andrew Zisserman. *VoxCeleb: A Large-Scale Speaker Identification Dataset*. Interspeech 2017. 2017.
- [73] Christop, Iwona. *NEMO: Dataset of Emotional Speech in Polish*. arXiv preprint arXiv:2404.06292. 2024.
- [74] Zhuoqian Yang and SenseTime Research and Robotics Institute, Carnegie Mellon University and Center for Research on Intelligent Perception and Computing, CASIA and University of Chinese Academy of Sciences and Shenzhen Institutes of Advanced Technology, Chinese Academy of Science and Nanyang Technological University. *MEAD: a large-scale audio-visual dataset for emotional talking-face generation*. arXiv preprint. 2021.
- [75] Unknown Author. *Linguistic Data Consortium (LDC)*. University of Pennsylvania. 2020.
- [76] Chung, Joon Son and Senior, Andrew and Vinyals, Oriol and Zisserman, Andrew. *The Oxford-BBC lip reading sentences in the Wild dataset*. Asian Conference on Computer Vision. 2017.
- [77] Diao, Xingjian and Cheng, Ming and Barrios, Wayner and Jin, SouYoung. *FT2TF: First-Person Statement Text-To-Talking Face Generation*. arXiv (Cornell University). 2023.
- [78] Kumar, Rithesh and Sotelo, Jose and Kumar, Kundan and De Brébisson, Alexandre and Bengio, Yoshua. *ObamaNet: Photo-realistic lip-sync from text*. arXiv (Cornell University). 2018.

- [79] Wang, Yuchi and Guo, Junliang and Bai, Jianhong and Yu, Runyi and He, Tianyu and Tan, Xu and Sun, Xu and Bian, Jiang. *InstructAvatar: Text-Guided Emotion and Motion Control for Avatar Generation*. arXiv (Cornell University). 2024.
- [80] Li, Lincheng and Wang, Suzhen and Zhang, Zhimeng and Ding, Yu and Zheng, Yixing and Yu, Xin and Fan, Changjie. *Write-a-Speaker: Text-based emotional and rhythmic talking-head generation*. arXiv (Cornell University). 2021.
- [81] Zhang, Sibo and Yuan, Jiahong and Liao, Miao and Zhang, Liangjun. *Text2Video: Text-driven Talking-head Video Synthesis with Personalized Phoneme-Pose Dictionary*. arXiv (Cornell University). 2021.
- [82] Jalalifar, Seyed Ali and Hasani, Hosein and Aghajan, Hamid. *Speech-Driven facial reenactment using conditional generative adversarial networks*. arXiv (Cornell University). 2018.
- [83] Wayne, Wu and Yunxuan, Zhang and Cheng, Li and Chen, Qian and Change, Loy Chen. *ReenactGAN: Learning to reenact faces via boundary transfer*. arXiv (Cornell University). 2018.
- [84] Kingma, Diederik P. and Ba, Jimmy Lei. *Adam: A method for stochastic optimization*. arXiv (Cornell University). 2014.
- [85] Shin, Ah-Hyung and Lee, Jae-Ho and Hwang, Jiwon and Kim, Yoonhyung and Park, Gyeong-Moon. *Wav2NeRF: Audio-driven realistic talking head generation via wavelet-based NeRF*. Image and Vision Computing. 2024.
- [86] Meng, Ming and Zhao, Yufei and Zhang, Bo and Zhu, Yonggui and Shi, Weimin and Wen, Maxwell and Fan, Zhaoxin. *A comprehensive taxonomy and analysis of talking head synthesis: techniques for portrait generation, driving mechanisms, and editing*. arXiv (Cornell University). 2024.
- [87] Gowda, Shreyank N and Pandey, Dheeraj and Gowda, Shashank Narayana. *From Pixels to Portraits: A comprehensive survey of talking head generation techniques and applications*. arXiv (Cornell University). 2023.
- [88] Liu, Ziwei. *Review of talking head synthesis for driving mechanisms and portrait rendering*. Applied and Computational Engineering. 2024.
- [89] Thies, Justus and Elgharib, Mohamed and Tewari, Ayush and Theobalt, Christian and Nießner, Matthias. *Neural Voice Puppetry: audio-driven facial reenactment*. arXiv (Cornell University). 2019.
- [90] Mildenhall, Ben and Srinivasan, Pratul P. and Tancik, Matthew and Barron, Jonathan T. and Ramamoorthi, Ravi and Ng, Ren. *NeRF: Representing scenes as neural radiance fields for view synthesis*. arXiv (Cornell University). 2020.
- [91] Terven, Juan and Cordova-Esparza, Diana M. and Ramirez-Pedraza, Alfonzo and Chavez-Urbiola, Edgar A.. *Loss functions and metrics in deep learning*. arXiv (Cornell University). 2023.
- [92] Chen, Lele and Cui, Guofeng and Kou, Ziyi and Zheng, Haitian and Xu, Chenliang. *What comprises a good talking-head video generation?: A Survey and Benchmark*. arXiv (Cornell University). 2020.
- [93] Joon Son Chung and Andrew Zisserman. *Out of time: automated lip sync in the wild*. Workshop on Multi-view Lip-reading, ACCV. 2016.
- [94] Heusel, Martin and Ramsauer, Hubert and Unterthiner, Thomas and Nessler, Bernhard and Hochreiter, Sepp. *GANs trained by a two Time-Scale update rule converge to a local Nash equilibrium*. arXiv (Cornell University). 2017.
- [95] Foschini, G.J. and Greenstein, L.J. and Vannucci, G.. *Noncoherent detection of coherent lightwave signals corrupted by phase noise*. IEEE Transactions on Communications. 1988.
- [96] Mukawa, N. and Kuroda, H. and Matsuoka, T.. *An Interframe Coding System for Video Teleconferencing Signal Transmission at a 1.5 Mbit/s Rate*. IEEE Transactions on Communications. 1984.
- [97] Zhou Wang and Bovik, A.C. and Sheikh, H.R. and Simoncelli, E.P.. *Image quality assessment: from error visibility to structural similarity*. IEEE Transactions on Image Processing. 2004.
- [98] Xie, Liangbin and Wang, Xintao and Zhang, Honglun and Dong, Chao and Shan, Ying. *VFHQ: A High-Quality Dataset and Benchmark for Video Face Super-Resolution*. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). 2022.
- [99] Cui, Jiahao and Li, Hui and Zhan, Yun and Shang, Hanlin and Cheng, Kaihui and Ma, Yuqi and Mu, Shan and Zhou, Hang and Wang, Jingdong and Zhu, Siyu. *Hallo3: Highly Dynamic and Realistic Portrait Image Animation with Diffusion Transformer Networks*. arXiv (Cornell University). 2024.

- [100] Aaron van den Oord and Oriol Vinyals and Koray Kavukcuoglu. *Neural Discrete Representation Learning*. arXiv preprint arXiv:1711.00937. 2018.
- [101] Li, Tianye and Bolkart, Timo and Black, Michael. J. and Li, Hao and Romero, Javier. *Learning a model of facial shape and expression from 4D scans*. ACM Transactions on Graphics, (Proc. SIGGRAPH Asia). 2017.
- [102] Jackson, Philip and Haq, Sanaul. *Surrey Audio-Visual Expressed Emotion (SAVEE) database*. University of Surrey. 2015.
- [103] Harte, Naomi and Gillen, Eoin. *TCD-TIMIT: An Audio-Visual Corpus of Continuous Speech*. IEEE Transactions on Multimedia. 2015.
- [104] Rainio, Oona and Teuho, Jarmo and Klén, Riku. *Evaluation metrics and statistical tests for machine learning*. Scientific Reports. 2024.
- [105] Karras, Tero and Laine, Samuli and Aila, Timo. *FFHQ*. arXiv preprint arXiv:1812.04948. 2018.
- [106] Zhang, Zhimeng and Li, Lincheng and Ding, Yu and Fan, Changjie and Virtual Human Group, Netease Fuxi AI Lab. *HDTF*. arXiv preprint. 2020.
- [107] Linnainmaa, Seppo. *Taylor expansion of the accumulated rounding error*. BIT Numerical Mathematics. 1976.
- [108] Huang, Guang-Bin and Zhu, Qin-Yu and Siew, Chee-Kheong. *Extreme learning machine: Theory and applications*. Neurocomputing. 2006.
- [109] Widrow, Bernard and Greenblatt, Aaron and Kim, Youngsik and Park, Dookun. *The No-Prop algorithm: A new learning algorithm for multilayer neural networks*. Neural Networks. 2012.
- [110] Ollivier, Yann and Charpiat, Guillaume. *Training recurrent networks online without backtracking*. arXiv (Cornell University). 2015.
- [111] Mienye, Ibomoiye Domor and Swart, Theo G.. *A comprehensive review of deep learning: architectures, recent advances, and applications*. Information. 2024.
- [112] Mienye, Ibomoiye Domor and Jere, Nobert. *Deep Learning for Credit Card Fraud Detection: A review of Algorithms, challenges, and solutions*. IEEE Access. 2024.
- [113] Krizhevsky, Alex and Sutskever, Ilya and Hinton, Geoffrey E.. *ImageNet classification with deep convolutional neural networks*. Communications of the ACM. 2017.
- [114] Sherstinsky, Alex. *Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) network*. Physica D Nonlinear Phenomena. 2020.
- [115] Chan, Eric R. and Lin, Connor Z. and Chan, Matthew A. and Nagano, Koki and Pan, Boxiao and De Mello, Shalini and Gallo, Orazio and Guibas, Leonidas and Tremblay, Jonathan and Khamis, Sameh and Karras, Tero and Wetzstein, Gordon. *Efficient Geometry-aware 3D Generative Adversarial Networks*. arXiv preprint arXiv:2112.07945. 2022.
- [116] Eric R. Chan and Connor Z. Lin and Matthew A. Chan and Koki Nagano and Boxiao Pan and Shalini De Mello and Orazio Gallo and Leonidas Guibas and Jonathan Tremblay and Sameh Khamis and Tero Karras and Gordon Wetzstein. *Efficient Geometry-aware 3D Generative Adversarial Networks*. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2022.
- [117] Zakharov, Egor and Ivakhnenko, Aleksei and Shysheya, Aliaksandra and Lempitsky, Victor. *Fast Bi-Layer Neural Synthesis of One-Shot Realistic Head Avatars*. Computer Vision – ECCV 2020. Springer. 2020.
- [118] Zhou, Hang and Sun, Yasheng and Wu, Wayne and Loy, Chen Change and Wang, Xiaogang and Liu, Ziwei. *Pose-Controllable Talking Face Generation by Implicitly Modularized Audio-Visual Representation*. arXiv preprint arXiv:2104.11116. 2021.
- [119] Bahmani, Sherwin and Park, Jeong Joon and Paschalidou, Despoina and Tang, Hao and Wetzstein, Gordon and Guibas, Leonidas and Luc, Van Gool and Timofte, Radu. *3D-Aware video Generation*. arXiv (Cornell University). 2022.
- [120] Hornik, Kurt and Stinchcombe, Maxwell and White, Halbert. *Multilayer feedforward networks are universal approximators*. Neural Networks. 1989.
- [121] Zhang, Chenshuang and Zhang, Chaoning and Zhang, Mengchun and Kweon, In So. *Text-to-image diffusion models in Generative AI: a survey*. arXiv (Cornell University). 2023.

- [122] Hagos, Desta Haileselassie and Battle, Rick and Rawat, Danda B.. *Recent advances in generative AI and large language models: current status, challenges, and perspectives*. IEEE Transactions on Artificial Intelligence. 2024.
- [123] McCulloch, Warren S. and Pitts, Walter. *A logical calculus of the ideas immanent in nervous activity*. The Bulletin of Mathematical Biophysics. 1943.
- [124] Harrell, Frank E. and Lee, Kerry L. and Califf, Robert M. and Pryor, David B. and Rosati, Robert A.. *Regression modelling strategies for improved prognostic prediction*. Statistics in Medicine. 1984.
- [125] Huang, Ricong and Lai, Peiwen and Qin, Yipeng and Li, Guanbin. *Parametric implicit face representation for Audio-Driven facial reenactment*. arXiv (Cornell University). 2023.
- [126] Wang, Qianyun and Fan, Zhenfeng and Xia, Shihong. *3D-TalkEMO: Learning to synthesize 3D emotional talking head*. arXiv (Cornell University). 2021.
- [127] Guo, Yudong and Chen, Keyu and Liang, Sen and Liu, Yong-Jin and Bao, Hujun and Zhang, Juyong. *AD-NeRF: Audio Driven Neural Radiance Fields for Talking Head Synthesis*. 2021 IEEE/CVF International Conference on Computer Vision (ICCV). 2021.
- [128] Song, Linsen and Wu, Wayne and Fu, Chaoyou and Loy, Chen Change and He, Ran. *Audio-Driven Dubbing for User Generated Contents via Style-Aware Semi-Parametric Synthesis*. IEEE Transactions on Circuits and Systems for Video Technology. 2023.
- [129] Hui Fang and Dongdong Weng and Zeyu Tian and Yin Ma and Xiangju Lu. *Audio-to-Deep-Lip: Speaking lip synthesis based on 3D landmarks*. Computers & Graphics. 2024.
- [130] Li, Dongze and Zhao, Kang and Wang, Wei and Peng, Bo and Zhang, Yingya and Dong, Jing and Tan, Tieniu. *AE-NERF: Audio Enhanced Neural radiance Field for few shot talking head synthesis*. arXiv (Cornell University). 2023.
- [131] Shaoxu Li and Ye Pan. *AniArtAvatar: Animatable 3D art avatar from a single image*. Neurocomputing. 2025.
- [132] Fanelli, Gabriele and Dantone, Matthias and Gall, Juergen and Fossati, Andrea and Van Gool, Luc. *Random Forests for Real Time 3D Face Analysis*. Int. J. Comput. Vision. 2013.
- [133] LI, Guo-Qiang and DU, Li-Min and XU, Yan-Jun and HOU, Zi-Qiang. *Maximum likelihood smoothes and predictions for fast speaker adaptation*. 100080.
- [134] Cao, Houwei and Cooper, David G. and Keutmann, Michael K. and Gur, Ruben C. and Nenkova, Ani and Verma, Ragini. *CREMA-D: Crowd-Sourced Emotional Multimodal Actors Dataset*. IEEE Transactions on Affective Computing. 2014.
- [135] Xuanyuan, Meidai and Wang, Yuwang and Guo, Honglei and Ma, Xiao and Guo, Yuchen and Yu, Tao and Dai, Qionghai. *Deep personalized characters from TV shows*. arXiv (Cornell University). 2023.
- [136] Andreas Rössler and Davide Cozzolino and Luisa Verdoliva and Christian Riess and Justus Thies and Matthias Nießner. *FaceForensics++: Learning to Detect Manipulated Facial Images*. International Conference on Computer Vision (ICCV). 2019.
- [137] Andreas Rössler and Davide Cozzolino and Luisa Verdoliva and Christian Riess and Justus Thies and Matthias Nießner. *FaceForensics++: Learning to Detect Manipulated Facial Images*. International Conference on Computer Vision (ICCV). 2019.
- [138] Zhang, Zhimeng and Li, Lincheng and Ding, Yu and Fan, Changjie. *Flow-guided One-shot Talking Face Generation with a High-resolution Audio-visual Dataset*. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2021.
- [139] Afouras, Triantafyllos and Chung, Joon Son and Zisserman, Andrew. *LRS3-TED: a large-scale dataset for visual speech recognition*. arXiv (Cornell University). 2018.
- [140] Yang, Shuang and Zhang, Yuanhang and Feng, Dalu and Yang, Mingmin and Wang, Chenhao and Xiao, Jingyun and Long, Keyu and Shan, Shiguang and Chen, Xilin. *LRW-1000: A Naturally-Distributed Large-Scale Benchmark for Lip Reading in the Wild*. 2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019). 2019.
- [141] Eskander, George S. and Sabourin, Robert and Granger, Eric. *Offline Signature-Based Fuzzy Vault (OSFV): review and new results*. arXiv (Cornell University). 2014.

- [142] Wang, Suzhen and Li, Lincheng and Ding, Yu and Fan, Changjie and Yu, Xin. *Audio2Head: Audio-driven One-shot Talking-head Generation with Natural Head Motion*. Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21. International Joint Conferences on Artificial Intelligence Organization. 2021. Main Track.
- [143] Liang, Borong and Pan, Yan and Guo, Zhiyi and Zhou, Hang and Hong, Zhibin and Han, Xiaoguang and Han, Junyu and Liu, Jingtuo and Ding, Errui and Wang, Jingdong. *Expressive Talking Head Generation with Granular Audio-Visual Control*. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2022.
- [144] Ji, Xinya and Zhou, Hang and Wang, Kaisiyuan and Wu, Qianyi and Wu, Wayne and Xu, Feng and Cao, Xun. *EAMM: One-Shot Emotional Talking Face via Audio-Based Emotion-Aware Motion Model*. arXiv (Cornell University). 2022.
- [145] Ma, Yifeng and Wang, Suzhen and Hu, Zhipeng and Fan, Changjie and Lv, Tangjie and Ding, Yu and Deng, Zhidong and Yu, Xin. *StyleTalk: One-shot Talking Head Generation with Controllable Speaking Styles*. arXiv (Cornell University). 2023.
- [146] Zhang, Wenxuan and Cun, Xiaodong and Wang, Xuan and Zhang, Yong and Shen, Xi and Guo, Yu and Shan, Ying and Wang, Fei. *SadTalker: Learning Realistic 3D Motion Coefficients for Stylized Audio-Driven Single Image Talking Face Animation*. arXiv preprint arXiv:2211.12194. 2023.
- [147] Bounareli, Stella and Tzelepis, Christos and Argyriou, Vasileios and Patras, Ioannis and Tzimiropoulos, Georgios. *StyleMask: Disentangling the Style Space of StyleGAN2 for Neural Face Reenactment*. 2023 IEEE 17th International Conference on Automatic Face and Gesture Recognition (FG). IEEE Press. 2023.
- [148] Tao, Jiale and Wang, Biao and Xu, Borun and Ge, Tiezheng and Jiang, Yuning and Li, Wen and Duan, Lixin. *Structure-Aware Motion Transfer with Deformable Anchor Model*. arXiv preprint arXiv:2204.05018. 2022.
- [149] Hong, Fa-Ting and Xu, Dan. *Implicit Identity Representation Conditioned Memory Compensation Network for Talking Head video Generation*. arXiv (Cornell University). 2023.
- [150] Yin, Fei and Zhang, Yong and Cun, Xiaodong and Cao, Mingdeng and Fan, Yanbo and Wang, Xuan and Bai, Qingyan and Wu, Baoyuan and Wang, Jue and Yang, Yujiu. *StyleHEAT: One-Shot High-Resolution Editable Talking Face Generation via Pre-trained StyleGAN*. arXiv preprint arXiv:2203.04036. 2022.
- [151] Zhao, Jian and Zhang, Hui. *Thin-Plate Spline Motion Model for Image Animation*. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2022.
- [152] Zhao, Shuling and Hong, Fa-Ting and Huang, Xiaoshui and Xu, Dan. *Synergizing Motion and Appearance: Multi-Scale Compensatory Codebooks for Talking Head Video Generation*. arXiv preprint arXiv:2412.00719. 2025.
- [153] Guo, Jianzhu and Zhang, Dingyun and Liu, Xiaoqiang and Zhong, Zhizhou and Zhang, Yuan and Wan, Pengfei and Zhang, Di. *LivePortrait: Efficient Portrait Animation with Stitching and Retargeting Control*. arXiv preprint arXiv:2407.03168. 2025.
- [154] Drobyshev, Nikita and Chelishev, Jenya and Khakhulin, Taras and Ivakhnenko, Aleksei and Lempitsky, Victor and Zakharov, Egor. *MegaPortraits: One-shot Megapixel Neural Head Avatars*. Proceedings of the 30th ACM International Conference on Multimedia. Association for Computing Machinery. 2022.
- [155] Karras, Tero and Laine, Samuli and Aila, Timo. *A Style-Based generator architecture for generative adversarial networks*. arXiv (Cornell University). 2018.
- [156] Agarwal, Madhav and Mukhopadhyay, Rudrabha and Namboodiri, Vinay and Jawahar, C V. *Audio-Visual Face Reenactment*. 2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). 2023.
- [157] Xie, You and Xu, Hongyi and Song, Guoxian and Wang, Chao and Shi, Yichun and Luo, Linjie. *X-Portrait: Expressive Portrait Animation with Hierarchical Motion Attention*. arXiv (Cornell University). 2024.
- [158] Xu, Zunnan and Yu, Zhentao and Zhou, Zixiang and Zhou, Jun and Jin, Xiaoyu and Hong, Fa-Ting and Ji, Xiaozhong and Zhu, Junwei and Cai, Chengfei and Tang, Shiyu and Lin, Qin and Li, Xiu and Lu, Qinglin. *HunyuanPortrait: Implicit Condition Control for Enhanced Portrait Animation*. arXiv preprint arXiv:2503.18860. 2025.
- [159] Wang, Kaisiyuan and Wu, Qianyi and Song, Lisen and Yang, Zhuoqian and Wu, Wayne and Qian, Chen and He, Ran and Qiao, Yu and Loy, Chen Change. *MEAD: A Large-scale Audio-visual Dataset for Emotional Talking-face Generation*. ECCV. 2020.

- [160] Tian, Linrui and Wang, Qi and Zhang, Bang and Bo, Liefeng. *EMO: Emote Portrait Alive – Generating Expressive Portrait Videos with Audio2Video Diffusion Model under Weak Conditions*. arXiv (Cornell University). 2024.
- [161] Zhang, Lin and Zhang, Lei and Mou, Xuanqin and Zhang, David. *FSIM: A Feature Similarity Index for Image Quality Assessment*. IEEE Transactions on Image Processing. 2011.
- [162] Poria, Soujanya and Hazarika, Devamanyu and Majumder, Navonil and Naik, Gautam and Cambria, Erik and Mihalcea, Rada. *MELD: a multimodal Multi-Party dataset for emotion recognition in conversations*. arXiv (Cornell University). 2018.
- [163] Unterthiner, Thomas and Sjoerd, Van Steenkiste and Kurach, Karol and Marinier, Raphael and Michalski, Marcin and Gelly, Sylvain. *Towards Accurate Generative Models of Video: A New Metric & Challenges*. arXiv (Cornell University). 2018.
- [164] Huang, Po-Hsiang and Yang, Fu-En and Wang, Yu-Chiang Frank. *Learning Identity-Invariant Motion Representations for Cross-ID Face Reenactment*. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2020.
- [165] Wu, Haozhe and Jia, Jia and Xing, Junliang and Xu, Hongwei and Wang, Xiangyuan and Wang, Jelo. *MMFace4D: a Large-Scale Multi-Modal 4D face dataset for Audio-Driven 3D face animation*. arXiv (Cornell University). 2023.
- [166] Wuu, Cheng-Hsin and Zheng, Ningyuan and Ardisson, Scott and Bali, Rohan and Belko, Danielle and Brockmeyer, Eric and Evans, Lucas and Godisart, Timothy and Ha, Hyowon and Hypes, Alexander and Koska, Taylor and Krenn, Steven and Lombardi, Stephen and Luo, Xiaomin and McPhail, Kevyn and Millerschoen, Laura and Perdoch, Michal and Pitts, Mark and Richard, Alexander and Saragih, Jason and Saragih, Junko and Shiratori, Takaaki and Simon, Tomas and Stewart, Matt and Trimble, Autumn and Weng, Xinshuo and Whitewolf, David and Wu, Chenglei and Yu, Shouou-I and Sheikh, Yaser. *Multiface: a dataset for neural face rendering*. arXiv (Cornell University). 2022.
- [167] Ren, Yurui and Li, Ge and Chen, Yuanqi and Li, Thomas H. and Liu, Shan . *PIRenderer: Controllable Portrait Image Generation via Semantic Neural Rendering* . 2021 IEEE/CVF International Conference on Computer Vision (ICCV) . IEEE Computer Society. 2021.
- [168] Zhou, Mohan and Bai, Yalong and Zhang, Wei and Yao, Ting and Zhao, Tiejun and Mei, Tao. *Responsive Listening Head Generation: a benchmark dataset and baseline*. arXiv (Cornell University). 2021.
- [169] Ting-Chun Wang and Arun Mallya and Ming-Yu Liu. *One-Shot Free-View Neural Talking-Head Synthesis for Video Conferencing (Talking Head-1KH Dataset)*. CVPR. 2021.
- [170] Cudeiro, Daniel and Bolkart, Timo and Laidlaw, Cassidy and Ranjan, Anurag and Black, Michael. *Capture, Learning, and Synthesis of 3D Speaking Styles*. Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). 2019.
- [171] Nagrani, Arsha and Chung, Joon Son and Zisserman, Andrew. *VoxCeleb1: a large-scale speaker identification dataset*. 2020.
- [172] Nagrani, Arsha and Chung, Joon Son and Xie, Weidi and Andrew Zisserman. *Voxceleb: Large-scale speaker verification in the wild*. 2019.
- [173] Chung, Joon Son and Nagrani, Arsha and Zisserman, Andrew and Visual Geometry Group. *VoxCeleb2: Deep Speaker Recognition*. 2018.
- [174] Richard, Alexander and Zollhoefer, Michael and Wen, Yandong and La Torre Fernando, De and Sheikh, Yaser. *MeshTalk: 3D Face Animation from Speech using Cross-Modality Disentanglement*. arXiv (Cornell University). 2021.
- [175] Zhu, Hao and Wu, Wayne and Zhu, Wentao and Jiang, Liming and Tang, Siwei and Zhang, Li and Liu, Ziwei and Loy, Chen Change. *CelebV-HQ: a Large-Scale Video Facial Attributes Dataset*. arXiv (Cornell University). 2022.
- [176] Alghamdi, Najwa and Maddock, Steve and Marixer, Ricard and Barker, Jon and Brown, Guy J.. *A corpus of audio-visual Lombard speech with frontal and profile views*. The Journal of the Acoustical Society of America. 2018.
- [177] Chung, J. S. and Senior, A. and Vinyals, O. and Zisserman, A.. *Lip Reading Sentences in the Wild*. IEEE Conference on Computer Vision and Pattern Recognition. 2017.
- [178] C. Busso and S. Parthasarathy and A. Burmania and M. AbdelWahab and N. Sadoughi and E. Mower Provost. *MSP-IMPROV: An Acted Corpus of Dyadic Interactions to Study Emotion Perception*. IEEE Transactions on Affective Computing. 2017.

- [179] Lin, Gaojie and Jiang, Jianwen and Yang, Jiaqi and Zheng, Zerong and Liang, Chao. *OmniHuman-1: Rethinking the Scaling-Up of One-Stage Conditioned Human animation Models*. arXiv (Cornell University). 2025.
- [180] Ma, Haoyu and Zhang, Tong and Sun, Shanlin and Yan, Xiangyi and Han, Kun and Xie, Xiaohui. *CVTHead: One-shot Controllable Head Avatar with Vertex-feature Transformer*. arXiv (Cornell University). 2023.
- [181] Du, Chenpeng and Chen, Qi and He, Tianyu and Tan, Xu and Chen, Xie and Yu, Kai and Zhao, Sheng and Bian, Jiang. *DAE-Talker: High Fidelity Speech-Driven Talking Face Generation with Diffusion Autoencoder*. Proceedings of the 31st ACM International Conference on Multimedia. ACM. 2023.
- [182] Deng, Yu and Yang, Jiaolong and Chen, Dong and Wen, Fang and Tong, Xin. *Disentangled and Controllable Face Image Generation via 3D Imitative-Contrastive Learning*. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2020.
- [183] Aaron Mir and Eduardo Alonso and Esther Mondragón. *DiT-Head: High Resolution Talking Head Synthesis Using Diffusion Transformers*. Proceedings of the 16th International Conference on Agents and Artificial Intelligence - Volume 3: ICAART. SciTePress. 2024.
- [184] Zhang, Bingyuan and Zhang, Xulong and Cheng, Ning and Yu, Jun and Xiao, Jing and Wang, Jianzong. *EmoTalker: Emotionally Editable Talking Face Generation via Diffusion Model*. ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2024.
- [185] Waibel, Alexander and Behr, Moritz and Eyiokur, Fevziye Irem and Yaman, Dogucan and Nguyen, Tuan-Nam and Mullov, Carlos and Demirtas, Mehmet Arif and Kantarcı, Alperen and Constantin, Stefan and Ekenel, Hazim Kemal. *Face-Dubbing++: Lip-Synchronous, Voice Preserving Translation of Videos*. arXiv preprint arXiv:2206.04523. 2022.
- [186] Thies, Justus and Zollhöfer, Michael and Stamminger, Marc and Theobalt, Christian and Nießner, Matthias. *Face2Face: Real-time face capture and reenactment of RGB videos*. arXiv (Cornell University). 2020.
- [187] Stan, Stefan and Haque, Kazi Injamamul and Yumak, Zerrin. *FaceDiffuser: Speech-Driven 3D Facial Animation Synthesis Using Diffusion*. Proceedings of the 16th ACM SIGGRAPH Conference on Motion, Interaction and Games. Association for Computing Machinery. 2023.
- [188] Lee, Haedeun and Oh, Bumjo and Kim, Seung-Chan. *Recognition of Forward Head Posture Through 3D Human Pose Estimation With a Graph Convolutional Network: Development and Feasibility Study*. JMIR Form Res. 2024.
- [189] Zhang, Zhimeng and Li, Lincheng and Ding, Yu and Fan, Changjie. *Flow-guided One-shot Talking Face Generation with a High-resolution Audio-visual Dataset*. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2021.
- [190] Zeng, Bohan and Liu, Boyu and Li, Hong and Liu, Xuhui and Liu, Jianzhuang and Chen, Dapeng and Peng, Wei and Zhang, Baochang. *FNeVR: neural volume rendering for face animation*. Proceedings of the 36th International Conference on Neural Information Processing Systems. Curran Associates Inc.. 2022.
- [191] Zakharov, Egor and Shysheya, Aliaksandra and Burkov, Egor and Lempitsky, Victor. *Few-Shot Adversarial Learning of Realistic Neural Talking Head Models*. 2019 IEEE/CVF International Conference on Computer Vision (ICCV). 2019.
- [192] Maalouf, Aldo and Larabi, Mohamed-Chaker and Fernandez-Maloigne, Christine. *A grouplet-based reduced reference image quality assessment*. 2009 International Workshop on Quality of Multimedia Experience. 2009.
- [193] Ni, Haomiao and Liu, Jiachen and Xue, Yuan and Huang, Sharon X.. *3D-Aware Talking-Head Video Motion Transfer*. 2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). 2024.
- [194] Hessel, Jack and Holtzman, Ari and Forbes, Maxwell and Bras, Ronan Le and Choi, Yejin. *CLIPScore: a reference-free evaluation metric for image captioning*. arXiv (Cornell University). 2021.
- [195] Li, Weichuang and Zhang, Longhao and Wang, Dong and Zhao, Bin and Wang, Zhigang and Chen, Mulin and Zhang, Bang and Wang, Zhongjian and Bo, Liefeng and Li, Xuelong. *One-Shot High-Fidelity Talking-Head Synthesis With Deformable Neural Radiance Field*. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2023.

- [196] Li, Weichuang and Zhang, Longhao and Wang, Dong and Zhao, Bin and Wang, Zhigang and Chen, Mulin and Zhang, Bang and Wang, Zhongjian and Bo, Liefeng and Li, Xuelong. *One-Shot High-Fidelity Talking-Head Synthesis with Deformable Neural Radiance Field*. arXiv (Cornell University). 2023.
- [197] Hong, Yang and Peng, Bo and Xiao, Haiyao and Liu, Ligang and Zhang, Juyong. *HeadNeRF: A Real-time NeRF-based Parametric Head Model*. arXiv preprint arXiv:2112.05637. 2022.
- [198] Zheng, Yufeng and Fernández Abrevaya, Victoria and Bühler, Marcel C. and Chen, Xu and Black, Michael J. and Hilliges, Otmar. *I M Avatar: Implicit Morphable Head Avatars from Videos*. arXiv preprint arXiv:2112.07471. 2022.
- [199] Wang, Yuchi and Guo, Junliang and Bai, Jianhong and Yu, Runyi and He, Tianyu and Tan, Xu and Sun, Xu and Bian, Jiang. *InstructAvatar: Text-Guided Emotion and Motion Control for Avatar Generation*. arXiv (Cornell University). 2024.
- [200] Provine, J.A. and Bruton, L.T.. *Lip synchronization in 3-D model based coding for video-conferencing*. Proceedings of ISCAS'95 - International Symposium on Circuits and Systems. 1995.
- [201] Zhou, Yang and Han, Xintong and Shechtman, Eli and Echevarria, Jose and Kalogerakis, Evangelos and Li, Dingzeyu. *MakeltTalk: speaker-aware talking-head animation*. ACM Transactions on Graphics (TOG). Association for Computing Machinery. 2020.
- [202] Ha, Sungjoo and Kersner, Martin and Kim, Beomsu and Seo, Seokjun and Kim, Dongyoung. *Mari-oNETte: Few-shot face reenactment Preserving identity of unseen targets*. arXiv (Cornell University). 2019.
- [203] Zhao, Qingcheng and Long, Pengyu and Zhang, Qixuan and Qin, Dafei and Liang, Han and Zhang, Longwen and Zhang, Yingliang and Yu, Jingyi and Xu, Lan. *Media2Face: Co-speech Facial Animation Generation With Multi-Modality Guidance*. arXiv preprint arXiv:2401.15687. 2024.
- [204] Zhang, Bowen and Qi, Chenyang and Zhang, Pan and Zhang, Bo and Wu, HsiangTao and Chen, Dong and Chen, Qifeng and Wang, Yong and Wen, Fang. *MetaPortrait: Identity-Preserving Talking Head Generation with Fast Personalized Adaptation*. arXiv preprint arXiv:2212.08062. 2023.
- [205] Ligong Han and Jian Ren and Hsin-Ying Lee and Francesco Barbieri and Kyle Olszewski and Shervin Minaee and Dimitris Metaxas and Sergey Tulyakov. *Show Me What and Tell Me How: Video Synthesis via Multimodal Conditioning*. arXiv (Cornell University). 2022.
- [206] Liu, Yunfei and Lin, Lijian and Yu, Fei and Zhou, Changyin and Li, Yu. *MODA: Mapping-Once Audio-driven Portrait Animation with Dual Attentions*. arXiv (Cornell University). 2023.
- [207] Kim, Seyeon and Jin, Siyoon and Park, Jihye and Kim, Kihong and Kim, Jiyoung and Nam, Jisu and Kim, Seungryong. *MODiTalker: Motion-Disentangled Diffusion Model for High-Fidelity Talking Head Generation*. arXiv (Cornell University). 2024.
- [208] Sung-Bin, Kim and Chae-Yeon, Lee and Son, Gihun and Hyun-Bin, Oh and Ju, Janghoon and Nam, Suekyeong and Oh, Tae-Hyun. *MultiTalk: Enhancing 3D Talking Head Generation Across Languages with Multilingual Video Dataset*. arXiv (Cornell University). 2024.
- [209] Kumar, Rithesh and Sotelo, Jose and Kumar, Kundan and De Brébisson, Alexandre and Bengio, Yoshua. *ObamaNet: Photo-realistic lip-sync from text*. arXiv (Cornell University). 2018.
- [210] Lin, Gaojie and Jiang, Jianwen and Yang, Jiaqi and Zheng, Zerong and Liang, Chao. *OmniHuman-1: Rethinking the Scaling-Up of One-Stage Conditioned Human animation Models*. arXiv (Cornell University). 2025.
- [211] Hsuan-Kai Huang and Joseph Kuo and Yang Zhang and Yousuf Aborahama and Manxiu Cui and Karteekya Sastry and Seonyeong Park and Umberto Villa and Lihong V. Wang and Mark A. Anastasio. *Fast aberration correction in 3D transcranial photoacoustic computed tomography via a learning-based image reconstruction method*. Photoacoustics. 2025.
- [212] Yu, Zhentao and Yin, Zixin and Zhou, Deyu and Wang, Duomin and Wong, Finn and Wang, Baoyuan. *Talking Head Generation with Probabilistic Audio-to-Visual Diffusion Priors*. 2023 IEEE/CVF International Conference on Computer Vision (ICCV). 2023.
- [213] Han, Tianshun and Gui, Shengnan and Huang, Yiqing and Li, Baihui and Liu, Lijian and Zhou, Benjia and Jiang, Ning and Lu, Quan and Zhi, Ruicong and Liang, Yanyan and Zhang, Du and Wan, Jun. *PMMTalk: Speech-Driven 3D Facial Animation from Complementary Pseudo Multi-modal Features*. arXiv (Cornell University). 2023.

- [214] Xu, Eric Zhongcong and Zhang, Jianfeng and Liew, Jun Hao and Zhang, Wenqing and Bai, Song and Feng, Jiashi and Shou, Mike Zheng. *PV3D: a 3D generative model for portrait video generation*. arXiv (Cornell University). 2022.
- [215] Wayne, Wu and Yunxuan, Zhang and Cheng, Li and Chen, Qian and Change, Loy Chen. *ReenactGAN: Learning to reenact faces via boundary transfer*. arXiv (Cornell University). 2018.
- [216] Wang, Yaohui and Yang, Di and Bremond, Francois and Dantcheva, Antitza. *Latent Image Animator: Learning to animate images via latent space navigation*. arXiv (Cornell University). 2022.
- [217] Mallya, Arun and Wang, Ting-Chun and Liu, Ming-Yu. *Implicit Warping for Animation with Image Sets*. arXiv (Cornell University). 2022.
- [218] Wu, Yue and Deng, Yu and Yang, Jiaolong and Wei, Fangyun and Chen, Qifeng and Tong, Xin. *AniFaceGAN: animatable 3D-aware face image generation for video avatars*. Proceedings of the 36th International Conference on Neural Information Processing Systems. Curran Associates Inc.. 2022.
- [219] Wang, Qiulin and Zhang, Lu and Li, Bo . *SAFA: Structure Aware Face Animation* . 2021 International Conference on 3D Vision (3DV) . IEEE Computer Society. 2021.
- [220] Jamaludin, Amir and Chung, Joon Son and Zisserman, Andrew. *You Said That?: Synthesising Talking Faces from Audio*. International Journal of Computer Vision. 2019.
- [221] Liu, Xian and Xu, Yinghao and Wu, Qianyi and Zhou, Hang and Wu, Wayne and Zhou, Bolei. *Semantic-Aware Implicit Neural Audio-Driven Video Portrait Generation*. arXiv (Cornell University). 2022.
- [222] Liu, Xian and Xu, Yinghao and Wu, Qianyi and Zhou, Hang and Wu, Wayne and Zhou, Bolei. *Semantic-Aware Implicit Neural Audio-Driven Video Portrait Generation*. Lecture notes in computer science. Springer. 2022.
- [223] Vougioukas, Konstantinos and Petridis, Stavros and Pantic, Maja. *Realistic Speech-Driven Facial Animation with GANs*. International Journal of Computer Vision. 2019.
- [224] Pham, Trong Thang and Do, Tuong and Le, Nhat and Le, Ngan and Nguyen, Hung and Tjiputra, Erman and Tran, Quang and Nguyen, Anh. *Style Transfer for 2D Talking Head Generation*. 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). 2024.
- [225] Suwajanakorn, Supasorn and Seitz, Steven M. and Kemelmacher-Shlizerman, Ira. *Synthesizing Obama: learning lip sync from audio*. ACM Trans. Graph.. Association for Computing Machinery. 2017.
- [226] Ko, Jaehoon and Cho, Kyusun and Lee, Joungbin and Yoon, Heeji and Lee, Sangmin and Ahn, Sangjun and Kim, Seungryong. *Talk3D: High-Fidelity talking Portrait synthesis via personalized 3D generative Prior*. arXiv (Cornell University). 2024.
- [227] Ma, Yifeng and Wang, Suzhen and Ding, Yu and Ma, Bowen and Lv, Tangjie and Fan, Changjie and Hu, Zhipeng and Deng, Zhidong and Yu, Xin. *TalkCLIP: Talking Head Generation with Text-Guided Expressive Speaking Styles*. arXiv (Cornell University). 2023.
- [228] Zhang, Sibo and Yuan, Jiahong and Liao, Miao and Zhang, Liangjun. *Text2video: Text-Driven Talking-Head Video Synthesis with Personalized Phoneme - Pose Dictionary*. ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2022.
- [229] Ni, Haomiao and Liu, Yihao and Huang, Sharon X. and Xue, Yuan. *Cross-identity Video Motion Retargeting with Joint Transformation and Synthesis*. 2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). 2023.
- [230] Bregler, Christoph and Covell, Michele and Slaney, Malcolm. *Video Rewrite: driving visual speech with audio*. Addison-Wesley Publishing Co.. ACM Press. 1997.
- [231] Li, Lincheng and Wang, Suzhen and Zhang, Zhimeng and Ding, Yu and Zheng, Yixing and Yu, Xin and Fan, Changjie. *Write-a-Speaker: Text-based emotional and rhythmic talking-head generation*. arXiv (Cornell University). 2021.
- [232] Wiles, Olivia and Koepke, A. Sophia and Zisserman, Andrew. *X2Face: A network for controlling face generation by using images, audio, and pose codes*. arXiv (Cornell University). 2018.
- [233] Fan, Bo and Xie, Lei and Yang, Shan and Wang, Lijuan and Soong, Frank K.. *A deep bidirectional LSTM approach for video-realistic talking head*. Multimedia Tools Appl.. Kluwer Academic Publishers. 2016.
- [234] Zhu, Hao and Huang, Huaibo and Li, Yi and Zheng, Aihua and He, Ran. *Arbitrary Talking Face Generation via Attentional Audio-Visual Coherence Learning*. arXiv preprint arXiv:1812.06589. 2020.

- [235] Tang, Junshu and Zhang, Bo and Yang, Binxin and Zhang, Ting and Chen, Dong and Ma, Lizhuang and Wen, Fang. *3DFaceShop: Explicitly Controllable 3D-Aware Portrait Generation*. IEEE Transactions on Visualization and Computer Graphics. 2024.
- [236] Sun, Jingxiang and Wang, Xuan and Wang, Lizhen and Li, Xiaoyu and Zhang, Yong and Zhang, Hongwen and Liu, Yebin. *Next3D: Generative Neural Texture Rasterization for 3D-Aware Head Avatars*. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2023.
- [237] Yin, Yu and Ghasedi, Kamran and Wu, HsiangTao and Yang, Jiaolong and Tong, Xin and Fu, Yun. *NeRFInvertor: High Fidelity NeRF-GAN Inversion for Single-Shot Real Image Animation*. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2023.
- [238] Shen, Shuai and Li, Wanhua and Zhu, Zheng and Duan, Yueqi and Zhou, Jie and Lu, Jiwen. *Learning Dynamic Facial Radiance Fields for Few-Shot Talking Head Synthesis*. Computer Vision – ECCV 2022. Springer Nature Switzerland. 2022.
- [239] Hong, Fa-Ting and Xu, Zunnan and Zhou, Zixiang and Zhou, Jun and Li, Xiu and Lin, Qin and Lu, Qinglin and Xu, Dan. *Audio-visual Controlled Video Diffusion with Masked Selective State Spaces Modeling for Natural Talking Head Generation*. arXiv (Cornell University). 2025.
- [240] Wei, Huawei and Yang, Zejun and Wang, Zhisheng. *AniPortrait: Audio-Driven Synthesis of Photorealistic Portrait Animation*. arXiv (Cornell University). 2024.
- [241] Tan, Shuai and Ji, Bin and Bi, Mengxiao and Pan, Ye. *EDTaLK: Efficient Disentanglement for Emotional Talking head synthesis*. arXiv (Cornell University). 2024.
- [242] Chen, Zhiyuan and Cao, Jiajiong and Chen, Zhiqian and Li, Yuming and Ma, Chenguang. *EchoMimic: Lifelike Audio-Driven Portrait Animations through Editable Landmark Conditions*. Proceedings of the AAAI Conference on Artificial Intelligence. 2025.
- [243] Liang, Jiadong and Lu, Feng. *Emotional Conversation: Empowering Talking Faces with Cohesive Expression, Gaze and Pose Generation*. arXiv (Cornell University). 2024.
- [244] Xu, Chao and Liu, Yang and Xing, Jiazheng and Wang, Weida and Sun, Mingze and Dan, Jun and Huang, Tianxin and Li, Siyuan and Cheng, Zhi-Qi and Tai, Ying and Sun, Baigui. *FaceChain-ImagineID: Freely Crafting High-Fidelity Diverse Talking Faces from Disentangled Audio*. arXiv (Cornell University). 2024.
- [245] Yao, Ziyu and Cheng, Xuxin and Huang, Zhiqi. *FD2Talk: Towards Generalized Talking Head Generation with Facial Decoupled Diffusion Model*. Proceedings of the 32nd ACM International Conference on Multimedia. Association for Computing Machinery. 2024.
- [246] Shuai Tan and Bin Ji and Ye Pan. *FlowVQTalker: High-Quality Emotional Talking Face Generation through Normalizing Flow and Quantization*. arXiv (Cornell University). 2024.
- [247] Xu, Mingwang and Li, Hui and Su, Qingkun and Shang, Hanlin and Zhang, Liwei and Liu, Ce and Wang, Jingdong and Yao, Yao and Zhu, Siyu. *Hallo: Hierarchical Audio-Driven Visual Synthesis for Portrait Image Animation*. arXiv preprint arXiv:2406.08801. 2024.
- [248] Yu, Runyi and He, Tianyu and Zeng, Ailing and Wang, Yuchi and Guo, Junliang and Tan, Xu and Liu, Chang and Chen, Jie and Bian, Jiang. *Make Your Actor Talk: Generalizable and High-Fidelity Lip Sync with Motion and Appearance Disentanglement*. arXiv (Cornell University). 2024.
- [249] Zhang, Yue and Zhong, Zhizhou and Liu, Minhao and Chen, Zhaokang and Wu, Bin and Zeng, Yubin and Zhan, Chao and He, Yingjie and Huang, Junxin and Zhou, Wenjiang. *MuseTalk: Real-Time High-Fidelity Video Dubbing via Spatio-Temporal Sampling*. arXiv preprint arXiv:2410.10122. 2025.
- [250] Guan, Jiazhi and Xu, Zhihong and Zhou, Hang and Wang, Kaisiyuan and He, Shengyi and Zhang, Zhanwang and Liang, Borong and Feng, Haocheng and Ding, Errui and Liu, Jingtuo and Wang, Jingdong and Zhao, Youjian and Liu, Ziwei. *ReSyncer: Rewiring Style-based Generator for Unified Audio-Visually Synced Facial Performer*. arXiv preprint arXiv:2408.03284. 2024.
- [251] Ye, Zhenhui and Zhong, Tianyun and Ren, Yi and Yang, Jiaqi and Li, Weichuang and Huang, Jiawei and Jiang, Ziyue and He, Jinzheng and Huang, Rongjie and Liu, Jinglin and Zhang, Chen and Yin, Xiang and Ma, Zejun and Zhao, Zhou. *Real3D-Portrait: One-shot Realistic 3D Talking Portrait Synthesis*. arXiv preprint arXiv:2401.08503. 2024.
- [252] Ji, Xiaozhong and Lin, Chuming and Ding, Zhonggan and Tai, Ying and Yang, Jian and Zhu, Junwei and Hu, Xiaobin and Zhang, Jiangning and Luo, Donghao and Wang, Chengjie. *RealTalk: Real-time and Realistic Audio-driven Face Generation with 3D Facial Prior-guided Identity Alignment Network*. arXiv (Cornell University). 2024.

- [253] Tan, Shuai and Ji, Bin and Ding, Yu and Pan, Ye. *Say Anything with Any Style*. arXiv (Cornell University). 2024.
- [254] Shuai Tan and Bin Ji and Ye Pan. *Style2Talker: High-Resolution Talking Head Generation with Emotion Style and Art Style*. arXiv (Cornell University). 2024.
- [255] Zhou, Yingjie and Zhang, Zicheng and Sun, Wei and Liu, Xiaohong and Min, Xiongkuo and Wang, Zhihua and Zhang, Xiao-Ping and Zhai, Guangtao. *Thqa: A Perceptual Quality Assessment Database for Talking Heads*. 2024 IEEE International Conference on Image Processing (ICIP). 2024.
- [256] Hochreiter, Sepp and Schmidhuber, Jürgen. *Long Short-Term Memory*. Neural Computation. 1997.
- [257] Sun, Keqiang and Jourabloo, Amin and Bhalodia, Riddhish and Meshry, Moustafa and Rong, Yu and Yang, Zhengyu and Nguyen-Phuoc, Thu and Haene, Christian and Xu, Jiu and Johnson, Sam and Li, Hongsheng and Bouaziz, Sofien. *GENCA: a text-conditioned generative model for realistic and drivable CODEC avatars*. arXiv (Cornell University). 2024.
- [258] Wenshuo Peng and Kaipeng Zhang and Sai Qian Zhang. *T3M: Text Guided 3D Human Motion Synthesis from Speech*. arXiv (Cornell University). 2024.
- [259] Chai, Zenghao and Tang, Chen and Wong, Yongkang and Kankanhalli, Mohan. *STAR: Skeleton-aware Text-based 4D Avatar Generation with In-Network Motion Retargeting*. arXiv (Cornell University). 2024.
- [260] Wang, Zhichao and Dai, Mengyu and Lundgaard, Keld. *Text-to-Video: a Two-stage Framework for Zero-shot Identity-agnostic Talking-head Generation*. arXiv (Cornell University). 2023.
- [261] Tandon, Pulkit and Chandak, Shubham and Pataranutaporn, Pat and Liu, Yimeng and Mapuranga, Anesu M. and Maes, Pattie and Weissman, Tsachy and Sra, Misha. *Txt2Vid: Ultra-Low Bitrate Compression of Talking-Head Videos via Text*. arXiv preprint arXiv:2106.14014. 2022.
- [262] Ma, Shengjie and Weng, Yanlin and Shao, Tianjia and Zhou, Kun. *3D Gaussian Blendshapes for Head Avatar Animation*. Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers '24. ACM. 2024.
- [263] Yu, Hongyun and Qu, Zhan and Yu, Qihang and Chen, Jianchuan and Jiang, Zhonghua and Chen, Zhiwen and Zhang, Shengyu and Xu, Jimin and Wu, Fei and Lv, Chengfei and Yu, Gang. *GaussianTalker: Speaker-specific Talking Head Synthesis via 3D Gaussian Splattting*. Proceedings of the 32nd ACM International Conference on Multimedia. ACM. 2024.
- [264] Li, Jiahe and Zhang, Jiawei and Bai, Xiao and Zheng, Jin and Ning, Xin and Zhou, Jun and Gu, Lin. *TalkingGaussian: Structure-Persistent 3D talking head synthesis via gaussian splatting*. arXiv (Cornell University). 2024.
- [265] Gerogiannis, Dimitrios and Papantoniou, Foivos Paraperas and Potamias, Rolandos Alexandros and Lattas, Alexandros and Moschoglou, Stylianos and Ploumpis, Stylianos and Zafeiriou, Stefanos. *AnimateMe: 4D Facial Expressions via Diffusion Models*. Lecture Notes in Computer Science. Springer. 2024.
- [266] Nocentini, Federico and Ferrari, Claudio and Berretti, Stefano. *EmoVOCA: Speech-Driven Emotional 3D Talking Heads*. arXiv preprint arXiv:2403.12886. 2024.
- [267] Shivangi Aneja and Justus Thies and Angela Dai and Matthias Nießner. *FaceTalk: Audio-Driven Motion Diffusion for Neural Parametric Head Models*. Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). 2024.
- [268] Sun, Yasheng and Chu, Wenqing and Zhou, Hang and Wang, Kaisiyuan and Koike, Hideki. *AVI-Talking: Learning Audio-Visual instructions for expressive 3D talking face Generation*. arXiv (Cornell University). 2024.
- [269] Daněček, Radek and Chhatre, Kiran and Tripathi, Shashank and Wen, Yandong and Black, Michael and Bolkart, Timo. *Emotional Speech-Driven Animation with Content-Emotion Disentanglement*. SIGGRAPH Asia 2023 Conference Papers. ACM. 2023.
- [270] Sun, Zhiyao and Lv, Tian and Ye, Sheng and Lin, Matthieu Gaetan and Sheng, Jenny and Wen, Yu-Hui and Yu, Minjing and Liu, Yong-Jin. *DiffPoseTalk: Speech-Driven stylistic 3D facial animation and head pose generation via diffusion models*. arXiv (Cornell University). 2023.
- [271] Thambiraja, Balamurugan and Habibie, Ikhwanul and Aliakbarian, Sadegh and Cosker, Darren and Theobalt, Christian and Thies, Justus. *Imitator: Personalized speech-driven 3D facial animation*. arXiv (Cornell University). 2023.

- [272] Haque, Kazi Injamamul and Yumak, Zerrin. *FaceXHuBERT: Text-less Speech-driven E(X)pressive 3D Facial Animation Synthesis Using Self-Supervised Speech Representation Learning*. arXiv (Cornell University). 2023.
- [273] Fang, H., Weng, D., Tian, Z. & Ma, Y. Manitalk: manipulable talking head generation from single image in the wild. *The Visual Computer*. **40**, 4913-4925 (2024,6), <https://doi.org/10.1007/s00371-024-03490-4>
- [274] Dong, B. & Zhang, L. Enhanced temporal representation and spatial alignment for High-Fidelity Talking Video Generation. *The Visual Computer*. (2025,5), <https://doi.org/10.1007/s00371-025-03999-2>
- [275] Rashid, M., Wu, S., Nie, Y. & Li, G. High-fidelity facial expression transfer using part-based local-global conditional gans. *The Visual Computer*. **39**, 3635-3646 (2023,7), <https://doi.org/10.1007/s00371-023-03035-1>
- [276] Cheng, W., Wang, X. & Mao, B. A multi-feature fusion algorithm for driver fatigue detection based on a lightweight convolutional neural network. *The Visual Computer*. **40**, 2419-2441 (2023,6), <https://doi.org/10.1007/s00371-023-02927-6>
- [277] Ying, T., Yazhi, L., Xiong, L. & Wei, L. Adaptive diffusion landmark dynamic rendering for realistic talking face video generation. *The Visual Computer*. (2025,5), <https://doi.org/10.1007/s00371-025-03907-8>