한국컴퓨터그래픽스학회
Korea Computer Graphics Society Vol. 32, No. 2, P. 47~56

포트레이트토커: 참조 이미지 없이 생성하는 텍스트-음성 기반 3D 말하는 얼굴

DU XIAN[1○] 유리[1,2*]

[1]아주대학교 인공지능학과
[2]아주대학교 소프트웨어학과
{duxian, riyu}@ajou.ac.kr

PortraitTalker: Reference-Free 3D Talking Head Generation from Text and Speech

Xian Du[1○] Ri Yu[1,2*]

[1]Department of Artificial Intelligence, Ajou University
[2]Department of Software and Computer Engineering, Ajou University

요 약

텍스트만으로 원하는 외형을 기술하고 음성만으로 자연스럽게 구동되는 디지털 아바타는 가상 에이전트, 교육 콘텐츠, 원격 커뮤니케이션, 디지털 휴먼 제작 자동화에 중요한 기반 기술이다. 그러나 기존 말하는 얼굴 생성 연구는 대개 참조 이미지, 인물별 리깅, 또는 수작업 3D 템플릿에 의존하므로 대규모 아바타 생성과 개인화에 한계가 있다. 본 논문에서는 텍스트 프롬프트와 음성 입력만으로 사실적인 3D 말하는 얼굴을 생성하는 end-to-end 프레임워크 PortraitTalker를 제안한다. 제안 방법은 SDS(score distillation sampling) 기반 텍스트-투-3D 합성 모듈, Transformer 기반 음성 인코더를 이용한 FLAME 파라미터 예측 모듈, 그리고 미분 가능 렌더링 모듈을 통합하여 외형 생성과 발화 애니메이션을 하나의 파이프라인으로 연결한다. HDTF 데이터셋 기반 실험에서 PortraitTalker는 Lip Sync Error Confidence(LSE-C) 7.230, Lip Sync Error Distance(LSE-D) 7.712, FID 21.997을 달성하였으며, 사용자 평가에서도 립싱크 정확도 68.13%, 모션 다양성 76.89%, 영상 선명도 74.06%, 전체 자연스러움 74.76%의 우세한 선호를 보였다. 본 연구는 참조 이미지와 리깅 없이도 확장 가능한 고품질 3D talking avatar 생성이 가능함을 보이며, 텍스트 기반 캐릭터 설계와 음성 구동 애니메이션을 통합하는 실용적 방향을 제시한다.

Abstract

Customizable digital avatars that can be specified by text and animated directly from speech are an important building block for virtual agents, educational media, telepresence, and scalable digital human production. Existing talking-head generation methods, however, usually depend on reference images, manual rigging, or subject-specific 3D templates, which limits scalability and personalization. We present PortraitTalker, an end-to-end framework that generates photorealistic 3D talking heads directly from text prompts and speech input. The proposed pipeline combines SDS-based text-to-3D synthesis, a transformer-based speech encoder that predicts FLAME expression and pose parameters, and a differentiable renderer that produces temporally coherent videos. Experiments on the HDTF dataset show that PortraitTalker achieves an LSE-C of 7.230, an LSE-D of 7.712, and an FID of 21.997. In a user study, the proposed method is preferred in terms of lip-sync accuracy (68.13%), motion diversity (76.89%), video sharpness (74.06%), and overall naturalness (74.76%). These results demonstrate that high-quality 3D talking avatars can be generated without reference images or manual rigging, providing a practical path toward scalable avatar creation.

키워드: 말하는 얼굴 생성, 텍스트-기반 3차원 생성, 음성 구동 애니메이션, 디지털 아바타, 미분가능 렌더링
Keywords: talking head generation, text-to-3D, speech-driven animation, digital avatar, differentiable rendering

*corresponding author: Ri Yu / Department of Artificial Intelligence, Department of Software and Computer Engineering, Ajou University (riyu@ajou.ac.kr)

Received : 2026.04.26./ Review completed : 1st 2026.05.06. / Accepted : 2026.05.14.
DOI : 10.15701/kcgs.2026.32.2.47
ISSN : 1975-7883(Print)/2383-529X(Online)

1 Introduction

Recent progress in generative artificial intelligence has rapidly improved the ability to create images, videos, and 3D assets directly from text prompts. In particular, text-to-3D studies such as DreamFusion, Magic3D, Fantasia3D, and SJC have significantly expanded the feasibility of prompt-driven 3D content creation by combining pretrained 2D diffusion models with differentiable rendering [1, 2, 3, 4]. Along with this trend, there is a growing demand for digital avatars that can not only be automatically created but also naturally animated for speech. Digital avatars are becoming key media interfaces in metaverse platforms, virtual customer service, educational tutoring systems, public guidance services, and character-driven media production [5, 6, 7]. Beyond static character design, practical applications increasingly require systems that can generate an avatar according to a user-specified text prompt and then animate it from arbitrary speech signals in real time or near real time [8, 9, 10].

However, most existing talking-face generation methods impose strong constraints on input conditions. Early landmark- or keypoint-based approaches [11, 12] and later methods such as MakeItTalk, Audio2Head, one-shot correlation learning, and SadTalker [5, 13, 14, 8] typically take a reference image and audio as input to synthesize a talking face video. While these approaches can generate high-quality animated frames with relatively simple inputs, they do not allow users to freely design a new identity, and their output quality is strongly affected by the quality and pose of the reference image. In contrast, text-to-3D methods can generate diverse appearances from prompts, but most of them focus on static 3D asset creation and do not directly address speech animation, facial dynamics, or temporal consistency across frames.

This gap becomes immediately apparent in realistic use scenarios where one wants to "design the appearance from text and animate it from speech." For example, a virtual customer service avatar should be created without requiring a portrait image while still reflecting a desired style, age, or overall visual tone. It should then speak naturally in multiple languages, including Korean, English, Chinese, and Japanese, while preserving consistent identity across the entire animation sequence. In addition, frame flickering, lip mismatch, and unstable facial dynamics must be minimized. These requirements indicate the need to combine 3D morphable face models [15] with 3D-aware rendering and talking head synthesis methods [7, 16, 17].

To address these issues, this paper presents PortraitTalker, a framework for 3D talking head generation from text prompts and speech input only. The key idea is to unify three components within a single pipeline. First, an SDS-based text-to-3D synthesis module creates a photorealistic appearance and texture without requiring any reference image. Second, a transformer-based speech encoder predicts frame-wise FLAME expression and pose parameters for speech-driven animation. Third, a differentiable renderer combines the generated 3D appearance with the time-varying facial parameters to produce temporally coherent videos.

PortraitTalker shows strong quantitative performance on the HDTF dataset and maintains consistent quality across a variety of languages, ages, and regional appearance conditions. The main strength of the work lies in integrating text-based avatar design and speech-driven facial animation into a coherent framework while also reporting both objective metrics and user preference studies. At the same time, computational efficiency, broader comparison with more recent baselines, and stronger support for real-time claims remain important directions for further improvement.

The main contributions of this paper are summarized as follows.

1. We present an integrated framework for generating 3D talking avatars from text and speech without requiring reference images or manual rigging.
2. We organize the pipeline around SDS-based 3D appearance synthesis, transformer-based FLAME parameter prediction, and differentiable rendering, and analyze the role of each component.
3. We report quantitative and user-study results on HDTF, showing strong performance in lip synchronization, visual quality, and perceptual naturalness.

2 Related Work

2.1 Text-to-3D Human and Portrait Generation

Research on text-conditioned 3D generation has grown rapidly with the development of diffusion priors and score distillation optimization. DreamFusion established a representative starting point by showing that the score of a pretrained text-to-image model can guide 3D optimization without paired 3D supervision [1]. Magic3D improved visual fidelity and efficiency through high-resolution supervision and a two-stage mesh optimization process [2]. Fantasia3D further improved geometric detail by disentangling geometry and appearance [3], while SJC interpreted 3D generation as lifting pretrained 2D diffusion models through Jacobian chaining [4]. For portrait-specific generation, Portrait3D introduced identity-aware supervision to improve 3D head quality and identity preservation [18]. Despite these advances, the primary goal of these methods is static 3D content creation rather than temporally coherent speech animation.

prompt. This general strategy has been widely used in DreamFusion, Magic3D, and Fantasia3D [1, 2, 3].

Conceptually, the text-to-3D optimization can be expressed as

$$\mathcal{L}_{\text{3D}} = \mathcal{L}_{\text{SDS}} + \lambda_{\text{reg}}\mathcal{L}_{\text{reg}}, \tag{1}$$

where $\mathcal{L}_{\text{SDS}}$ encourages prompt-consistent renderings and $\mathcal{L}_{\text{reg}}$ represents regularization terms that stabilize geometry and texture. Since the focus here is on the overall modeling framework, the objective is described at a conceptual level rather than through exhaustive implementation-specific hyperparameters.

PortraitTalker uses an animation-ready representation based on orthogonal feature planes or tri-grid structures. Such a representation stores appearance and geometry information compactly while allowing efficient access during the downstream animation stage. For example, a prompt such as "upper body photo, 25 y.o man in casual clothes, night, city street, soft lighting, high quality, film grain" can be translated into a coherent face and upper-body appearance with appropriate age, clothing, scene mood, and lighting style.

3.3 Transformer-Based Speech-Driven Animation

The second module addresses the challenge of capturing the fine-grained, temporally extended dynamics needed to drive a 3D face from speech. To model the nonlinear mapping between audio signals and realistic facial motion, PortraitTalker uses a transformer-based audio encoder to regress FLAME-based expression and pose parameters directly. This design is consistent with recent speech-driven animation approaches such as SadTalker, Neural Voice Puppetry, and Audio2Head, which also model the nonlinear mapping between audio and facial motion using temporally aware networks [8, 19, 13].

Given an input audio sequence $\mathbf{a}_{1:T}$, the model predicts FLAME parameters $\mathbf{y}_t$ at each time step:

$$\mathbf{y}_t = F_{\text{audio}}(\mathbf{a}_{1:T}, t), \tag{2}$$

where $\mathbf{y}_t$ may contain expression coefficients, jaw motion, and head pose. A FLAME-compatible control space makes it easier to interpret and manipulate lip, jaw, cheek, and eye-region motion than purely image-based warping, which in turn improves temporal stability.

Lip synchronization is not only a matter of matching mouth opening timings. It also requires phoneme-level mouth shapes, smooth transitions, and plausible global facial dynamics. Therefore, the audio encoder must capture both local acoustic cues and longer temporal context. A transformer architecture is well suited to this requirement through self-attention, and it is potentially more robust to multilingual speech, prosodic variation, and differences in speaking style.

3.4 Differentiable Rendering and Temporal Consistency

The third module combines the text-generated appearance with the speech-predicted dynamic facial parameters to render the final video. A differentiable renderer combines FLAME-based geometry with hierarchical tri-grid textures. The goal is not simply to render each frame independently, but to preserve a coherent identity throughout the sequence while preventing instability in lighting, texture, and facial outline. This viewpoint is closely related to 3D-aware and explicit 3D talking head methods such as AD-NeRF, face-vid2vid, GaussianTalker, and MGGTalk [17, 7, 9, 10].

The final frame $\mathbf{I}_t$ can be written conceptually as

$$\mathbf{I}_t = R(\mathcal{M}, \mathbf{y}_t, \Theta), \tag{3}$$

where $\mathcal{M}$ denotes the text-generated 3D appearance representation, $\mathbf{y}_t$ is the FLAME parameter vector at time t, and Θ represents camera and rendering settings. A differentiable renderer provides a practical mechanism for maintaining visual realism together with temporal coherence during synthesis.

3.5 Design Perspective

An important design choice of PortraitTalker is that it separates new identity creation from speech animation while still optimizing for the quality of the final video output. Traditional reference-image-based approaches limit the freedom of identity creation [5, 8, 14], whereas text-to-3D methods alone do not solve the animation problem [1, 2, 18]. PortraitTalker provides a practical middle ground between these two extremes.

4 Experimental Setup

4.1 Dataset and Comparison Setting

Experiments are conducted on the HDTF dataset. HDTF is a high-resolution audio-visual benchmark widely used for evaluating lip synchronization, temporal consistency, and perceptual quality in talking face generation [20]. Since PortraitTalker introduces a new task setting where both identity creation and animation are performed without any reference image, no existing method is directly comparable out of the box. To establish a fair and meaningful baseline, we construct a cascaded pipeline for each reference-

based prior method: we first generate a reference portrait image from the same text prompt using a pretrained text-to-image model, and then feed this synthetic image together with the speech input into MakeItTalk [5] and the method of Wang et al. [14]. This adaptation allows both baselines to operate under the same input conditions as PortraitTalker. More importantly, this cascaded setup exposes the structural weakness of delegating identity creation to a 2D image: any imperfection in the generated portrait, such as missing 3D cues, unstable identity, or limited pose coverage, propagates directly into the subsequent animation stage. PortraitTalker, by contrast, circumvents these issues by operating on an explicit 3D representation from the start.

The qualitative results further cover a range of prompt conditions involving language, region, and age. The examples include Korean news speech, a young European male, an Asian male, a muscular adult male, a child, a middle-aged man, and an elderly woman. These examples suggest that the method is not restricted to a narrow identity distribution and can support a broad design space.

4.2 Evaluation Metrics

Objective evaluation uses metrics related to lip synchronization and visual quality. The key metrics reported in the paper are as follows.

- **LSE-C**: Lip Sync Error Confidence, where higher values indicate better synchronization confidence.
- **LSE-D**: Lip Sync Error Distance, where lower values indicate better alignment between speech and mouth motion.
- **FID**: Fréchet Inception Distance, where lower values indicate better visual realism.

The user study was conducted with 20 participants. A total of 50 video clips were generated, covering samples from PortraitTalker, MakeItTalk, and Wang et al. across diverse prompts and speech inputs. Each participant watched the videos in a randomized order and was asked to rate each clip on a five-point Likert scale for four perceptual criteria: lip-sync accuracy, motion diversity, video sharpness, and overall naturalness. Table 2 reports the percentage of responses in which each method received the highest rating. This multidimensional protocol complements objective metrics by capturing perceptual quality that may not be fully reflected by a single numerical score.

Table 1: Quantitative comparison on the HDTF dataset [20]. MakeItTalk [5] and Wang et al. [14] are adapted to the text-and-speech setting via a cascaded text-to-image pipeline (see Section 4). PortraitTalker achieves the best lip-sync confidence (LSE-C), the lowest lip-sync distance (LSE-D), and the best visual quality (FID) among the compared methods.

Method	LSE-C↑	LSE-D↓	FID↓
MakeItTalk [5]	5.051	9.999	28.183
Wang et al. [14]	4.872	9.995	22.372
PortraitTalker (Ours)	**7.230**	**7.712**	**21.997**

5 Results

This section reports the performance of PortraitTalker from both quantitative and qualitative perspectives. We first compare the proposed method with prior talking-head baselines using objective synchronization and visual-quality metrics on the HDTF dataset [20]. We then analyze user preference results and representative visual examples to examine whether the proposed text-to-3D and speech-driven animation framework produces perceptually convincing and diverse talking portraits.

5.1 Quantitative Comparison

Table 1 summarizes the quantitative results. PortraitTalker achieves an LSE-C of 7.230, an LSE-D of 7.712, and an FID of 21.997 on HDTF. Compared against the adapted cascaded pipelines, The LSE-C score improves by about 43.2% over MakeItTalk [5], which reports 5.051, and by about 48.4% over the Wang et al. baseline [14], which reports 4.872. The LSE-D score is reduced from 9.999 and 9.995 to 7.712, indicating more accurate alignment between audio and lip motion. In terms of visual quality, PortraitTalker also outperforms the baselines in FID, improving from 28.183 and 22.372 to 21.997.

These results suggest that the proposed method does not merely generate visually plausible faces, but more accurately synchronizes facial motion with the speech signal. The fact that both LSE-C and LSE-D improve simultaneously indicates that the method achieves a balanced improvement in both temporal alignment and motion quality rather than optimizing only a single aspect of synchronization. The reported FID of 21.997 further indicates strong visual quality.

5.2 User Study

Table 2 summarizes the user study results based on 20 participants and 50 samples. PortraitTalker achieves preference scores of 68.13% for lip-sync accuracy, 76.89% for motion diversity, 74.06% for video sharpness, and 74.76% for overall naturalness.

[8] W. Zhang, X. Cun, X. Wang, Y. Zhang, X. Shen, Y. Guo, Y. Shan, and F. Wang, "Sadtalker: Learning realistic 3d motion coefficients for stylized audio-driven single image talking face animation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 8652–8661.

[9] K. Cho, J. Lee, H. Yoon, Y. Hong, J. Ko, S. Ahn, and S. Kim, "Gaussiantalker: Real-time high-fidelity talking head synthesis with audio-driven 3d gaussian splatting," *arXiv preprint arXiv:2404.16012*, 2024.

[10] S. Gong, H. Li, J. Tang, D. Hu, S. Huang, H. Chen, T. Chen, and Z. Liu, "Monocular and generalizable gaussian talking head animation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025, pp. 5523–5534.

[11] L. Chen, R. K. Maddox, Z. Duan, and C. Xu, "Hierarchical cross-modal talking face generation with dynamic pixel-wise loss," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7832–7841.

[12] A. Siarohin, S. Lathuiliere, S. Tulyakov, E. Ricci, and N. Sebe, "First order motion model for image animation," in *Advances in Neural Information Processing Systems*, vol. 32, 2019, pp. 7137–7147.

[13] S. Wang, L. Li, Y. Ding, C. Fan, and X. Yu, "Audio2head: Audio-driven one-shot talking-head generation with natural head motion," in *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, 2021, pp. 1098–1105.

[14] ——, "One-shot talking face generation from single-speaker audio-visual correlation learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 3, 2022, pp. 2531–2539.

[15] T. Li, T. Bolkart, M. J. Black, H. Li, and J. Romero, "Learning a model of facial shape and expression from 4d scans," *ACM Transactions on Graphics*, vol. 36, no. 6, pp. 194:1–194:17, 2017.

[16] F.-T. Hong, L. Zhang, L. Shen, and D. Xu, "Depth-aware generative adversarial network for talking head video generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 3397–3406.

[17] Y. Guo, K. Chen, S. Liang, Y.-J. Liu, H. Bao, and J. Zhang, "Ad-nerf: Audio driven neural radiance fields for talking head synthesis," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 5784–5794.

[18] J. Hao, J. Tang, J. Zhang, R. Yi, Y. Hong, M. Li, W. Cao, Y. Wang, and L. Ma, "Portrait3d: 3d head generation from single in-the-wild portrait image," *arXiv preprint arXiv:2406.16710*, 2024.

[19] J. Thies, M. Elgharib, A. Tewari, C. Theobalt, and M. Niessner, "Neural voice puppetry: Audio-driven facial reenactment," in *Computer Vision – ECCV 2020*, 2020, pp. 716–731.

[20] Z. Zhang, L. Li, Y. Ding, and C. Fan, "Flow-guided one-shot talking face generation with a high-resolution audio-visual dataset," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 3661–3670.

〈저자소개〉

DU XIAN
- 2022년~2024년 아주대학교 정보통신대학원 석사
- 2024년~현재 아주대학교 인공지능학과 박사과정
- 관심분야: Computer Graphics, Facial Animation, Character Animation
- https://orcid.org/0009-0000-9836-1346

유 리
- 2021년 서울대학교 컴퓨터공학 박사
- 2021년~2022년 서울대학교병원 의생명연구원 연구교수
- 2022년~현재 아주대학교 소프트웨어학과 조교수
- 관심분야: 컴퓨터 그래픽스, 컴퓨터 비전, 사람 동작 재건, 사람 동작 분석, 캐릭터 애니메이션, 사람-환경 상호작용, 디지털 휴먼
- https://orcid.org/0000-0002-2377-8654