ChatGPT

Timeline: Image-to-Video Avatar Generators (2022–Present)

- **June 2022 – DaGAN (CVPR 2022):** Introduction of a *Depth-Aware GAN* for one-shot talking head video generation from a single image [1]. By leveraging depth estimation, DaGAN produces more realistic head rotations and facial movements from still portraits. The authors released their code publicly, marking one of the first open-source talking-head models of 2022 [1].

- **June 2022 – Thin-Plate Spline Motion Model (CVPR 2022):** A warping-based approach using a **Thin-Plate Spline** framework to animate still images [1]. This model improved upon earlier motion-transfer methods by learning smoother 2D deformations, resulting in more natural head and torso movements for talking-head videos. Code was open-sourced, becoming a foundation for subsequent avatar animation pipelines [2].

- **October 2022 – StyleHEAT (ECCV 2022):** A one-shot, high-resolution talking-face generator built on StyleGAN's latent space [3]. StyleHEAT enabled editing of the generated talking head (e.g. changing appearance or expressions) while maintaining lip-sync, by leveraging a pretrained StyleGAN. The project's code was released, demonstrating state-of-the-art fidelity (even megapixel resolution) in 2022's talking-head benchmarks [4].

- **March 2023 – SadTalker (CVPR 2023):** Open-source release of *SadTalker*, an audio-driven single-image animation model producing realistic talking-head videos [5]. SadTalker learns **3D motion coefficients** (head pose and facial expression parameters of a 3DMM) from input audio [6]. It introduced an *ExpNet* to predict accurate facial expressions from audio and a *PoseVAE* to generate natural head movements with style control [6]. By mapping the generated 3D coefficients to a 3D-aware renderer, it achieves synchronized lip movements and expressive, stylized head motion while preserving the person's identity [7]. *(Code and a live demo were provided by the authors [5].)*

- **Mid-2023 – Diffusion-Based Talking Heads:** Emerging diffusion models began to overtake GAN-based methods. For example, **Diffused Heads** (2023 preprint) demonstrated that diffusion models can outperform GANs on talking-face generation [8], achieving higher fidelity and stability. These approaches generate frames via iterative denoising, improving lip-sync and visual quality at the cost of more computation. Open project pages (and later code releases) signaled a shift toward diffusion for avatar video synthesis [8].

- **August 2023 – GeneFace++ (ArXiv 2023):** An improved version of the GeneFace framework by Microsoft Research, designed for **generalized, stable, and real-time** audio-driven 3D talking head generation [9]. It built on prior work to enhance output stability and speed, enabling live talking-head animation. GeneFace++ was released with a research demo and code, offering a template for real-time lip-syncing avatars that maintain high fidelity [9].

- **April 2025 – OmniTalker (Alibaba Tongyi Lab):** Introduction of a unified **text-to-video talking head** model that generates both the speech audio *and* the synchronized talking-head video directly from text [10]. OmniTalker uses a dual-branch **diffusion Transformer** – one branch synthesizes speech (mel-spectrogram) from text, while the other predicts the corresponding head pose and facial motion – with a cross-modal fusion ensuring perfect lip-sync [10] [11]. In a zero-shot setting, it can take a single reference portrait + audio style sample and produce a video of that person speaking arbitrary text, preserving the voice and facial style. Notably, it achieves real-time performance (~25 FPS) [10]. *(Project page and model weights were made available, pioneering end-to-end talking avatar generation.)*

- **April 2025 – ACTalker (Tencent AI Lab):** Release of an audio-visual diffusion framework for talking-head generation that supports **multi-signal control** [12]. ACTalker can be driven by an audio clip, by explicit facial motion parameters, or by both simultaneously. It employs a parallel multi-branch ("mamba") architecture where each driving signal controls specific facial regions, coordinated through a gating mechanism [13]. This design allows mixing of controls (for example, using a speech clip *and* a target expression sequence) without interference, yielding natural synchronized results. The authors open-sourced the code on GitHub [14], making it a state-of-the-art toolkit for controllable talking-head video synthesis.

Timeline: Voice Cloning & Text-to-Speech (TTS) Tools (2022–Present)

- **May 2022 – TorToiSe TTS v2.1:** Open-source release of **TorToiSe** (by J. Betker), a groundbreaking zero-shot multi-voice TTS system. TorToiSe could clone voices from a few seconds of reference audio, producing speech with remarkably natural prosody and intonation. Version 2.1.0 (May 2, 2022) introduced key features like completely *random voice* generation and user-provided voice conditioning latents for one-shot cloning [15]. Under the hood, it combined an autoregressive transformer for coarse speech generation with a diffusion model decoder for high-quality audio output [16] [17]. TorToiSe set a new standard for open TTS, albeit with slower inference, and included an audio detector to mitigate misuse [18].

- **April 2023 – Bark by Suno:** The company Suno AI open-sourced **Bark**, a text-prompted generative audio model that produces highly realistic speech in multiple languages [19]. Bark uses a fully end-to-end transformer pipeline (text → semantic tokens → coarse audio tokens → fine audio) without relying on phoneme transcription [20]. It can capture voice characteristics (tone, accent, emotion) in a zero-shot manner and even generate **non-speech sounds** (music, background noise, laughter) as prompted [21]. Released under MIT License, Bark (April 2023) represented a major advance in open TTS, demonstrating near human-level expressiveness and multilingual support, though the authors initially limited user-provided voice cloning for safety [19] [21].

- **Mid-2023 – F5-TTS ("Fairytaler"):** Developed by researchers at Shanghai Jiao Tong University, **F5-TTS** is a fully non-autoregressive TTS model based on *flow matching* diffusion techniques [22]. First released as a preprint in late 2023, F5-TTS uses a Diffusion Transformer (DiT) to generate speech by treating text and audio as sequences of equal length (avoiding explicit alignments) [22]. It introduced a novel *Sway Sampling* strategy to improve convergence and speech quality, enabling fast inference (Real-Time Factor ≈0.15) [23]. Trained on 100k hours of multilingual data, F5-TTS achieves highly natural, expressive speech with zero-shot voice cloning and even seamless code-switching between languages [23]. The authors open-sourced the code and model checkpoints for the community [24], providing a cutting-edge diffusion-based TTS tool.

- **July 2024 – CosyVoice 1.0 (Alibaba):** Alibaba's Tongyi Lab launched **CosyVoice** as part of their FunAudioLLM open-source project [25] [26]. CosyVoice 1.0 is a large-scale multilingual TTS model with a focus on *zero-shot voice cloning* and expressive speech synthesis. It supports English, Chinese, Cantonese, Japanese, and Korean, among others [26]. With only 3–10 seconds of reference audio, CosyVoice can mimic a target speaker's timbre, rhythm, and even emotional tone – including cross-language cloning (e.g., speaking English in the style of a Chinese speaker) [27]. It also allows fine-grained control of output speech via rich text or natural-language prompts (to adjust emotion, pitch, speed, etc.) [28]. The initial release (summer 2024) provided model weights and code, giving researchers a powerful new multilingual voice cloning tool.

- **Dec 2024 – CosyVoice 2.0 (Alibaba):** An upgraded version of CosyVoice emphasizing **streaming TTS** and improved quality [29]. Announced on Dec 16, 2024, CosyVoice 2.0 introduced an integrated offline/streamable architecture, achieving bidirectional streaming synthesis with as low as 150 ms initial delay [30]. It greatly improved pronunciation accuracy – reducing error rates by 30–50% versus v1.0 – and reached the lowest character error rate on challenging test sets (e.g. tongue twisters, homophones) [31]. The new version also enhanced cross-language consistency (maintaining a speaker's accent/tone in zero-shot cloning across languages) and added more nuanced emotional and accent controls [32]. Speech naturalness saw a boost (MOS up from 5.40 to 5.53, approaching human-level) [33]. Like its predecessor, CosyVoice 2.0 was released openly, solidifying its status as a state-of-the-art open TTS model by late 2024.

- **Sept 2024 – FireRedTTS (Tencent):** Tencent researchers proposed **FireRedTTS**, an industry-scale TTS framework oriented toward *personalized and expressive* speech generation [34] [35]. It combines a *codec language model* approach (using a semantic audio tokenizer and LLM-based generation) with a two-stage neural vocoder [36]. FireRedTTS demonstrated strong *in-context learning* abilities: it can perform zero-shot voice cloning for user-generated content and few-shot adaptation for high-quality studio voices [35]. While initially a research paper (arXiv 2024) with demos, the system laid groundwork for robust open-source voice cloning, and an upgraded streamable version "FireRedTTS-1S" followed with code release in 2025 [37] [38].

- **Oct 2024 – Fish-Speech v1.4/v1.5 (FishAudio):** *Fish-Speech* emerged as a community-driven open TTS model targeting efficient voice cloning. Version 1.4 (late 2024) was trained on ~700,000 hours of multilingual audio and notably ran on consumer GPUs with only 4 GB VRAM [39]. Shortly after, **Fish-Speech v1.5** expanded training to 1+ million hours, further improving quality [40]. These models enable zero-shot voice cloning from 10–30 second samples [41], and support multiple languages out-of-the-box. With an easy setup and low resource requirements, Fish-Speech made high-quality AI voices accessible; its open releases in late 2024 gained popularity as a free alternative for voiceover and content creation.

- **Feb 2025 – IndexTTS (Bilibili AI):** Announcement of **IndexTTS**, an industrial-grade zero-shot TTS system built by Bilibili's AI team [42]. IndexTTS (released as open-source in Feb 2025) is based on large language model techniques and integrates ideas from prior systems (XTTS, Tortoise) with new improvements [43]. Notably, it uses a hybrid text representation (characters + pinyin) for Chinese, enabling controllable pronunciation of polyphonic characters [44]. It also employs a conformer-based acoustic encoder for stronger voice cloning, and replaces the vocoder with a **BigVGAN2** model for high-fidelity output [45]. Compared to contemporary open models (Fish-Speech, CosyVoice2, FireRedTTS, F5-TTS), IndexTTS achieved superior naturalness, content consistency, and faster inference [46]. Demos and code were provided [47], making IndexTTS a cutting-edge open tool for multilingual voice cloning as of 2025.

- **Feb 2025 – Zonos v0.1 Beta (Zyphra):** The startup Zyphra released **Zonos-v0.1** as a pair of 1.6 billion-parameter TTS models under an Apache-2.0 license [48] [49]. Unveiled on Feb 10, 2025, Zonos includes one pure Transformer model and one SSM-based hybrid model, both trained on 200k+ hours of speech to deliver ultra-high quality speech and cloning [48] [50]. The key capability of Zonos is **high-fidelity voice cloning in real-time**, with expressiveness and naturalness matching or exceeding top commercial systems (e.g. ElevenLabs) according to the team's tests [49]. All model weights were openly released on Hugging Face along with inference code [51]. Zonos' release significantly advanced open TTS, providing a ready-to-use, state-of-the-art voice cloning solution to the community.

Sources: The information above is compiled from academic papers, model release notes, and project pages for each tool or model. Key references include conference papers (e.g., CVPR, ECCV, SIGGRAPH) for talking-head methods [5] [4], as well as arXiv preprints and official GitHub repositories for TTS models [23] [45]. Each timeline entry cites representative sources that document the release date, features, and availability of the model (e.g., open-source code or weights).

---

[1] [2] [3] [4] [5] [8] [9] GitHub - harlanhong/awesome-talking-head-generation
https://github.com/harlanhong/awesome-talking-head-generation

[6] [7] SadTalker
https://sadtalker.github.io

[10] [11] OmniTalker: Real-Time Text-Driven Talking Head Generation with In-Context Audio-Visual Style Replication
https://arxiv.org/html/2504.02433v1

[12] [13] Audio-visual Controlled Video Diffusion with Masked Selective State Spaces Modeling for Natural Talking Head Generation
https://arxiv.org/html/2504.02542v2

[14] ACTalker: an end-to-end video diffusion framework for talking head ...
https://github.com/harlanhong/ACTalker

[15] [16] [17] [18] Tortoise TTS | Open Laboratory
https://openlaboratory.ai/models/tortoise

[19] [20] [21] Bark | Open Laboratory
https://openlaboratory.ai/models/bark

[22] [23] [24] F5-TTS: A Fairytaler that Fakes Fluent and Faithful Speech with Flow Matching
https://arxiv.org/html/2410.06885v1

[25] [26] [27] [28] The Latest in Open Source AI from Alibaba's Tongyi Lab: FunAudioLLM - DEV Community
https://dev.to/xidaisme/the-latest-in-open-source-ai-from-alibabas-tongyi-lab-funaudiollm-3ebd

[29] [30] [31] [32] [33] Alibaba Tongyi Laboratory Voice Generation Model CosyVoice Upgraded to Version 2.0
https://www.aibase.com/news/www.aibase.com/news/13976

[34] [35] [36] [2409.03283] FireRedTTS: A Foundation Text-To-Speech Framework for Industry-Level Generative Speech Applications
https://arxiv.org/abs/2409.03283

[37] FireRedTeam/FireRedTTS: An Open-Sourced LLM ... - GitHub
https://github.com/FireRedTeam/FireRedTTS

[38] FireRedTTS-1S: An Upgraded Streamable Foundation Text-to ...
https://arxiv.org/abs/2503.20499

[39] New Open Text-to-Speech Model: Fish Speech v1.4 - Reddit
https://www.reddit.com/r/LocalLLaMA/comments/1fe7fz7/new_open_texttospeech_model_fish_speech_v14/

[40] fishaudio/fish-speech-1.5 - Hugging Face
https://huggingface.co/fishaudio/fish-speech-1.5

[41] fishaudio/fish-speech: SOTA Open Source TTS - GitHub
https://github.com/fishaudio/fish-speech

[42] [43] [44] [45] [46] [47] IndexTTS: An Industrial-Level Controllable and Efficient Zero-Shot Text-To-Speech System | Papers With Code
https://paperswithcode.com/paper/indextts-an-industrial-level-controllable-and

[48] [49] [50] [51] Zyphra
https://www.zyphra.com/post/beta-release-of-zonos-v0-1