ChatGPT

Benchmark Data for Generative AI Tools (2022–2025)

Image-to-Video Avatar Generators

Tool	**Speed** (FPS)	**Inference Latency**	**Hardware (GPU/VRAM)**	**Model Size**	**Output Video Res.**	**Quality Metrics**
SadTalker (2023)	~10 FPS at 512×512 (RTX 4090) [1]	≈0.1 s per frame (optimized pipeline) [1]	NVIDIA GPU (e.g. RTX 4090); model ~1 GB, supports 256 or 512 px [2]	~1 GB on disk (safetensors) [2]	256×256 or 512×512 [2]	High-fidelity frames but less natural motion dynamics [3] (fidelity ↑, realism of head movement ↓)
DaGAN (2022)	*Not reported* (single-pass GAN; likely real-time)	*N/R* (feed-forward per frame)	NVIDIA GPU recommended (trained on 8× RTX 3090) [4]	N/R (CNN + depth network)	~256×256 (face crops)	**CSIM**≈0.723, **PRMSE**≈2.33, **AUCON**≈0.873 on test set (identity similarity ↑, pose error ↓, expression sync ↑ vs prior work) [5]
GeneFace++ (2023)	~23.5 FPS (3090 Ti) [6]	≈0.042 s per frame (real-time) [6]	1× RTX 3090 Ti (24 GB) used in tests [7]	N/R (NeRF-based multi-module)	~512×512 (head/torso)	**FID**≈29.1 (best, ↓), **LMD**≈3.78, **SyncNet**≈6.11 (lip-sync close to GT) [6]; PSNR~31.2 (high fidelity)

Voice Cloning / TTS Models

Model	**Speed** (RTF)	**Inference Latency**	**Hardware (GPU/ VRAM)**	**Model Size**	**Output**	**Quality Metrics**
TorToiSe TTS (2022)	0.25–0.30× (faster than real-time) [8]	<0.5 s latency with streaming enabled [8]	NVIDIA GPU (≥4 GB VRAM) required [9]	~420 M (AR model) + diffusion/ vocoder (>500 M total) [10]	22 kHz audio	Near-human speech fidelity (outperforms prior TTS in realism) [11], albeit originally very slow (now greatly optimized)
CosyVoice 2 (2024)	*Supports streaming* (bidirectional); ~150 ms first audio [12]	~0.15 s initial response (stream mode) [12]	Tested on NVIDIA L40S (data center GPU); ~5 s per sentence on that hardware [13]	≈0.5 B parameters (LLM-based) [14]	24 kHz audio (multi-lingual)	MOS ≈5.53 (↑ from 5.4, 1–10 scale) [15]; lowest CER on challenging dataset (high intelligibility) [16]; high speaker similarity [17]
IndexTTS 1.5 (2025)	N/R (2× faster than CosyVoice2 in eval) [18]	– (200 samples in ~397 s vs 805 s for CosyVoice2) [18]	Single GPU (28% utilization in tests vs ~48% for CosyVoice2) [18]	N/R (GPT-style multi-module pipeline)	24 kHz audio (EN/ZH)	MOS ~4.01/5.0 (zero-shot voice cloning; vs ~3.81 for CosyVoice2) [19]; SOTA content consistency & speaker similarity (objective WER/SS) [17] [20]

Model	Speed (RTF)	Inference Latency	Hardware (GPU/ VRAM)	Model Size	Output	Quality Metrics
Zonos v0.1 (2023)	*Real-time* (≈1.0× RTF, on GPU) [21]	Low latency (designed for instantaneous TTS)	Runs on ≥8 GB VRAM (GPU) [22]	~1.6 B parameters [21] (Transformer or hybrid)	24 kHz audio (multi-lingual)	High-quality, expressive speech generation (voice cloning from 10 s sample) [23]; results described as "studio-quality" and convincing by users [23]
Bark (2023)	~1.25× (slower than real-time) [24]	~1.25 s per 1 s audio (autoregressive) [25]	1 B param model; requires GPU for reasonable performance [24] [26]	~1 B parameters [24]	24 kHz audio (multi-modal, incl. music & FX)	Very natural and expressive speech (MOS≈3.3/5 for naturalness) [27] but sometimes inconsistent voice identity (speaker similarity can drift) [28] [27]

Combined Comparison Table

Category	Tool/Model	Speed	Latency	Model Size	Hardware (GPU/VRAM)	Output	Quality (Key Metrics)
Avatar Gen	**SadTalker** (2023)	~10 FPS @512px [1]	~0.1 s/ frame [1]	~1 GB (weights) [2]	NVIDIA GPU (tested on RTX 4090) [2]	256–512² video	High fidelity video, but less natural head dynamics [3].

Category	Tool/Model	Speed	Latency	Model Size	Hardware (GPU/VRAM)	Output	Quality (Key Metrics)
Avatar Gen	**DaGAN** (2022)	N/R (≈real-time on GPU)	– (feed-forward GAN)	N/R (CNN + depth)	NVIDIA GPU (used 8× RTX 3090 train) [4]	~256² video	CSIM 0.723, PRMSE 2.33, AUCON 0.873 (strong identity & expression accuracy) [5].
Avatar Gen	**GeneFace++** (2023)	~23.5 FPS [6]	~0.04 s/frame [6]	N/R (NeRF-based)	1× RTX 3090 Ti (24 GB) [7]	~512² video	FID 29.1 (best) & LMD 3.78 (low) [6]; sync score 6.11 (near GT) [6].
Voice TTS	**TorToiSe** (2023)	0.25–0.30 RTF [8]	<0.5 s (streaming) [8]	>500 M params (AR+diff.) [10]	NVIDIA GPU (≥4 GB VRAM) [9]	22 kHz audio	Near-human speech realism (outperforms prior models) [11].
Voice TTS	**CosyVoice 2** (2024)	Streaming (150 ms init) [12]	~0.15 s initial [12]	~0.5 B params [14]	GPU (e.g. L40S: ~5 s per sentence) [13]	24 kHz audio	MOS 5.53 (↑ from 5.4) [15]; lowest CER (high intelligibility) [16].
Voice TTS	**IndexTTS** (2025)	N/R (2× faster vs CosyV2) [18]	– (200 utts in 397 s) [18]	N/R (~multi-module)	Single GPU (28% util vs CosyV2 48%) [18]	24 kHz audio	MOS ~4.01/5 (zero-shot cloning, tops peers) [19]; SOTA speaker sim.
Voice TTS	**Zonos** (2023)	~1.0 × RTF (real-time) [21]	Low (interactive)	~1.6 B params [21]	≥8 GB VRAM (GPU) [22]	24 kHz audio	“Studio-quality” expressive TTS; high similarity voice cloning [23].

Category	Tool/Model	Speed	Latency	Model Size	Hardware (GPU/VRAM)	Output	Quality (Key Metrics)
Voice TTS	**Bark** (2023)	~1.25 × RTF (slow AR) [24]	~1.25 s/sec audio [25]	~1 B params [24]	High-end GPU (required for speed) [26]	24 kHz audio	Very natural prosody (MOS ≈3.3/5) [27] but occasional speaker drift [28].

<small>**Legends:** FPS – frames per second; RTF – real-time factor (synthesis time / audio duration); CSIM – cosine similarity (identity); PRMSE – pose error; AUCON – Action Unit accuracy; FID – Fréchet Inception Distance; LMD – Landmark Dist. (lip sync error); MOS – Mean Opinion Score; CER – Character Error Rate. Lower ↓ = better, higher ↑ = better.</small>

1 I can generate 512x512 images in 95ms per image.(4090) · OpenTalker SadTalker · Discussion #685 · GitHub
https://github.com/OpenTalker/SadTalker/discussions/685

2 Releases · OpenTalker/SadTalker · GitHub
https://github.com/OpenTalker/SadTalker/releases

3 ecva.net
https://www.ecva.net/papers/eccv_2024/papers_ECCV/papers/05783.pdf

4 GitHub - harlanhong/CVPR2022-DaGAN: Official code for CVPR2022 paper: Depth-Aware Generative Adversarial Network for Talking Head Video Generation
https://github.com/harlanhong/CVPR2022-DaGAN

5 Depth-Aware Generative Adversarial Network for Talking Head Video Generation
https://openaccess.thecvf.com/content/CVPR2022/papers/Hong_Depth-Aware_Generative_Adversarial_Network_for_Talking_Head_Video_Generation_CVPR_2022_paper.pdf

6 7 arxiv.org
https://arxiv.org/pdf/2305.00787

8 9 GitHub - neonbjb/tortoise-tts: A multi-voice TTS system trained with an emphasis on quality
https://github.com/neonbjb/tortoise-tts

10 Tortoise TTS decoded. I have been using Tortoise TTS for... | by shashank Jain | Medium
https://medium.com/@jain.sm/tortoise-tts-decoded-ff12871be432

11 [2305.07243] Better speech synthesis through scaling
https://ar5iv.org/abs/2305.07243

12 13 15 16 chenxwh/cosyvoice2-0.5b – Run with an API on Replicate
https://replicate.com/chenxwh/cosyvoice2-0.5b

14 EmoVoice: LLM-based Emotional Text-To-Speech Model ... - arXiv
https://arxiv.org/html/2504.12867

17 [2407.05407] CosyVoice: A Scalable Multilingual Zero-shot Text-to-speech Synthesizer based on Supervised Semantic Tokens
https://ar5iv.org/pdf/2407.05407

18 19 20 [2502.05512] IndexTTS: An Industrial-Level Controllable and Efficient Zero-Shot Text-To-Speech System
https://ar5iv.org/abs/2502.05512

21 22 23 Zonos, the easy to use, 1.6B, open weight, text-to-speech model that creates new speech or clones voices from 10 second clips : r/LocalLLaMA
https://www.reddit.com/r/LocalLLaMA/comments/1irhttv/zonos_the_easy_to_use_16b_open_weight/

24 25 26 27 28 Evaluating Text-to-Speech Synthesis from a Large Discrete Token-based Speech Language Model
https://arxiv.org/html/2405.09768v1