Interactive Real-Time Visual Avatar Generation: A Survey

Arnab Dey Marcel Alcoverro Itziar Zabaleta Carla Fernández
Unith Research Lab - research@unith.ai

Figure 1. Conceptual illustration of the five methodological families surveyed (left to right): parametric mesh, GAN-based portrait synthesis, NeRF/triplane based neural rendering, 3D Gaussian Splatting, and diffusion/flow-matching denoising.

Abstract

Photorealistic, audio-driven digital avatars have moved from offline research prototypes to real-time interactive production systems in the last few years, driven by the convergence of neural rendering, generative modelling, and parametric body representations. We review 100 methods published between 2023 and 2026, grouped into five families: GAN-based synthesis, Neural Radiance Fields (NeRF) and triplane representations, 3D Gaussian Splatting (3DGS), diffusion and flow-matching models, and hybrid parametric full-body representations. Methods are compared along four production axes rather than within a single representation family: rendering latency, anatomical coverage (head-only *vs.* upper-body *vs.* full-body), personalisation paradigm (generalizable *vs.* subject-specific), and input modality. A central focus is the *head-body integration problem*: how to combine the FLAME head model and the SMPLX body model despite their geometric and topological mismatch, and recent hybrid representations that claim to resolve it, in particular the Expressive Human Model (EHM) [1]. Three findings stand out. 3DGS coupled with 3DMM-driven expression control is now the dominant paradigm for real-time production deployment, reaching sub-50 ms frame latency with near-photorealistic quality. Flow matching is emerging as a preferred alternative to score-based diffusion for quality-critical, offline generation, with growing adoption among recent state-of-the-art systems. And generalizable one-shot methods are closing the quality gap with subject-specific fine-tuning. We close by investigating the open challenges that remain temporal identity drift, co-speech gesture synthesis, and scalable personalisation from limited data and by providing concrete research directions toward production-deployable interactive avatars.

Keywords: Talking head generation, 3D Gaussian splatting, audio-driven avatars, neural rendering, flow matching, parametric body models

1. Introduction

Generating photorealistic interactive digital avatars is a long-standing goal at the intersection of computer vision, computer graphics, and speech processing. Early work in this area focused on lip synchronisation from audio, treating the problem as a texture-based in-painting task confined to the lower part of the face. Over the past decade, advances in generative adversarial networks (GANs), neural radiance fields, differentiable rendering, and large-scale video pre-training have steadily expanded the scope of what is possible, from static talking heads to fully animatable, full-body avatars that can be driven in real time.

The practical significance of this progress is considerable. Applications now span virtual assistants, AI-powered presenters, video dubbing, telepresence, immersive gaming, digital healthcare companions, and interactive customer-service agents. In each of these settings the underlying system must satisfy three hard constraints simultaneously: **real-time rendering** (typically <50 ms per frame to maintain conversational responsiveness), faithful **identity preservation** across extended sequences so that the avatar remains recognisably the same person, and **expressive naturalness**, encompassing not only accurate lip synchronisation but also head micro-movements, eye gaze, blinking, brow dynamics, emotional affect, and upper-body gesture which are essential for engaging and realistic avatar interactions. Meeting all three at once remains an open research problem: optimising for speed typically sacrifices visual fidelity, while high-quality generative models often exceed real-time budgets by an order of magnitude. We compare these competing demands across three axes, speed, identity fidelity, and expressiveness, in Figure 2, which reveals how each method family navigates the trade-off space: 3DGS currently achieves the broadest coverage, while

Figure 2. Radar chart comparing four talking head synthesis paradigms, 3D Gaussian Splatting (3DGS), Diffusion/Flow, NeRF/Triplane, and GAN-based methods, across identity preservation, expressiveness (lip sync, gaze, blink, gesture), and rendering speed. Each polygon represents a method family; larger area indicates stronger overall performance.

diffusion and flow-matching models lead on expressiveness at the cost of real-time throughput.

Many production systems still rely on GAN-based mouth synthesis in the style of Wav2Lip [2], which composites a generated mouth region onto a source video frame. This paradigm is lightweight, easy to integrate, and temporally stable, and has carried the bulk of deployed conversational avatars to date. Documented limitations, such as boundary blurriness, limited holistic facial dynamics, and no upper-body coverage, motivate the active migration of mature products toward 3DGS- and diffusion-based pipelines that this survey charts.

The research community has responded along three complementary fronts (Figure 4). **(i) Explicit 3D representations.** 3D Gaussian Splatting (3DGS) [3] represents scenes as collections of explicit Gaussian primitives that can be rasterised in real time at 67–150 FPS on consumer GPUs. By anchoring Gaussians to the vertices of parametric face models such as FLAME [4], recent methods achieve expression-controllable rendering with sub-50 ms latency, making 3DGS the fastest-growing backbone for interactive avatar deployment. **(ii) High-fidelity generative models.** Diffusion-based architectures and their flow-matching successors produce fine-grained detail (skin pores, hair strands, subtle lighting) that explicit representations miss. The iterative denoising step is the bottleneck: current systems sit at offline or quality-critical applications such as film post-production and dubbing. **(iii) Unified parametric body models.** Hybrid representations that combine FLAME heads with SMPLX [5] bodies address a critical yet under-discussed bottleneck, namely the geometric and topological incompatibility between the two meshes at the neck seam, enabling both facial fidelity and full-body coverage within a single coherent model. This line of work, exemplified by the Expressive Human Model (EHM) [1], opens the door to avatars that gesture, emote, and speak as a unified whole rather than a stitched composition of independently animated parts.

Table 1. Relation to prior surveys. ✓/×/∼denote covered/not covered/partial coverage. RT: real-time deployment as a primary evaluation axis. 3DGS: explicit treatment of 3D Gaussian Splatting. H–B: head–body integration (*e.g.* FLAME–SMPLX seam).

Survey	Year	Scope	RT	3DGS	H–B
Gowda *et al.* [6]	2023	Talking-head taxonomy	×	×	×
Nyatsanga *et al.* [12]	2023	Co-speech gesture	×	×	∼
Chen & Wang [10]	2024	3DGS fundamentals	✓	✓	×
Pei *et al.* [9]	2024	Deepfake gen. & det.	×	×	×
Jiang *et al.* [7]	2024	Audio facial animation	×	×	×
Wang *et al.* [11]	2024	3D human avatars	×	✓	∼
Rakesh *et al.* [8]	2025	Multi-modal THG	×	×	×
Ours	2026	Audio avatars, 5 families, EHM	✓	✓	✓

Relation to prior surveys. Recent surveys cover slices of this literature but not the full trade-off between speed, fidelity, and expressiveness across all five representation families (illustrated in Figure 1). Gowda *et al.* [6] classify talking-head methods by input modality, predating the 3DGS wave; Jiang *et al.* [7] focus on audio-driven facial image and mesh animation; Rakesh *et al.* [8] catalogue methodologies, datasets, and losses but do not target real-time deployment; Pei *et al.* [9] include talking-face as one sub-problem within a deepfake benchmark. On the representation side, Chen and Wang [10] cover 3DGS fundamentals without audio conditioning, and Wang *et al.* [11] include 3DGS in a broader human-avatar survey that again skips speech-driven deployment. Co-speech gesture is reviewed by Nyatsanga *et al.* [12], treating the body in isolation from the face.

Three gaps remain (Table 1): real-time deployment is rarely a primary axis; 3DGS is absent from the talking-face surveys that predate it; and head–body integration (Section 4) falls between the head- and body-only literatures rather than addressing the FLAME–SMPLX seam itself. This survey covers 2023–2026 across all five families, evaluates each method against real-time constraints, and devotes a dedicated section to EHM-based head–body integration.

Our contributions are as follows:

- We present a **systematic taxonomy** of approximately 50 key methods, drawn from ∼100 screened candidates, organised into five methodological families: GAN-based mouth synthesis, NeRF and triplane representations, 3D Gaussian Splatting, diffusion and flow-matching models, and hybrid

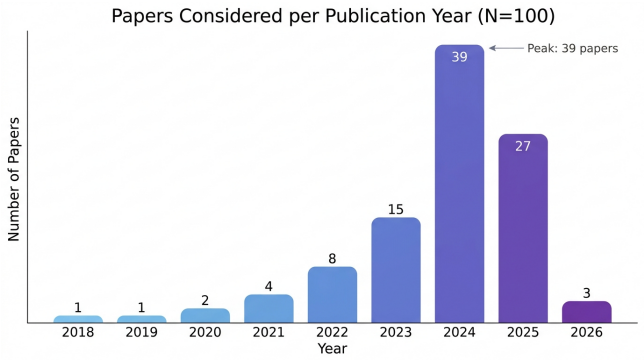

Figure 3. Distribution of the ∼100 candidate papers screened for this survey by publication year; approximately 50 are covered in depth. The sharp increase from 2023 onward reflects the rapid growth of audio-driven avatar research, with 2024–2025 accounting for over 65% of all screened works.

parametric full-body systems. For each family we analyse the core architectural design, training methodology, and characteristic trade-offs.

- We provide the first **formal treatment of the head-body integration problem**: the geometric, topological, and rendering discontinuities that arise when combining separate head and body models, and present an architectural analysis of recent solutions, centred on the Expressive Human Model (EHM) introduced by GUAVA [1].
- We compile a **fact-corrected comparative analysis** covering real-time capability, anatomical coverage (head-only *vs.* upper-body *vs.* full-body), personalisation paradigm (generalizable *vs.* subject-specific), and code availability for 30 representative methods, enabling direct cross-family comparison along axes that govern production viability.
- We identify **open challenges** including temporal identity drift, co-speech gesture synthesis, scalable personalisation from limited data, and the absence of standardised evaluation protocols, and outline concrete research directions toward the next generation of production-deployable interactive avatars.

2. Preliminaries

2.1. Problem Definition

The core task is speech-driven portrait animation: given a speech audio waveform $\mathbf{A} \in \mathbb{R}^{T \times F}$ and a reference identity $\mathcal{I}$ (a short video or single image), synthesise a video sequence $\hat{\mathbf{V}} = \{\hat{I}_t\}_{t=1}^{T}$ whose lip motion, facial expression, and head pose are temporally synchronised with the audio and the context of the speech:

$$G : (\mathbf{A}, \mathcal{I}) \longrightarrow \hat{\mathbf{V}}. \tag{1}$$

A deployable generator faces three competing constraints, the radar chart of Figure 2: (i) ***speed***: each frame within $<$40 ms ($\geq$25 FPS), and typically $<$20 ms for conversational applications; (ii) ***identity fidelity***: the output preserves the appearance of $\mathcal{I}$ across all frames without drift; and (iii) *expressiveness*: motion goes beyond lip sync to include natural blinks, brow raises, and head dynamics. 3DGS currently leads across all three axes, though each family brings distinct strengths; the remainder of the survey is organised around how each navigates these trade-offs.

2.2. Input Modalities

The choice of input modality determines which axes of the performance radar are easiest to satisfy. **Audio-driven** systems, the backbone of virtual assistant and dubbing applications, synthesise lip movements, facial expressions, and head motion directly from a speech waveform, offering the most natural path to expressiveness but requiring learned audio-to-motion mappings that are difficult to generalise across speakers and inherently lack rich expression information. **Text-driven** frameworks use a TTS model to generate an audio clip, which is then used to drive a standard talking head generation model, adding 100–300 ms of synthesis latency in exchange for fully automated text-to-video pipelines.

Video-driven (reenactment) methods transfer the full performance of a driving actor to a target identity, achieving high expressiveness by construction but demanding a source performance at inference time which is not available for most applications. **Image-driven** (one-shot) approaches animate a single portrait from an audio or motion signal: the most deployment-friendly setting, yet the hardest for identity preservation, since the model must fill in unseen views and expressions from a single reference. This survey focuses primarily on audio- and image-driven methods, as their combination defines the most demanding operating point in the radar chart.

2.3. Audio Representations

The choice of audio encoder determines what information is available to drive lip and expression motion. Three families dominate the surveyed methods.

Spectral features (mel spectrograms, MFCCs) are computed directly from the raw waveform without learned speech structure. They are fast and language-agnostic but carry no semantic phoneme identity, making generalisation across speakers and accents harder.

Self-supervised speech representations (HuBERT [13], Wav2Vec 2.0 [14]) learn discrete pseudo-phoneme clusters from unlabelled speech. Their rich content representations transfer well to unseen speakers and languages, and most 3DGS and flow-matching methods in this survey use HuBERT features as the audio backbone. The trade-off is a fixed encoder frozen at training that cannot be fine-tuned without additional speech data.

Audio-visual representations (AV-HuBERT [15]) learn jointly from lip video and audio, providing synchrony-aware features that are more robust to varying head pose than audio-only encoders. Recent analysis [16] confirms that AV-HuBERT-based sync metrics are more stable than SyncNet-derived LSE-D/LSE-C scores, which degrade under non-frontal pose. Production systems that require low LSE-D at arbitrary head angles should prefer AV-HuBERT condition-

ing.

The lack of a standardised audio encoder choice across published methods complicates cross-method comparison: two methods reporting similar FPS and PSNR may use Mel-Spec vs. HuBERT conditioning, making their expressiveness results incompatible without controlling for encoder quality.

2.4. Parametric Face and Body Models

Most methods surveyed here do not generate pixels directly from audio; instead they predict the parameters of a *parametric face model* that provides a compact, disentangled geometric scaffold for rendering. The classical 3D Morphable Model (3DMM) represents a face as a linear combination of identity and expression basis vectors learned from 3D scans, enabling lightweight fitting and real-time coefficient regression, but its fixed linear basis limits the range of reproducible expressions.

FLAME [4] is the parametric head model used by most audio-driven avatar work. A FLAME face has 5,023 vertices, 4 articulated joints (neck, jaw, two eyeballs), and up to 100 principal expression components, enough for fine-grained lip, brow, and eyelid control. The vertex set is small enough that 3DGS methods (Section 3.3) bind Gaussian primitives directly to it.

Extending generation beyond the head requires **SMPLX** [5], a full-body model with 10,475 vertices and 54 joints that covers the torso, arms, and articulated hands. However, SMPLX's facial expression space is coarser than FLAME's, having been trained on whole-body rather than head-specific scans. The two models were designed independently and are geometrically incompatible: naively compositing a FLAME-driven head onto a SMPLX body produces visible seam artefacts and disconnected head motion. Section 4 treats this incompatibility in detail.

2.5. Evaluation Metrics

Quantitative evaluation maps onto the three radar axes. **Visual quality** (fidelity axis) is measured by PSNR for pixel-level reconstruction (production targets $>$30 dB), SSIM for structural similarity (0.85+ for high quality), LPIPS for perceptual distance (lower is better; $<$0.15 is near-photorealistic), and FID for distributional realism.

Lip synchronisation (expressiveness, mouth-level) is captured by Landmark Distance (LMD) over 20 lip keypoints and the SyncNet-based Lip Sync Error distance and confidence (LSE-D, LSE-C) [2]. Recent analysis [16] shows the SyncNet metrics are unstable under varying head pose and audio encoding; AV-HuBERT-based [15] alternatives are more robust. **Identity preservation** uses ArcFace cosine similarity (CSIM) [17]; high-quality methods reach CSIM$>$0.85. **Expression accuracy** uses Action Unit Hamming Distance (AU-H) and Average Expression Distance (AED) between predicted and ground-truth 3DMM coefficients. **Efficiency** (speed axis) covers FPS, pipeline latency, peak VRAM, and per-subject preprocessing time.

The **THEval framework** [16] addresses holistic evaluation: it benchmarks 17 models on 85,000 generated videos and produces a composite score with Spearman $\rho = 0.870$ against human ratings, substantially higher than any single metric. The **THQA database** [18] adds the first dedicated perceptual benchmark for AI talking heads (800 videos with subjective MOS scores). Standard test sets include HDTF [19], VoxCeleb2, and MEAD [20].

Table 2. General pain points in audio-driven avatar generation, mapped to the performance radar axes.

Pain Point	Triangle Axis	Technical Challenge
Visual Quality	Fidelity	Blurring, temporal flicker, teeth/tongue artefacts
Real-Time Latency	Speed	Per-frame budget $>$40 ms on consumer GPUs
Lip Sync Accuracy	Expressiveness	Audio-visual misalignment, especially on unseen speakers
Beyond-Lip Expr.	Expressiveness	No brow, blink, gaze, or emotional affect
Identity Preserv.	Fidelity	Drift across long sequences and unseen poses
Head & Body Motion	Expressiveness	Static/looping head; no torso or gestures
Generalisation	Speed + Fidelity	Per-subject training (30 min–hours); one-shot quality gap

2.6. Technical Pain Points

The performance radar's three-way tension manifests as concrete production failures when deployed systems fall short on any axis. Synthesising the failure modes reported in the literature surveyed in Section 3 with insights from operating interactive avatars in production, we identify seven recurring technical pain points that the field as a whole continues to address (Table 2), each traceable to a gap on one or more radar axes: visual quality and personalisation failures reflect insufficient identity fidelity; limited expressiveness and static avatars reflect a narrow motion space; and latency violations directly breach the speed constraint. The absence of hand gestures and full-body rendering represents the frontier beyond the radar, where the field is expanding from head-only to holistic avatar generation.

3. Methodology Families

As illustrated in the performance radar (Figure 2), each family has a distinct performance profile: 3DGS achieves the broadest coverage across all three axes, while diffusion and flow-matching models lead on expressiveness, NeRF and triplane methods on identity, and GAN-based approaches offer the lightest computational footprint.

3.1. GAN-Based and Classical Methods

In the performance radar (Figure 2), GAN-based methods occupy the smallest polygon overall: they offer real-time inference at low computational cost, but trail on both identity fidelity and holistic expressiveness.

The earliest talking head systems used Generative Adversarial Networks (GANs) to synthesise the mouth region in sync with audio, compositing the result onto a source video. Wav2Lip [2] (ACM Multimedia 2020) remains the most

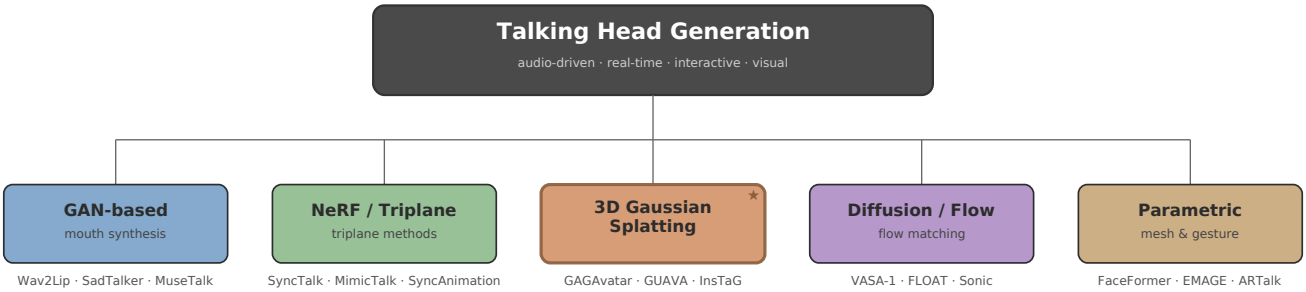

Figure 4. Taxonomy of the five methodological families for audio-driven interactive avatar generation, with representative methods per family. The 3DGS branch has emerged as the dominant paradigm for real-time deployment (2024–2025).

widely deployed baseline, though architecturally obsolete: it generates only a 96×96 mouth region with no head motion, no blinking, and no expression variation, and requires post-processing through face-restoration networks such as GFPGAN and CodeFormer for acceptable quality.

SadTalker [21] (CVPR 2023) represented the first meaningful upgrade, using 3DMM coefficients to decouple head pose from expression and producing full-face animation with head motion from a single portrait, though its linear 3DMM basis still yields stiff expressions and is not real-time. PC-AVS [22] (CVPR 2021) introduced implicit modular audio-visual representations with disentangled pose control. VideoReTalking [23] (SIGGRAPH Asia 2022) proposed a three-stage pipeline of expression canonicalisation, lip-sync, and face enhancement. MuseTalk [24] adopts a VAE-based encoder achieving $\sim$30 FPS at 256×256, though it retains the mouth-region-only paradigm.

Overall, GANs are increasingly retained as discriminators within newer 3DGS-based frameworks rather than as primary generators. This trend is exemplified by Tavus's Phoenix-2 system, which performed a drop-in replacement of its NeRF backbone with 3DGS rendering and, on the vendor's own report, crossed the real-time threshold [25].

Training objectives.. GAN-based talking-head systems are trained with three loss families. An *adversarial loss* ($\mathcal{L}_{\text{adv}}$) — typically a patch-GAN or multi-scale discriminator — drives perceptual realism but is the primary source of training instability. A *sync loss* ($\mathcal{L}_{\text{sync}}$) penalises audio-visual misalignment, most commonly via a frozen SyncNet discriminator as in Wav2Lip [2]. A *perceptual/identity loss* ($\mathcal{L}_{\text{id}}$) uses a pretrained face recognition network (*e.g.* ArcFace [17]) to preserve appearance across frames. Most GAN methods combine all three, weighting sync and identity losses to counter the adversarial term's tendency to hallucinate plausible-but-wrong mouth shapes. This three-way loss balance was a key architectural insight that later families (3DGS, diffusion) absorbed in modified form.

3.2. NeRF and Triplane Methods

NeRF and triplane methods excel on the *identity* axis of the performance radar (Figure 2): they preserve novel-view photorealism and fine facial detail, but pay for it in render time. NeRF itself takes seconds per frame; triplane representations (implicit features on three orthogonal planes) retain that quality at lower cost and underlie most of the NeRF-derived talking-head work below.

AD-NeRF [26] (ICCV 2021) pioneered the direct coupling of NeRF to audio. SyncTalk [27] (CVPR 2024) introduced a three-component synchronisation framework (Face-Sync, Head-Sync, Portrait-Sync) combining NeRF with 3DMM guidance. SyncAnimation [28] jointly synthesises audio-synchronised upper-body, head, and lip shapes at 41 FPS on RTX 4090, making it the first real-time triplane method with upper-body coverage. MimicTalk [29] (NeurIPS 2024) exploits a pre-trained person-agnostic triplane model and adapts to a specific identity in $\sim$15 minutes from $\sim$20 seconds of video, $47\times$ faster than traditional per-subject methods, via an In-Context Stylised Audio-to-Motion (ICS-A2M) module.

Volumetric rendering cost is the structural ceiling on this family; the explicit-representation methods in Section 3.3 bypass it.

3.3. 3D Gaussian Splatting Methods

3DGS methods represent the best current balance in the performance radar (Figure 2), combining real-time speed with strong identity preservation; expressiveness, while competitive, remains the axis with the most headroom for improvement.

3DGS [3] represents scenes as explicit anisotropic 3D Gaussian primitives rendered via differentiable splatting (Figure 5), providing GPU-accelerated rasterisation at 60–150 FPS with photorealistic quality. Within the talking head community, 3DGS methods split into two sub-families distinguished by their driving signal and personalisation strategy.

3.3.1. 3DMM-Driven 3DGS (Generalizable)

These methods bind Gaussian primitives to FLAME [4] (5,023 vertices; 4 joints: neck, jaw, two eyeballs; 100 expression components). FLAME's structured expression space makes the methods *generalizable*: one trained model animates any identity from a single image with no per-subject optimisation.

GAGAvatar [30] (NeurIPS 2024, MIT licence) established

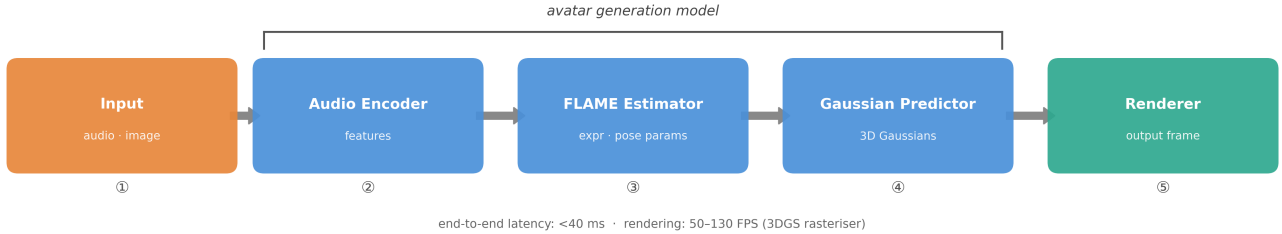

Figure 5. End-to-end audio-driven 3DGS avatar generation pipeline. Audio features are encoded and fed to a FLAME expression estimator; the resulting parameters drive a Gaussian predictor whose output is rendered at interactive frame rates ($\geq$25 FPS).

this sub-family by introducing *dual-lifting*: source image DINOv2 features are simultaneously lifted forward and backward from the image plane to produce $\sim$175,232 static Gaussians capturing identity, while per-frame expression Gaussians are computed from FLAME parameters. Inference requires only one portrait (512×512), running at 67 FPS on A100 with 2.5 GB VRAM.

ARTalk [31] (SIGGRAPH Asia 2025) generates FLAME expression and pose codes directly from audio via a causal Transformer with temporal multi-scale Vector Quantisation (VQ) autoencoding, removing the need for a driving video entirely. A style encoder extracts speaking style from example clips, enabling personalised motion without retraining.

LoRAvatar [32] (NeurIPS 2025) extends GAGAvatar with LoRA fine-tuning and a *Register Module* (inspired by ViT register tokens) that learns high-frequency identity details (wrinkles, tattoos, moles) during short adaptation from video, then deactivates at inference, preserving 67 FPS while improving identity fidelity.

Two methods target emotional control. EmoTalk3D [33] (European Conference on Computer Vision, ECCV 2024) introduces a Speech-to-Geometry-to-Appearance pipeline with canonical/dynamic Gaussian decomposition controlling eight emotion categories. EmoVOCA [34] (WACV 2025) addresses emotional 3D talking heads with a synthetic dataset of entangled speech and emotion motions, enabling continuous emotion intensity control.

3.3.2. *Video-Driven 3DGS (Person-Specific)*

Video-driven methods learn audio-to-Gaussian deformation fields from monocular video, achieving higher photorealism at the cost of per-identity training.

Two distinct papers both titled *GaussianTalker* appeared at ACM Multimedia 2024 with different designs:

- Cho *et al.* [35] encode canonical Gaussian attributes in a multi-resolution triplane hash, merging spatial features with audio via spatial-audio attention *without* any 3DMM dependency (120 FPS on RTX 3090).
- Yu *et al.* [36] bind Gaussians to FLAME and add speaker-specific BlendShapes with timbre- transformation contrastive learning (130 FPS on RTX 4090).

Table 3. 3DGS talking-head methods. †: SyncTalk++ code is for v1 (NeRF) only; no v2 code released. 1-shot: no per-subject training. 3DM: FLAME-based parametric control.

Method	Yr	FPS	3DM	1-shot	Code
GT-Cho [35]	24	120	×	×	✓
GT-Yu [36]	24	130	✓	×	×
TalkingGauss [37]	24	>100	✓	×	✓
SyncTalk++ [38]	25	101	✓	×	×†
InsTaG [39]	25	82.5	✓	×	✓
GAGAvatar [30]	24	67	✓	✓	✓
LoRAvatar [32]	25	67	✓	✓	✓
GUAVA [1]	25	∼50	EHM	✓	✓
MimicTalk [29]	24	∼30	✓	×	∼

TalkingGaussian [37] (ECCV 2024) uses persistent Gaussian primitives undergoing smooth continuous deformations; its Face-Mouth Decomposition (FMD) module separately optimises face and inside-mouth branches to resolve motion inconsistency at the lip boundary.

SyncTalk++ [38] (IEEE TPAMI 2025) upgrades SyncTalk to 3DGS, achieving 101 FPS at 512×512 with an Expression Generator for out-of-distribution audio robustness and a motion codebook. Note: the public repository (`ZiqiaoPeng/SyncTalk`) contains code for SyncTalk v1 (NeRF-based); no SyncTalk++ code has been released.

InsTaG [39] (CVPR 2025) pre-trains a Universal Motion Field on identity-free video corpora, then adapts to a new identity from as little as 5 s of video, achieving 82.5 FPS on RTX 3080 Ti.

Table 3 summarises the 3DGS family.

3.4. Diffusion and Flow-Matching Methods

Diffusion and flow-matching models lead on the *expressiveness* axis of the performance radar (Figure 2), producing the highest perceptual quality at the cost of real-time speed, though flow matching is rapidly closing this gap.

Diffusion models denoise random noise into frames conditioned on audio and identity embeddings. They generalise zero-shot (a single image is enough to animate a new face) at higher compute cost than explicit rendering.

3.4.1. Diffusion-Based Portrait Animation

VASA-1 [40] (NeurIPS 2024, Microsoft Research) set a quality benchmark by operating in a *disentangled face latent space* that separates appearance from dynamics (expression, gaze, head pose). A conditional diffusion model generates holistic facial dynamics in this compact space, achieving 40 FPS at 512×512 (45 FPS offline) on RTX 4090. VASA-1 has not been publicly released.

Several methods explore disentanglement along different axes. Sonic [41] (CVPR 2025) proposes a pure audio-driven paradigm separating intra-clip from inter-clip audio perception. DICE-Talk [42] (ACM Multimedia 2025) disentangles identity from emotion through identity-agnostic Gaussian emotion embeddings and a learnable emotion bank built on Stable Video Diffusion, though output quality does not yet match production requirements. EchoMimic [43] (AAAI 2025, Ant Group) evolved from head-only generation (V1) to semi-body animation (V2, CVPR 2025), demonstrating that diffusion methods can extend to upper-body coverage.

Real-time throughput with diffusion is still hard. MoDA [44] pairs a diffusion backbone with a Transformer for facial expressiveness and head dynamics; deployment is straightforward but artefacts remain visible. READ [45] pipelines asynchronous diffusion to $\sim$27.4 FPS effective throughput, throughput-real-time but not low-latency streaming. StableAvatar [46] traces long-form quality degradation to audio-induced latent distribution drift and adds a timestep-aware audio adapter for infinite-length generation.

At the extreme scale, LiveAvatar [47] (Alibaba, December 2024) targets full-body streaming at 20 FPS but requires $5\times$ H800 GPUs ($\sim$400 GB VRAM total) via timestep-forcing pipeline parallelism; a subsequent single-GPU mode ($\geq$80 GB VRAM) runs at considerably lower frame rates, and the multi-GPU requirement makes cloud inference expensive.

DAWN (Dynamic frame Avatar with Non-autoregressive diffusion) [48] (ICLR 2025) takes a different approach with non-autoregressive diffusion that generates dynamic-length video all-at-once, avoiding error accumulation. JoyAvatar [49] achieves 16 FPS through causal autoregressive diffusion on a single GPU.

3.4.2. Flow Matching

Flow matching learns a deterministic velocity field defining an ODE that transports noise to data, enabling generation in far fewer function evaluations than standard diffusion. Three foundational techniques underpin the speedup: OT-CFM [50] (ICLR 2024) uses minibatch optimal transport couplings; Rectified Flow [51] (ICLR 2023) iteratively straightens trajectories; and MeanFlow [52] (NeurIPS 2025 Oral) achieves FID 3.43 at a single neural function evaluation (NFE) by directly learning average velocity fields.

Applied to talking portraits, FLOAT [53] (ICCV 2025, DeepBrain AI/KAIST) constructs an orthogonal motion latent space via a motion VAE. A transformer-based vector field predictor with frame-wise Adaptive Layer Normalisation generates diverse head motion in $\sim$10 NFE *vs.* 20–50 for standard diffusion, achieving the fastest per-GPU inference among compared methods.

OmniTalker [54] combines diffusion and flow matching for text-driven talking head generation at 25 FPS, supporting multilingual generation and emotion control. Livatar-1 [55] achieves 141 FPS throughput via chunk-based flow matching with 24-frame parallel generation and 0.17 s end-to-end latency on a single A10 GPU. These results suggest that flow matching is converging toward the speed of explicit rendering methods while retaining diffusion-level quality.

3.5. Speech-Driven 3D Facial Animation

Speech-driven mesh animation sits outside the performance radar: the output is FLAME or VOCA vertex sequences rather than pixels. These methods are not renderers themselves but supply motion to any of the renderers above.

FaceFormer [56] (CVPR 2022) established the transformer paradigm for this task, though global self-attention over full sequences precludes real-time use. CodeTalker [57] (CVPR 2023) improved motion fidelity through a discrete codebook representation. DiffPoseTalk [58] (SIGGRAPH 2024) applies diffusion to generate stylistically diverse pose and expression sequences. EmoVOCA [59] targets emotional mesh animation with continuous emotion intensity control.

Extending beyond the face, two methods address co-speech body gesture. SHOW [60] (CVPR 2023) generates holistic SMPLX motion from speech with separate face and body/hand generators. EMAGE [61] (CVPR 2024) unifies holistic co-speech gesture generation via masked audio gesture modelling on the 60-hour BEAT2 dataset. These mesh-level outputs serve as the motion backbone for the head-body integration systems discussed next.

4. The Head-Body Integration Problem

Extending from head-only to full-body avatars introduces anatomical coverage as a fourth dimension beyond the performance radar (Figure 2), compounding the difficulty of satisfying speed, fidelity, and expressiveness simultaneously.

4.1. The Geometric Incompatibility

Deploying body-present avatars requires reconciling two parametric models that were designed independently:

- **FLAME** [4]: 5,023 vertices, 4 joints (neck, jaw, two eyeballs), 100 expression principal components: excellent facial detail, head region only.
- **SMPLX** [5]: 10,475 vertices, 54 joints: full-body coverage including hands, but with a coarser facial expression space trained on whole-body scans.

Naively compositing a FLAME-driven head onto a body video fails because the audio-driven head pose changes independently of the torso, producing a visible geometric disconnection qualitatively different from legacy mouth-pasting.

4.2. Full-Body Alternatives

SHOW [60] drives SMPLX body and hands from speech audio with a separate FLAME encoder for the face, producing co-speech gesture synthesis with natural hand-arm dynamics.

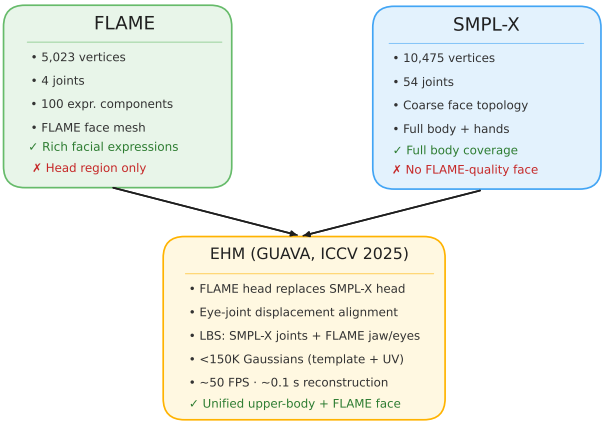

Figure 6. The head-body integration problem and GUAVA's Expressive Human Model (EHM) solution. FLAME provides rich facial dynamics but covers the head only; SMPLX provides full-body coverage but with coarse facial expression. EHM fuses both via eye-joint displacement alignment, yielding a unified deformable mesh animatable at ∼50 FPS.

SyncAnimation [28] jointly generates audio-synchronised upper-body, head, and lip animation from a triplane at 41 FPS, the most complete audio-to-body real-time system available. LiveAvatar [47] targets full-body streaming via a large DiT but requires 5×H800 GPUs for interactive frame rates. Each of these systems addresses parts of the problem; the EHM approach described next offers a more principled geometric solution.

4.3. GUAVA: The EHM Solution

GUAVA [1] (ICCV 2025) introduces the Expressive Human Model (EHM, Figure 6), a composite parametric representation that unifies SMPLX and FLAME into a single coherent deformable mesh. The fusion uses eye-joint displacement alignment: FLAME head vertices are translated so that FLAME's eye joints align with SMPLX's eye joints via a computed displacement vector; those vertices then replace the SMPLX head region. The fused mesh is deformed via Linear Blend Skinning using SMPLX body and hand pose parameters, ensuring kinematic head-body consistency:

$$\begin{aligned} \text{EHM} = \text{LBS}\big(\mathcal{U}[M_b(\boldsymbol{\beta}_b),\ M_f(\boldsymbol{\beta}_f, \boldsymbol{\psi}_f, \theta_j)], \\ \theta_b,\ \theta_h,\ J(\boldsymbol{\beta}_b) + \Delta J\big), \end{aligned} \tag{2}$$

where $\mathcal{U}[\cdot]$ denotes the vertex-replacement union, M_b is SMPLX body, M_f is FLAME, $\boldsymbol{\psi}_f$ are expression parameters, θ_j is jaw pose, and ΔJ are learned joint offsets for compatibility.

3D Gaussians are predicted through a two-branch architecture: *template Gaussians* from EHM vertex positions with projected DINOv2 image features, and *UV Gaussians* via inverse texture mapping of dense surface features. Both are refined by a StyleUNet neural decoder. The total Gaussians number fewer than 150,000 (template + UV combined). The system achieves ∼0.1 s reconstruction from a single image and ∼50 FPS rendering at 512×512, outperforming all evaluated 2D and 3D baselines on PSNR, SSIM, LPIPS, and ArcFace CSIM.

A current limitation is that GUAVA is animated by pose parameters from a reference video rather than audio. A promising deployment path is an ARTalk + GUAVA hybrid: ARTalk generates FLAME expression and pose from audio to drive EHM's head component; a separate body motion model (*e.g.* SHOW) drives SMPLX body; and GUAVA's EHM deformation pipeline produces the final upper-body Gaussian cloud. Realising this requires disentangling GUAVA's pipeline to accept independent face and body driving signals.

Table 4. SMPLX-based body 3DGS reconstruction. "1-img": single image input; "Trn": training required per subject.

Method	FPS	Code	1-img	Notes
LHM [62]	fast	✓	✓	Single image, easy to run
HGM [63]	300	✓	✓	ICLR 2025; SOTA novel-view synthesis
GST [64]	47–50	✓	✓	Transformer-based
Anim. Gauss. [65]	—	✓	×	Per-subject, multi-view
ExAvatar [66]	—	✓	×	ECCV 2024, SMPLX mesh

4.4. Body 3DGS Reconstruction

Independent of the head-body fusion, a body reconstruction pipeline is needed to build the SMPLX-based Gaussian representation. Table 4 summarises current approaches.

The Large Animatable Human reconstruction Model (LHM) [62] is the most practical option: it accepts a single image, runs in seconds, and produces a usable body Gaussian without per-subject training. The Human Gaussians Model (HGM) [63] (ICLR 2025) reports strong novel-view synthesis quality at 300 FPS using a SMPLX dual-branch architecture with ControlNet back-view refinement.

5. Comparative Analysis

The performance radar (Figure 2) provides the organising lens for this section: we evaluate methods along speed, identity fidelity, and expressiveness to reveal where each family excels and where gaps remain.

5.1. Comprehensive Comparison

Table 5 compares 30 key methods against production criteria. Methods are ordered by a production-viability score (0–40) derived from evaluation against the seven pain points of Table 2. The score is a subjective composite assigned by the authors weighting: real-time capability (10 pts), expressiveness range (10 pts), body coverage (8 pts; head-only = 3, upper-body = 6, full-body = 8), personalisation flexibility (7 pts; zero-shot = 7, fast fine-tuning = 4, per-subject = 0), and code availability (5 pts). It is an authorial assessment, not a standardised metric; future work with a unified benchmark (Section 7) should replace it with measured values.

Table 5. Comprehensive comparison of 30 key methods. RT: real-time (≥25 FPS). Body: H = head-only, U = upper-body, F = full-body. Expr.: L = low, M = medium, H = high. Gen.: generalizable (no per-subject training). Code: ✓ = open, ∼ = partial, × = unavailable. † SyncTalk++ code is v1 only. ‡ LiveAvatar requires 5×H800 GPUs at 20 FPS.

Method	Family	Input	RT	FPS	Body	Expr.	Gen.	Code	Score
InsTaG [39]	3DGS	Vid+Aud	✓	82.5	H	M	×	✓	40
GT-Cho [35]	3DGS	Vid+Aud	✓	120	H	M	×	✓	35
GT-Yu [36]	3DGS	Vid+Aud	✓	130	H	M	×	×	35
GUAVA [1]	3DGS-EHM	Img+3DM	✓	∼50	U	H	✓	✓	35
ARTalk [31]	Tfmr+3DM	Audio	✓	RT	H	M	✓	✓	35
SyncTalk++ [38]	3DGS	Vid+Aud	✓	101	H	M	×	×†	30
MoDA [44]	Diff+Tfmr	Img+Aud	✓	RT	H	H	✓	✓	30
RealTalk [67]	Triplane	Vid+Aud	×	—	H	H	×	×	30
FLOAT [53]	Flow Match	Img+Aud	✓	fastest	H	M	✓	✓	28
GAGAvatar [30]	3DGS	Image	✓	67	H	M	✓	✓	26
READ [45]	Diffusion	Img+Aud	✓	∼27	H	L	✓	×	26
OmniTalker [54]	Flow+Diff	Img+Txt	✓	25	H	H	✓	×	26
LiveAvatar [47]	14B DiT	All	✓‡	20‡	F	H	✓	✓	26
SadTalker [21]	Diff+VAE	Img+Aud	×	—	H	M	✓	✓	26
DICE-Talk [42]	Diffusion	Img+Aud	×	—	H	H	✓	✓	25
SyncAnim. [28]	Triplane	Vid+Aud	✓	41	U	M	×	∼	25
LoRAvatar [32]	3DGS+-LoRA	Img+3DM	✓	67	H	M	✓	✓	25
VASA-1 [40]	Diffusion	Vid+Aud	✓	40	H	H	✓	×	21
TalkingGauss [37]	3DGS	Vid+Aud	✓	>100	H	M	×	✓	21
EmoTalk3D [33]	3DGS	—	—	—	H	H	×	∼	21
MANGO [68]	3DGS+Diff	—	—	—	H	M	✓	✓	21
StableAv. [46]	Diffusion	Vid+Aud	×	—	H	M	✓	✓	21
TalkVerse [69]	Diffusion	All	×	—	H	H	✓	✓	21
MimicTalk [29]	Triplane	Img+Aud	×	—	H	H	×	∼	20
AvatarF. [70]	Diffusion	Vid+Aud	✓	—	H	H	×	soon	17
Livatar-1 [55]	FlowMatch	Img+Aud	✓	141	H	M	✓	×	13
EmoVOCA [59]	3D mesh	Aud+3DM	✓	RT	H	H	✓	✓	12
DAWN [48]	Diffusion	Img+Aud	✓	—	H	H	✓	✓	—
LHM [62] (body)	3DGS-SMPLX	Image	fast	—	F	L	✓	✓	—
SHOW [60] (gesture)	SMPLX mesh	Audio	—	—	F	M	✓	✓	—

5.2. Analysis by Production Axis

Real-time performance.. The highest frame rates are exclusively achieved by 3DGS methods (Figure 7). GaussianTalker variants (120–130 FPS) and SyncTalk++ (101 FPS) provide streaming headroom after per-identity training. Generalizable 3DGS methods (GAGAvatar 67 FPS, GUAVA ∼50 FPS) achieve real-time performance without per-subject training. Flow-matching methods (Livatar-1 141 FPS throughput, FLOAT fastest per-GPU, OmniTalker 25 FPS) are closing the gap. The speed advantage of 3DGS stems from explicit rasterisation of Gaussian primitives, which avoids the iterative denoising steps required by diffusion and flow models; per-subject methods push throughput further because their learned deformation fields are already specialised, eliminating generalisation overhead at inference.

Body coverage.. Only five systems address upper or full-body: GUAVA (upper, real-time, single image), SyncAnimation (upper, real-time), LiveAvatar (full, 20 FPS on 5×H800), LHM (full reconstruction, not audio-driven), SHOW (full SMPLX, not rendered). The scarcity reflects two structural barriers: the geometric incompatibility between FLAME and SMPLX (Section 4) increases modelling complexity, while adding body Gaussians multiplies the per-frame rendering budget, making joint real-time head-body generation non-trivial.

Expressiveness vs. identity consistency.. Diffusion methods produce more diverse, naturalistic motion but can hallucinate appearance details across frames. 3DGS methods preserve identity deterministically but are more constrained in expression range. This trade-off arises because diffusion models operate in a stochastic latent space that enables diverse motion sampling but offers no structural guarantee on appearance constancy, whereas 3DGS methods bind appearance to fixed Gaussian attributes anchored to a parametric mesh. Bridging this gap, that is, expressive like diffusion and consistent like 3DGS, is the defining research challenge.

Personalisation paradigm.. Three paradigms exist: (1) *zero-shot/single-image* (GAGAvatar, GUAVA, VASA-1): no per-

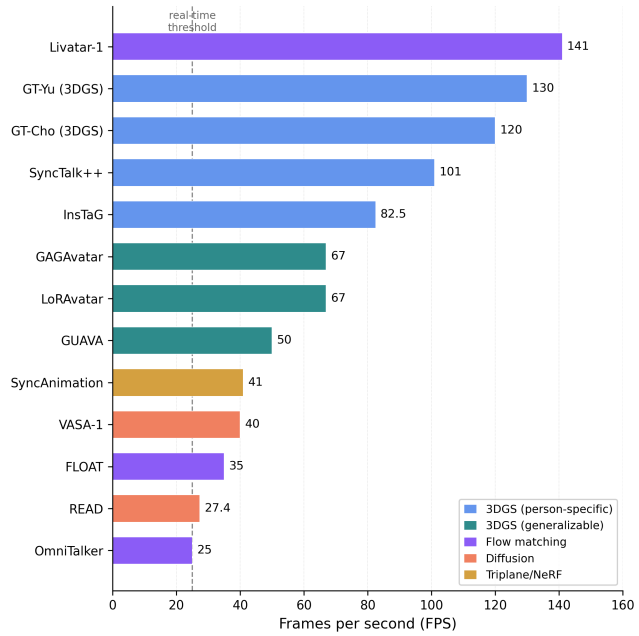

Figure 7. Rendering speed comparison across 13 key methods, grouped by methodological family. 3DGS variants (blue/green) dominate the high-FPS regime; flow-matching methods (purple) are closing the gap. The real-time threshold of 25 FPS is exceeded by all shown systems.

subject training, highest deployment flexibility; (2) *fast fine-tuning* (MimicTalk ∼15 min, LoRAvatar): improved identity fidelity at modest compute cost; (3) *full per-subject training* (GaussianTalker variants): maximum photorealism after hours. Production deployment typically requires paradigm (1) or (2). The quality gap between paradigms (1) and (3) narrows as architectural capacity grows: LoRAvatar's register-based LoRA adaptation recovers high-frequency identity details with minutes of fine-tuning, suggesting that lightweight adaptation may converge toward per-subject quality without per-subject cost.

6. Commercial Landscape

Commercial platforms occupy distinct positions within the constraint triangle (Figure 2), reflecting deliberate product strategy rather than technical limitation alone. Platforms targeting live, interactive use cases such as customer service bots, real-time co-presence prioritise low latency and throughput at the cost of expressiveness. Platforms serving asynchronous or near-synchronous scenarios accept higher latency (often 500 ms to several seconds) in exchange for photorealistic fidelity and richer motion quality. This bifurcation means the commercial landscape does not converge on a single operating point: speed-first and quality-first strategies coexist, each optimised for different deployment contexts.

Table 6 compares 18 commercial interactive avatar platforms based on hands-on evaluation by the authors in January 2026. For each platform we accessed the publicly available demo or product trial on the vendor's own website and probed conversational latency, rendering quality, and feature coverage with short scripted English prompts; advertised technology stack was taken from the same vendor pages at the time of evaluation.

No platform currently delivers real-time full-body interactive avatars with high emotional expressiveness from arbitrary identities. The competitive frontier is contested between Tavus (Phoenix-4, Gaussian-diffusion; ∼500 ms), Anam (diffusion CARA II; sub-1 s), and Akool (FastAvatar/READ; real-time; up to 16K). The quality gap between commercial offerings and leading academic methods is narrowing, suggesting production-ready upper-body interactive avatars are within reach.

7. Open Challenges and Research Directions

Each of the gaps below blocks a different axis of the performance radar (Figure 2). We list them in the order a system builder hits them: per-frame generative quality first, then head–body and hand integration, and finally the evaluation and compute layers that determine whether a method is deployable at all.

Temporal identity consistency.. Diffusion and flow-matching models drift in identity across frames, while 3DGS methods stay identity-stable but cannot reach extreme expressions (Section 5.2). The natural composition, per-frame 3DGS renderings conditioning a diffusion branch, supplies appearance constancy from the explicit pathway and motion diversity from the stochastic one. No published method occupies this configuration, and we view it as the most promising near-term route to closing the consistency–expressiveness gap.

Audio-driven full-body integration.. GUAVA's EHM resolves head–body integration architecturally (Section 4), but no end-to-end audio-driven path through it has been demonstrated. The near-term hybrid (ARTalk for the FLAME head, SHOW for the SMPLX body, GUAVA for 3DGS rasterisation) inherits three obstacles from its disaggregation: face–body boundary continuity is uncontrolled under independent driving signals, the neck–body junction has no shared loss to enforce kinematic plausibility, and audio-to-EHM training data does not yet exist at scale.

Hand gesture generation.. Co-speech hand generation lags head animation on every axis. EMAGE [61], SHOW [60], and the Language of Motion [71] (CVPR 2025) each tackle a slice (semantic alignment with speech, articulated SMPLX wrists, or learned style priors), but none renders hands at production quality in real time. Three issues compound: speech-to-gesture mapping is many-to-many and weakly supervised, SMPLX hand articulation is coarse relative to forearm visibility in upper-body framing, and adding hand-resolution Gaussians inflates a per-frame primitive count that is already tight.

Table 6. Commercial interactive avatar platforms (2025–2026). Column abbreviations: Cv = conversational avatar, VG = video generation, Em = emotion control, Bd = body coverage (F = face only, T = torso, U = upper body, Fu = full body), Ph = custom photo upload, Vi = custom video upload, Speed = end-to-end latency or generation time, Res = output resolution (pixels on shortest axis unless stated). ✓ = yes, × = no, ∼ = partial/coming soon. Source: hands-on evaluation by the authors in January 2026 using each vendor's public demo or product trial.

Platform	Cv	VG	Em	Bd	Ph	Vi	Technology	Speed	Res
Hedra	×	✓	∼	F	∼	∼	—	<3 min	540–1080
Akool	✓	×	∼	U	✓	✓	FastAvatar / READ	Real-time	Up to 16K
Synthesia	∼	✓	×	T	✓	✓	3DGS/NeRF	Offline	—
Simli	✓	×	∼	F	✓	×	—	<300ms	∼950
Runway	∼	✓	×	—	∼	∼	—	Real-time	—
bitHuman	✓	✓	✓	F	✓	∼	—	Real-time†	500–800
Tavus	✓	✓	∼	T	×	✓	Phoenix-4 3DGS+diff.	∼500ms	—
Colossyan	∼	✓	×	—	✓	✓	—	Minutes	—
LemonSlice	✓	✓	✓	F	✓	∼	Diffusion	—	656×978
D-ID	✓	✓	∼	T	✓	✓	—	Stream API	—
Atmanity	✓	×	✓	F	×	×	3DGS+LoRA	—	∼850
HeyGen	✓	✓	∼	T	✓	✓	GAN+diffusion+Tfmr	∼5 min	720–4K
DeeVid AI	×	✓	×	—	✓	✓	—	∼1 min	—
Anam	✓	✓	✓	T	✓	×	Diffusion CARA II	<1 s	720×480
Soul Machines	✓	×	✓	—	×	×	—	Real-time	—
BeyondPresence	✓	∼	∼	T	∼	∼	—	—	—
LiveAvatar	✓	×	×	T	∼	∼	GAN+diffusion	Real-time	720–1080
Elai.io	×	✓	×	—	✓	∼	—	Minutes	—

†Platform instability reported in testing.

Scalable one-shot personalisation.. The deployment ideal, a high-quality personalised avatar from a single photograph with no per-subject training, is approached from two sides. Generalizable methods (GAGAvatar, GUAVA) reach zero training cost but trail subject-specific systems on identity sharpness, while LoRA-based personalisation (LoRAvatar) recovers high-frequency detail in minutes. Extending LoRAvatar's register-based adaptation from FLAME-only heads to the full EHM joint set is technically open: the larger parameter pool risks overfitting under minutes-scale fine-tuning, and no published work characterises the tradeoff.

Evaluation standards.. The standard benchmarks (HDTF, VoxCeleb2, MEAD, LRS3) were assembled for talking-head systems and predate the upper-body, emotion-rich, identity-diverse settings most current methods target. THEval [16] ($\rho = 0.870$ against human ratings) and the THQA-NTIRE benchmark (12,247 videos, CVPR 2025 challenge) close part of this gap with holistic perceptual scoring and broader subject coverage, but neither yet provides upper-body ground truth or controlled emotion annotations. SyncNet-based lip-sync metrics (LSE-D/LSE-C) drift with head pose and audio encoding; AV-HuBERT- [15] and StableSyncNet-based alternatives should be reported alongside, not instead of, the legacy numbers.

Compute and infrastructure efficiency.. Cloud deployment is gated by VRAM as much as by frame rate. GAGAvatar's 2.5 GB inference footprint sets a deployable lower bound for one-shot methods; LiveAvatar's multi-H800 requirement restricts that system to high-end clusters. Among the standard compression toolkit, distillation from DiT teachers to flow-matching or 3DGS students is more promising than uniform quantisation: quantisation degrades each step of the iterative denoising trajectory, whereas distillation shortens the trajectory itself, which is where most of the cost lives.

8. Conclusion

We surveyed 100+ methods published between 2023 and 2026 for interactive real-time visual avatar generation, organised by the speed–identity–expressiveness performance radar (Figure 2). Three observations follow. 3DGS with 3DMM-driven expression control is the prevailing real-time choice, achieving 67–130 FPS with generalizable single-image personalisation, and GUAVA's EHM (SMPLX+FLAME, ICCV 2025) provides an explicit architecture for head–body integration, though without an audio-driven expression controller. Flow matching has largely supplanted score-based diffusion in quality-critical pipelines at an order-of-magnitude lower function-evaluation cost. Diffusion-based upper-body methods such as VASA-1 retain the quality lead but remain too compute-intensive for low-cost interactive deployment.

A near-term route to upper-body interactive avatars combines an audio-to-expression model such as ARTalk with a co-speech gesture model such as SHOW, rendered through GUAVA's 3DGS pipeline; the remaining obstacles are catalogued in Section 7. Resolving audio-driven FLAME+SMPLX integration and scaling one-shot personalisation to EHM are the gating problems for production-quality interactive avatars over the next 12–18 months.

The same advances enable misuse, including non-consensual identity replication and synthetic media abuse; identity verification and provenance watermarking warrant treatment as core pipeline components rather than post-hoc

additions.

References

[1] Dongbin Zhang, Yunhao Li, Shijie Tang, Wen Zhou, Pan He, Jian Li, and Binbin Yan. GUAVA: Generalizable upper body 3D Gaussian avatar. In *ICCV*, 2025. URL https://arxiv.org/abs/2505.03351. 1, 2, 3, 6, 8, 9

[2] K. R. Prajwal, Rudrabha Mukhopadhyay, Vinay Namboodiri, and C. V. Jawahar. A lip sync expert is all you need for speech to lip generation in the wild. In *ACM Multimedia*, 2020. URL https://arxiv.org/abs/2008.10010. 2, 4, 5

[3] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3D Gaussian splatting for real-time radiance field rendering. In *ACM SIGGRAPH*, 2023. 2, 5

[4] Tianye Li, Timo Bolkart, Michael J. Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4D scans. In *ACM SIGGRAPH Asia*, 2017. 2, 4, 5, 7

[5] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3D hands, face, and body from a single image. In *CVPR*, 2019. 2, 4, 7

[6] Shreyank N. Gowda, Dheeraj Pandey, and Shashank Narayana Gowda. From pixels to portraits: A comprehensive survey of talking head generation techniques and applications. *arXiv preprint arXiv:2308.16041*, 2023. URL https://arxiv.org/abs/2308.16041. 2

[7] Diqiong Jiang, Jian Chang, Lihua You, Shaojun Bian, Robert Kosk, and Greg Maguire. Audio-driven facial animation with deep learning: A survey. *Information*, 15(11):675, 2024. doi: 10.3390/info15110675. 2

[8] Vineet Kumar Rakesh, Soumya Mazumdar, Research Pratim Maity, Sarbajit Pal, Amitabha Das, and Tapas Samanta. Advancing talking head generation: A comprehensive survey of multi-modal methodologies, datasets, evaluation metrics, and loss functions. *arXiv preprint arXiv:2507.02900*, 2025. URL https://arxiv.org/abs/2507.02900. 2

[9] Gan Pei, Jiangning Zhang, Menghan Hu, Zhenyu Zhang, Chengjie Wang, Yunsheng Wu, Guangtao Zhai, Jian Yang, and Dacheng Tao. Deepfake generation and detection: A benchmark and survey. *ACM Computing Surveys*, 2024. URL https://arxiv.org/abs/2403.17881. arXiv:2403.17881. 2

[10] Guikun Chen and Wenguan Wang. A survey on 3D Gaussian splatting. *arXiv preprint arXiv:2401.03890*, 2024. URL https://arxiv.org/abs/2401.03890. 2

[11] Ruihe Wang, Yukang Cao, Kai Han, and Kwan-Yee K. Wong. A survey on 3D human avatar modeling – from reconstruction to generation. *arXiv preprint arXiv:2406.04253*, 2024. URL https://arxiv.org/abs/2406.04253. 2

[12] Simbarashe Nyatsanga, Taras Kucherenko, Chaitanya Ahuja, Gustav Eje Henter, and Michael Neff. A comprehensive review of data-driven co-speech gesture generation. *Computer Graphics Forum (Eurographics)*, 2023. URL https://arxiv.org/abs/2301.05339. 2

[13] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. HuBERT: Self-supervised speech representation learning by masked prediction of hidden units. In *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2021. URL https://arxiv.org/abs/2106.07447. 3

[14] Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *NeurIPS*, 2020. URL https://arxiv.org/abs/2006.11477. 3

[15] Bowen Shi, Wei-Ning Hsu, Kushal Lakhotia, and Abdelrahman Mohamed. Learning audio-visual speech representation by masked multimodal cluster prediction. In *ICLR*, 2022. URL https://arxiv.org/abs/2201.02184. 3, 4, 11

[16] others. THEval: Evaluation framework for talking head video generation. *arXiv*, 2025. 3, 4, 11

[17] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. ArcFace: Additive angular margin loss for deep face recognition. In *CVPR*, 2019. 4, 5

[18] others Zhou. THQA: A perceptual quality assessment database for talking heads. In *ICIP*, 2024. URL https://arxiv.org/abs/2404.09003. 4

[19] Zhimeng Zhang, Lincheng Li, Yu Ding, and Changjie Fan. Flow-guided one-shot talking face generation with a high-resolution audio-visual dataset. In *CVPR*, 2021. 4

[20] Kaisiyuan Wang et al. MEAD: A large-scale audio-visual dataset for emotional talking-face generation. In *ECCV*, 2020. 4

[21] Wenxuan Zhang, Xiaodong Cun, Xuan Wang, Yong Zhang, Xi Shen, Yu Guo, Ying Shan, and Fei Wang. SadTalker: Learning realistic 3D motion coefficients for stylized audio-driven single image talking face animation. In *CVPR*, 2023. URL https://arxiv.org/abs/2211.12194. 5, 9

[22] Hang Zhou, Ziwei Liu, Xudong Xu, Ping Luo, and Xiaogang Wang. Pose-controllable talking face generation by implicitly modularized audio-visual representation. In *CVPR*, 2021. 5

[23] Kun Cheng, Xiaodong He, Linya Li, Menglei Shen, Ping Luo, Hujun Bao, and Weidong Zhang. VideoReTalking: Audio-based lip synchronization for talking head video editing in the wild. In *SIGGRAPH Asia*, 2022. 5

[24] Yue Zhang, Minhao Liu, Zhaokang Chen, Bin Huang, Yubin Zheng, Chao Zhan, and Wenjiang Jiang. MuseTalk: Real-time high quality lip synchorization with latent space inpainting. *arXiv*, 2024. 5

[25] Christian Safka and Tavus. Phoenix-2: Advanced techniques in talking head generation — 3D Gaussian Splatting. Tavus engineering blog, July 2024. URL https://www.tavus.io/post/advanced-techniques-in-talking-head-generation-3d-gaussian-splatting. Accessed 2026-04. 5

[26] Yudong Guo, Keyu Chen, Sen Liang, Yongjin Liu, Hujun Bao, and Juyong Zhang. AD-NeRF: Audio driven neural radiance fields for talking head synthesis. In *ICCV*, 2021. URL https://arxiv.org/abs/2103.11078. 5

[27] Ziqiao Peng, Wentao Hu, Yue Shi, Xiangyu Zhu, Xiaomei Zhang, Zhengqi Zhao, Jun He, Hongyan Liu, and Zhaoxin Fan. SyncTalk: The devil is in the synchronization for talking head synthesis. In *CVPR*, 2024. 5

[28] Yi Wang, Jian Liang, et al. SyncAnimation: A real-time end-to-end framework for audio-driven human pose and talking head animation. *arXiv*, 2025. 5, 8, 9

[29] Zhenhui Ye, Tianyun Zhong, Yi Ren, Jiaqi Yang, Weichuang Li, Jiawei Huang, Ziyue Jiang, Jinzheng He, Rongjie Huang, Jinglin Liu, Chen Zhang, Pengfei Liu, Xiang Yin, Zejun Ma, and Zhou Zhao. MimicTalk: Mimicking a personalized and expressive 3D talking face in minutes. In *NeurIPS*, 2024. URL https://arxiv.org/abs/2410.06734. 5, 6, 9

[30] Xuangeng Chu and Tatsuya Harada. GAGAvatar: Generalizable and animatable Gaussian head avatar. In *NeurIPS*, 2024. URL https://arxiv.org/abs/2410.07971. 5, 6, 9

[31] Xuangeng Chu, Jianfeng Wang, Jing Yang, and Tatsuya Harada. ARTalk: Speech-driven 3D head animation via autoregressive model. In *SIGGRAPH Asia*, 2025. URL https://arxiv.org/abs/2502.20323. 6, 9

[32] Sai Tanmay Reddy Chakkera, Xuangeng Chu, and Tatsuya Harada. Low-rank head avatar personalization with registers. In *NeurIPS*, 2025. URL https://arxiv.org/abs/2506.01935. 6, 9

[33] Qiangeng He et al. EmoTalk3D: High-fidelity free-view synthesis of emotional 3D talking head. In *ECCV*, 2024. URL https://arxiv.org/abs/2408.00297. 6, 9

[34] Federico Nocentini, Claudio Thomas, Federico Becattini, Lorenzo Seidenari, and Alberto Del Bimbo. EmoVOCA: Speech-driven emotional 3D talking heads. In *WACV*, 2025. URL https://arxiv.org/abs/2403.12886. 6

[35] Kyusun Cho, Joungbin Lee, Heeji Shin, Sangjun Kim, Chaehun Yang, Minh-Tuan Kim, and Sung-Eui Yoon. GaussianTalker: Real-time high-fidelity talking head synthesis with audio-driven 3D Gaussian splatting. In *ACM Multimedia*, 2024. URL https://arxiv.org/abs/2404.16012. 6, 9

[36] Hongyun Yu, Zhan Zhan, Mingdeng Sun, Yihua Jiang, Feilong Zhu, Zhihong Fan, Changan Han, Yupeng Quan, and Yuanyuan Chen. GaussianTalker: Speaker-specific talking head synthesis via 3D Gaussian splatting. In *ACM Multimedia*, 2024. URL https://arxiv.org/abs/2404.14037. 6, 9

[37] Longwen Li, Yifan Zhang, Hang Zhang, Xin Wang, Fan Zhong, Zhan Yang, Jialing Song, Kunpeng Chen, Mengcheng Xu, Zhiyong Zhang, and Wei Xu. TalkingGaussian: Structure-persistent 3D talking head synthesis via Gaussian splatting. In *ECCV*, 2024. URL https://arxiv.org/abs/2404.15264. 6, 9

[38] Ziqiao Peng, Wentao Hu, Yue Shi, Xiangyu Zhu, Xiaomei Zhang, Jun He, Hongyan Liu, and Zhaoxin Fan. SyncTalk++: High-fidelity and efficient synchronized talking heads synthesis using gaussian splatting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025. URL https://arxiv.org/abs/2506.14742. 6, 9

[39] Longwen Li et al. InsTaG: Learning personalized 3D talking head from few-second video. In *CVPR*, 2025. URL https://arxiv.org/abs/2502.20387. 6, 9

[40] Sicheng Xu, Guojun Chen, Yu-Xiao Guo, Jiaolong Yang, Chong Li, Zhenyu Zang, Yizhong Zhang, Xin Liu, Yichong Chen, Jianfeng Wang, Sheng Liu, Fang-Lue He, Li Wen, Dong Chen, Jianfeng Wang, and Baining Liu. VASA-1: Lifelike audio-driven talking faces generated in real time. In *NeurIPS*, 2024. URL https://arxiv.org/abs/2404.10667. 7, 9

[41] Xiaozhong Ji et al. Sonic: Shifting focus to global audio perception in portrait animation. In *CVPR*, 2025. URL https://arxiv.org/abs/2411.16331. 7

[42] others. Disentangle identity, cooperate emotion: Correlation-aware emotional talking portrait generation. In *ACM Multimedia*, 2025. URL https://arxiv.org/abs/2504.18087. 7, 9

[43] Zhiyuan Chen et al. EchoMimic: Lifelike audio-driven portrait animations through editable landmark conditions. In *AAAI*, 2025. 7

[44] others. MoDA: Multi-modal diffusion architecture for talking head generation. *arXiv*, 2025. 7, 9

[45] others. READ: Real-time and efficient asynchronous diffusion for audio-driven talking head generation. *arXiv*, 2025. 7, 9

[46] others. StableAvatar: Infinite-length audio-driven avatar video generation. *arXiv*, 2025. 7, 9

[47] Alibaba-Quark. Live Avatar: Streaming real-time audio-driven avatar generation with infinite length. *arXiv*, 2024. 7, 8, 9

[48] others. DAWN: Dynamic frame avatar with non-autoregressive diffusion framework for talking head video generation. In *ICLR*, 2025. URL https://arxiv.org/abs/2410.13726. 7, 9

[49] others. JoyAvatar: Real-time and infinite audio-driven avatar generation with autoregressive diffusion. *arXiv*, 2024. 7

[50] Alexander Tong, Nikolay Malkin, Guillaume Huguet, Yanlei Zhang, Jarrid Rector-Brooks, Kilian Fatras, Guy Wolf, and Yoshua Bengio. Improving and generalizing flow-matching with minibatch optimal transport. In *ICLR*, 2024. URL https://arxiv.org/abs/2302.00482. 7

[51] Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *ICLR*, 2023. URL https://arxiv.org/abs/2209.03003. 7

[52] Zhengyang Geng, Ashwini Pokle, and J. Zico Kolter. Mean flows for one-step generative modeling. In *NeurIPS*, 2025. URL https://arxiv.org/abs/2505.13447. 7

[53] Taekyung Ki, Dongchan Min, Gyeongsu Im, Jaeyoon Kim, Gyungtaek Kim, and Jung-Woo Ha Kim. FLOAT: Generative motion latent flow matching for audio-driven talking portrait. In *ICCV*, 2025. URL https://arxiv.org/abs/2412.01064. 7, 9

[54] others. OmniTalker: Real-time text-driven talking head generation with in-context audio-visual style replication. *arXiv*, 2025. 7, 9

[55] Haiyang Liu et al. Livatar-1: Real-time talking heads generation with tailored flow matching. *arXiv*, 2025. 7, 9

[56] Evelyn Fan, Qi Pan, Zi-Yi Liu, Deana Coupland, and Anish Chandna. FaceFormer: Speech-driven 3D facial animation with transformers. In *CVPR*, 2022. URL https://arxiv.org/abs/2112.05329. 7

[57] Jinbo Xing, Menghan Xia, Yuechen Zhang, Xiaodong Cun, Jue Wang, and Tien-Tsin Wong. CodeTalker: Speech-driven 3D facial animation with discrete motion prior. In *CVPR*, 2023. 7

[58] Zhiyao Sun, Tian Lv, Sheng Ye, Matthieu Lin, Jenny Shao, Yu-Hui Wen, Meng Wang, and Yong-Jin Liu. DiffPoseTalk: Speech-driven stylistic 3D facial animation and head pose generation via diffusion models. In *ACM SIGGRAPH*, 2024. 7

[59] Federico Nocentini, Claudio Thomas, Federico Becattini, Lorenzo Seidenari, and Alberto Del Bimbo. EmoVOCA: Speech-driven emotional 3D talking heads. In *WACV*, 2025. 7, 9

[60] Hongwei Yi, Hualin Liang, Yifei Liu, Qiong Cao, Yandong Wen, Timo Bolkart, Dacheng Tao, and Michael J. Black. Generating holistic 3D human motion from speech. In *CVPR*, 2023. 7, 9, 10

[61] Haiyang Liu et al. EMAGE: Towards unified holistic co-speech gesture generation via expressive masked audio gesture modeling. In *CVPR*, 2024. 7, 10

[62] Lingteng Qiu et al. LHM: Large animatable human reconstruction model for single image to 3D in seconds. In *CVPR*, 2025. URL https://arxiv.org/abs/2503.10625. 8, 9

[63] Jinnan Chen et al. Generalizable human Gaussians from single-view image. In *ICLR*, 2025. 8

[64] Abdullah Hamdi et al. GST: Precise 3D human body from a single image with Gaussian splatting transformers. In *CVPR Workshop*, 2025. URL https://abdullahamdi.com/gst/. 8

[65] Zhe Li, Zerong Zheng, Lizhen Wang, and Yebin Liu. Animatable Gaussians: Learning pose-dependent Gaussian maps for high-fidelity human avatar modeling. In *CVPR*, 2024. 8

[66] Zerong Zheng et al. Expressive whole-body 3D Gaussian avatar. In *ECCV*, 2024. doi: 10.1007/978-3-031-72940-9_2. 8

[67] others Wang. RealTalk: Realistic emotion-aware lifelike talking-head synthesis. In *ICCVW*, 2025. 9

[68] Feilong Zhu et al. MANGO: Natural multi-speaker 3D talking head generation via 2D-lifted enhancement. *arXiv*, 2026. 9

[69] Zhengzhi Wang et al. TalkVerse: Democratizing minute-long audio-driven video generation. *arXiv*, 2024. 9

[70] Taekyung Ki et al. Avatar Forcing: Real-time interactive head avatar generation for natural conversation. *arXiv*, 2026. 9

[71] Juze Chen et al. The language of motion: Unifying verbal and non-verbal language of 3D human motion. In *CVPR*, 2025. 10