Hey Gen | Research

Avatar 6: Talking Avatars That Are More Like You

HeyGen Research

Research page · Product · Developer docs

1 Introduction

Audio-driven avatar generation synthesises video of a specific person delivering new speech. It has become a standard component of video production for training material, multilingual communication, outreach and creator content, and the field has advanced quickly. Early methods re-rendered only the mouth region of an existing video to match new audio [4], or animated a single portrait through intermediate 3D or latent motion representations [5, 6]. Diffusion models [11], built on latent image and video backbones [12, 13], then made it possible to generate the face directly from audio, with far richer expression and head motion [7, 8, 9, 10, 14], and were soon extended to the half body [15] and to longer, higher-resolution output [16]. Built on large video diffusion transformers [17, 18, 19], many of them trained with flow matching [20], recent work extends generation to the full body and the surrounding scene [21, 22, 23, 24, 25, 28], from one speaker to several [26, 27], and to long or unbounded durations [29, 30, 31]. In parallel, general video generation models such as Sora [32], Kling [46], Veo 3 [33], Seedance [34, 35, 36] and Wan [19] now produce convincing talking people from an image and a prompt, increasingly with synchronised speech generated jointly with the video.

HeyGen has developed audio-driven avatar models over several generations, from Avatar III and Avatar IV to Avatar V [1], which introduced the video-reference formulation, together with HeyGen Voice for speech in the speaker's own voice [2] and composable control of avatar behaviour [3]. These models are deployed in HeyGen's products for businesses and individual creators.

As visual quality has matured, the central requirement has shifted from realism to identity fidelity. Most of the methods above condition on a single image of the person: they reproduce how the person looks in one frame, but must invent how the person speaks, gestures and moves. An avatar today usually stands in for a specific person, and its output is useful only if people who know that person would accept it as theirs. This requires reproducing not only appearance but also manner of speaking and habits of motion, in the person's own setting and over videos of arbitrary length. Avatar 6 is built around this objective.

Avatar 6 is an audio-driven avatar video generation model that produces video of a specific person that not only looks like them but also speaks and moves like them, at 1080p and 2K and with no limit on duration. It extends the video-reference formulation of Avatar V into a unified framework: the person is specified by a short identity recording or by a single photograph, and generation is conditioned on speech and an optional written motion prompt. The system comprises the generation model, which includes a super-resolution stage and outputs 1080p and 2K natively, a motion planner for speech- and prompt-driven performance, and HeyGen Voice; all components are developed in-house. Compared with Avatar IV and Avatar V, Avatar 6 improves likeness, motion, resolution and long-duration stability; it is consistently preferred in large-scale online preference tests and comes close to real footage in a Turing test with external raters (Section 2).

The same framework extends beyond a single talking head to multi-speaker dialogue, fine-grained motion control from written prompts, and word-level editing of real recordings. Avatar 6 is offered in two configurations: *Avatar 6 Quality* and *Avatar 6 Turbo*, a lower-latency variant served by our in-house accelerated inference stack.

Its key capabilities are as follows.

- **Identity fidelity from a video reference.** From a short identity recording, Avatar 6 learns a person's appearance, including skin texture, hair and clothing, together with their manner of speaking and the motion of their head, face, hands and body, rather than imposing a single learned presenter style. In a Turing test, Avatar 6 clips shown alone are judged real in 65.1% of answers, against 78.9% for genuine recordings, and next to the real recording it is 84.4% indistinguishable from real footage, against 44.4% for Avatar V (Section 2.5).
- **Generalisation across scenes and styles.** Avatar 6 is not restricted to a presenter in front of a plain backdrop. It generalises to studios, offices, stages and concerts, and from photorealistic footage to stylised characters, preserving the scene, lighting and framing of the reference.
- **High resolution and unbounded duration.** Avatar 6 natively generates 1080p and 2K video. Identity, framing and sharpness remain stable over videos of any length: a 30-minute sequence is generated end to end in a single pass, without identity drift or late-stage loss of sharpness.
- **Diverse applications.** The same model supports **(1) multi-speaker dialogue** within one frame; **(2) fine-grained motion control**, placing actions such as a wave, a nod or a pointing gesture at a specified moment; and **(3) talking video editing**, which replaces words or adds gestures within a real recording while leaving the surrounding footage intact (Section 3).

Avatar 6 is available in the HeyGen web app, the mobile app and the API from mid-November 2026.

From the video-reference formulation introduced by Avatar V [1] to the unified framework of Avatar 6, the HeyGen avatar series rests on a single premise: a person's own footage is the most faithful description of their appearance, manner of speaking and motion.

Looking ahead, we will continue to scale the reference itself. Avatar 6 learns a person from a short recording; we aim to extend its capacity and memory to tens of minutes and ultimately hours of a person's footage, so that identity fidelity keeps improving with the amount of footage a person provides.

Because an avatar reproduces a real person, it is gated by that person's consent. Before an avatar created from a person's footage can generate any video, the person must record a short on-camera consent statement that includes a one-time code generated for that session, so footage recorded in advance or taken from elsewhere cannot satisfy the check; HeyGen then verifies that the person in the consent recording matches the person in the training footage, and the avatar cannot render until this check passes. The verified consent recording also becomes the identity reference against which later looks of the avatar are checked. All people shown in this report are HeyGen employees or contracted actors who consented to the use of their likeness in this report. The AI avatars in Section 3.1 are synthetic characters, and the real-person avatars there are HeyGen public stock avatars.

2 Evaluation

2.1 Overview

We evaluate Avatar 6 against three references: state-of-the-art general video generation models; avatar-centric video generation models, namely our previous models and commercial avatar products; and real footage of the same people. Section 3 then presents applications.

Comparison with general video generation models. Avatar 6 outperforms state-of-the-art general video generation models on avatar video. HeyGen employees judging videos of themselves prefer Avatar 6 Quality over Seedance 2.5 in 86.5% and over Wan 3.0 in 87.9% of comparisons (Section 2.3).

Comparison with avatar-centric video generation models. Avatar 6 Quality is preferred over Avatar V in 75.0% of votes, and Avatar 6 Turbo over the Turbo tiers of Avatar IV and Avatar V in 71.4% and 68.2% (Section 2.4). Against commercial avatar products, Avatar 6 reaches the highest lip-sync confidence of all methods, ranks first on every criterion in the absolute ratings by external raters, whether it starts from a single photo or from the person's own footage, and is preferred side by side over every product that renders only its own avatars or voices, except Wan 3.0 from a photo, where raters split evenly.

Comparison with real footage. Shown alone, Avatar 6 clips are judged real in 65.1% of answers, against 78.9% for real recordings of the same people; side by side with the real recording, raters pick the Avatar 6 clip as the real one in 42.2% of their choices, which makes it 84.4% indistinguishable from real footage, against 44.4% for Avatar V (Section 2.5).

Talking video editing. On word and motion edits of real recordings, Avatar 6 leads on every metric against MiniMax H3-max, Wan 3.0, Seedance 2.5 and HappyHorse Edit. It keeps the untouched footage at 48.3 dB while every other method regenerates the whole clip, and it has the best lip sync on the new words (Section 3.4).

2.2 Evaluation Framework

Baselines. We compare against general video generation models (Wan 3.0 [47] and Seedance 2.5 [48]); avatar-centric models, namely our previous Avatar IV and Avatar V and commercial avatar products (OmniHuman v2 [37], Hedra Character 3 [38], VEED Fabric 1.0 [39], Creatify Aurora [40], D-ID V2 Talks [41], Colossyan NEO 1 Instant avatar [42], Akool Talking Photo [43], Argil image avatar [44], Synthesia Express-2 photo avatar [45] and Kling 4.0 [46] from a photo; Akool Talking Avatar, D-ID V3 Express avatar, Colossyan NEO 2 / NEO 1, Synthesia Personal video avatar and Stock avatars v3 / v1, VEED Preset avatars and Creatify AI Avatar v1 from footage); and real footage of the same people. The comparisons with commercial avatar products, the ratings against Seedance 2.5, Wan 3.0 and Kling 4.0 by external raters, the Turing test and the video editing results use Avatar 6 Quality (ten sampling steps, 2K output).

Metrics. Against our previous models, raters on HeyGen's in-house pairwise evaluation platform choose the better of two videos rendered from the same request. HeyGen employees judge videos of themselves against general video generation models. Against commercial products we report SyncNet lip-sync confidence [49], HeyGen's in-house Talking Style, Gesture, Appearance and Visual Quality scores, and human ratings [53]: an absolute category rating (ACR) where a product accepts our inputs and, for products that render only from their own scripts and voices or their own stock avatars, a side-by-side comparison category rating (CCR). A Turing test with external raters compares Avatar 6 with real footage.

2.3 Comparison with General Video Generation Models

General video generation models now produce convincing talking people from an image and an audio track. We compare Avatar 6 Quality with two of the strongest, Seedance 2.5 and Wan 3.0, in studies where HeyGen employees judge videos of themselves by which looks most like them (Figure 1). Because the judges are the people shown in the videos, these studies isolate identity fidelity, and Avatar 6 is preferred in 86.5% of comparisons against Seedance 2.5 and 87.9% against Wan 3.0 (Figure 1). The external ratings in Section 2.4 measure something different: raters who have never met the person judge overall quality, lip sync and motion in a general one-shot setting (Table 2 and Figure 4). In this setting the margin over Seedance 2.5 and Wan 3.0 is smaller: Avatar 6 wins 50.1% to 74.1% of the answers with a preference.

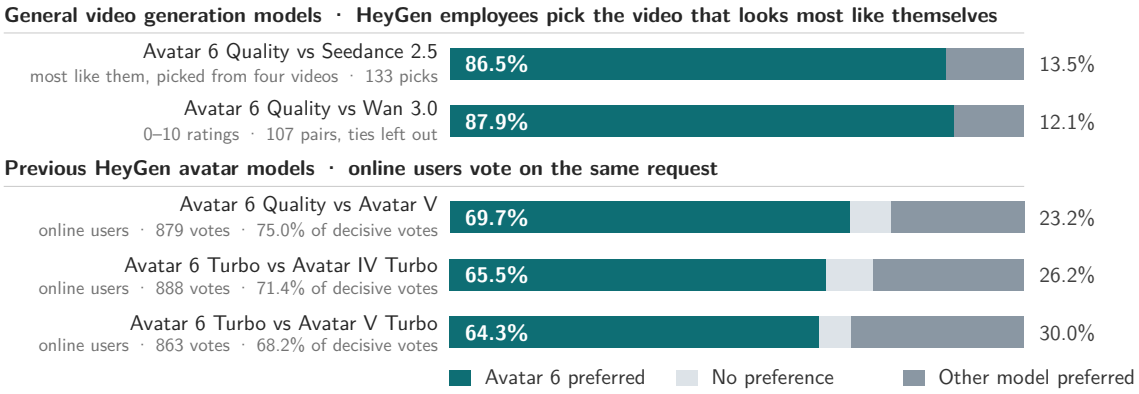


Figure 1 Pairwise preference for Avatar 6. Likeness, judged by the people in the videos. Top: HeyGen employees judge videos of themselves against state-of-the-art general video generation models; against Seedance 2.5 each employee picks which of four videos (three Avatar 6 settings and Seedance 2.5) looks most like them, and against Wan 3.0 each rates the videos from 0 to 10, a pair counting when the two scores differ. Bottom: online votes on HeyGen's pairwise evaluation platform against our previous avatar models, with both videos rendered from the same request; each Avatar 6 configuration is compared with the earlier model serving the same tier, Avatar 6 Quality with Avatar V and Avatar 6 Turbo with the Turbo tiers of Avatar IV and Avatar V.

2.4 Comparison with Avatar-Centric Video Generation Models

We compare Avatar 6 with our previous avatar models and with commercial avatar products.

Previous HeyGen avatar models. The lower part of Figure 1 compares Avatar 6 with our previous avatar models. Both configurations are preferred over the earlier model in the same tier: Avatar 6 Quality over Avatar V in 75.0% of votes, and Avatar 6 Turbo over the Turbo tiers of Avatar IV and Avatar V in 71.4% and 68.2%. On avatars whose reference footage shows little motion, Avatar 6 Quality is preferred over Avatar V in 69.4% of votes.

Commercial avatar products. Avatar 6 reaches a mean lip-sync confidence of 9.25, the highest of all methods, and ranks first on all five objective scores (Figure 2).

Human ratings. External raters score the commercial products in two settings (Tables 1 and 2, Figures 3 and 4). Where a product accepts our inputs, every method renders from the same photo, or the same footage of the person, and the same speech, and raters score each video on its own on a 0–5 scale. Avatar 6 ranks first on all six criteria, and on the overall score, in both groups and, scored by the same rater on the same input, beats every product in most comparisons. Its overall score leads the closest product by 0.32 from a photo and by more than a full point from footage, and 71.5% of its videos from a photo score 4 or more, against 61.2% for the closest product, VEED Fabric 1.0. Synthesia, Kling, VEED Preset avatars and Creatify AI Avatar v1 either take only a script and synthesise their own voice, or offer only their own stock avatars, so raters compare their videos side by side with Avatar 6. Against Kling 4.0, the closest of these, raters prefer Avatar 6 in 66.5% of answers and Kling in 24.0%; against the other products they prefer Avatar 6 in 78.3% to 87.1% of answers. Seedance 2.5 and Wan 3.0, and Kling 4.0 working from the person's footage, also speak the script in their own voice and are compared the same way in a separate session. Of the answers with a preference, Avatar 6 wins 59.6% against Seedance 2.5 and 50.1% against Wan 3.0 from a photo, a tie, and 54.9% against Kling 4.0, 72.4% against Seedance 2.5 and 74.1% against Wan 3.0 from footage.

Third-party leaderboard. On Avatar Arena [52], an independent leaderboard on which people blindly compare two providers' avatar videos made from the same script, Avatar 6 ranks first at the time of writing both on the Twins board, where each person's own twin is compared, and on the Stock avatars board, where providers' ready-made presenters are compared.

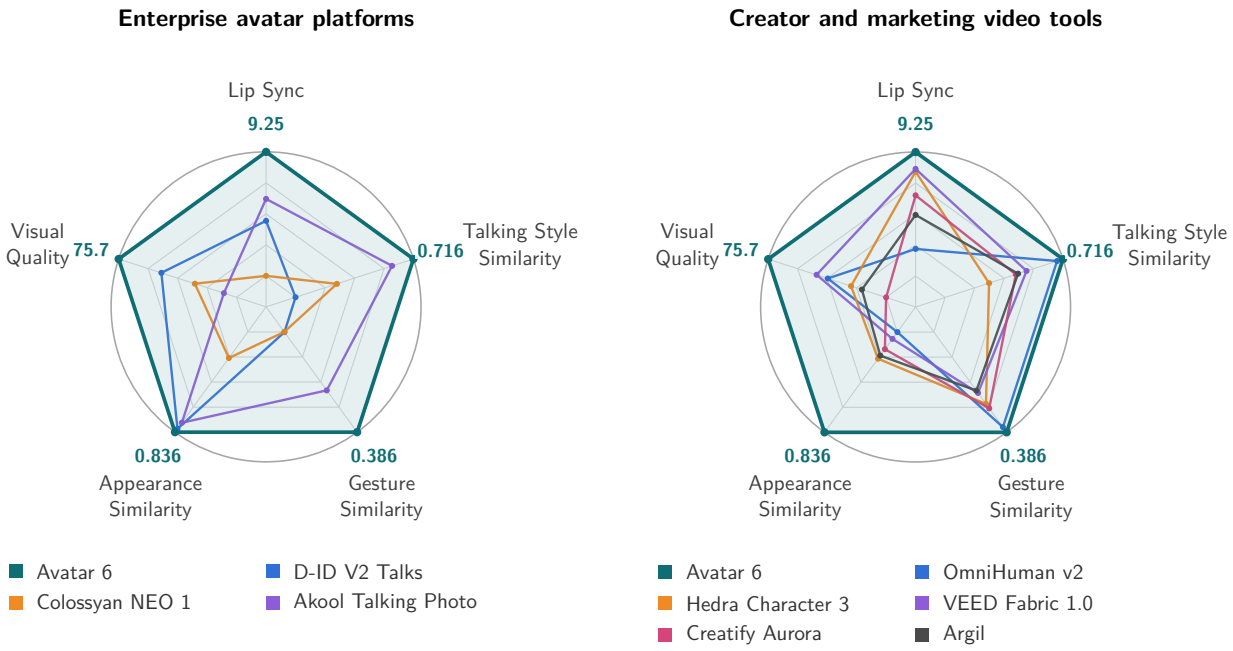


Figure 2 Objective scores against enterprise avatar platforms (left) and creator and marketing video tools (right); each axis spans the lowest to the highest score across methods, with Avatar 6 values labelled.

Table 1 Absolute category rating (ACR) against commercial avatar products on shared inputs, by 20 raters for the photo group and 23 for the footage group. Mean opinion score (0–5). Avatar 6 wins: share of comparisons, by the same rater on the same input, that Avatar 6 wins when the two get different scores. Products are ordered by this share, closest first.

Method	Overall	Visual quality	Identity	Lip sync	Motion naturalness	Gesture–speech	Artifacts, stability	Avatar 6 wins
From a single photo and speech								
VEED Fabric 1.0	3.68	3.96	4.20	3.71	3.74	3.95	3.70	69.6%
OmniHuman v2	3.32	3.70	3.97	3.44	3.41	3.48	3.36	80.8%
Argil image avatar	3.46	3.76	4.32	3.52	3.55	3.64	3.58	80.9%
Creatify Aurora	3.29	3.33	3.99	3.48	3.40	3.58	3.32	84.3%
Akool Talking Photo	3.19	3.52	4.11	3.28	3.33	3.35	3.40	86.4%
Hedra Character 3	3.16	3.62	4.05	3.24	3.21	3.28	3.33	88.2%
D-ID V2 Talks	1.48	3.43	3.07	2.18	0.94	0.71	2.42	97.9%
Colossyan NEO 1 Instant avatar	1.59	3.40	3.48	1.76	1.19	0.78	2.57	98.1%
Avatar 6	**4.00**	**4.33**	**4.39**	**4.12**	**4.01**	**4.07**	**4.08**	–
From the person's own footage								
Akool Talking Avatar	2.93	3.70	3.89	2.78	2.85	2.72	3.29	92.9%
D-ID V3 Express avatar	2.85	3.74	3.83	2.60	2.76	2.71	3.23	95.3%
Colossyan NEO 2 / NEO 1	2.68	3.59	3.65	2.56	2.69	2.48	3.21	97.3%
NEO 2, scripts 1–3	2.90	3.59	3.58	3.24	2.86	2.58	3.35	96.2%
NEO 1, scripts 4–5	2.35	3.60	3.74	1.55	2.43	2.34	3.01	98.9%
Avatar 6	**3.98**	**4.24**	**4.19**	**4.10**	**4.04**	**4.08**	**3.93**	–

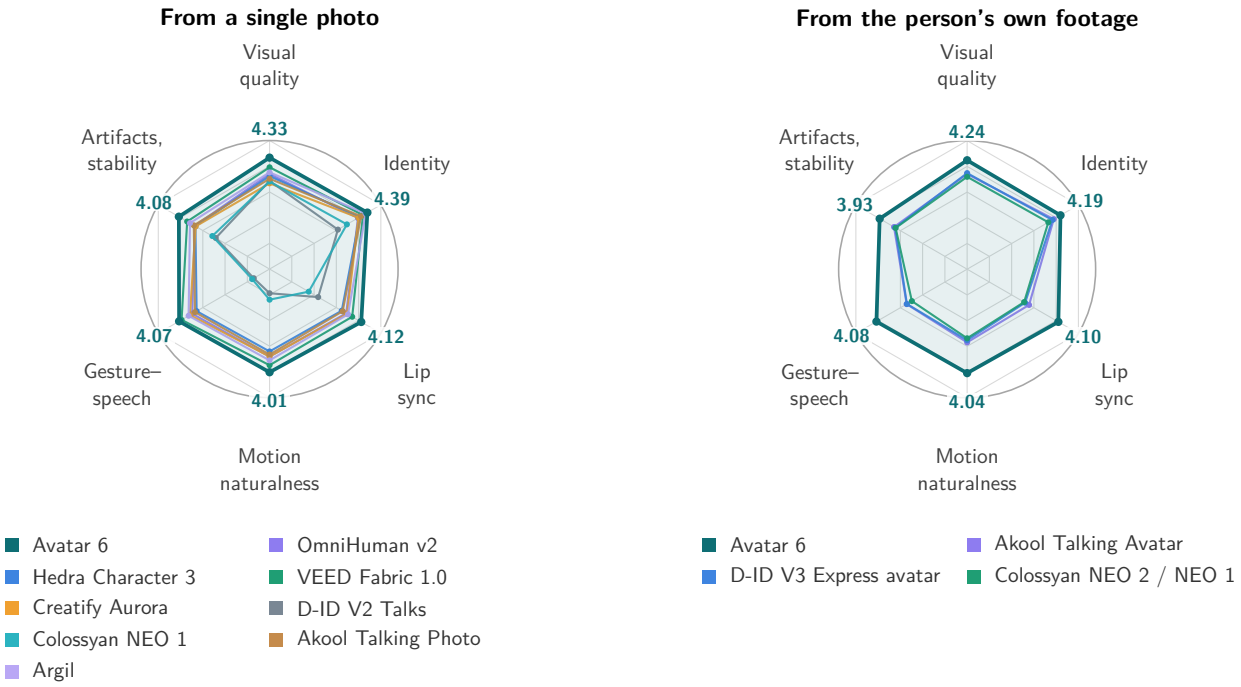


Figure 3 Absolute category rating against commercial avatar products on the six criteria, from a single photo (left) and from the person's own footage (right), on a 0–5 scale (Table 1).

Table 2 Comparison category rating (CCR) against products that render only from their own scripts and voices or their own avatars, by external raters who have never met the people shown and judge overall quality, unlike the self-judged likeness study of Figure 1. Scale −3 to +3; positive values favour Avatar 6. Avatar 6 wins: share of answers with a preference that favour Avatar 6; products ordered as in Table 1. †Custom avatar built from the person's footage, driven by the product's own script and voice. ‡The product's public stock avatars. §Given our photo, or the person's footage as the reference video plus a reference image, with the script in the prompt spoken in the model's own voice; Avatar 6 is driven by that audio. Rated in a separate session by 20 raters.

Avatar 6 vs	Overall	Visual quality	Identity	Lip sync	Motion naturalness	Gesture–speech	Artifacts, stability	Avatar 6 wins
Photo input								
Wan 3.0§	−0.01	+0.55	+0.09	+0.01	−0.18	−0.19	+0.04	50.1%
Seedance 2.5§	+0.30	+1.05	+0.08	+0.03	+0.12	+0.10	−0.01	59.6%
Kling 4.0	+0.79	+0.66	+0.14	+0.42	+0.69	+0.81	+0.30	73.5%
Synthesia Express-2 photo avatar	+1.50	+0.96	+0.55	+1.21	+1.56	+1.51	+0.58	87.0%
Footage input								
Kling 4.0 reference-to-video§	+0.17	+0.73	−0.08	+0.04	+0.01	−0.02	+0.02	54.9%
Seedance 2.5 reference-to-video§	+0.61	+1.10	+0.40	+0.14	+0.18	+0.21	+0.03	72.4%
Wan 3.0 reference-to-video§	+0.75	+0.54	+0.88	+0.31	+0.31	+0.35	+0.29	74.1%
VEED Preset avatars‡	+1.15	+0.71	+0.17	+0.94	+1.12	+1.18	+0.58	87.6%
Creatify AI Avatar v1‡	+1.31	+0.84	+0.25	+1.12	+1.30	+1.37	+0.70	93.3%
Synthesia Personal video avatar†	+1.68	+1.64	+0.79	+1.38	+1.51	+1.56	+1.16	93.6%
Synthesia Stock avatars v3 / v1‡	+1.28	+0.92	+0.24	+1.01	+1.26	+1.37	+0.65	93.7%
v3, four avatars	+1.29	+0.88	+0.23	+1.02	+1.30	+1.39	+0.62	92.8%
v1, Cooper	+1.26	+1.05	+0.28	+1.00	+1.12	+1.27	+0.75	97.5%

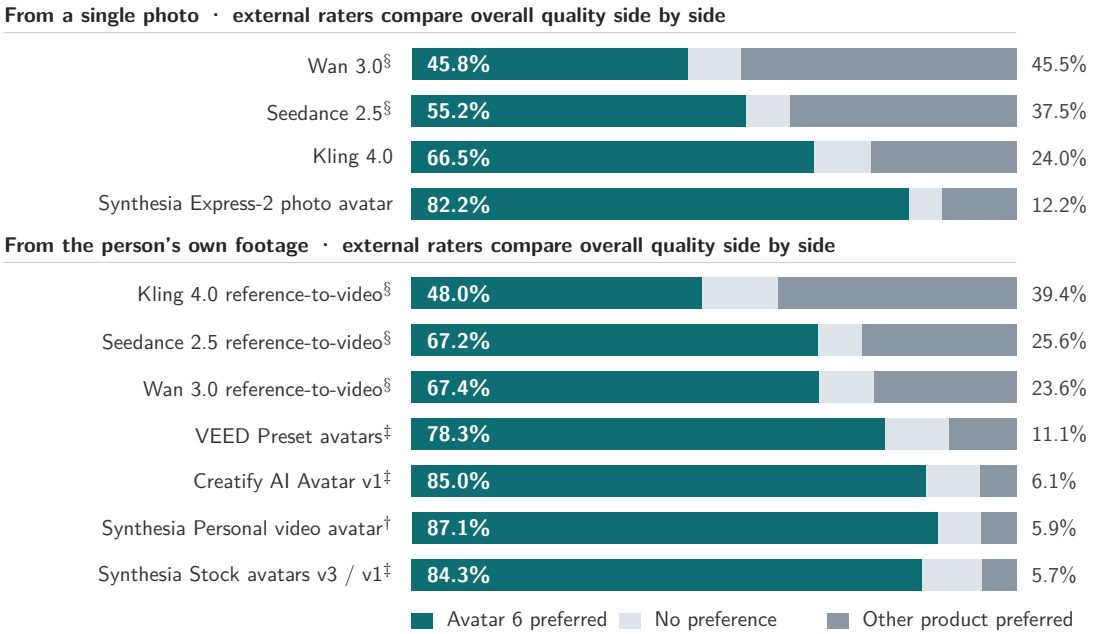


Figure 4 Side-by-side comparison with products that render from their own script, voice or avatars: share of answers preferring Avatar 6, neither, or the other product (Table 2). Unlike Figure 1, where HeyGen employees judge which video looks most like themselves, these raters have never met the people shown and judge overall quality, lip sync and motion; the margins over Seedance 2.5 and Wan 3.0 are therefore smaller.

2.5 Comparison with Real Footage: Turing Test

The strictest test of likeness is whether Avatar 6 can pass for the person next to their own footage. We use real recordings of HeyGen employees; each recording is split in two, one part serving as the Avatar 6 reference and the other held out as the real comparison (Figure 5, Table 3). In the single-clip test, raters label each clip, real or generated, as real or AI, with Avatar 6 driven either by the person's own voice or by a HeyGen Voice clone of it [2]. In the side-by-side test, the real clip and the Avatar 6 clip say the same sentence over the same audio, and raters pick the real one. The single-clip test covers 77 people and the side-by-side test 76, with 20 external raters and about four raters per clip or pair. To our knowledge, this is the largest Turing test of a talking-avatar model in a public technical report.

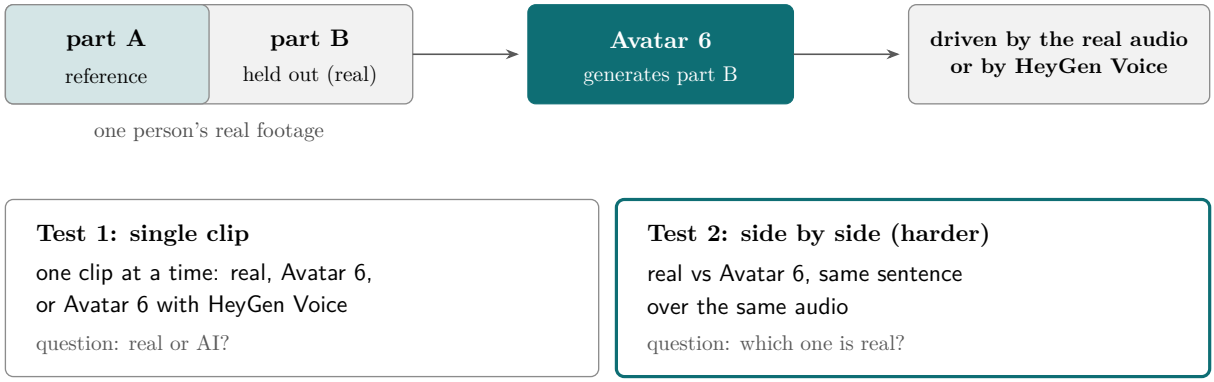


Figure 5 Turing test design. Each person's footage is split; Avatar 6 sees only the reference part and generates the held-out part. Raters judge single clips (test 1) and side-by-side pairs of the same sentence over the same audio (test 2).

Shown alone, Avatar 6 with the person's own voice is judged real in 65.1% of answers, against 78.9% for real recordings (Table 4); even with a HeyGen Voice clone, where neither picture nor sound comes from the

Table 3 Turing test footage per person.

	Range	Median
Recording per person	32–239 s	66 s
Evaluated segment (real recording and Avatar 6)	11.6–44.9 s	28.6 s
Avatar 6 clip with HeyGen Voice	8.2–39.3 s	22.2 s

Table 4 Turing test results from 20 external raters. Side by side, 50% means the two cannot be told apart (indistinguishability 100%); a further 11.2% of the 304 side-by-side answers were "can't tell".

Single clip, shown alone	**Judged real**	**Answers**
Real recording	78.9%	612
Avatar 6, person's own voice	65.1%	304
Avatar 6, HeyGen Voice	56.5%	308
Side by side, same sentence and audio	**Real picked**	**Answers**
Real vs Avatar 6	57.8%	270
Avatar 6 picked as real	42.2%	270
Indistinguishability (2× the above)	84.4%	270

person, 56.5% still judge it real. Side by side with the real person saying the same sentence, raters pick the Avatar 6 clip as the real one in 42.2% of their choices (114 of 270), against 50% if the two could not be told apart. We report this as the indistinguishability rate, twice the share of choices that pick the Avatar 6 clip: 100% when raters cannot tell the two apart and 0% when they always spot it. Avatar 6 reaches 84.4%; measured the same way, Avatar V reaches 44.4% (22.2% picked as real) in its technical report [1].

3 Applications

At its core, Avatar 6 is a talking avatar video generator supporting diverse identities, native output up to 2K and unlimited duration (Section 3.1). This foundation supports a range of advanced applications, including motion prompt control (Section 3.2), multi-speaker dialogue and podcasts (Section 3.3), and talking video editing (Section 3.4), for which we also give a quantitative comparison with other methods.

3.1 Unlimited Talking Avatar Video Generation

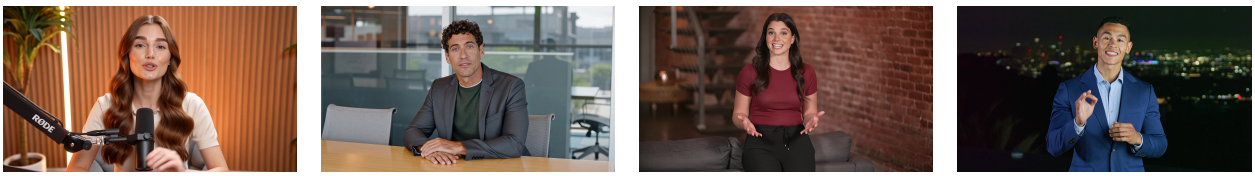
Avatar 6 generates talking videos from a single photo, optionally with a short recording as an identity reference, across diverse identities and visual styles, at up to 2K and with no fixed duration limit.

Any identity. Avatar 6 places no restriction on who appears in the reference: real people in studios, offices, on stage or at concerts, as well as stylised and cartoon characters. In every case it keeps not only the appearance but the subject's own manner of speaking, with the expressions, gestures and postures that make them recognisable.

Native 2K. Avatar 6 supports a range of resolutions and aspect ratios, from portrait to landscape, and natively generates 1080p and 2K: its super-resolution stage, developed in-house jointly with the generator, is part of the model and restores identity-specific detail such as skin texture and hair.

Unlimited duration. Avatar 6 imposes no fixed limit on video length. A 30-minute talking video is generated in a single pass, with identity, framing and this personal repertoire of motion maintained over the full duration.

Figure 6 shows all three at once: four avatars, each created from a single photo, AI-generated characters and real people. Every frame is generated at 2880×1620, and each video is 25 to 28 minutes of one continuous generation from one audio track, with no loops. The length follows the audio.



26 min · 2880×1620 **26** min · 2880×1620 **25** min · 2880×1620 **28** min · 2880×1620

Figure 6 Any identity. Native 2K. Unlimited duration. *Any avatar.* Each created from a single photo: AI-generated characters and real people. *Native 2K.* Every frame generated at 2880×1620. *Unlimited duration.* 25–28 min per video here, each one continuous generation from one audio track, no loops. Length follows the audio. Full videos: https://www.heygen.com/research/avatar-6-model#identity.

3.2 Motion Prompt Control

Avatar 6 accepts an optional written motion prompt that can request gestures on specific phrases, counted gestures such as holding up three fingers, facial expressions, and head or body actions. An in-house motion planner aligns each requested action with the words it refers to and blends it into the person's own motion style. On requests with a motion prompt, Avatar 6 is preferred over Avatar V in 78.1% of votes in online pairwise preference tests.

Figure 7 shows Avatar 6 following written motion prompts, from a thumbs up or a wave to pointing, cupping an ear or counting on the fingers. Each action lands on its words, while identity, setting, lip sync and the person's own motion style between actions are preserved.

3.3 Multi-Speaker Dialogue and Podcasts

Avatar 6 generates two speakers within a single frame, supporting interview, dialogue and podcast formats. Figure 8 shows frames from a two-minute podcast between a host and a guest, alternating between wide shots with both people in frame and closer views of each speaker.

3.4 Talking Video Editing

Avatar 6 edits existing recordings at the level of words and motion. A word or phrase can be replaced, a gesture can be added on a chosen word, or both at once. The new speech is synthesised in the person's own voice, and only the span around the edit is regenerated, so the rest of the recording is preserved as filmed. General video generation models instead regenerate the whole clip. We compare Avatar 6 with MiniMax H3-max, Wan 3.0, Seedance 2.5 and HappyHorse Edit on word, motion and combined edits of real recordings.

Avatar 6 leads on every metric (Table 5). It is the only method that keeps the original recording: frames outside the edit match the original at 48.3 dB, the level of video encoding noise, while every other method alters 99.7% to 100.0% of the frames. On the new words its lip sync exceeds that of the original recording (LSE-C 7.67 against 6.25), and the new speech matches the speaker's own voice. In motion and combined edits, the requested gesture appears on the chosen word while the footage around it is preserved (Figure 9).



"Hello" thumbs up | *"the counter on the left"* point to the side | *"always happy to chat"* hand on heart | *"some water"* wave

Prompt: Friendly pharmacist. On 'Hello', give a thumbs up. On 'the counter on the left', point to the side with your index finger. On 'always happy to chat', place your right hand on your heart. On 'Take a seat', open both arms wide. On 'some water', wave your right hand.



"Welcome" arms open, palms up | *"grandparents"* thumbs up | *"front desk"* wave | *"Thanks for trusting us"* hand on heart

Prompt: Warm clinic host. On 'Welcome', open both arms wide, palms up. On 'grandparents', give a thumbs up. On 'front desk', wave your right hand. On 'Thanks for trusting us', place your right hand on your heart.



"not the melody" wag index finger | *"hear it"* cup an ear | *"no, slower"* palm to the camera | *"the taste, the choices and the feeling"* count to three

Prompt: Playful storyteller. On 'not the melody', wag your index finger. On 'hear it', cup a hand behind your ear. On 'no, slower', raise a flat palm to the camera. On 'rain on a window', flutter your fingers downward. On 'the taste, the choices and the feeling', count to three on your fingers.

Figure 7 Motion prompt following. Each row shows frames from one generated video, labelled with the spoken words each action is tied to; the written prompt is shown below the row.



Figure 8 A two-speaker podcast generated by Avatar 6. Full video: https://www.heygen.com/research/avatar-6-model#podcast.

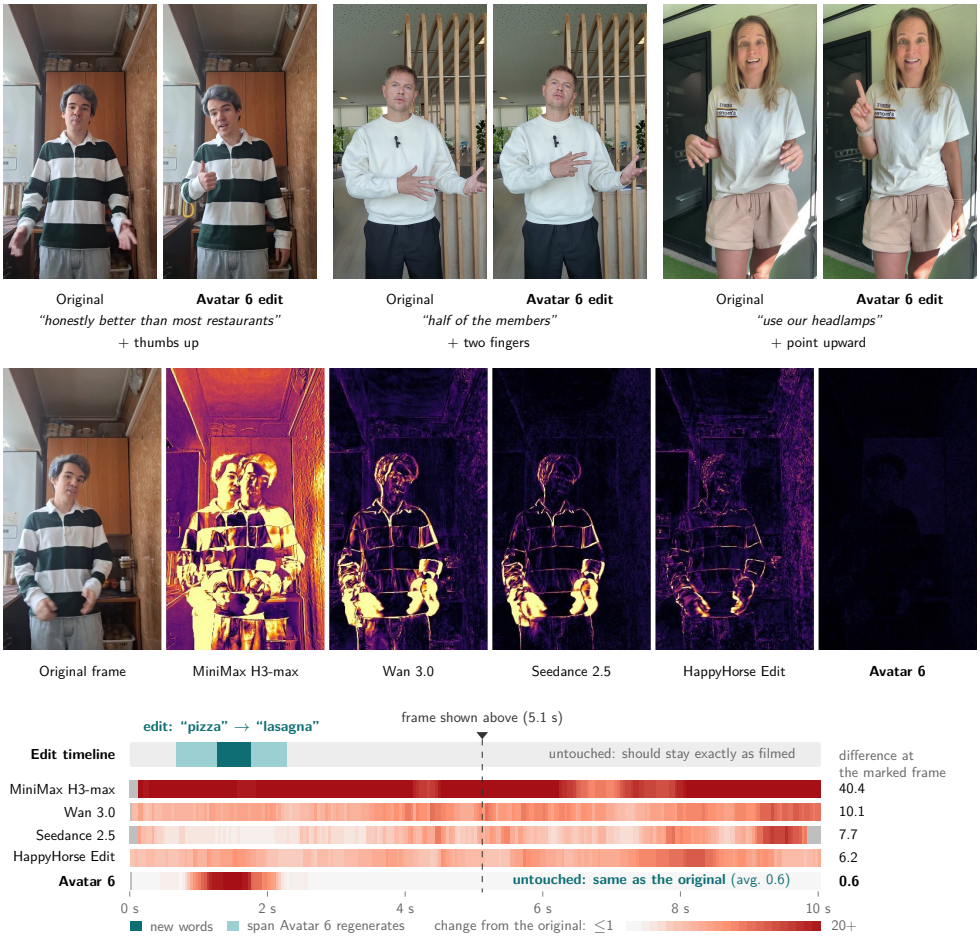


Figure 9 Talking video editing. Top: combined edits that change the words and add a gesture on a chosen word. Bottom: a word edit ("made a pizza" → "made a lasagna"), showing the difference from the original on a frame outside the edit (brighter means a larger change). Avatar 6 leaves the frame visually unchanged, while the other methods regenerate the person and the scene. The timeline below places that frame on the clip: it marks the replaced words and the span Avatar 6 regenerates, and shows each method's difference from the original frame by frame. Outside that span Avatar 6 stays at the level of video encoding noise (0.6 on a 0–255 scale at the marked frame), while the other methods change every frame of the clip.

Table 5 Talking video editing on real recordings. Untouched PSNR compares frames outside the edit with the original; face similarity uses ArcFace [50], voice similarity a speaker embedding [51], and lip sync SyncNet [49] on the new words. Frames changed is the share of frames whose mean absolute luma difference from the original exceeds 2 (of 255), averaged over the 20 edits.

Method	Untouched PSNR (dB) ↑	Frames changed ↓	Face similarity ↑	Voice similarity ↑	Lip sync LSE-C ↑	Lip sync LSE-D ↓
MiniMax H3-max	19.5	100.0%	0.907	0.717	7.25	7.75
Wan 3.0	25.8	100.0%	0.835	0.727	5.70	8.89
Seedance 2.5	29.3	99.7%	0.883	0.739	3.46	10.89
HappyHorse Edit	22.8	100.0%	0.751	0.543	2.47	10.36
Avatar 6	**48.3**	**34.4%**	**0.913**	**0.763**	**7.67**	**7.19**

References

[1] HeyGen Research. Avatar V: Scaling Video-Reference Avatar Generation. Technical report, April 2026. https://www.heygen.com/research/avatar-v-model

[2] HeyGen Research. HeyGen Voice: Scene-Aware Voice Cloning for Talking Videos. Technical report, October 2026. https://www.heygen.com/research/heygen-voice-scene-aware-voice-cloning

[3] HeyGen Research. One Model, Multiple Controllability: Composable Behavior LoRAs for Video Avatars. Technical report, July 2026. https://www.heygen.com/research/composable-behavior-loras-for-video-avatars

[4] K. R. Prajwal, R. Mukhopadhyay, V. P. Namboodiri and C. V. Jawahar. A Lip Sync Expert Is All You Need for Speech to Lip Generation in the Wild. *ACM Multimedia*, 2020.

[5] 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. *CVPR*, 2023.

[6] S. Xu, G. Chen, Y.-X. Guo, J. Yang, C. Li, Z. Zang, Y. Zhang, X. Tong and B. Guo. VASA-1: Lifelike Audio-Driven Talking Faces Generated in Real Time. *NeurIPS*, 2024.

[7] L. Tian, Q. Wang, B. Zhang and L. Bo. EMO: Emote Portrait Alive – Generating Expressive Portrait Videos with Audio2Video Diffusion Model under Weak Conditions. *ECCV*, 2024.

[8] M. Xu, H. Li, Q. Su, H. Shang, L. Zhang, C. Liu, J. Wang, Y. Yao and S. Zhu. Hallo: Hierarchical Audio-Driven Visual Synthesis for Portrait Image Animation. arXiv:2406.08801, 2024.

[9] Z. Chen, J. Cao, Z. Chen, Y. Li and C. Ma. EchoMimic: Lifelike Audio-Driven Portrait Animations through Editable Landmark Conditions. *AAAI*, 2025.

[10] J. Jiang, C. Liang, J. Yang, G. Lin, T. Zhong and Y. Zheng. Loopy: Taming Audio-Driven Portrait Avatar with Long-Term Motion Dependency. *ICLR*, 2025.

[11] J. Ho, A. Jain and P. Abbeel. Denoising Diffusion Probabilistic Models. *NeurIPS*, 2020.

[12] R. Rombach, A. Blattmann, D. Lorenz, P. Esser and B. Ommer. High-Resolution Image Synthesis with Latent Diffusion Models. *CVPR*, 2022.

[13] A. Blattmann et al. Stable Video Diffusion: Scaling Latent Video Diffusion Models to Large Datasets. arXiv:2311.15127, 2023.

[14] X. Ji et al. Sonic: Shifting Focus to Global Audio Perception in Portrait Animation. *CVPR*, 2025.

[15] R. Meng, X. Zhang, Y. Li and C. Ma. EchoMimicV2: Towards Striking, Simplified, and Semi-Body Human Animation. *CVPR*, 2025.

[16] J. Cui et al. Hallo2: Long-Duration and High-Resolution Audio-Driven Portrait Image Animation. *ICLR*, 2025.

[17] W. Peebles and S. Xie. Scalable Diffusion Models with Transformers. *ICCV*, 2023.

[18] W. Kong et al. HunyuanVideo: A Systematic Framework for Large Video Generative Models. arXiv:2412.03603, 2024.

[19] Wan Team. Wan: Open and Advanced Large-Scale Video Generative Models. arXiv:2503.20314, 2025.

[20] Y. Lipman, R. T. Q. Chen, H. Ben-Hamu, M. Nickel and M. Le. Flow Matching for Generative Modeling. *ICLR*, 2023.

[21] J. Cui et al. Hallo3: Highly Dynamic and Realistic Portrait Image Animation with Video Diffusion Transformer. *CVPR*, 2025.

[22] G. Lin, J. Jiang, J. Yang, Z. Zheng and C. Liang. OmniHuman-1: Rethinking the Scaling-Up of One-Stage Conditioned Human Animation Models. *ICCV*, 2025.

[23] J. Jiang et al. OmniHuman-1.5: Instilling an Active Mind in Avatars via Cognitive Simulation. arXiv:2508.19209, 2025.

[24] M. Wang et al. FantasyTalking: Realistic Talking Portrait Generation via Coherent Motion Synthesis. *ACM Multimedia*, 2025.

[25] Q. Gan, R. Yang, J. Zhu, S. Xue and S. Hoi. OmniAvatar: Efficient Audio-Driven Avatar Video Generation with Adaptive Body Animation. arXiv:2506.18866, 2025.

[26] Y. Chen et al. HunyuanVideo-Avatar: High-Fidelity Audio-Driven Human Animation for Multiple Characters. arXiv:2505.20156, 2025.

[27] Z. Kong et al. Let Them Talk: Audio-Driven Multi-Person Conversational Video Generation. *NeurIPS*, 2025.

[28] X. Gao et al. Wan-S2V: Audio-Driven Cinematic Video Generation. arXiv:2508.18621, 2025.

[29] S. Tu et al. StableAvatar: Infinite-Length Audio-Driven Avatar Video Generation. arXiv:2508.08248, 2025.

[30] S. Yang et al. InfiniteTalk: Audio-driven Video Generation for Sparse-Frame Video Dubbing. arXiv:2508.14033, 2025.

[31] Y. Ding et al. Kling-Avatar: Grounding Multimodal Instructions for Cascaded Long-Duration Avatar Animation Synthesis. arXiv:2509.09595, 2025.

[32] T. Brooks et al. Video Generation Models as World Simulators. Technical report, OpenAI, 2024. https://openai.com/research/video-generation-models-as-world-simulators

[33] Google DeepMind. Veo 3 Tech Report. 2025. https://storage.googleapis.com/deepmind-media/veo/Veo-3-Tech-Report.pdf

[34] Y. Gao et al. Seedance 1.0: Exploring the Boundaries of Video Generation Models. arXiv:2506.09113, 2025.

[35] Seedance Team, ByteDance Seed. Seedance 1.5 pro: A Native Audio-Visual Joint Generation Foundation Model. arXiv:2512.13507, 2025.

[36] Seedance Team, ByteDance Seed. Seedance 2.0: Advancing Video Generation for World Complexity. arXiv:2604.14148, 2026.

[37] ByteDance. OmniHuman. https://omnihuman-lab.github.io

[38] Hedra. Character 3. https://www.hedra.com

[39] VEED. Fabric 1.0. https://www.veed.io

[40] Creatify. Aurora. https://creatify.ai

[41] D-ID. https://www.d-id.com

[42] Colossyan. https://www.colossyan.com

[43] Akool. https://akool.com

[44] Argil. https://www.argil.ai

[45] Synthesia. https://www.synthesia.io

[46] Kuaishou. Kling. https://klingai.com

[47] Alibaba. Wan. https://wan.video

[48] ByteDance Seed. Seedance. https://seed.bytedance.com/en/seedance

[49] J. S. Chung and A. Zisserman. Out of Time: Automated Lip Sync in the Wild. *ACCV Workshops*, 2016.

[50] J. Deng, J. Guo, N. Xue and S. Zafeiriou. ArcFace: Additive Angular Margin Loss for Deep Face Recognition. *CVPR*, 2019.

[51] L. Wan, Q. Wang, A. Papir and I. L. Moreno. Generalized End-to-End Loss for Speaker Verification. *ICASSP*, 2018.

[52] Avatar Arena. Leaderboard. https://avatar.newmeasure.ai/leaderboard

[53] ITU-T. Recommendation P.800: Methods for subjective determination of transmission quality. International Telecommunication Union, 1996.

Contributions and Acknowledgments

All authors of Avatar 6 are listed in alphabetical order by their last names.

[TBD: contributor names.]