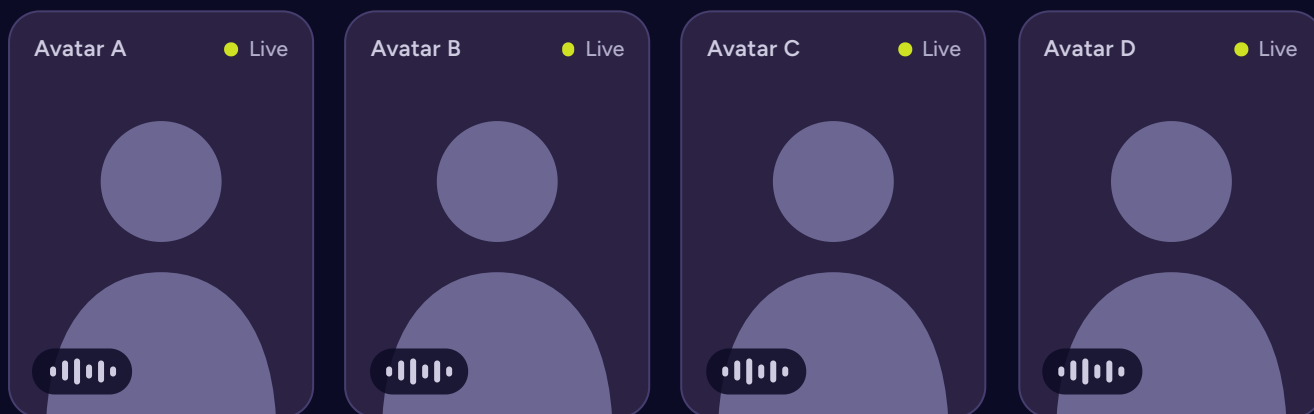


Real-Time Avatar Comparison Study

Which real-time avatar do people **rate highest**?



100+ participants held live conversations with four real-time avatars, without knowing which provider built each one, and rated them on five criteria.

100+ Participants	4 Real-time avatar providers	5 Criteria rated
-----------------------------	--	----------------------------

Contents

Part 1 · Study design

1	Study introduction and materials	4
2	Agents and tools	5
3	Methodology	6
4	Understanding the measures	7

Part 2 · Results

5	Results overview	10
6	Visual quality	11
7	Expressions	13
8	Facial movement	15
9	Overall experience	17
10	Naturalness	19
11	Comparison with Tavus	21
12	Complete mean ratings and study conclusions	23

Part 3 · Appendices

A	Appendix A · Recording guidelines	25
B	Appendix B · Avatar information	25
C	Appendix C · Integration references	26
D	Appendix D · Conversation prompt	26

About this report

This report covers Beyond Presence Genesis 2, HeyGen LiveAvatar, Tavus Phoenix 4 and D-ID V3 Pro.

The results present mean opinion scores, participant rating comparisons and Elo estimates, all calculated from the raw ratings of the retained participants. The methodology and measurement notes explain the scope, uncertainty and calculation of each measure.

Part 1

Study design

How the experiment was set up: the avatars, the shared agent configuration, the task and evaluation criteria, and how each measure is calculated.

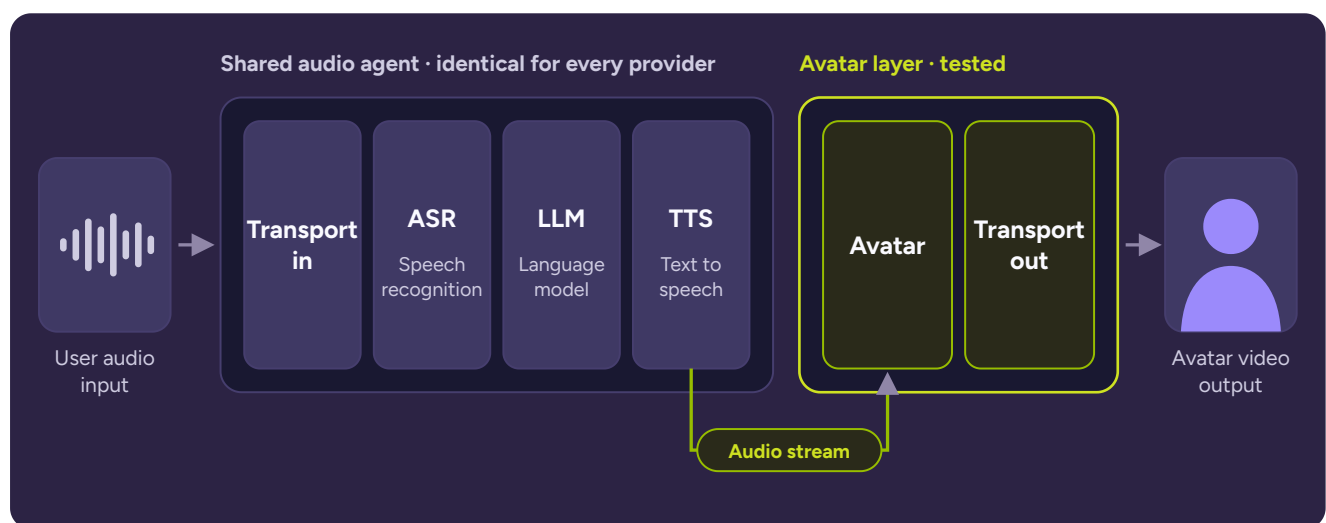
1	Study introduction and materials	04
2	Agents and tools	05
3	Methodology	06
4	Understanding the measures	07

1) Study introduction and materials

Purpose of the study

This study evaluates how people experience photorealistic conversational avatars. The comparison covers four model versions as tested in August 2026: Beyond Presence Genesis 2, HeyGen LiveAvatar, Tavus Phoenix 4 and D-ID V3 Pro. An independent researcher conducted the experiment, commissioned by Beyond Presence.

Conversational video agents combine automatic speech recognition, a language model, text to speech and an avatar. Each of these can affect how a conversation feels, so the study held everything except the avatar constant. The test isolates the video avatar layer.



Components of a conversational video agent, redrawn from the original study. The audio agent was identical for every provider; only the avatar layer changed. Beyond Presence is shown as the example avatar.

Avatars

Custom avatars represented the same contracted human model. Training videos followed each provider's recording guidelines, and the custom avatars were created between July 22 and August 4, 2026. The source materials, model versions and creation dates are listed in the appendices.

Comparison scope

This report covers four providers, each using a custom avatar trained from video: Beyond Presence Genesis 2, HeyGen LiveAvatar, Tavus Phoenix 4 and D-ID V3 Pro. The findings apply to those model versions, and to the identity, task and configurations tested, rather than to every available conversational avatar product or to later releases. Model details are listed in Appendix B.

2) Agents and tools

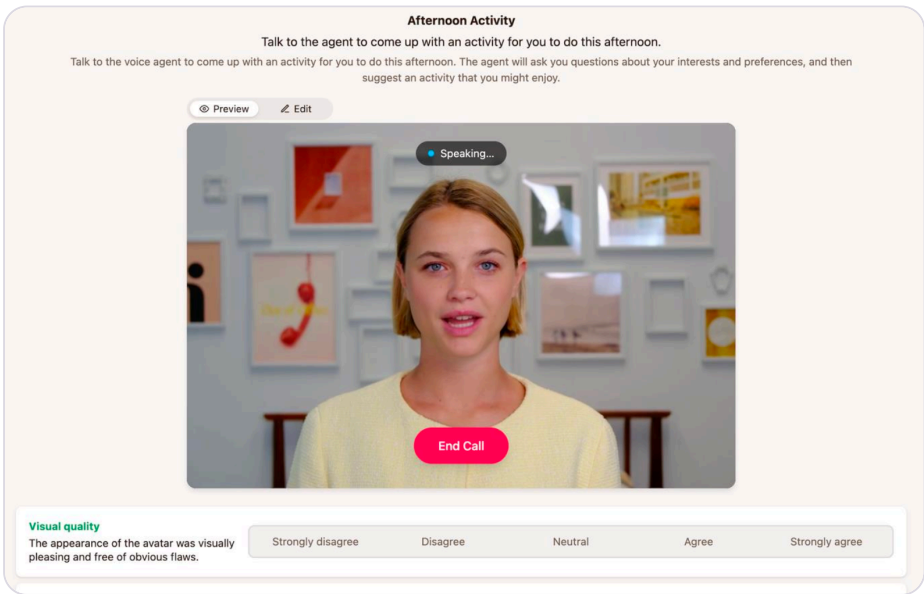
Agents

All four providers were tested inside the same agent. Speech recognition, language model, text to speech, voice and conversation prompt were identical in every condition, so the avatar layer was the only component that differed between providers. Beyond Presence, HeyGen LiveAvatar and Tavus ran on the same LiveKit agent base. D-ID used a different integration path because the avatar model tested could not be integrated through LiveKit.

Shared component	Configuration
Speech recognition	Deepgram Nova 3
Language model	OpenAI GPT 5.4 Mini
Text to speech	Cartesia Sonic 3.5
Voice identifier	25d7abcb-4d6d-4aca-adce-8a1c85620c8b
Conversation prompt	Same prompt for every provider (Appendix D)
Agent framework	LiveKit (D-ID: separate integration path)

Experiment platform

The experiment ran on Mabyduck. Avatars appeared in a 640 by 480 pixel embedded window. Participants provided ratings on a five point agreement scale from minus 2 to plus 2 after each interaction.



The experiment interface. Every condition used the same display window size; the study did not compare output at each provider's native resolution.

3) Methodology

The study ran from August 5 to 13, 2026. If a participant reported any technical issue, such as an avatar failing to load, all data from that participant were discarded. The retained analysis included 119 participants.

Each participant evaluated one interaction per provider and scored five criteria after each interaction. Provider presentation order was randomised. Participants were not told which company they were evaluating.

Task

Before each session, participants completed a practice conversation to become familiar with the interface and task. They then discussed possible activities for a free afternoon with each agent. Conversations averaged 98.4 seconds and were limited to two minutes. The shared prompt appears in Appendix D.

Evaluation criteria

Criterion	Statement rated by participants
Visual quality	The appearance of the avatar was visually pleasing and free of obvious flaws.
Expressions	The avatar's facial expressions matched the emotional tone of the spoken audio.
Facial movement	The avatar's mouth and facial movements matched the spoken audio.
Overall experience	I enjoyed talking to this avatar and found the overall experience positive.
Naturalness	The avatar's behavior seemed natural and lifelike.

Response scale

-2	-1	0	+1	+2
Strongly disagree	Disagree	Neutral	Agree	Strongly agree

Higher values indicate stronger agreement with the statement. Zero represents a neutral response. Overall experience was rated through its own question and is not the average of the other four criteria.

4) Understanding the measures

Mean opinion scores

Mean opinion scores, or MOS, average the questionnaire ratings for each provider. The scale runs from minus 2 to plus 2, with zero representing neutral. MOS whiskers reproduce two standard errors of the mean from the original figures.

Participant rating comparisons

For each provider pair, the same participant's ratings are compared. A higher Beyond Presence score is counted as a higher rating, the same score as an equal rating and a lower score as a lower rating. These preferences are inferred from separate ratings, rather than collected as direct choices between two avatars.

"Same or higher" combines the equal and higher shares. A rating ratio divides the higher share by the lower share. For visual quality against Tavus, the approximate shares are 33% higher, 54% equal and 13% lower, giving about 87% same or higher and a ratio of about 2.5 to 1.

Elo estimates and what they mean

Elo estimates use a Bayesian Bradley Terry model, the approach used by the Mabyduck platform, to estimate relative strengths from the head to head comparisons contained in each participant's ratings. Only the gaps between providers carry meaning: the scores are relative to the providers included, and are centred on 2000 across the four providers compared.

A gap can be read as a win share. It answers one practical question: out of every 100 head to head comparisons between two providers, how many does the model expect Beyond Presence to win, counting an equal rating as half a win for each side? A gap of 0 points gives 50 out of 100, 50 points gives about 57, and 100 points about 64. The formula below converts a score difference Δ , equal to Beyond Presence minus the competitor, into that win share and the matching odds.

$$P(\text{BP}) = \frac{1}{1 + 10^{-\Delta/400}} \quad \text{Odds}(\text{BP}) = 10^{\Delta/400}$$

Understanding the measures

For visual quality against Tavus, the 56.7 point Elo gap gives a win share of approximately 58 in 100 and odds of 1.39 to 1. This is a model output, not a count of participants: it differs from the observed shares, where 33% rated Beyond Presence higher, 54% rated the two the same and 13% rated Tavus higher. It is also not the probability that one product is truly better.

The transformations use posterior mean estimates without propagating uncertainty. Elo whiskers show 95% credible intervals, which is the Bayesian counterpart of a confidence interval and the correct term for estimates from this model; the platform dashboard labels the equivalent bars confidence intervals. The calculation was carried out independently of that dashboard, so its estimates can differ slightly from platform output.

Figure sources

MOS values are taken from the study platform. Participant shares and interval endpoints are approximate readings of the original figures. Shares are rounded to total 100, and rating ratios use those rounded shares. Elo estimates and their intervals are calculated from the raw ratings exported from the platform. Shares and MOS intervals have not been recalculated from participant data.

Method reference: Caron and Doucet (2012), Efficient Bayesian inference for generalized Bradley-Terry models, https://www.stats.ox.ac.uk/~doucet/caron_doucet_bayesianbradleyterry.pdf

Part 2

Results

The overall picture, then mean opinion scores, Elo estimates and participant rating comparisons for each criterion, the head to head comparison with Tavus, and the study conclusions.

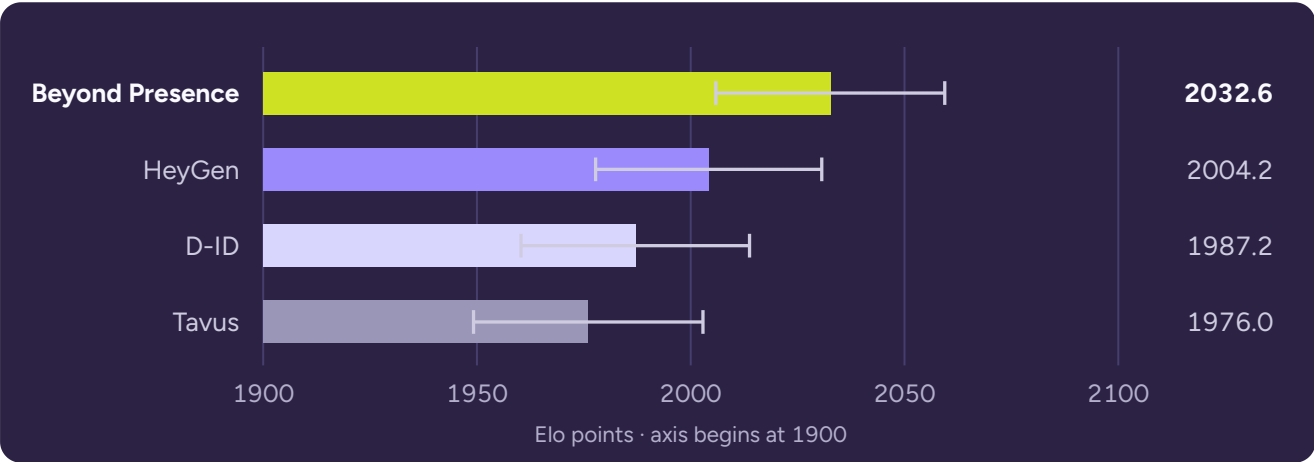
5	Results overview	10
6	Visual quality	11
7	Expressions	13
8	Facial movement	15
9	Overall experience	17
10	Naturalness	19
11	Comparison with Tavus	21
12	Complete mean ratings and study conclusions	23

5) Results overview

Beyond Presence recorded the highest visual quality scores among the four providers compared. Participant rating comparisons favoured Beyond Presence over Tavus across all five categories, and its expression Elo estimate was nearly identical to HeyGen LiveAvatar.

Finding 01

Highest visual quality scores in this comparison



Posterior mean Elo estimates calculated for the four providers compared. Whiskers show 95% credible intervals. The Elo axis begins at 1900; bar lengths do not represent ratios of quality.

Finding 02

More higher ratings than Tavus in every category

Beyond Presence received more higher than lower participant ratings against Tavus for visual quality, expressions, naturalness, facial movement and overall experience. For visual quality, approximately 2.5 times as many participants rated Beyond Presence higher as rated Tavus higher, based on the rounded chart shares.

Finding 03

Close expression scores alongside HeyGen

Beyond Presence scored 2009.9 Elo for expressions, compared with 2014.0 for HeyGen, a difference of 4.1 points. Beyond Presence also had the higher visual quality point estimate.

These findings describe the tested configurations and point estimates. They apply to the model versions tested in August 2026: Beyond Presence Genesis 2, HeyGen LiveAvatar, Tavus Phoenix 4 and D-ID V3 Pro. Approximate participant shares are reconstructed from the original figures. Elo estimates were calculated from the raw ratings for the four providers compared.

6) Visual quality

Statement rated by participants

The appearance of the avatar was visually pleasing and free of obvious flaws.

Highest visual quality scores among the four providers

Beyond Presence recorded the highest visual quality mean and Elo point estimates in this comparison. Approximately 87% rated its visual quality the same as or higher than Tavus, including 33% who rated it higher. Against HeyGen, about 24% rated Beyond Presence higher and 15% rated HeyGen higher. Against HeyGen, about 24% rated Beyond Presence higher and 15% rated HeyGen higher.

+0.76

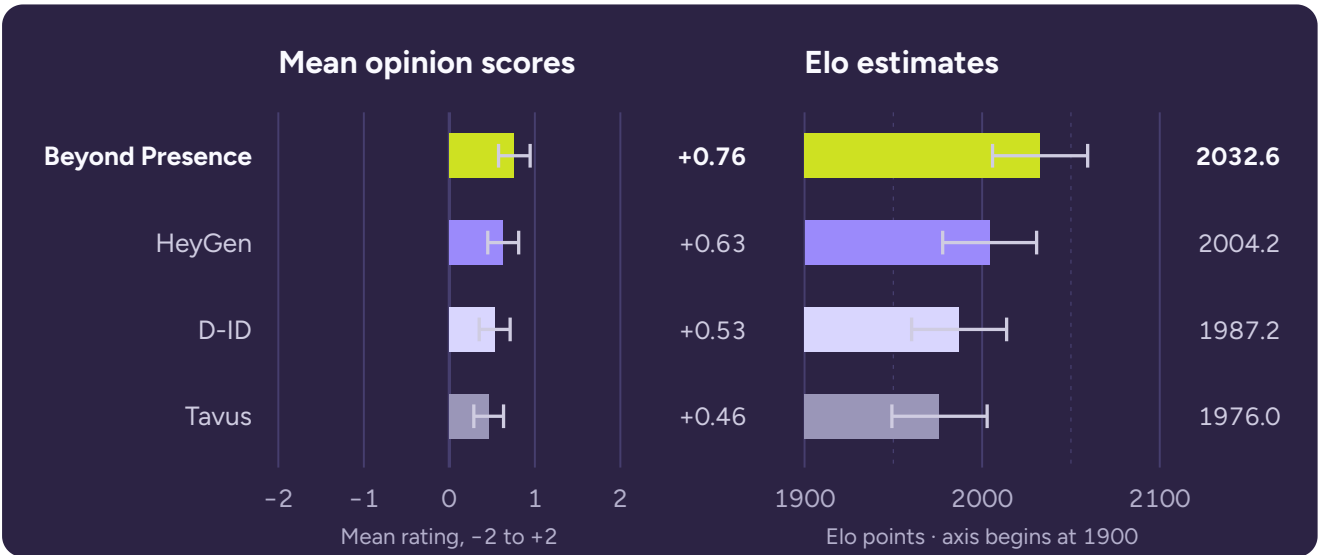
Mean opinion score

2032.6

Elo estimate

≈87%

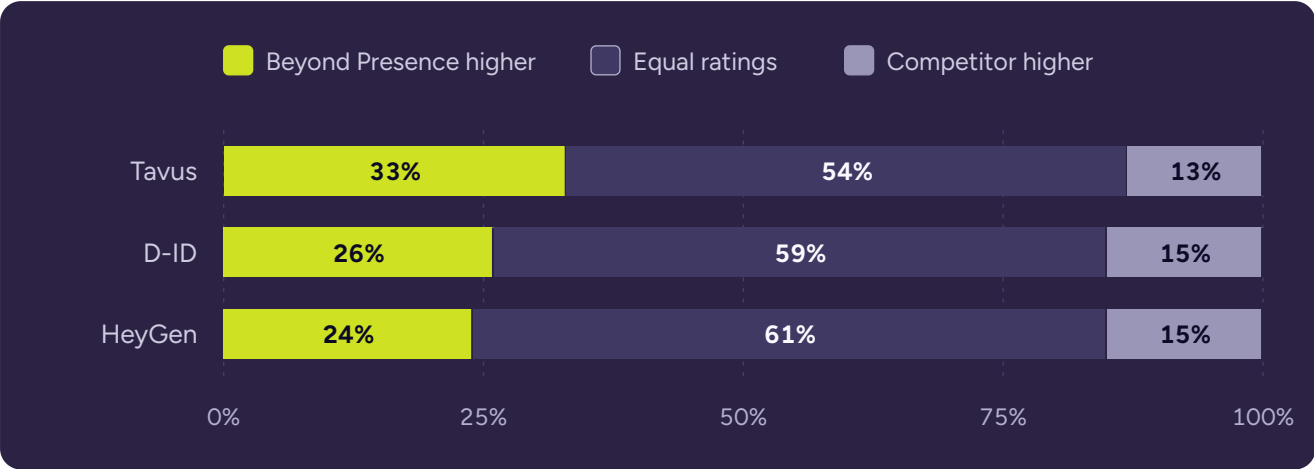
Same or higher than Tavus



Versions tested: Beyond Presence Genesis 2, HeyGen LiveAvatar, Tavus Phoenix 4, D-ID V3 Pro. MOS values are from the study platform. MOS whiskers are approximate readings of the original figure and show two standard errors. Elo estimates are posterior means calculated for the four providers compared; whiskers show 95% credible intervals. Elo bar lengths are not quality ratios.

Visual quality

Participant rating comparisons



Versions tested: Genesis 2, LiveAvatar, Phoenix 4, V3 Pro. Approximate participant shares from the original chart, rounded to whole percentages that total 100. Equal ratings mean the same questionnaire response.

Comparison summary

Compared with	BP same or higher	BP higher ratings per lower rating	Modelled Elo win share for BP
Tavus	≈87%	≈2.5 to 1	58.1%
D-ID	≈85%	≈1.7 to 1	56.5%
HeyGen	≈85%	≈1.6 to 1	54.1%

BP denotes Beyond Presence. Rating ratios compare the higher and lower shares, excluding equal ratings. The Elo win share is modelled from the score difference, not an observed participant share. See Understanding the measures for interpretation.

7) Expressions

Statement rated by participants

The avatar's facial expressions matched the emotional tone of the spoken audio.

Expression estimates closely match HeyGen

Beyond Presence and HeyGen recorded expression Elo estimates of 2009.9 and 2014.0. The 4.1 point gap translates to a modelled comparison probability of approximately 49.4% for Beyond Presence. Against Tavus, around 32% rated Beyond Presence higher and 21% rated Tavus higher.

2009.9

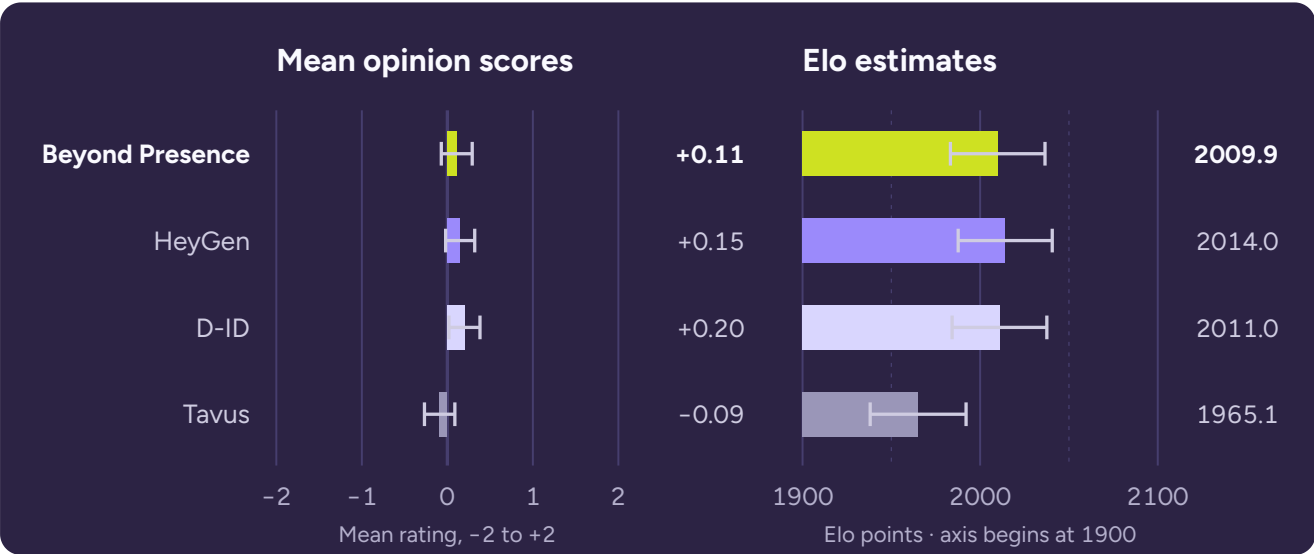
Beyond Presence Elo

2014.0

HeyGen Elo

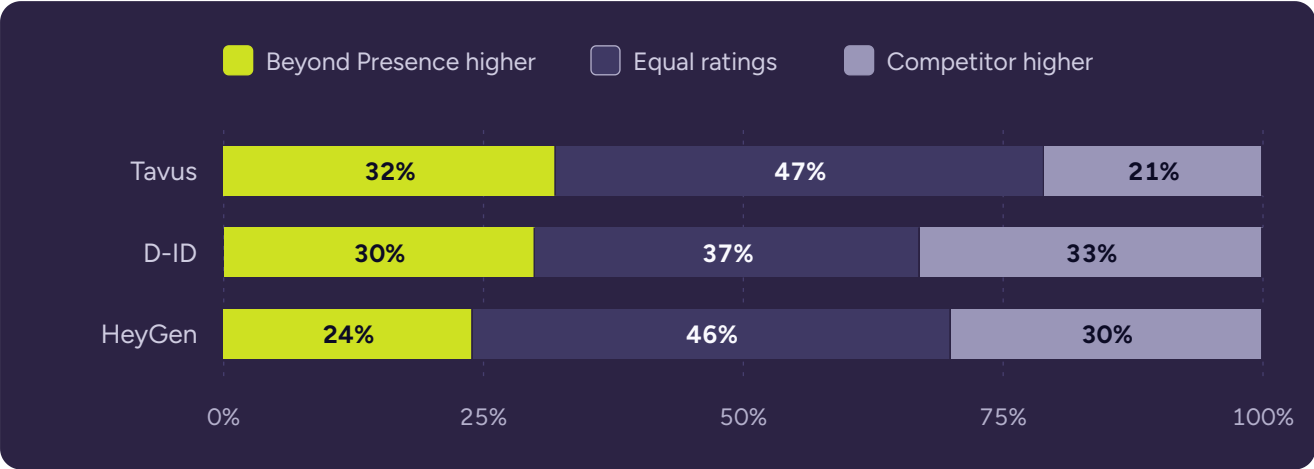
4.1

Point gap



Versions tested: Beyond Presence Genesis 2, HeyGen LiveAvatar, Tavus Phoenix 4, D-ID V3 Pro. MOS values are from the study platform. MOS whiskers are approximate readings of the original figure and show two standard errors. Elo estimates are posterior means calculated for the four providers compared; whiskers show 95% credible intervals. Elo bar lengths are not quality ratios.

Participant rating comparisons



Versions tested: Genesis 2, LiveAvatar, Phoenix 4, V3 Pro. Approximate participant shares from the original chart, rounded to whole percentages that total 100. Equal ratings mean the same questionnaire response.

Comparison summary

Compared with	BP same or higher	BP higher ratings per lower rating	Modelled Elo win share for BP
Tavus	≈79%	≈1.5 to 1	56.4%
D-ID	≈67%	≈0.9 to 1	49.8%
HeyGen	≈70%	≈0.8 to 1	49.4%

BP denotes Beyond Presence. Rating ratios compare the higher and lower shares, excluding equal ratings. The Elo win share is modelled from the score difference, not an observed participant share. See Understanding the measures for interpretation.

8) Facial movement

Statement rated by participants

The avatar's mouth and facial movements matched the spoken audio.

More than twice as many higher ratings against Tavus

Approximately 35% rated Beyond Presence higher than Tavus for facial movement, compared with 15% who rated Tavus higher. This is about 2.3 higher ratings for Beyond Presence for every higher rating for Tavus. The criterion covers mouth and facial movements matching the spoken audio, including lip sync.

≈35%

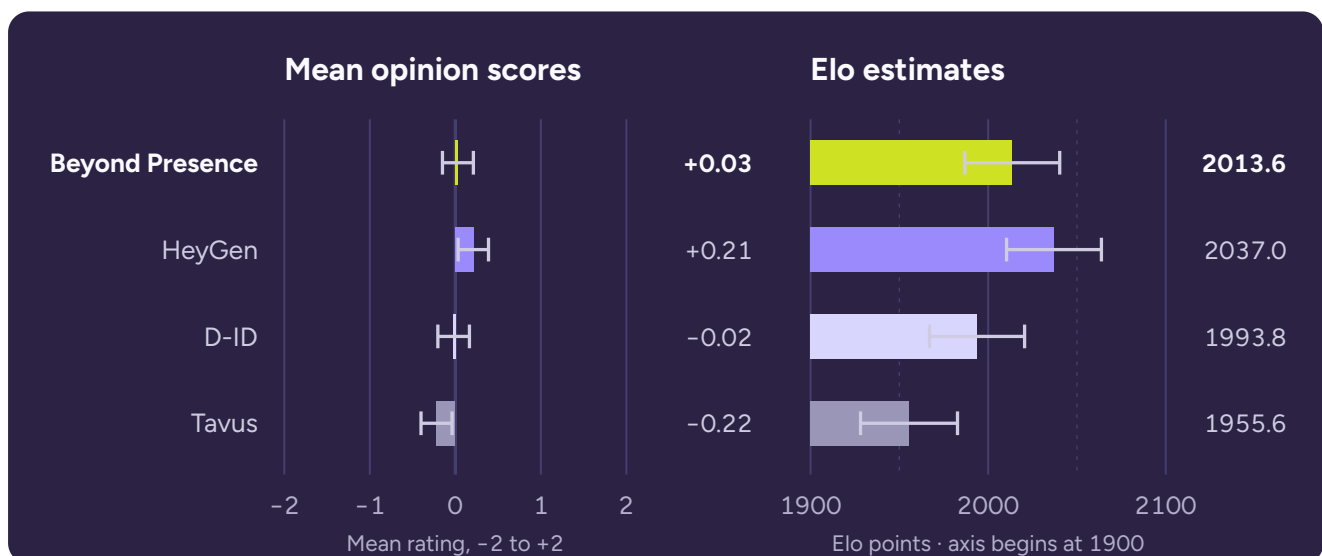
Rated BP higher than Tavus

≈15%

Rated Tavus higher

≈2.3 to 1

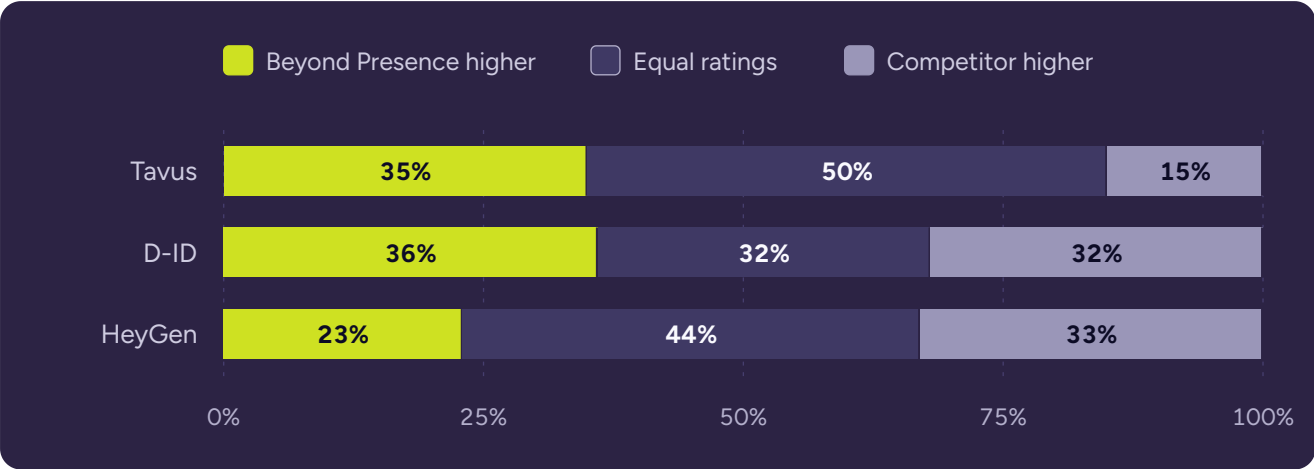
Higher per lower rating



Versions tested: Beyond Presence Genesis 2, HeyGen LiveAvatar, Tavus Phoenix 4, D-ID V3 Pro. MOS values are from the study platform. MOS whiskers are approximate readings of the original figure and show two standard errors. Elo estimates are posterior means calculated for the four providers compared; whiskers show 95% credible intervals. Elo bar lengths are not quality ratios.

Facial movement

Participant rating comparisons



Versions tested: Genesis 2, LiveAvatar, Phoenix 4, V3 Pro. Approximate participant shares from the original chart, rounded to whole percentages that total 100. Equal ratings mean the same questionnaire response.

Comparison summary

Compared with	BP same or higher	BP higher ratings per lower rating	Modelled Elo win share for BP
Tavus	≈85%	≈2.3 to 1	58.3%
D-ID	≈68%	≈1.1 to 1	52.9%
HeyGen	≈67%	≈0.7 to 1	46.6%

BP denotes Beyond Presence. Rating ratios compare the higher and lower shares, excluding equal ratings. The Elo win share is modelled from the score difference, not an observed participant share. See Understanding the measures for interpretation.

9) Overall experience

Statement rated by participants

I enjoyed talking to this avatar and found the overall experience positive.

Positive overall comparisons with Tavus and D-ID

Beyond Presence received more higher than lower ratings against Tavus and D-ID for overall experience. Approximately 79% rated it the same as or higher than Tavus, and 76% did so against D-ID. Beyond Presence placed second among the four providers on the overall mean rating and Elo point estimates.

≈79%

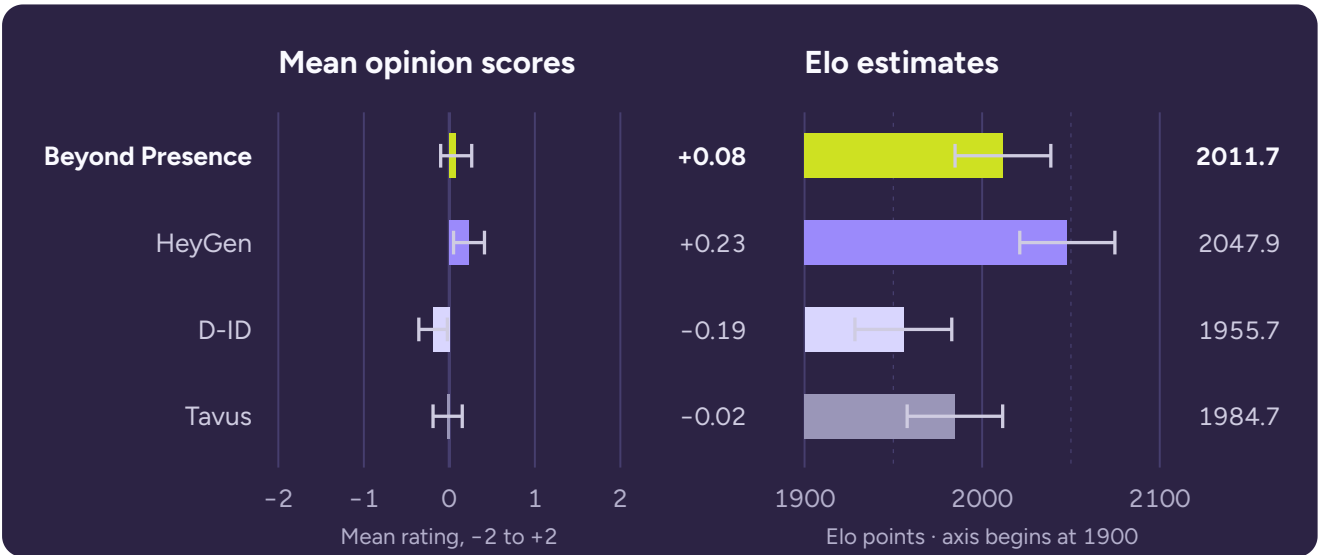
Same or higher than Tavus

≈76%

Same or higher than D-ID

2nd

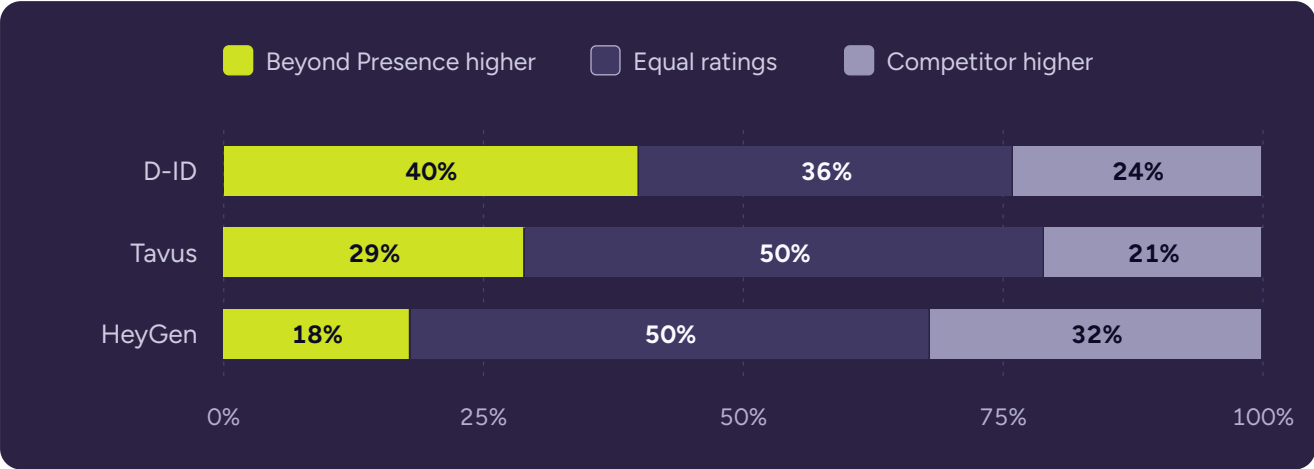
Of four on mean and Elo



Versions tested: Beyond Presence Genesis 2, HeyGen LiveAvatar, Tavus Phoenix 4, D-ID V3 Pro. MOS values are from the study platform. MOS whiskers are approximate readings of the original figure and show two standard errors. Elo estimates are posterior means calculated for the four providers compared; whiskers show 95% credible intervals. Elo bar lengths are not quality ratios.

Overall experience

Participant rating comparisons



Versions tested: Genesis 2, LiveAvatar, Phoenix 4, V3 Pro. Approximate participant shares from the original chart, rounded to whole percentages that total 100. Equal ratings mean the same questionnaire response.

Comparison summary

Compared with	BP same or higher	BP higher ratings per lower rating	Modelled Elo win share for BP
D-ID	≈76%	≈1.7 to 1	58.0%
Tavus	≈79%	≈1.4 to 1	53.9%
HeyGen	≈68%	≈0.6 to 1	44.8%

BP denotes Beyond Presence. Rating ratios compare the higher and lower shares, excluding equal ratings. The Elo win share is modelled from the score difference, not an observed participant share. See Understanding the measures for interpretation.

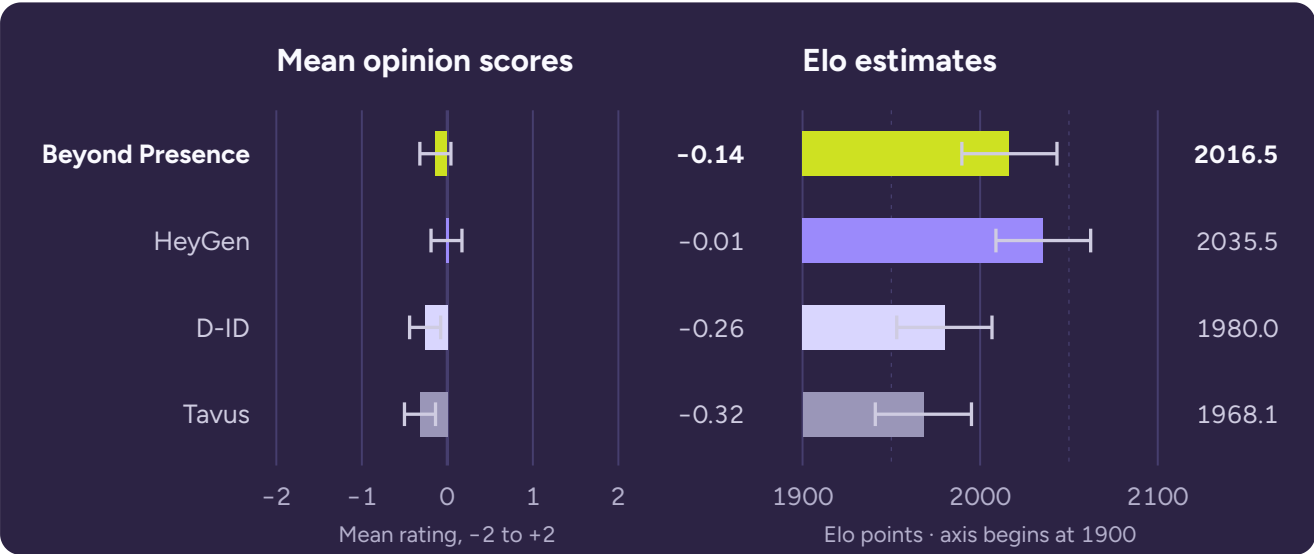
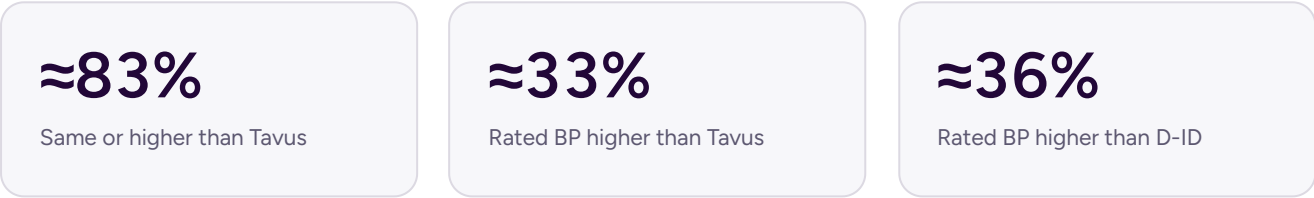
10) Naturalness

Statement rated by participants

The avatar's behavior seemed natural and lifelike.

Favourable rating comparisons with Tavus and D-ID

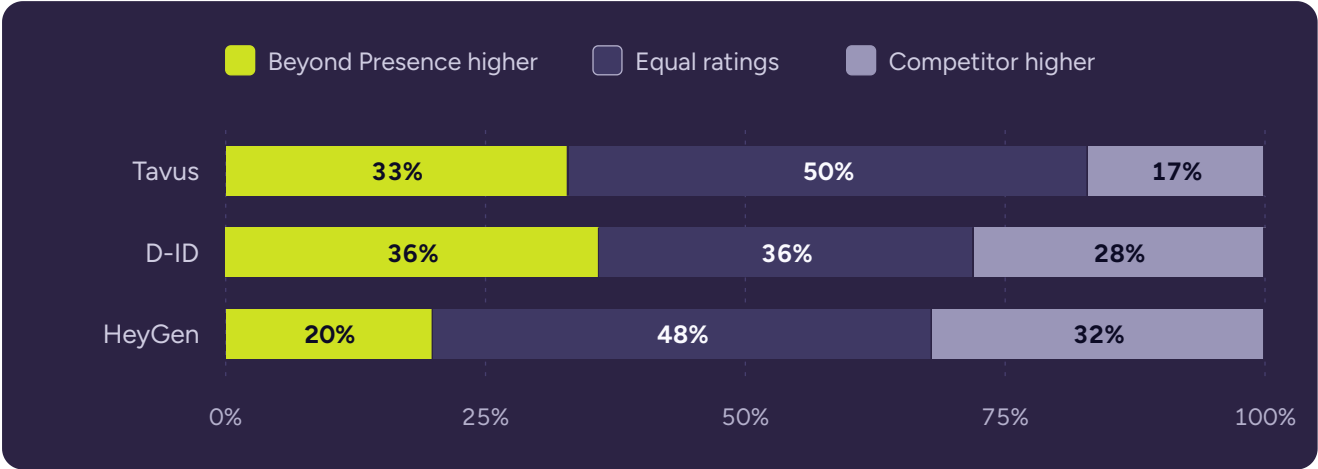
Approximately 83% rated Beyond Presence the same as or higher than Tavus for naturalness, including about 33% who rated it higher. Against D-ID, about 36% rated Beyond Presence higher and 28% rated D-ID higher. The complete mean ratings and uncertainty estimates are shown below.



Versions tested: Beyond Presence Genesis 2, HeyGen LiveAvatar, Tavus Phoenix 4, D-ID V3 Pro. MOS values are from the study platform. MOS whiskers are approximate readings of the original figure and show two standard errors. Elo estimates are posterior means calculated for the four providers compared; whiskers show 95% credible intervals. Elo bar lengths are not quality ratios.

Naturalness

Participant rating comparisons



Versions tested: Genesis 2, LiveAvatar, Phoenix 4, V3 Pro. Approximate participant shares from the original chart, rounded to whole percentages that total 100. Equal ratings mean the same questionnaire response.

Comparison summary

Compared with	BP same or higher	BP higher ratings per lower rating	Modelled Elo win share for BP
Tavus	≈83%	≈1.9 to 1	56.9%
D-ID	≈72%	≈1.3 to 1	55.2%
HeyGen	≈68%	≈0.6 to 1	47.3%

BP denotes Beyond Presence. Rating ratios compare the higher and lower shares, excluding equal ratings. The Elo win share is modelled from the score difference, not an observed participant share. See Understanding the measures for interpretation.

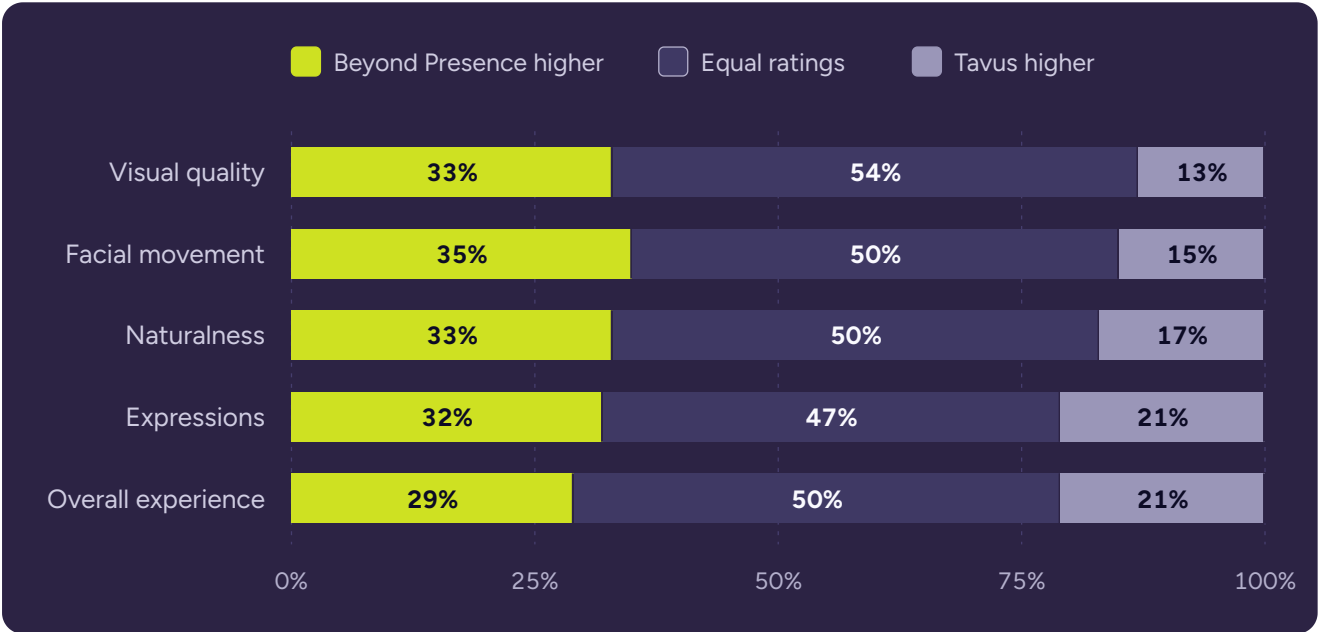
11) Comparison with Tavus

Favourable rating comparisons across all five categories

In every evaluated category, more participants rated Beyond Presence higher than Tavus than rated Tavus higher. Visual quality and facial movement showed the strongest ratios, at approximately 2.5 to 1 and 2.3 to 1 respectively.

Visual quality draws the clearest distinction

For visual quality, about 33% rated Beyond Presence higher, 54% gave equal ratings and 13% rated Tavus higher. Among those who gave different ratings, approximately 72% rated Beyond Presence higher, calculated from the rounded chart shares.



Shares are reconstructed from the original figures and rounded. Rows are ordered by the ratio of higher to lower ratings for Beyond Presence. The green segment shows higher ratings; the grey segments show equal and lower ratings.

Comparison with Tavus

The comparison in numbers

Criterion	BP same or higher	BP higher ratings per lower rating	Modelled Elo win share for BP
Visual quality	≈87%	≈2.5 to 1	58.1%
Facial movement	≈85%	≈2.3 to 1	58.3%
Naturalness	≈83%	≈1.9 to 1	56.9%
Expressions	≈79%	≈1.5 to 1	56.4%
Overall experience	≈79%	≈1.4 to 1	53.9%

Reading the ratio

The 2.5 to 1 ratio describes how many ratings favoured each provider. It does not measure how much better one product is. The Elo transformation gives a separate modelled comparison probability of 58.1%. Results apply to Beyond Presence Genesis 2 and Tavus Phoenix 4 as tested in August 2026.

12) Complete mean ratings and study conclusions

The full mean rating comparison provides context for the category findings. Beyond Presence's strongest mean result was visual quality. Its point estimates also exceeded Tavus across all five questions.

Criterion	Beyond Presence	HeyGen	D-ID	Tavus
Visual quality	+0.76	+0.63	+0.53	+0.46
Expressions	+0.11	+0.15	+0.20	-0.09
Facial movement	+0.03	+0.21	-0.02	-0.22
Overall experience	+0.08	+0.23	-0.19	-0.02
Naturalness	-0.14	-0.01	-0.26	-0.32

Mean opinion scores from the study platform, on the scale from -2 to +2. Error bars and participant distributions appear on the individual category pages.

Study conclusions

Beyond Presence led this comparison on visual quality mean and Elo point estimates. The direction of participant ratings favoured Beyond Presence over Tavus in every category. Its expression Elo estimate was within 4.1 points of HeyGen, supporting a description of closely matched expression estimates.

Results apply to the model versions and integrations tested in August 2026: Beyond Presence Genesis 2, HeyGen LiveAvatar, Tavus Phoenix 4 and D-ID V3 Pro.

Part 3

Appendices

The reference material behind the experiment: recording guidelines, avatar versions, integration references and the conversation prompt.

A	Appendix A · Recording guidelines	25
B	Appendix B · Avatar information	25
C	Appendix C · Integration references	26
D	Appendix D · Conversation prompt	26

Appendix A Recording guidelines

Source recordings followed each provider's recommended requirements at the time of the study.

Recording categories

Active listening was silent and showed engagement through expressions such as raised eyebrows or a smile. Speaking was clear, confident and concise, with fewer fillers and a fully closed mouth during pauses. Idling was silent, attentive and neutral, with occasional nodding.

Provider	Recording sequence
Beyond Presence	20 seconds idling, then 3 to 4 minutes speaking
HeyGen LiveAvatar	15 seconds listening, 90 seconds speaking, then 15 seconds idling
Tavus	30 seconds speaking, then 30 seconds listening
D-ID	A 2 to 3 second pause at the start and end; 3.5 minutes speaking and 30 seconds listening

Appendix B Avatar information

Model names, sources and output resolutions reflect the original report. All avatars appeared in a 640 by 480 pixel window during the experiment.

Provider	Model	Source	Resolution	Created in 2026
Beyond Presence	Genesis 2	4 minute video	1080p	August 4
HeyGen LiveAvatar	Not specified	2 minute video	720p	July 25
D-ID	V3 Pro	4 minute video	1080p	July 22
Tavus	Phoenix 4	1 minute video	1080p	July 22

These versions define the comparison. Findings should not be interpreted as a benchmark of subsequent model releases. Tavus has since released Phoenix 4.5, which was not part of this study.

Appendix C Integration references

LiveKit integration references recorded in the study. D-ID used a different integration path.

Beyond Presence	https://docs.livekit.io/agents/models/avatar/plugins/bey/
HeyGen LiveAvatar	https://docs.livekit.io/agents/models/avatar/plugins/liveavatar/
Tavus	https://docs.livekit.io/agents/models/avatar/plugins/tavus/

Appendix D Conversation prompt

Prompt content from the original study, with numeric ranges written out for readability.

Prompt

You are participating in a conversational benchmark. Your goal is to help the user decide how to spend a free afternoon. The user has approximately 3 to 4 hours available and is open to any of the following activities:

- Going for a hike or walk outdoors
- Visiting a museum or cultural attraction
- Watching movies or TV at home
- Trying a new café or restaurant
- Reading a book in a relaxing location

Your task is to learn enough about the user's preferences and current situation to recommend one of these activities.

Rules

- Begin by presenting the five activity options and asking which ones initially sound most appealing.
- Ask at most one question per turn.
- Build on the user's previous answers rather than asking generic survey questions.
- Use complete, natural sentences.
- Most responses should be 2 to 3 sentences long.
- Demonstrate memory of information the user has shared.
- Explain your reasoning when discussing options.
- Avoid listing pros and cons for every option after every turn.
- Do not ask more than five questions total.
- Provide a final recommendation no later than your 7th response.

Conversation structure

Response 1	Present options and ask the opening question.
-------------------	---

Responses 2 to 6	Explore preferences, energy level, mood and constraints.
-------------------------	--

Response 7	Give a clear recommendation and explain why it fits the user.
-------------------	---

Response 8	Wish the user a fun time on whichever activity you have recommended and say goodbye.
-------------------	--

Sources (Appendix B)

Avatar Comparison Study, September 1, 2026, methods and original result figures. Mabyduck experiment Afternoon Activity Benchmark, dashboard and raw rating export, used to calculate the Elo estimates. The Elo method reference is listed in Understanding the measures.