

Evaluating interactive avatars

LT

Lucas Theis
Founder

ZQ

Zihao Qi
Research intern

IG

Ivan Gerov
Software engineer

20 Questions

Think of a word related to professions.

Examples: doctor, teacher, firefighter, pilot, chef, artist, police officer, musician



Visual quality

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

Strongly disagreeDisagreeNeutralAgreeStrongly agree

Naturalness

The avatar's behavior seemed natural and lifelike.

Strongly disagreeDisagreeNeutralAgreeStrongly agree

Lip sync quality

About this report. This evaluation was designed and run by [Mabyduck](#) and commissioned by [Anam](#), one of the four providers evaluated. The study was performed in October 2025 using the avatars and personas available at the time, and was originally published on 19 January 2026 at mabyduck.com/blog/evaluating-interactive-avatars. For Anam's write-up of the results, see anam.ai/blog/cara-3-interactive-avatars.

Interactive avatars are virtual personas that interact with users in real time through voice, gestures, and visual animations. As interactive avatars become more widely deployed in high-stakes domains such as education, customer support, and healthcare, rigorous evaluations are becoming essential to understand where systems excel and where they fall short. We conducted an independent evaluation of real-time interactive avatars and compared several leading solutions on the market. The study was commissioned by [Anam](#), one of the providers included in the evaluation.

Challenges and study design

Evaluating interactive avatars is challenging due to limitations in study design, gaps in applicable research, and the significant technical complexity inherent in running a larger-scale study that integrates the services of multiple companies.

One of the limitations of interactive study design is that participants generally lack intrinsic motivation to engage in sustained interaction. To address this, we asked participants to play an abridged version of the game [20 Questions](#). We reduced the number of questions to five and limited the choice of words to a given topic to keep each interaction to 2 minutes or less. The topics were: animals, clothing, food, furniture, professions, sports, transportation, and water. For each topic, participants were provided with example words, though participants were free to make up their own words.

We considered alternative scenarios but settled on 20 Questions for the following reasons. First, the game is easy to explain and requires minimal instruction. Second, it presents a low barrier to interaction. Third, the cognitive load of the task is minimal, allowing participants to attend to the avatar rather than the task itself. Finally, the avatar speaks for the majority of the interaction, providing participants with ample opportunity to assess the avatar.

Participants played 8 rounds of the game. Following each round, they were asked to agree or disagree with the following statements on a 5-point Likert scale from "Strongly disagree" to "Strongly agree".

- **Visual quality:** "The appearance of the avatar was visually pleasing and free of obvious flaws."
- **Responsiveness:** "The avatar's response time felt natural and as expected in conversation."
- **Lip sync quality:** "The mouth and facial movements matched the spoken audio."
- **Naturalness:** "The avatar's behavior seemed natural and lifelike."
- **Interruptibility:** "I could interrupt the avatar as I would in a natural conversation."
- **Overall experience:** "I enjoyed talking to this avatar and found the overall experience positive."

Implementation



We implemented interactive avatars offered by four companies: [Anam](#), [HeyGen](#), [Tavus](#), and [D-ID](#), while trying to follow the documentation and recommendations provided by each company's website as closely as possible. To implement the study, we used Mabyduck's [embedded experiments](#), which support custom content inside an iframe that is controlled via JavaScript. Avatars were rendered in an iframe of 720×480 pixels resolution.

For each provider, we used 6 different avatars (also known as "personas"). Where providers offered avatars of varying quality, we tried to choose their best-performing avatars*. For example, for D-ID, we only used their "Premium+ agents". We chose 3 female and 3 male personas with an office-like background. Out of the remaining avatars available on each platform, we chose avatars at the top of each list.

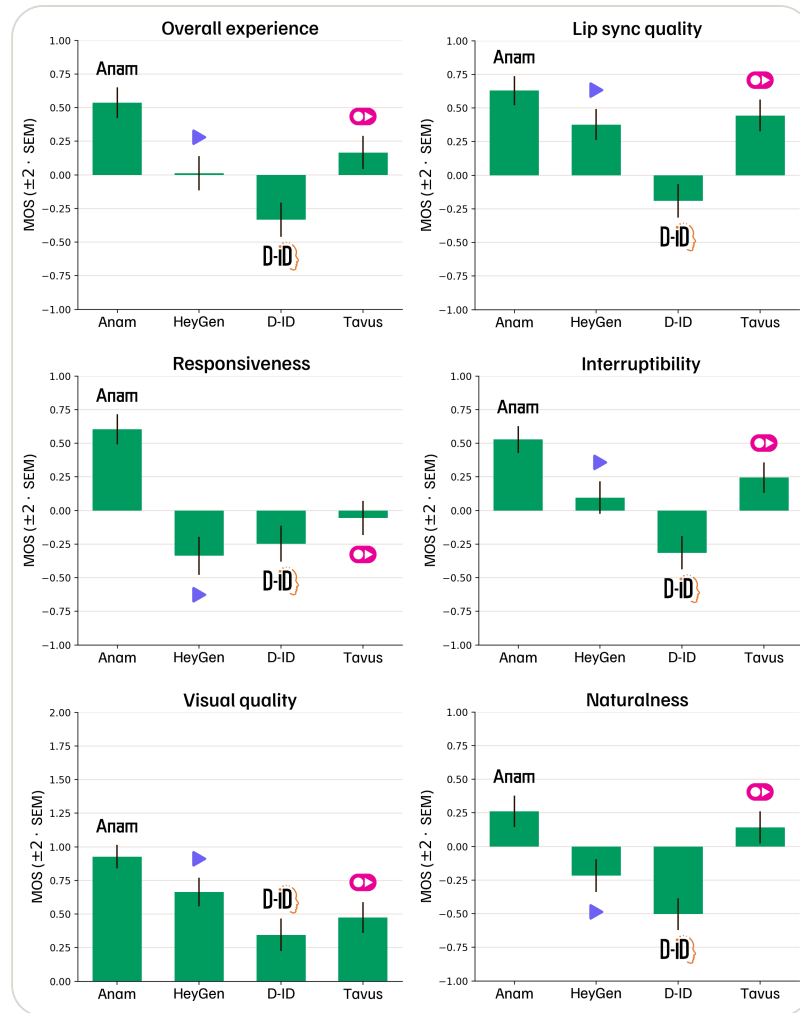
While Anam, HeyGen, and Tavus support voice interactions out of the box and as part of their APIs or SDKs, D-ID did not as of October 2025. To enable voice interactions for D-ID's avatars, we therefore relied on [Cartesia](#)'s Ink-Whisper streaming speech-to-text service.

All avatars received the same text prompt as input, which instructed them to play a game of 20 Questions on a given topic, but to try to guess the word after 5 questions.

Participants

Participants took part in the experiment online and came from two different pre-screened rater pools. One group of participants was pre-screened for hearing ability and audio equipment (86 participants), while another group was pre-screened for visual ability (e.g., color blindness) and monitor quality (e.g., contrast sensitivity; 92 participants). We found no meaningful difference between the two groups' perception of the avatars. Before each session, we further measured each participant's internet speed and screened out participants with speeds below 50 Mbps.

Results



On average, participants significantly preferred interactions with Anam's avatars over all other avatars ("Overall experience", $p < 0.001$). We find that the score for the overall experience was most strongly correlated with the responsiveness of the avatar (Spearman rank correlation of 0.697), suggesting that responsiveness is crucial.

We also investigated the quality of different personas. For Anam's avatars, the persona had no significant effect on the perceived responsiveness or interruptibility of the avatar, but the measured lip sync quality, naturalness, and visual quality of the personas were significantly different ($p < 0.001$).

Conclusion

Recent research by [Microsoft](#) found that the realism of avatar videos correlates strongly with perceived levels of trust and affinity in a non-interactive setting. In an interactive setting, we find that while the visual quality still had a high positive correlation with overall experience (Spearman rank correlation of 0.473), it was lower than that of all other aspects we considered.

As AI is becoming more interactive, evaluations are becoming more demanding. Mabyduck's goal is to simplify even the most complex studies. Our embedded experiments feature allowed us to focus on the implementation of the avatars themselves, and makes it easy to ensure that only participants of a study receive access to the API endpoints being evaluated.

* The study was performed in October 2025 and used avatars and personas available at the time.