Perceptual Judgments of Video Authenticity: An Examination of Viewing Duration, Confidence, Content, and Strategies

Catherine E. Davodi, Sarah Barrington, Hany Farid, Emily A. Cooper
University of California, Berkeley
Berkeley CA, USA
{cdavodi, sbarrington, hfarid, emilycooper}@berkeley.edu

Abstract

The rapid spread of AI-generated videos on social media has made judging video authenticity an everyday task. At the same time, computational techniques that can accurately and quickly distinguish natural from synthetic content at internet scale have proven elusive. We investigated people's ability to distinguish authentic videos and AI-generated videos produced by state-of-the-art text-to-video models while varying video duration (from a single frame to 8s). Detection of AI-generated content improved with duration (58% to 75%), but accuracy for real videos remained near constant (∼60%). Signal-detection and confidence analyses revealed a key asymmetry: additional evidence increased rejection of fakes without increasing trust in real videos. We also examined factors influencing judgments: detection was higher for AI-generated videos depicting humans, and reported strategies emphasized motion, body parts (e.g., hands), and image quality. These findings suggest that the improvements in the realism of AI-generated videos may reduce the basis for trusting any video, rather than only making fakes harder to detect.

1. Introduction

During the recent U.S. government shutdown, federal food assistance was disrupted for millions of Americans [12, 16, 25]. The ensuing events were covered by Fox News in a story titled "SNAP beneficiaries threaten to ransack stores over government shutdown" [10, 24]. Soon after, it was revealed that the associated video footage was AI-generated, having originated from a spate of racially-charged videos shared on social media [20].

This is just one example of how the proliferation of AI-generated videos shared on social media is causing confusion about what is happening in the world. Just in the past year (2025) for example, social media posts have also included a mixture of authentic and AI-generated videos of ICE agents clashing with local police officers [11] and of the damage in Jamaica following Hurricane Melissa [2]. This mixture of content raises concerns that AI-generated videos will be mistaken as real, and the inverse problem: that important authentic content can be dismissed as AI-generated [3, 22].

Current computational techniques for detecting and alerting users about AI-generated videos do not yet operate at the efficiency or accuracy to be practical at internet scale [6] nor are they readily available to the general public. As such, consumers often have to fall back on their own perceptual and subjective judgments of video authenticity. As AI technology continues to advance rapidly, our understanding of whether we can trust these judgments has become a moving target.

When it comes to still image and audio-only media, there is converging evidence that people can no longer reliably distinguish between what is real and what is AI generated. Research conducted in 2022 and 2023, for example, found that participants were on average near or below chance at detecting that an image of a face was AI generated [14, 18]. Contemporaneous studies on detection of AI-generated voices reported performance well above chance [15, 23], but a more recent study in 2025 suggests that this gap is closing [1].

While the realism of AI-generated or AI-manipulated videos is advancing rapidly, it has tended to lag behind image- and audio-only content [26]. This lag is likely due to the additional cues available in videos, such as audio-visual misalignment and motion artifacts. For example, recent studies examining videos of people talking suggest that the combination of audio and visual cues can boost people's detection of AI-generated and AI-modified content [5, 9]. For visual-only content, the presence of violations of the laws of physics or deviations from typical human motion can provide salient cues in AI fakes [8, 17]. However, a recent evaluation by Runway AI (a company producing tools for creating AI-generated video) reported that over 90% of viewers could not reliably distinguish their state-of-the-

Figure 1. Initial frames from example videos used in the study. The left-most frames are authentic; the rest of the frames are generated based on text descriptions of the authentic videos (from left to right: Veo 3.1, Sora 2, and Seedance 1). Participants judged whether each video was real or AI-generated. Authentic videos were all sourced from Pixabay. From top to bottom, the Pixabay creators are: AlexKopeykin, AlexKopeykin, Sayakulov, and McPix22.

art AI-generated videos from real footage [21]. Thus, at present, people's ability to detect that a video is AI generated likely hinges on the interaction between the realism of the multimodal video dynamics and the sensitivity of the human visual system to pick up on any physical violations.

There are many different techniques available for using AI to generate and manipulate videos. Modern text-to-video techniques, which are the focus of our study, create videos from a descriptive text prompt (e.g., "An outdoor seaside promenade or boardwalk in the late afternoon with a cloudy sky."). Because the text-to-video models that are currently popular online, like OpenAI's Sora [19] and Google's Veo [4], are being updated constantly, the realism of their most recent iterations has yet to be systematically studied. Here we focus on state-of-the-art text-to-video content and report the results of an online study addressing some basic questions: How well can people reliably distinguish authentic from AI-generated visual content? How does their performance depend on the length of the video and the type of motion? How confident are people in their judgments and is their confidence warranted?

2. Methods

2.1. Participants

We recruited a total of 400 participants through the online research platform Prolific. To ensure data quality and linguistic comprehension, eligibility was restricted to individuals located in the United States who spoke English as a first language and maintained a prior approval rating of at least 75% on the platform. Participants were also required to have normal or corrected-to-normal vision, which was self-reported via a screening questionnaire. The study was approved by the Institutional Review Board (IRB) at University of California, Berkeley. All participants provided informed consent prior to starting the study and were compensated for their time. After applying exclusion criteria (see Section 2.4.1), the final sample consisted of 376 participants.

2.2. Stimuli

2.2.1. Authentic Videos

Authentic videos were sourced from Pixabay, a stock video platform. To ensure diverse content, we used GPT-5.2 to generate 24 search phrases: 12 targeting videos featuring

people (e.g., "person walking through a busy market") and 12 targeting videos without people (e.g., "waterfall in a forest," "traffic crossing a bridge"). The non-people phrases were balanced between urban/indoor scenes and natural outdoor scenes. For each search phrase, we selected the first video meeting the following criteria: (1) natural motion without slow-motion, time-lapse, or other speed effects; (2) single continuous shot; (3) horizontal format; (4) minimum duration of 12 seconds; (5) no obvious unnatural artifacts; (6) not flagged as AI generated by Pixabay; and (7) full color. All videos were posted in 2024 or earlier. Videos were downloaded at the highest available resolution.

2.2.2. AI Video Synthesis

For each authentic video, we generated a corresponding text prompt using GPT-5.2. Each video was uploaded with the instruction: "Write a prompt for a text-to-video generator that will lead it to create a video with content similar to this one." The resulting prompts were reviewed and lightly edited for accuracy, brevity, and clarity. These 24 prompts were then used to generate synthetic videos from three state-of-the-art text-to-video models: Google Veo 3.1, ByteDance Seedance Pro 1, and Sora 2, which we refer to as Veo, Seedance, and Sora. These models were selected based on their technical advancement and recent release windows (May–November 2025). This ensured that our stimuli represented the current state-of-the-art diffusion models. Although some models (Sora and Veo) generate accompanying audio tracks, we focused exclusively on visual realism and removed all audio from the stimuli. This procedure yielded content-matched stimulus sets: for each of 24 content themes, we had one authentic video and three AI-generated versions depicting the same scene, enabling direct comparison of detection performance across sources while controlling for content.

2.2.3. Video Processing and Standardization

All stimuli were processed using ffmpeg. We generated five duration conditions for each unique video: a static image extracted from the first frame (0s), and video clips of 2s, 4s, 6s, and 8s. The maximum duration was capped by the generation limits of the Veo model. Source videos ranged in resolution from 960×540 to 1920×1080 pixels and in frames rate from 24 to 30 frames per second. To control for resolution differences, all videos were down-sampled to a consistent width of 800 pixels while preserving the original aspect ratio, and a frame rate of 24 frames per second. The complete stimulus pool comprised 480 items, structured as a fully crossed set: 24 content themes $\times$ 4 sources (1 authentic, 3 AI models) $\times$ 5 durations. Example first frames from four content themes are shown in Figure 1.

2.3. Procedure

To prevent memory effects while ensuring balanced exposure, the experiment employed a within-subjects design with randomized sampling. Each participant completed 48 experimental trials (plus four catch trials). For each content theme, a participant viewed two instantiations: one authentic and one AI generated. The authentic trial stimulus was selected at random from the five possible durations. The AI-generated trial was randomly selected from the pool of 15 options (3 models $\times$ 5 durations). The presentation order of the 48 trials, and the timing of the catch trials, was fully randomized. This design ensured that each authentic/AI-generated video was only seen at one duration, and that the balance of authentic and AI-generated content was 1:1. As a result of this design, more responses were collected for the authentic stimuli at each duration (which had a 1/5 chance of being presented) than for the AI-generated stimuli from each model (which had a 1/15 chance of being presented). Given the large number of research participants, this still resulted in sufficient data to analyze differences between models. Participants were not told what the ratio of authentic to AI-generated content would be.

The experiment was hosted on the Qualtrics platform and all trials were presented through the participant's web browser on their personal device. There was no restriction placed on device type: it could be a desktop computer, laptop computer, tablet, or mobile phone. On each trial, a single stimulus was presented in the browser window. For the non-static conditions (2s, 4s, 6s, and 8s), participants were forced to play the video at least once all the way through before submitting their responses, and they were allowed to replay the video as many times as they wanted. Following the stimulus presentation, participants responded to two items: "Is this video real or AI-generated?" and a 4-point Likert scale asking "How confident are you in your choice?" Response options were: 0 (Pure guess), 1 (Mostly guessing), 2 (Somewhat confident), or 3 (Fully confident).

To monitor participant attention, we interspersed four catch trials throughout the session. Catch-trial stimuli were constructed manually: each was 4s long, in which the first half displayed a video, and the remaining half consisted of a static image with explicit response instructions (e.g., "Select Real and 3 - Fully confident"). To preserve the overall balance of authentic and AI-generated stimuli in the survey, two catch trials used authentic clips and two used AI-generated clips.

At the end of the survey, participants were asked the following questions: What strategies did you use to decide whether something was real or AI-generated? What parts of the content did you focus on? When you were highly confident in an AI-generated piece of content, what information were you using?

2.4. Data Analysis

2.4.1. Data Preprocessing

Based on a pilot study, we established two a priori exclusion criteria: participants were removed if they failed $\geq$1 catch trial or took less than 10 minutes to complete the survey. Of 400 initial responses, we excluded 2 participants who completed in under 10 minutes, 19 participants who failed at least one catch trial, and 3 incomplete responses, yielding a final sample of N=376.

2.4.2. Statistical Analysis

Responses were analyzed using a mixed-effects logistic regression model fit using MATLAB's `fitglme` (maximum pseudo-likelihood method). The dependent variable was whether participants responded "real" on each trial (1 = real, 0 = AI-generated). Fixed effects included stimulus type (real, Veo, Sora, Seedance), video duration (0s, 2s, 4s, 6s, or 8s), and their interaction. Stimulus type and duration were coded such that model coefficients could be interpreted relative to the real stimulus type at 0s (baseline). Random intercepts and random slopes for duration were included per stimulus item (N=96, defined as unique combinations of model and video content) and random intercepts only were included for participants (N=376).

Additionally, we quantified response bias using signal detection theory metrics [13], collapsing across all participants and stimuli. For each duration, we computed d' (sensitivity) from hit rates (proportion of real videos judged "real") and false-alarm rates (proportion of AI-generated videos judged "real"). We also computed the response criterion, where positive values indicate a conservative bias (tendency to respond "AI generated") and negative values indicate a liberal bias (tendency to respond "real").

To examine the relationship between confidence and correctness, we computed a confidence calibration score. For each unique combination of stimulus type and duration, we fit the trial by trial response accuracy as a function of the confidence rating with a linear regression. Higher positive regression line slopes indicate a stronger confidence calibration: greater confidence was associated with better performance. To assess statistical significance of this relationship, we also fit a second logistic regression model to the raw responses. This model was fit to the response correctness (1=correct, 0=incorrect), and an additional fixed effect was added for self-reported confidence (z scored). All two- and three-way interactions among the fixed effects (confidence, stimulus type, duration) were included. Random effects were the same as the first model.

For both logistic regression models, effect sizes are reported in terms of odds ratios and statistical significance was determined from t-statistics using a threshold of $p < 0.05$.

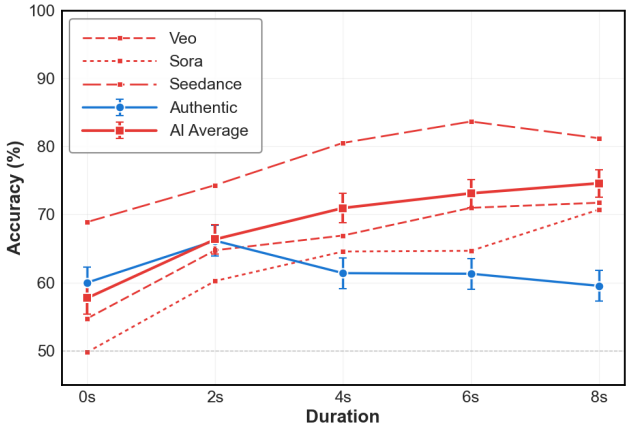

Figure 2. Classification accuracy in terms of percent correct (blue for authentic, red for AI generated). Each data point was computed by averaging across all participants and stimuli. Error bars indicate 95% confidence intervals, and are omitted from the individual AI model lines for clarity. The horizontal dashed line indicates chance performance of 50%.

2.4.3. Exploratory Analyses

We conducted additional exploratory analyses to characterize content-type and participant-level trends. We examined whether performance differed between videos featuring human motion versus non-human motion, and whether device type (mobile vs. other device types) or completion time affected accuracy. Finally, we performed a word frequency analysis on participants' open-ended responses to identify commonly reported strategies and cues.

3. Results

3.1. Duration Improved Detection of AI-Generated Videos but Not Authentic Videos

When all people saw was the first video frame, they performed modestly but consistently above chance at detecting authentic and AI-generated content. Across all participants and stimuli, the hit rate for authentic content at 0s was 60.0% (95% CI: [57.7%, 62.2%]) and for AI-generated content it was 57.7% (95% CI: [55.4%, 60.0%]).

The additional motion information provided by watching *authentic* videos did not produce sustained improvement: after a brief rise at 2 seconds, accuracy for authentic content returned to baseline levels (Figure 2, blue line). This conclusion is supported by the regression model: the odds ratio associated with duration for "real" responses was near one and not statistically significant (Table 1).

Video duration did, however, notably help people detect fakes. For all three AI models, detection accuracy systematically increased with duration (Figure 2, red lines). Averaging across all models, this increase was almost 17 per-

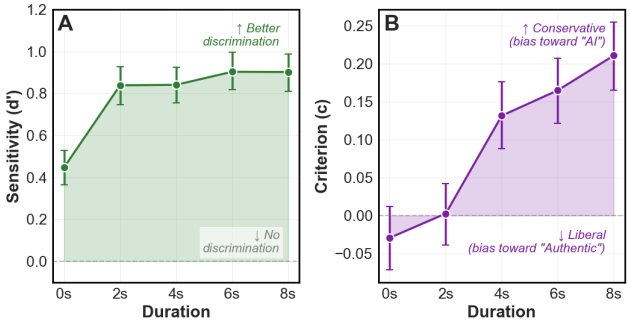

Figure 3. (A) Sensitivity (d') and (B) Criterion for authentic versus AI-generated classification. Dashed line in B indicates neutral criterion. Each data point was computed by averaging across all participants and stimuli. Error bars indicate bootstrapped 95% confidence intervals.

centage points (from 57.7% at 0s to 74.6% at 8s). In the regression analysis, this increase is indicated by significant interactions with duration for each AI model, showing decreases in "real" responses with increasing viewing duration (Table 1).

A follow up signal detection analysis suggests that this pattern reflects both increased sensitivity (d') (Figure 3A), and a criterion shift towards skepticism of authenticity (Figure 3B). These changes together support the selective improvement in detection of AI-generated videos with duration: while classification sensitivity improves with duration, the accompanying criterion shift toward skepticism helps detect AI-generated content but does not improve trust in authentic content.

Although the effect of duration was similar across AI models, overall accuracy for AI-generated video differed markedly across models. Across all durations, people were best at detecting that Seedance videos were AI-generated (77.8%), followed by Veo (65.7%) and Sora (62.0%).

Taken together, we infer that unrealistic dynamics in videos help identify fakes, but realistic dynamics do not necessarily improve the ability to confirm authenticity.

3.2. Duration Improved Metacognition for AI-Generated Videos, but Not Authentic Ones

In addition to performing the real/fake classification, participants reported their confidence on each trial. The median confidence rating was 2 (Somewhat confident) for all four stimulus types (real, Veo, Sora, and Seedance), indicating similar confidence levels for authentic and AI-generated videos.

We next examined whether participants' confidence tracked their actual accuracy, a measure of metacognitive awareness [7]. For authentic videos, the relationship between confidence and accuracy was relatively flat across durations, indicating stable (if modest) metacognition (Figure 4, blue line). However, for all AI-generated videos, confidence calibration improved consistently with duration. In the logistic regression analysis, this pattern is reflected in a set of significant three-way interactions indicating that, across the AI models, duration is associated with improved accuracy more strongly when confidence levels are high (Table 2).

Thus, our results suggest that unrealistic dynamics in videos also selectively improve metacognitive awareness when identifying fakes.

Table 1. Logistic Regression Model for P(Real) Responses.

Predictor	β	OR	95% CI	t	p
(Intercept)	0.590	1.80	[1.40, 2.33]	4.52	<0.001
Duration	−0.014	0.99	[0.95, 1.03]	−0.62	0.534
Mdl:Veo	−0.914	0.40	[0.28, 0.58]	−4.88	<0.001
Mdl:Sora	−0.695	0.50	[0.35, 0.72]	−3.73	<0.001
Mdl:Seedance	−1.542	0.21	[0.15, 0.31]	−8.12	<0.001
Mdl:Veo × Duration	−0.102	0.90	[0.85, 0.96]	−3.05	0.002
Mdl:Sora × Duration	−0.109	0.90	[0.84, 0.96]	−3.30	<0.001
Mdl:Seedance × Duration	−0.112	0.89	[0.84, 0.96]	−3.25	0.001

Note. Model includes random intercepts for participant (std. dev.=0.662) and stimulus (std. dev.=0.583), and random slopes of duration by stimulus (std. dev.=0.098, r=-0.19). Reference category=Real videos. Mdl = model; OR=odds ratio.

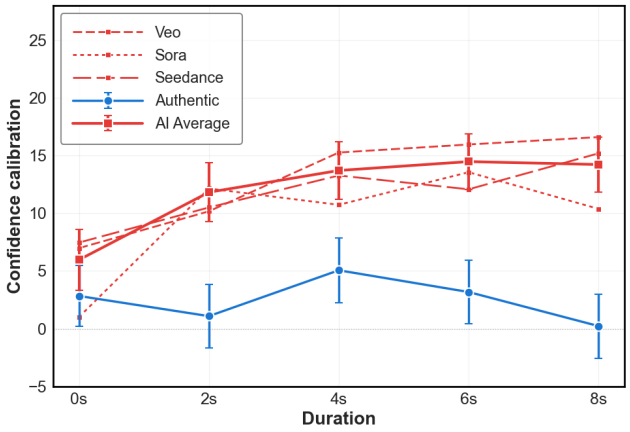

Figure 4. Confidence calibration across viewing durations. Slopes from linear regressions of accuracy with confidence for each model at each duration. Positive slopes indicate that higher confidence is associated with higher accuracy. Error bars indicate 95% confidence intervals, and are omitted from the individual AI model lines for clarity. The horizontal dashed line indicates zero calibration.

3.3. Stimulus and Participant Variability

Individual videos varied notably in the associated response accuracy, with performance ranging from from 19% to 95% across all individual videos (Figure 5). Collapsing across the 24 unique content themes, variation in accuracy was

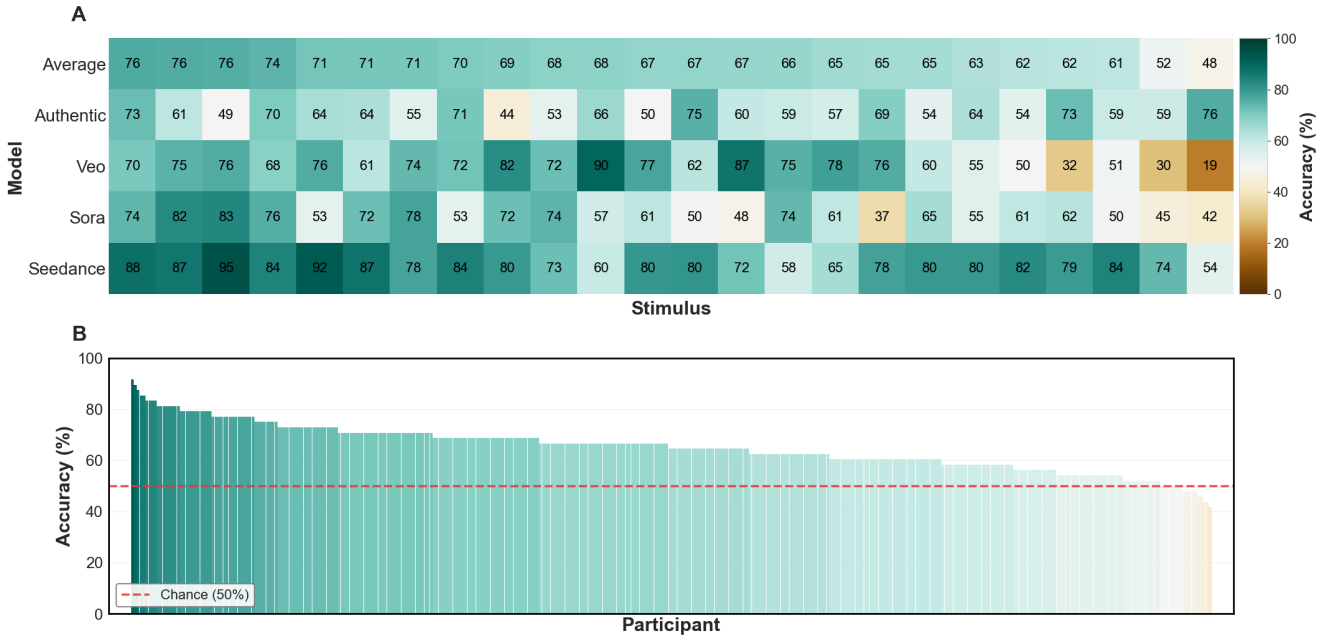

Figure 5. (A) Accuracy for each of the 24 content themes across the four stimulus types and on average. (B) Accuracy for each study participant. Color indicates accuracy (teal=high, white = chance, brown=low).

Table 2. Logistic Mixed-Effects Model for P(Correct) with Confidence Included.

Predictor	β	OR	95% CI	t	p
(Intercept)	0.572	1.77	[1.41, 2.23]	4.90	<0.001
Duration	−0.014	0.99	[0.95, 1.02]	−0.73	0.464
Mdl:Veo	−0.237	0.79	[0.56, 1.11]	−1.37	0.172
Mdl:Sora	−0.434	0.65	[0.46, 0.91]	−2.51	0.012
Mdl:Seedance	0.336	1.40	[0.99, 1.97]	1.91	0.057
Confidence (z)	0.148	1.16	[1.07, 1.25]	3.76	<0.001
Two-Way Interactions					
Mdl:Veo × Duration	0.105	1.11	[1.05, 1.18]	3.53	<0.001
Mdl:Sora × Duration	0.115	1.12	[1.06, 1.19]	3.96	<0.001
Mdl:Seedance × Duration	0.111	1.12	[1.05, 1.19]	3.62	<0.001
Mdl:Veo × Conf	0.097	1.10	[0.94, 1.29]	1.22	0.221
Mdl:Sora × Conf	0.028	1.03	[0.89, 1.19]	0.37	0.713
Mdl:Seedance × Conf	0.164	1.18	[0.99, 1.40]	1.88	0.061
Dur × Conf	−0.012	0.99	[0.97, 1.00]	−1.45	0.147
Three-Way Interactions					
Mdl:Veo × Dur × Conf	0.057	1.06	[1.02, 1.10]	3.25	0.001
Mdl:Sora × Dur × Conf	0.047	1.05	[1.02, 1.08]	2.90	0.004
Mdl:Seedance × Dur × Conf	0.067	1.07	[1.03, 1.11]	3.51	<0.001

Note. Model includes random intercepts by participant (SD=0.281) and stimulus (SD=0.532), and random slopes for duration by stimulus (SD=0.083; r=-0.25). Reference category=Real videos. Mdl = model; OR=odds ratio.

lower but still notable (from 48% to 76%). We considered whether some of this content-based variability might derive from differences between videos featuring human motion versus non-human motion. Indeed, across all three AI models and all video durations, detection accuracy was consistently higher for videos depicting human motion than non-human motion (Figure 6, red lines). Interestingly, this human advantage was similar across video durations, and it was not present for the authentic videos (Figure 6, blue lines). Overall, this is consistent with the hypothesis that visual artifacts in the appearance of people in AI-generated imagery can provide key evidence of fakeness. However, these results suggest the key artifact may be in the basic appearance of people rather than their patterns of motion.

Participants also differed substantially in their accuracy, from 42% to 92% (Figure 5B). To assess whether participants who performed poorly may have simply rushed through the task, we examined the relationship between completion time and accuracy. We found no significant correlation between completion time and accuracy among the remaining participants ($r = -0.01$, $p = 0.781$).

Another potential source of individual differences could be device usage. We examined whether performance differed for participants using mobile phones (N=38) versus other devices (N=338). Mobile users showed slightly lower accuracy on average (62%) compared to non-mobile users (65%), but this variation is small in comparison to the full range of performance. Since social media postings may be likely to be viewed on mobile devices, this suggests that the associated small screens do not meaningfully obscure authenticity cues that people are currently relying on.

In responding to the open-ended strategy questions, participants tended to use keywords that emphasized motion, body parts/people, and image quality (Figure 7). However, we did not observe any notable differences in the pattern of

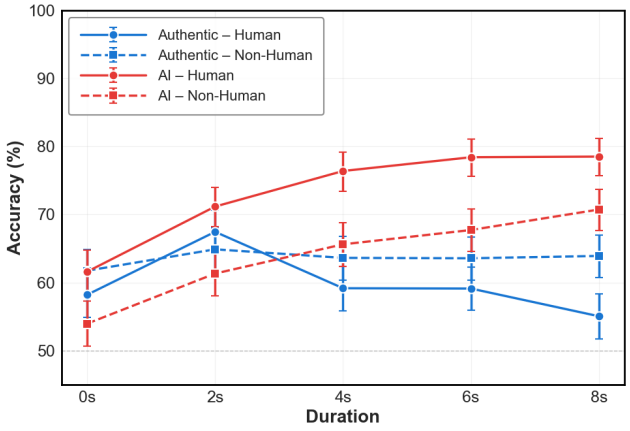

Figure 6. Accuracy for content as a function of duration, separated by motion type. Solid lines indicate videos featuring human motion; dashed lines indicate non-human motion. Error bars indicate 95% confidence intervals.

keywords used by participants with different performance levels. Understanding the contributors to individual differences in performance on this task would be an interesting direction for future work, since it may inform the design of training or educational materials.

4. Discussion

Our findings reveal an asymmetry in how visual information supports authenticity judgments: video dynamics help viewers more confidently and accurately detect AI-generated content but do not help them confirm that something is real. This pattern suggests that participants were not verifying authenticity per se, but instead searching for violations of expected physical structure and motion. As temporal information accumulated, videos could be identified as AI generated when their dynamics conflicted with learned expectations about how real scenes evolve (e.g., an extra arm appears on a person). In contrast, authentic videos did not appear to contain perceptual features that uniquely certify realness: plausible content (e.g., a person with just two arms) was not treated as proof of authenticity.

This interpretation carries an important implication: if observers rely primarily on detecting violations rather than verifying authenticity, then eliminating artifacts in AI-generated videos may not cause fake videos to be accepted as real so much as it may remove the basis for trusting any video at all. Our confidence calibration results support this interpretation: as violations in AI-generated videos accumulated over time, observers not only became more accurate but also more aware of why they were correct. Confidence in authentic videos, however, did not benefit from additional viewing because they provide little confirmatory evidence, only the continued absence of detectable errors.

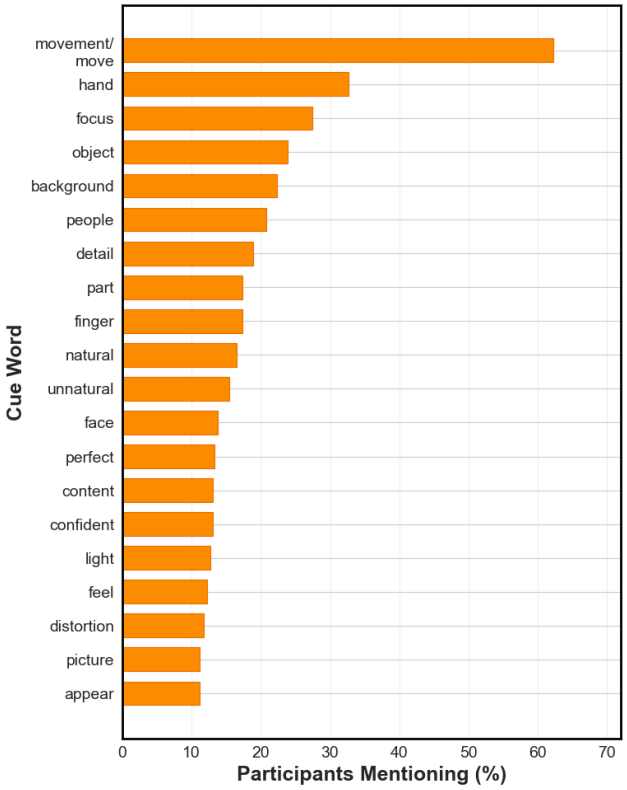

Figure 7. Values show percentage of participants who used each word(s) in response to the three open-ended questions at the end of the study.

The substantial variation across AI models, with Sora producing the most convincing content and Veo the least in this study, suggests that the detectability of fakes needs to be re-evaluated as models improve. Importantly, the effect of duration on performance was common across all models, suggesting that even the best current models contain visual artifacts that people can use to accumulate evidence against authenticity over time. As text-to-video models become both more realistic and capable of producing longer videos, it remains to be seen whether increasing fidelity will eliminate the artifacts that currently allow viewers to gather evidence against authenticity as viewing time increases.

Several limitations in the current study warrant highlighting. First, participants viewed videos in a controlled experimental setting rather than in social media contexts where attention and motivation differ. Second, our sample was restricted to US-based English speakers; cross-cultural differences in media literacy may moderate these effects. Lastly, we used a relatively small sample of unique videos, which may limit generalizability to video content themes that are not captured in our dataset.

Future work should test whether the duration-driven benefits observed here persist under realistic viewing condi-

tions (scrolling feeds, audio present, social context) and whether targeted instruction can amplify them. More broadly, our results suggest that temporal coherence may be a key remaining failure mode for text-to-video models, and that human performance under prolonged viewing provides a practical behavioral benchmark for tracking progress.

References

[1] Sarah Barrington, Emily A. Cooper, and Hany Farid. People are poorly equipped to detect AI-powered voice clones. *Scientific Reports*, 15:11004, 2025. 1

[2] Kevin Chan and Taimoor Sobhan. Phony AI videos of hurricane Melissa flood social media. PBS NewsHour, 2025. https://www.pbs.org/newshour/world/phony-ai-videos-of-hurricane-melissa-flood-social-media. 1

[3] Robert Chesney and Danielle Keats Citron. Deep fakes: A looming challenge for privacy, democracy, and national security. *California Law Review*, 107:1753–1819, 2019. 1

[4] Eli Collins and Douglas Eck. New generative media models and tools, built with and for creators. Google DeepMind, 2024. https://blog.google/technology/ai/google-generative-ai-veo-imagen-3. 2

[5] Di Cooke, Abigail Edwards, Sophia Barkoff, and Kathryn Kelly. As good as a coin toss: Human detection of AI-generated content. *Communications of the ACM*, 68(10): 100–109, 2025. 1

[6] Hany Farid. Mitigating the harms of manipulated media: Confronting deepfakes and digital deception. *PNAS Nexus*, 4(7):pgaf194, 2025. 1

[7] Stephen M. Fleming and Hakwan C. Lau. How to measure metacognition. *Frontiers in Human Neuroscience*, 8: 443, 2014. 5

[8] Xingyu Fu, Siyi Liu, Yinuo Xu, Pan Lu, Guangqiuse Hu, Tianbo Yang, Taran Anantasagar, Christopher Shen, Yikai Mao, Yuanzhe Liu, Keyush Shah, Chung Un Lee, Yejin Choi, James Zou, Dan Roth, and Chris Callison-Burch. Learning human-perceived fakeness in AI-generated videos via multimodal LLMs. arXiv:2509.22646, 2025. 1

[9] Matthew Groh, Aruna Sankaranarayanan, Nikhil Singh, Dong Young Kim, Andrew Lippman, and Rosalind Picard. Human detection of political speech deepfakes across transcripts, audio, and video. *Nature Communications*, 15(1): 7629, 2024. 1

[10] Peter Kafka. Now AI fakes are fooling news outlets — and maybe AI pros? Business Insider, 2025. https://www.businessinsider.com/fox-ai-fake-snap-meta-yann-lecun-newsmax-2025-11. 1

[11] Jason Koebler. AI-generated Sora videos of ICE raids are wildly viral on Facebook. 404 Media, 2025. https://www.404media.co/ai-generated-videos-of-ice-raids-are-wildly-viral-on-facebook. 1

[12] Lorie Konish. Where SNAP benefits stand amid negotiations to end the government shutdown. CNBC, 2025. https://www.cnbc.com/2025/11/12/snap-benefits-government-shutdown-negotiations.html. 1

[13] Neil A. Macmillan and C. Douglas Creelman. *Detection Theory: A User's Guide*. Lawrence Erlbaum Associates, Mahwah, NJ, 2nd edition, 2005. 4

[14] Elizabeth J. Miller, Ben A. Steward, Zak Witkower, Clare A. Sutherland, Eva G. Krumhuber, and Amy Dawel. AI hyperrealism: Why AI faces are perceived as more real than human ones. *Psychological Science*, 34(12):1390–1403, 2023. 1

[15] Nicolas M. Müller, Karla Pizzi, and Jennifer Williams. Human perception of audio deepfakes. In *Proceedings of the 1st International Workshop on Deepfake Detection for Audio Multimedia*, page 85–91, 2022. 1

[16] Geoff Mulvihill and David Lieb. States scramble to send full SNAP food benefits to millions of people after government shutdown ends. AP News, 2025. https://apnews.com/article/government-shutdown-snap-food-states-6cef598c92000bdff8384a9da1bfd23c. 1

[17] Peter Neri, M. Concetta Morrone, and David C. Burr. Seeing biological motion. *Nature*, 395:894–896, 1998. 1

[18] Sophie J. Nightingale and Hany Farid. AI-synthesized faces are indistinguishable from real faces and more trustworthy. *Proceedings of the National Academy of Sciences*, 119(8): e2120481119, 2022. 1

[19] OpenAI. Sora is here. OpenAI, 2024. https://openai.com/index/sora-is-here. 2

[20] Christopher Rhodes. Fox News called out for reporting on AI video of black woman with '7 different baby daddies' complaining of SNAP cuts. Yahoo News, 2025. https://www.yahoo.com/news/articles/fox-news-called-reporting-ai-210804379.html. 1

[21] Runway. The Turing reel, 2025. https://runwayml.com/research/theturingreel. 2

[22] Cristian Vaccari and Andrew Chadwick. Deepfakes and disinformation: Exploring the impact of synthetic political video on deception, uncertainty, and trust in news. *Social Media + Society*, 6(1), 2020. 1

[23] Kevin Warren, Tyler Tucker, Anna Crowder, Daniel Olszewski, Allison Lu, Caroline Fedele, Magdalena Pasternak, Seth Layton, Kevin Butler, Carrie Gates, and Patrick Traynor. Better be computer or I'm dumb: A large-scale evaluation of humans as audio deepfake detectors. In *ACM SIGSAC Conference on Computer and Communications Security*, pages 2696–2710, 2024. 1

[24] Yahoo News. 'gobsmacking': Media industry observers call out Fox News's website for getting duped by AI-generated video, then attempting to sweep it under the rug. Yahoo News, 2025. https://www.yahoo.com/news/articles/gobsmacking-media-industry-observers-call-155829667.html. 1

[25] Grace Yarrow and Marcia Brown. Nearly 42 million Americans lose their food stamp benefits. Politico, 2025. https://www.politico.com/news/2025/11/01/millions-lose-food-aid-snap-trump-shutdown-00632404. 1

[26] YouTube. Will smith eating spaghetti AI video - (2023 vs 2024). Just A Happy Troll, 2024. https://www.youtube.com/watch?v=vbWe5k4fFWE. 1