https://doi.org/10.2352/EI.2026.38.10.HVEI-232

Real or synthetic? An evaluation of AI-generated audio-visual content

Devi Klein; Dolby Laboratories Inc.; Sunnyvale, CA, USA
Anustup Choudhury; Dolby Laboratories Inc.; Sunnyvale, CA, USA
Evan Gitterman; Dolby Laboratories Inc.; Sunnyvale, CA, USA
Jaclyn Pytlarz; Dolby Laboratories Inc.; Sunnyvale, CA, USA
Scott Daly; Dolby Laboratories Inc.; Sunnyvale, CA, USA

Abstract

Generative AI (GenAI) models enable scalable multimedia content creation but can introduce artifacts that lack perceived realism. We conducted a perceptual study to assess how audio-visual cues impact people's ability to discriminate real user-generated content (UGC) from synthetic AI-generated content. Observers (N=36) participated in a two-interval forced-choice task across conditions that manipulated audiovisual consistency. They reliably identified synthetic content, achieving the highest accuracy when visual cues were available and the lowest when having to solely rely on audio content/quality issues. Our eye-tracking analysis indicated that biological motion inconsistencies were salient, while lower-level, texture-related distortions received less attention. Our proposed taxonomy of audio content and quality issues did not significantly predict task performance. However, these findings highlight the dominant role of visual artifacts in the decision-making process and the relative robustness of GenAI audio. Our work provides guidance for improving the perceptual quality of future, edge-deployed GenAI models.

Introduction

State-of-the-art Generative AI (GenAI) algorithms produce high-quality synthetic video with audio. Commercial models like Sora 2, Veo 3.1, and Kling 3.0 can take text prompts, single/multiple images, or a combination of the two modalities to produce imagery that is hard to distinguish from User-Generated Content (UGC) or studio film. These models enable scalable automated multimedia content creation with significant cost reductions in business marketing and advertising [1], and a lower barrier to entry for content creators without traditional production equipment.

GenAI models typically require cloud infrastructure due to their high computational cost and inefficiency [2]. For example, diffusion models iteratively denoise latent representations over many steps and generate video frames via joint spatiotemporal self-attention [3]. Recent industry trends point to the deployment of GenAI models on edge devices (e.g., smartphones, laptops) to reduce latency and network load, as well as improve inference speed and privacy [4], [5], [6]. This shift requires smaller models with fewer parameters to meet hardware constraints, often at the cost of robustness and generality for synthesizing varied audiovisual content. Consequently, these more efficient models will be tuned for specific use cases rather than being an all-purpose content creation engine. Moreover, such models may produce visual artifacts and audio-video inconsistencies, similar to first-generation cloud-based models [7].

The objective of this work is two-fold. First, we conduct a perceptual study with eye tracking to delineate how different

Table 1: Taxonomy of visual artifacts and audio content/quality issues observed in GenAI videos.

Visual Artifacts	Audio content/quality issues
Merging objects and/or object emergence	Logic issues - Categorically incorrect audio
Perspective - multiple vanishing points or "tunnel vision"	Context cues – missing important context cues to the nature of the sound sources
Mannequin effect - uncanny valley	A/V synchronization – poor synchronization with motion speed of visual elements or overly responsive
Violations of biological motion - shape shifting, extra or missing limbs	Stasis – non-varying audio across duration of video
Violations of physics laws	Nonsensical music/dialogue
Depth of field inconsistencies – lack of Bokeh, melting/distorted backgrounds, inconsistent "sharp" zones	Hum/whine – undesirable hum/whine/resonance/ringing
Small complex objects – incorrect rending of small or detailed objects	Fragmented sound – garbled/glitchy/fragmented sound
Eyes – slow motion eye blinking, inaccurate saccades and gaze orientation	Clipping – clipping associated with excessive loudness
Textures – intermingling of real-world textures with data compression artifacts	Band-limiting – band-limited or extremely muffled/tinny sound
Arrow of time – lack of unidirectional progression of causality	Background noise - Excessive or harsh background noise
Object permanence – lack of	
Shadows – lack of understanding of 3D scene geometry and light source	
Articulated motion – e.g., unnatural bending or movement of limbs/fingers and machinery	
Reflections – inaccurate emulation of real-world objects reflecting off water or mirrors	
Non-sensical text	

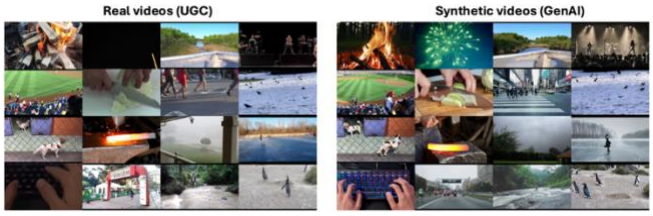


Figure 1. *Example video frames taken at the 1-second mark from UGC videos (left) and AI-generated videos (right). Each cell index of the left and right plots denotes a unique UGC-synthetic video pair.*

combinations of synthetic imagery and audio influence an observer's ability to distinguish UGC from GenAI videos. Specifically, observers viewed a sequence of two 5-second video clips, one containing real-world UGC and the other synthetic. They were tasked to select the more realistic video, enabling interval-scale measurement of perceived GenAI quality. Second, we propose a taxonomy of video artifacts and audio content/quality issues we expect to be present in future GenAI Edge models (see Table 1). Together, these objectives aim to improve the perceived realism and quality of synthetic content by helping inform future industry research on which visual artifacts and audio issues are most conspicuous and should be minimized during the development of new GenAI Edge models.

Background and Related Work

There are many ways of defining image/video or audio quality. Historically, objective full-reference metrics, like PSNR for images, compared a degraded signal (e.g., due to lossy compression) to a pristine reference, providing a measure of distortion, but poorly correlated with the human visual system (HVS) quantified via Mean Opinion Scores (MOS) [8]. Further improvements have been made with the introduction of the Structural Similarity Index (SSIM) to model local changes in luminance, contrast, and structural information [9], the Visible Difference Predictor (VDP) that accounts for the contrast sensitivity function of the HVS as well as viewing conditions [10], and extensions of VDP [11].

For GenAI videos, full-reference metrics are not applicable because the video outputs lack a pristine reference. No-reference metrics have been introduced [12], often based on deep-learning approaches that account for semantic alignment between a text prompt input and video output [13], or capturing visual fidelity and temporal coherence across synthetic video frames [14]. However, pure deep learning approaches operate solely on pixels, neglecting important viewing considerations like display resolution and dynamic range, as well as ambient lighting and viewing distance. Moreover, these approaches do not explicitly evaluate high-level aspects such as geometric, structural, and biological consistency, which the HVS is particularly sensitive to.

Recent work by Ghildyal et al. [15] has pushed to bridge this gap by aggregating a large set of human-annotated visual artifacts to enable benchmarking of perceptual and temporal errors produced by text-to-video GenAI models. They created a high-level taxonomy of issues related to GenAI imagery: (1) visual artifacts, (2) shape, form, and geometry, (3) semantic mismatch with text, (4) physics, (5) motion, and (6) "other". Furthermore, through online crowdsourcing, they obtained bounding box locations for these different categories. We take a similar approach in this work by tagging the visual artifacts listed in the left column of Table 1 for each video in our GenAI dataset.

The evaluation of GenAI audio quality introduces its own challenges. MOS studies remain the gold standard for the assessment of speech and music quality, with older objective metrics like PEAQ (1998) and POLQA (ITU-T P.863) designed for evaluating compression. New audio objective metrics are primarily deep-learning-based [16], like for video.

More recent perceptual studies have focused on deepfake speech detection [17], [18] and have helped bridge the gap between MOS and deep-learning-based approaches. Like our work, these studies use tight experimental control to allow for a better understanding of how GenAI audio impacts human perception. For example, Barrington et al. [17] measured quality in terms of how well people could identify real versus AI-cloned voices. Moreover, they created a taxonomy of audio cues such as voice inflection, accents, speech pace, pauses, and breathing. Our work takes a similar approach (a perceptual study and creation of a taxonomy of GenAI audio giveaways/cues) but focuses on a wider array of GenAI audiovisual content by intentionally excluding speech. We also frame our findings in terms of "quality of experience" rather than deepfake detection and the corresponding societal impact.

Methods

We have described the methods of this work in detail previously [7]. We provide a condensed version below and refer the reader to [7] for additional details.

Participants

Thirty-six observers participated in this study, of whom thirty-three completed a post-experiment survey (nineteen male and fourteen female). There were four audio experts, twelve image/video/color experts, and seventeen who reported being neither an image nor an audio expert. All observers were naive to the study hypotheses. Visual acuity was assessed at the beginning of the experiment.

Datasets/Stimuli

Two datasets were assembled for this experiment, each containing sixty-six five-second video clips. The first dataset comprised unique UGC videos, while the second consisted of GenAI videos synthesized to correspond to the same visual events as in the first dataset. Specifically, for each UGC video clip, we wrote text prompts that described the scene dynamics (e.g., object interactions, character intent, location, camera recording style, and camera motion). We used the text prompts as input to various off-the-shelf GenAI models, with or without one or more frames extracted from the original UGC clips. This approach enforced a one-to-one pairing between every synthetic video and its corresponding UGC video from the first dataset (Figure 1). In sum, there were sixty-six unique pairs of videos.

Dataset 1: UGC Videos

The UGC videos in the first dataset were recorded on the authors' personal mobile devices, on YouTube, on Pexels.com, and from an internal archive. The videos spanned a variety of everyday scenes, including human activities, animals moving within natural habitats, and physical interactions with objects or environments (Figure 1, left).

The recordings were captured under uncontrolled conditions using unique devices and settings. Thus, there is a considerable amount of variability in terms of audiovisual compression artifacts, frame rate and resolution for video, and sampling rate and frequency

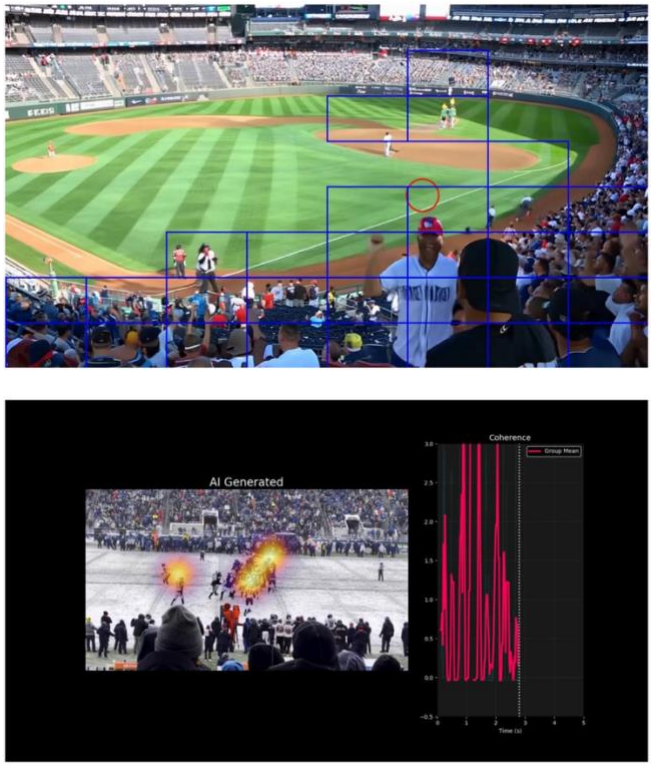


Figure 2. *Visual analysis. (Top) Example video frame from a GenAI video with ground truth boxes (blue rectangles) for violations of biological motion superimposed over the frame, and a single observer's recorded momentary fixation position (red circle). (Bottom) The NSS analysis for assessing inter-observer visual attention coherence.*

response for the microphone audio. Importantly, the audio tracks also contained incidental environmental sounds not directly related to the visual action, such as wind noise, crowd ambience, or distant background events (e.g., a car driving on the road).

Dataset 2: GenAI Videos

The imagery in the video clips in the second dataset was synthesized using several off-the-shelf text-to-video models: Runway Gen-3 Alpha/Turbo, Luma Dream Machine, Pika AI, and Meta Movia Gen (Figure 1, right). We utilized three prompting strategies to guide the generation process: (1) a text prompt, (2) a text prompt and a single image frame from the corresponding UGC video, or (3) a text prompt and two video frames. Several additional clips were obtained from publicly released examples from Meta Movie Gen and Pika AI.

Excluding the publicly released videos, which contained their own synthetic audio, we utilized MaskVAT [19], a video-to-audio model, to generate synthetic audio from the GenAI imagery. Lastly, we applied MaskVAT to the videos in Dataset 1. Specifically, we fed in the optically captured video frames to the model to produce a third, hybrid dataset of videos that contained real imagery and synthetic audio.

Apparatus

Stimuli were presented in a dark room on a 55-inch 4K LG OLED display operating in SDR at 60 Hz. Observers sat at a distance of ~1.5 picture heights from the screen (44"), yielding 61.4 pixels per degree of visual angle. Eye movements were recorded with a Tobii Fusion Pro tracker at 60 Hz. Participants listened to the audio through Sennheiser HD 650 headphones. The experiment was controlled via Psychopy, an open-source Python library [20].

Task Procedure and Experiment Design

A two-interval forced-choice (2IFC) paradigm was used. Specifically, two video clips, from a single pairing as described in the ***Datasets/Stimuli*** section, were presented sequentially on each trial. Observers had to indicate which interval contained the more realistic video. The 2IFC format was chosen to reduce internal criterion drift [21] and operationally define "quality" in terms of task performance, measured on an interval scale.

The eye tracker was calibrated and validated per observer at the beginning of the experimental session, and a custom drift check was included at the beginning of every trial. On each trial, one clip in the video pair was randomly assigned to Interval A and played for 5 seconds, followed by a 1-second segment of achromatic spatio-temporal 1/f noise to mask visual persistence. Interval B then presented the second 5-second clip, followed by a second noise video. Participants pressed the spacebar to end the sequence and were asked to report which interval, A or B, contained the more realistic video.

Four audiovisual conditions were defined to probe the relative contribution of visual and auditory cues in observers' ability to discriminate UGC from GenAI content. For the acronyms below, the first term before the dash denotes the video source and the second the audio source. The expression "vs." denotes the two stimuli presented within a trial; the first corresponds to the "real" video, and the second to the GenAI video. Thus, selecting the first stimulus in a given trial represents the correct response.

1. **OC**-* vs. **AI**-*: **O**ptically **c**aptured versus **AI**-generated video with audio removed (Silent)
2. OC-**MA** vs. AI-AI: Optically captured video with **m**icrophone **a**udio versus fully AI-generated audiovisual content (Full Match)
3. OC-AI vs. AI-AI: Optically captured video with AI-generated audio versus fully AI-generated audiovisual content, emphasizing visual cues (AI Audio Swap)
4. OC-MA vs. OC-AI: Identical optically captured video paired with microphone audio or AI-generated audio, focusing on audio cues (Audio Only Differ)

The study used a within-subjects design in which each participant completed trials from all four conditions across two blocks. The first block corresponded to condition 1 (~33 trials). The second block contained the remaining trials, evenly split across the remaining three conditions. Video pairs were counterbalanced across conditions. Thus, different participant subsets encountered the same pair under different audiovisual configurations, avoiding repeated exposure to modified versions of the same content for any given observer.

Eye Movement Analysis

We conducted two types of image analysis. The first focused on defining ground truth bounding boxes for the various visual artifacts listed in Table 1 and correlating observers' gaze with these artifacts. The second analysis implemented the Normalized Scanpath Saliency metric (NSS) from Dorr et al. [22] to assess coherence—the similarity in gaze patterns across a group of observers who saw the same stimulus within the same audiovisual condition.

GenAI videos were annotated by partitioning each frame into an 8x8 grid of rectangles. For each artifact-video combination, we

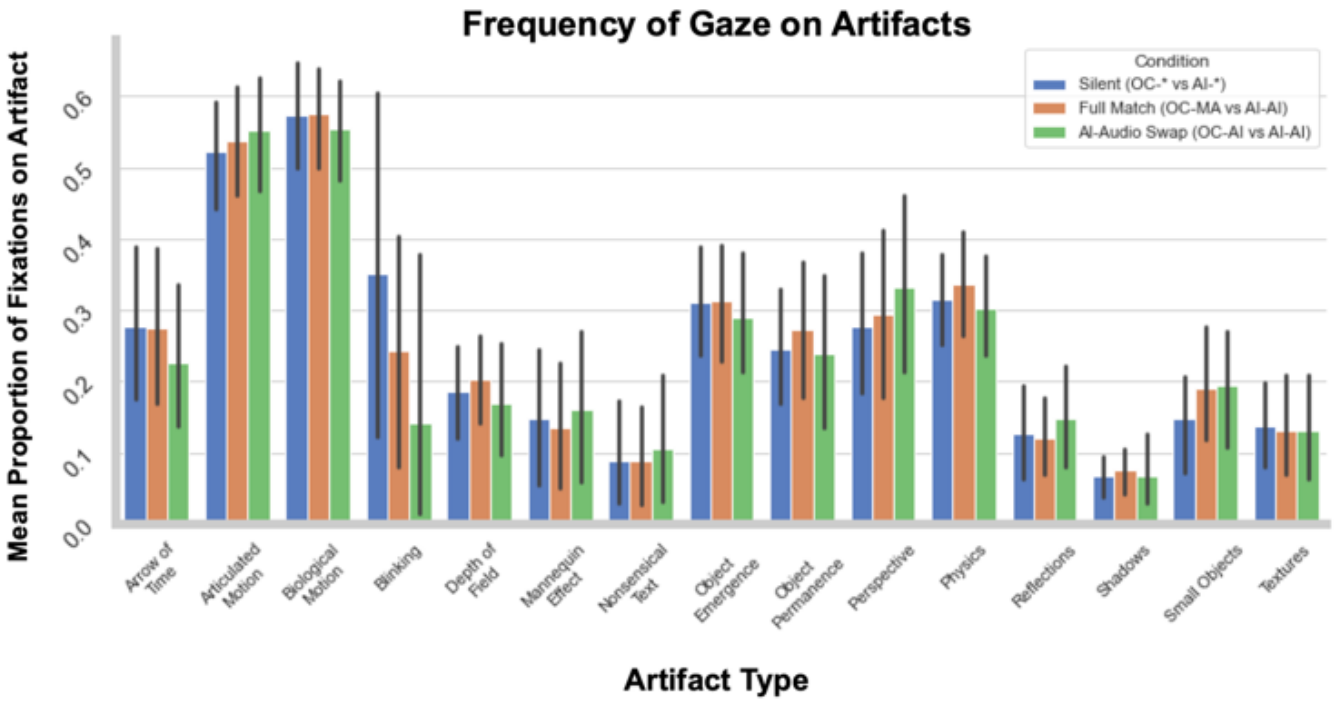


Figure 3. *Proportion of fixations directed at the various visual artifacts listed in Table 1, left.*

labeled the rectangles containing the visual artifact across frames, defining spatiotemporal artifact ROIs. Fixations recorded from the eye tracker were mapped to this grid (Figure 2, top), and the primary metric was fixation frequency on the various artifacts. Additional measures included cumulative fixation duration and the first artifact fixated in each trial. Statistical significance was assessed using a permutation test in which gaze data were randomly reassigned to artifact annotations across videos to generate a null distribution. For each artifact type and condition, 5,000 permutations were computed to estimate z-scores and p-values for fixation frequency, cumulative dwell time, and reaction time to detect the artifact.

Inter-subject gaze similarity (coherence) was quantified using the NSS metric (Figure 2, bottom). Specifically, spatiotemporal fixation maps were constructed from the gaze data of all observers except one (leave-one-out) by centering Gaussian kernels ($\sigma_x = \sigma_{xy} = 1.2°$ visual angle, $\sigma_t = 26$ ms) at the recorded fixation locations in a 3D (x, y, t) volume. The volume was subsequently normalized to zero mean and unit variance. NSS was computed as the mean normalized *salience* value at the held-out observer's gaze locations within sliding temporal windows of the volume (225 ms, 25 ms stride), as shown in Figure 2, bottom right. To quantify overall gaze coherence, we computed inter-subject correlation (ISC). ISC was defined as the correlation of each observer's NSS time series on a given video with the mean of all others' NSS time series for that same video and applying a Fisher z-transform to that correlation coefficient. A Wilcoxon Signed-Rank test was computed between ISC for real and synthetic videos within each condition.

Audio Analysis

Like the visual artifact analysis, we evaluated each video with synthetic audio for the presence of audio content or quality issues. For each audio waveform-issue pair, we rated the degree to which the issue (see Table 1, right) was present using a 5-point scale. A rating of 1 indicated strong disagreement that the issue was present, whereas a rating of 5 indicated strong agreement.

Audio issues were stratified into two categories: content issues and quality issues. Audio content issues focused on high-level aspects of the audio waveform and its relation to the imagery in the video. For example, *logic* issues refer to the notion that the generated audio is categorically incorrect (i.e., there is a categorical mismatch with objects/actions shown in the video). As another example, *Statsis* refers to the fact that the generated audio does not vary enough over the course of the video clip. Audio quality issues focus on the low-level features of the audio waveform. For example, if the audio is band-limited or the sound is extremely muffled/tiny, then this can be readily identified by examining a spectrogram.

Table 2: Statistical model of observer performance

Condition	Odds ratio (rel. to Silent)	Detection Prob.	P-value
Silent	1.00 (Ref.)	0.932	**<0.001**
Full Match	1.05	0.934	0.834
AI-Audio Swap	0.40	0.845	**<0.001**
Audio Only	0.33	0.820	**<0.001**

Results

Table 2 highlights the group performance within each of the 4 conditions. A generalized linear mixed-effects model, or GLMM, (binomial, logit link) predicted detection probabilities with condition as a fixed effect. We included random effects for observers (Variance = 0.18) and video pair IDs (Variance = 0.70).

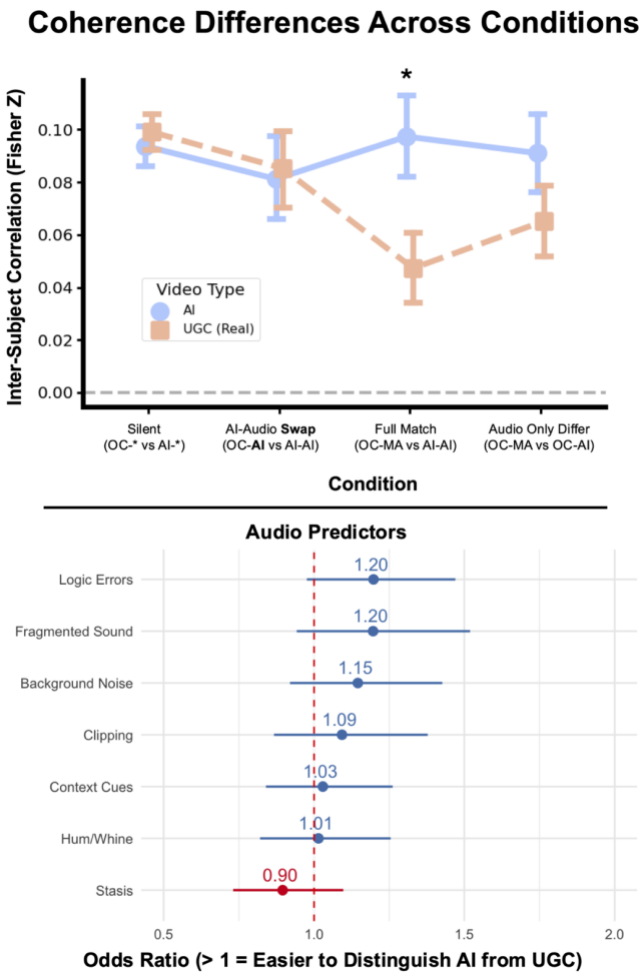


Figure 4. *Visual Coherence (top), quantified via the NSS metric and a subset of the GLMM audio cue predictors for decision accuracy (bottom), * = p-val < 0.05.*

The random effects indicated consistent observer responses (i.e., low estimated variance) but varying difficulty across the sixty-six unique video pairs. Observer performance was highest in the Silent and Full Match conditions and lowest when AI audio was paired with real video (AI Audio Swap and Audio Only Differ).

Figure 3 depicts the frequency of fixations on the various artifacts across all videos and observers, for each condition containing GenAI imagery. Our permutation test revealed that violations of biological motion and physics, object emergence /mergence, and inaccurate representation of small complex objects were the artifacts gazed at most frequently. Our supplementary analysis on cumulative gaze duration and first artifact fixated (not shown) revealed that only violations of biological motion were statistically significant (p-val < 0.05) across all three metrics. Taken together, violations of biological motion were the most conspicuous visual artifact.

Table 3: Statistical model of audio issues, a subset of predictors.

Audio Issue (cue)	Prob. of correct decision (at +1 SD)	Change from Baseline	P-value
Logic Errors	0.883	+ 0.020	0.084
Fragmented Sound	0.883	+ 0.020	0.141
Background Noise	0.878	+ 0.015	0.223
Clipping	0.873	+ 0.010	0.452
Stasis (non-varying audio)	0.849	-0.014	0.228

Figure 4, top, exemplifies the mean inter-observer Coherence, or the "collective attention" of the group of observers for real (orange) and synthetic (blue) videos across the 4 conditions. We predicted that the Coherence would be higher for synthetic videos than UGC videos because the visual artifacts would draw multiple observers' attention to the same locations while viewing the video clips. For UGC videos lacking visual artifacts, observers' scan paths would be more idiosyncratic. That is, their attention would be drawn to different locations based on their unique cognitive schema or top-down attention. Although the correlation in NSS between observers was low across video types and conditions, we observed a significant difference in the Full Match condition, with synthetic videos having a higher correlation in NSS than UGC videos (p-val < 0.01). This suggests the simultaneous occurrence of an audio issue and a visual artifact that together guide multiple observers' gaze to similar artifact ROIs.

For our audio analysis, we ran a GLMM, the same statistical model structure described for detection performance across conditions. Rather than using conditions as fixed effects, we treated the audio content/quality issues as fixed effects (Figure 4, bottom). Random effects for observer ID (Variance = 0.06) and video ID (Variance = 0.5) were included. Baseline accuracy across the three conditions with audio was 86%. Neither audio content nor quality issues were significant predictors of decision accuracy. Audio logic issues (i.e., categorically incorrect audio) trended toward significance (Odds ratio = 1.2, p-val = 0.084), but only marginally improved discriminability performance by 2% at one standard deviation above the mean rating. These results suggest that observers have a high "noise floor". That is, they do not rely on audio fidelity cues to discriminate real from synthetic videos.

Discussion

In this work, we assessed how well observers are at discriminating real UGC videos from synthetic ones created using off-the-shelf text-to-video GenAI models. Not surprisingly, people were adept at identifying synthetic content (Table 2). In the absence of conflicting audiovisual cues (Silent and Full Match conditions), the visuals were the primary giveaway. When synthetic audio was remuxed with optically captured imagery (AI-Audio Swap), people found the task more difficult, suggesting cross-modal confusion. Interestingly, when people had to focus on just the audio (Audio Only Differ), they performed worse, meaning the AI audio cues were harder to spot than visual artifact cues. Alternatively, people could have been too focused on the imagery in this condition since it was matched across the two videos in each trial, and that is why performance was lowest. To fully assess the relative weighting of

the auditory and visual modalities in the decision-making process [23], [24], future work would need to include an audio-only condition.

Artifact conspicuity was dominated by violations of biological motion (i.e., it attracted the greatest attention across observers). Notably, observers were not instructed to search for specific artifacts. Instead, they freely viewed the clips, and their gaze was correlated with artifact ROIs. A more systematic assessment of artifact salience and its impact on quality of experience would require isolating each artifact type as an independent variable [25]. Nevertheless, these findings offer practical guidance for GenAI models deployed on edge devices. For applications or use cases involving the generation of synthetic human or animal subjects, high-level violations of biological motion should be prioritized for mitigation. They are more perceptually salient during free viewing than lower-level artifacts such as the blending of real-world textures with compression distortions.

The taxonomy of audio content and quality issues we developed were not significant predictors for discriminating real from GenAI audio. As discussed earlier, the visual component of the task could have masked the audio issues. Additionally, we mainly tested a single video-to-audio GenAI model, MaskVAT. Future work could use our taxonomy and rating procedure to either (1) assess multiple audio models or (2) a single model, but trained with a varying number of parameters to systematically degrade the audio. These approaches could improve the SNR for raters assessing each audio issue and help tease apart which issues are most apparent.

Conclusion

Our work addresses the growing need to not only predict but also explain GenAI audiovisual quality in a systematic way. We propose a taxonomy of visual artifacts and audio content and quality issues to help guide the development of future GenAI models, particularly as they are deployed on edge devices. In addition, our task-based definition of GenAI "quality", grounded in discriminability performance, provides a more reliable and reproducible alternative to traditional MOS studies, which can often face consistency challenges due to the online nature of how the data is collected.

References

[1] J. C. Madathil, "Generative AI advertisements and Human–AI collaboration: The role of humans as gatekeepers of humanity," *Journal of Retailing and Consumer Services*, vol. 87, p. 104381, Oct. 2025, doi: 10.1016/j.jretconser.2025.104381.

[2] M. A. Cusumano, "Generative AI as a New Innovation Platform," *Commun. ACM*, vol. 66, no. 10, pp. 18–21, Oct. 2023, doi: 10.1145/3615859.

[3] H. Ding *et al.*, "Efficient-vDiT: Efficient Video Diffusion Transformers With Attention Tile," 2025, *arXiv*. doi: 10.48550/ARXIV.2502.06155.

[4] L. Ale, N. Zhang, S. A. King, and D. Chen, "Empowering generative AI through mobile edge computing," *Nat Rev Electr Eng*, vol. 1, no. 7, pp. 478–486, Jun. 2024, doi: 10.1038/s44287-024-00053-6.

[5] T. Isobe, H. Cui, D. Zhou, M. Ge, D. Li, and E. Barsoum, "AMD-Hummingbird: Towards an Efficient Text-to-Video Model," 2025, *arXiv*. doi: 10.48550/ARXIV.2503.18559.

[6] B. R. Cowley, P. L. Stan, J. W. Pillow, and M. A. Smith, "Compact deep neural network models of the visual cortex," *Nature*, Feb. 2026, doi: 10.1038/s41586-026-10150-1.

[7] D. Klein, A. Choudhury, J. Pytlarz, and S. Daly, "Perceptual Realness of Generative Audio-Visual Content," in *2025 IEEE International Conference on Image Processing Workshops (ICIPW)*, Anchorage, AK, USA: IEEE, Sep. 2025, pp. 120–125. doi: 10.1109/ICIPW68931.2025.11386083.

[8] Zhou Wang and A. C. Bovik, "A universal image quality index," *IEEE Signal Process. Lett.*, vol. 9, no. 3, pp. 81–84, Mar. 2002, doi: 10.1109/97.995823.

[9] Zhou Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Trans. on Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004, doi: 10.1109/TIP.2003.819861.

[10] S. J. Daly, "Visible differences predictor: an algorithm for the assessment of image fidelity," presented at the SPIE/IS&T 1992 Symposium on Electronic Imaging: Science and Technology, B. E. Rogowitz, Ed., San Jose, CA, Aug. 1992, p. 2. doi: 10.1117/12.135952.

[11] R. K. Mantiuk, P. Hanji, M. Ashraf, Y. Asano, and A. Chapiro, "ColorVideoVDP: A visual difference predictor for image, video and display distortions," *ACM Trans. Graph.*, vol. 43, no. 4, pp. 1–20, Jul. 2024, doi: 10.1145/3658144.

[12] A. Ghildyal, Y. Chen, S. Zadtootaghaj, N. Barman, and A. C. Bovik, "Quality Prediction of AI Generated Images and Videos: Emerging Trends and Opportunities," Oct. 20, 2024, *arXiv*: arXiv:2410.08534. doi: 10.48550/arXiv.2410.08534.

[13] A. Radford *et al.*, "Learning Transferable Visual Models From Natural Language Supervision," Feb. 26, 2021, *arXiv*: arXiv:2103.00020. doi: 10.48550/arXiv.2103.00020.

[14] T. Unterthiner, S. van Steenkiste, K. Kurach, R. Marinier, M. Michalski, and S. Gelly, "Towards Accurate Generative Models of Video: A New Metric & Challenges," Mar. 27, 2019, *arXiv*: arXiv:1812.01717. doi: 10.48550/arXiv.1812.01717.

[15] J. Kang, M. Silva, P. Sangkloy, K. Chen, N. Williams, and Q. Sun, "GeneVA: A Dataset of Human Annotations for Generative Text to Video Artifacts," Sep. 10, 2025, *arXiv*: arXiv:2509.08818. doi: 10.48550/arXiv.2509.08818.

[16] B. Patton, Y. Agiomyrgiannakis, M. Terry, K. Wilson, R. A. Saurous, and D. Sculley, "AutoMOS: Learning a non-intrusive assessor of naturalness-of-speech," 2016, *arXiv*. doi: 10.48550/ARXIV.1611.09207.

[17] S. Ahmed and H. W. Chua, "Perception and deception: Exploring individual responses to deepfakes across different modalities," *Heliyon*, vol. 9, no. 10, p. e20383, Oct. 2023, doi: 10.1016/j.heliyon.2023.e20383.

[18] S. Barrington, E. A. Cooper, and H. Farid, "People are poorly equipped to detect AI-powered voice clones," *Sci Rep*, vol. 15, no. 1, p. 11004, Mar. 2025, doi: 10.1038/s41598-025-94170-3.

[19] S. Pascual, C. Yeh, I. Tsiamas, and J. Serrà, "Masked Generative Video-to-Audio Transformers with Enhanced Synchronicity," in *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part LXXXVII*,

Berlin, Heidelberg: Springer-Verlag, Nov. 2024, pp. 247–264. doi: 10.1007/978-3-031-73021-4_15.

[20] J. Peirce *et al.*, "PsychoPy2: Experiments in behavior made easy," *Behav Res Methods*, vol. 51, no. 1, pp. 195–203, Feb. 2019, doi: 10.3758/s13428-018-01193-y.

[21] H. Levitt, "Transformed up-down methods in psychoacoustics," *J Acoust Soc Am*, vol. 49, no. 2, p. Suppl 2:467+, Feb. 1971.

[22] M. Dorr, T. Martinetz, K. R. Gegenfurtner, and E. Barth, "Variability of eye movements when viewing dynamic natural scenes," *Journal of Vision*, vol. 10, no. 10, pp. 28–28, Aug. 2010, doi: 10.1167/10.10.28.

[23] M. O. Ernst and M. S. Banks, "Humans integrate visual and haptic information in a statistically optimal fashion," *Nature*, vol. 415, no. 6870, pp. 429–433, 2002, doi: 10.1038/415429a.

[24] D. Alais and D. Burr, "The Ventriloquist Effect Results from Near-Optimal Bimodal Integration," *Current Biology*, vol. 14, no. 3, pp. 257–262, Feb. 2004, doi: 10.1016/j.cub.2004.01.029.

[25] N. Ponomarenko *et al.*, "Image database TID2013: Peculiarities, results and perspectives," *Signal Processing: Image Communication*, vol. 30, pp. 57–77, Jan. 2015, doi: 10.1016/j.image.2014.10.009.

Author Biography

Devi Klein received his BS in psychology from the University of Washington and his PhD in cognitive neuroscience from the University of California, Santa Barbara. His research interests lie in the trade space of human perception and hardware constraints for new technologies.

Anustup Choudhury is a Researcher at Dolby Laboratories, and his research interests include image/video analysis, machine learning, computational photography and computer vision. He received his PhD and M.S. in Computer Science from the University of Southern California.

Evan Gitterman received his BS in Symbolic Systems from Stanford University. His background is in neuroscience, perception, music, and audio engineering.

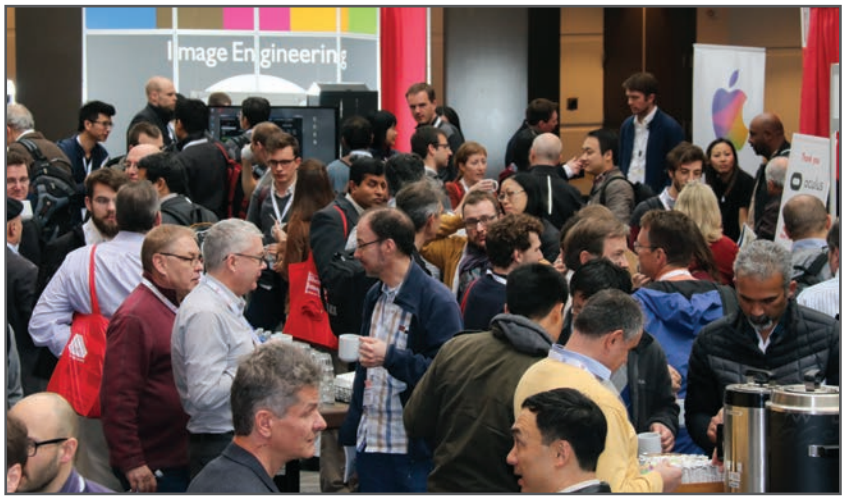
Jaclyn Pytlarz is a Senior Staff Researcher at Dolby Laboratories where she leads Dolby's Vision Science research team. Her research interests lie in applying human perception research to the media entertainment industry. She holds a B.S. degree in Motion Picture Science from Rochester Institute of Technology and an M.S. in Computational and Mathematical Engineering from Stanford University.

Scott Daly is an applied perception scientist at Dolby Laboratories, with specialties in spatio-chromatic-temporal vision, auditory-visual interactions, AI generative media, and contemporary art history. He is an EE with an MS in bioengineering from the University of Utah. Past accomplishments led to the Otto Schade award from SID in 2011, a team technical Emmy in 1990, and completing the 100-patent dash in just under 30 years.

JOIN US AT THE NEXT EI!

electronic IMAGING

Imaging across applications . . . Where industry and academia meet!



• SHORT COURSES • EXHIBITS • DEMONSTRATION SESSION • PLENARY TALKS •
• INTERACTIVE PAPER SESSION • SPECIAL EVENTS • TECHNICAL SESSIONS •

www.electronicimaging.org

ADVANCING IMAGING SCIENCE 75 1947 IS&T 2022 imaging.org