Evaluating Human Aesthetic and Emotional Aspects of 3D generated content through eXtended Reality

Lorenzo Stacchio[1], Claudia Scorolli[2] and Gustavo Marfia[3,*]

[1]University of Bologna, Department for Life Quality Studies
[2]University of Bologna, Department of Philosophy and Communication Studies
[3]University of Bologna, Department of the Arts

Abstract

The Metaverse era is rapidly shaping novel and effective tools particularly useful in the entertainment and creative industry. A fundamental role is played by modern generative deep learning models, that can be used to provide varied and high-quality multimedia content, considerably lowering costs while increasing production efficiency. The goodness of such models is usually evaluated quantitatively with established metrics on data and humans using simple constructs such as the Mean Opinion Score. However, these scales and scores don't take into account the aesthetical and emotional components, which could play a role in positively controlling the automatic generation of multimedia content while at the same time introducing novel forms of human-in-the-loop in generative deep learning. Furthermore, considering data such as 3D models/scenes, and 360° panorama images and videos, conventional display hardware may not be the most effective means for human evaluation. A first solution to such a problem could consist of employing eXtendend Reality paradigms and devices. Considering all such aspects, we here discuss a recent contribution that adopted a well-known scale to evaluate the aesthetic and emotional experience of watching a 360° video of a musical concert in Virtual Reality (VR) compared to a classical 2D webstream, showing that adopting fully immersive VR experience could be a possible path to follow.

Keywords

Generative Artificial Intelligence, eXtended Reality, aesthetic evaluation, human-in-the-loop

1. Introduction

With the advancements in technologies such as eXtended reality (XR), Artificial Intelligence (AI), Cloud Computing (CC), and Digital Twins (DT) the Metaverse is stepping into an upcoming reality [1]. From industrial, healthcare, and research applications to entertainment, tourism, and gaming, Metaverse-related technologies are shaping new tools to improve the way we interact with both the digital and physical worlds [1, 2]. Several works studied how AI-aided paradigms could be applied and integrated into virtual worlds for the aforementioned fields, with

CREAI 2023: Second Workshop on Artificial Intelligence and Creativity, Rome, Italy
*Corresponding author.
lorenzo.stacchio2@unibo.it (L. Stacchio); claudia.scorolli@unibo.it (C. Scorolli); gustavo.marfia@unibo.it (G. Marfia)
https://www.unibo.it/sitoweb/lorenzo.stacchio2/en (L. Stacchio); https://www.unibo.it/sitoweb/claudia.scorolli/en (C. Scorolli); https://www.unibo.it/sitoweb/gustavo.marfia/en (G. Marfia)
0000-0002-9341-7651 (L. Stacchio); 0000-0003-3058-8004 (C. Scorolli); 0000-0003-3058-8004 (G. Marfia)
© 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).
CEUR Workshop Proceedings (CEUR-WS.org)

CEUR-WS.org/Vol-3519/paper4.pdf

a particular focus on entertainment and creative industry [3, 4, 5, 1, 6, 7, 2, 8, 9]. Considering the latter, a plethora of scientific contributions aimed at designing and developing Machine Learning (ML), especially Deep learning (DL), models to improve user experiences in XR environments. ChatBots, Virtual Agents, 2D/3D visual restoration, and generative models are just examples of tools that we can nowadays employ in XR experiences [1, 5, 6, 7, 2, 8, 10].

In particular, Generative Content Models (GCM) architectures, such as Generative pre-trained Transformers (GPT), Generative Adversarial Networks (GAN), Variational Autoencoders (VAEs), Neural Radiance Fields (NeRFs), and Diffusion Models (DM) are crucial to providing varied and high-quality multimedia content, including images, videos, music, and 3D assets, considerably lowering costs while increasing production efficiency. Those DL architectures were successfully trained to perform tasks such as Text-To-Image (e.g., Imagen, Stable Diffusion, DALL·E) or Text-to-Video tools (e.g., Gen-1, VideoFusion), 2D-to-3D (e.g., NeRFs) and Text-to-Panorama (e.g., DiffCollage) on several datasets, providing novel tools that could be used to easily create customized and personalized content [11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23]. Recently, also multi-modal diffusion models were employed to generate a joint audio-video generation framework to create engaging watching and listening experiences simultaneously, with high-quality realistic videos [22, 24].

By using such technologies, it is possible to envision entire entertainment products created with automatic or nearly automatic approaches. 3D environments generated with basic descriptions, digital art paintings, and fashion 3D garments are just examples of virtual elements that could be automatically generated for the benefit of the entertainment and creative industry in virtual experiences. Given these capabilities it is also important to consider how such models are evaluated and if the existing evaluation can fulfill human needs in the target scenarios (i.e., entertainment and creative industry in this study). The goodness of such models is usually evaluated quantitatively with metrics such as Learned Perceptual Image Patch Similarity (LPIPS), Fréchet Inception Distance (FID), Structural Similarity (SSIM), Kullback–Leibler Divergence (KL), Mover's distance (EMD), Chamfer (pseudo)-distance (CD) and Minimum Matching Distance (MMD) [25, 26, 27, 28, 29, 30, 31]. Considering instead human evaluation of such models, the majority of those contributions adopt the Mean Opinion Score (MOS), which corresponds to a 5-point Likert question, or the classical Turing Test delivered in controlled labs or with services such as Amazon Mechanical Turk [32, 11, 15, 12, 16, 17, 18, 19, 20, 22].

To the best of our knowledge, there is in fact a lack of work to establish standardized scales and metrics that allow humans to evaluate such generated content based on different and more complex factors that go beyond the simple MOS and Turing test, like the aesthetics and emotional ones [33, 34]. As reported in [33] emotions and aesthetics bear high-level semantics that could be bonded to low-level computable visual features to create novel and reliable inferring systems with positive steering the perception of digital multimedia material, even if challenging considering human subjectivity. However, those constructs could be put to good use to positive control of the design and automatic generation of multimedia content while at the same time introducing novel forms of human-in-the-loop in generative deep learning, particularly related to (i) creative fields like art, music, and literature, where human artists collaborate with generative models to maintain artistic vision and control to induce certain moods or emotions; (ii) healthcare applications where customized positive emotional effects can be used for mental health treatment, rehabilitation, and stress reduction [35, 33, 36]. Furthermore,

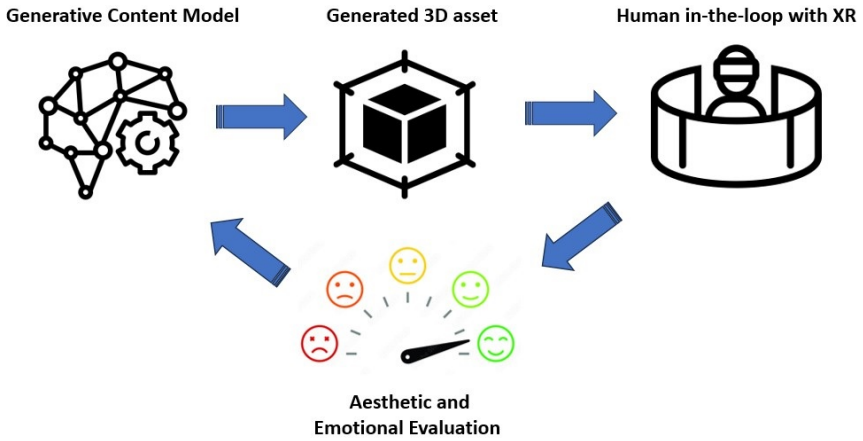

Figure 1: Schema of the proposed Aesthetic and Emotional Evaluation framework for 3D generated content with XR.

when extending our perspective to encompass digital content that goes beyond traditional 2D images, videos, and audio, such as 3D models/scenes, and 360° panorama images and videos, it becomes clear that conventional display hardware may not be the most effective means for human evaluation, as noted in previous research studies [37, 36, 38]. A first solution to such a problem could consist of employing XR paradigms and devices to find better and more effective ways to present such visual/auditory stimuli to humans and ease their evaluation. A side effect of adopting XR paradigms for this purpose could be the improvement of the role of humans in driving generative models to specific outcomes, from a human-in-the-loop perspective [39]. The overall framework is visually depicted in Figure 1.

Considering all the mentioned aspects, we will first discuss whether XR paradigms could be employed to evaluate multimedia content, particularly referring to 3D models and scenes, and 360° panorama images/videos. Then, we will report the main insights from a recent related work [40] that adopted a well-known scale to evaluate the aesthetic and emotional experience of watching a 360° video of a musical concert in VR compared to living the real experience and enjoying it from classical 2D displays. This contribution introduced an evaluation framework that could be deployed to evaluate generative deep learning from aesthetic and emotional perspectives.

2. Evaluating multimedia content from an Aesthetic and Emotional perspective with eXtended Reality

XR integrates digital and physical to various degrees, e.g., augmented reality (AR), mixed reality (MR), and virtual reality (VR). Different degrees of virtuality can be exploited to furnish multimedia digital assets to be enjoyed and evaluated by humans [41]. VR is a technology that permits the immersion of the user within a digital layer of information, making them part of the simulated scenario, completely replacing the physical world with devices such as HMDs (Head Mounted Displays) and CAVEs (Cave Automatic Virtual Environments)[42, 43, 38]. On

the other hand, AR overlays digital information onto the physical world using screen-based interfaces (e.g., mobile devices) or dedicated glasses (e.g., Magic Leap) [44]. Finally, MR merges the physical and the virtual world, providing user-interactable digital content responsive to the surrounding physical environment, [1, 38, 7, 2].

For the particular aim of evaluating immersive and/or 3D digital multimedia content, MR and VR represent potential tools to evaluate responses to visual conditions addressing limitations of the other experimental methodologies [38]. Referring to MR, Spacedesign is one of the first frameworks introducing a mixed virtual environment to evaluate the aesthetic of 3D models [42], focusing on free-form curves and surfaces. The authors reported that the introduced system has been tested by experienced industrial designers who appreciated the 3D visualization and navigation, real-time editing, and intuitive interaction. In this scenario, designers and stylists play a significant role in the development process, enabling them to effectively steer their vision from start to finish, resulting in the final product.

Authors of [45] proposed using VR paradigms to emphasize aesthetic and emotional abilities in students for 3D design. Through a statistical assessment process, the authors reported that VR could have a positive impact on creative thinking as well as students' academic performance. This means that virtual technologies are able to highlight these human factors while interacting with such multimedia data and so impact their evaluation. On a similar line, some works have shown that VR can be effectively used while evaluating the aesthetic and emotional effect of 3D scenes or 360° videos [36, 40].

For example, [46] studied the impact of VR in awe emotions. They considered 360° immersive videos, examining 42 participants who watched immersive and normal 2D videos displaying awe or neutral content, rating their level of awe and sense of presence after the experiment. Results indicated that immersive videos significantly enhanced the self-reported intensity of awe as well as the sense of presence.

To the best of our knowledge, there is no prior work that considered both aesthetic and emotional components in evaluating immersive 3D models or 360° videos by adopting XR paradigms. Considering this, in the following, we discuss one of the most recent works on the topic [40] which represents a first attempt at comparing aesthetic and emotional constructs among real and VR experiences, focusing on 360° videos. The introduced framework could put the basis to create a novel generative DL human evaluation pipeline for synthetic multimedia data [13, 47].

3. Evaluating digital content from an Aesthetic and Emotional perspective through eXtended Reality: use case of 360° panorama Tango Concert video in Virtual Reality

In evaluating digital multimedia content with VR, some validated scales could be adopted to assess the Aesthetic and Emotional dimensions related to it. To elaborate on this topic, we reference a recent study conducted in [40]. In this work, authors presented initial findings from a project whose primary objective was to compare the aesthetic and emotional experiences of a musical concert lived in presence with respect to virtual ones. They also explored the social dimension, which won't be discussed here to highlight their contribution to aesthetic

and emotional dimensions. This multidisciplinary project has been questioning whether virtual environments could recreate live aesthetic experiences, considering the lack of related works, circumscribing the analysis to the musical ones. This could demonstrate the ability of virtual environments to provide novel ways and tools of enjoying not only digital assets but also compare the enjoyment and their evaluation with respect to a realistic situation. To understand whether and how these aesthetic and emotional experiences are systematically modulated by differently immersive settings, authors selected physical and social environments for which the degree of immersiveness is incrementally increased.

Consistently, authors contrasted the classic experience of a live concert (LC, in presence) with three "remote" conditions, not simultaneous with respect to the event: the experience with audiovisual musical files (MV, the classic viewing of a concert on YouTube) and two experiences lived thanks to a less or more sophisticated eye-mounted apparatus for virtual reality (VR), i.e., google cardboard (CVR) and an HTC-Vive (HVR). The CVR and the HVR allow respectively for a basic and easily accessible experience vs. a less affordable but more immersive experience. Both the devices permitted a three-dimensional vision: by moving the head or the whole body, participants could have a 360° view, therefore an overall vision of the concert venue, including musicians and audience – together with their possible reactions to a virtuosity or a false note played by the performers (see Figure 2).

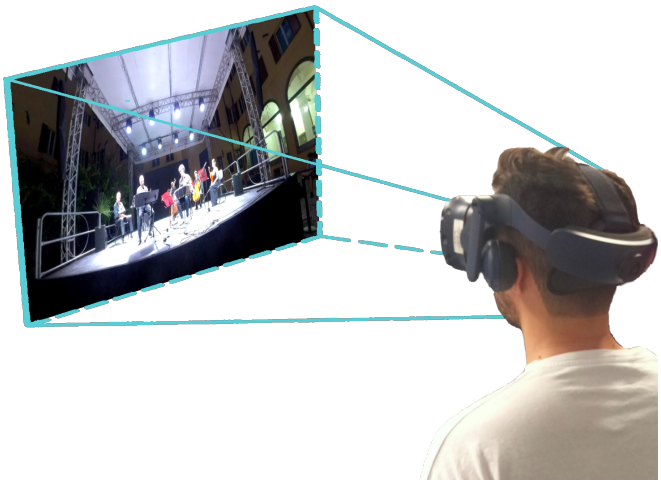

Figure 2: Exemplar visualization of a user experiencing the virtual concert with the HTC-Vive (HVR). The virtual concert is provided as a 360-degree video.

For this reason, authors take as a use case a theater experience. From an experimental perspective, authors intended to exploit VR technologies as a methodological boost to empirical aesthetics: virtual environments provide an excellent compromise between ecological validity and experimental control. Here, authors compared different devices able to convey an aesthetic-musical experience, including VR devices with varying degrees of immersiveness, in order to investigate their ability to engage, more or less powerfully, the participant with respect to that experience, typically enjoyed in a theater. To investigate so human aesthetic and emotional

experience, the authors stated that a rethinking of classic laboratory experiments, using VR paradigms to overcome traditional and reductive approaches in aesthetic and emotional. Thus, the authors addressed the aesthetics and emotions evoked in the four conditions (LC, MV, CVR, HVR) through the administration of a validated questionnaire: the Aesthetic Emotions Scale [48], structured in 21 subscales covering prototypical aesthetic emotions, epistemic emotions, and emotions indicative of amusement.

As mentioned, these experimental perspectives were adopted and applied for a particular use case: evaluating how Virtual Experiences could be put to good use within a musical aesthetic experience. Specifically, they used VR technologies to increase the interest and engagement of a population (young students) that is not interested in attending a certain kind of cultural experience, namely a Tango music concert in a theatre, but that is prone to use Virtual Experience technologies [49, 50] and comparing them with a strong baseline: passionate adult people, which is usually the target audience for this kind of cultural activities [51]. In practice, authors used LC participants' survey scores as a benchmark to compare the magnitude of interest of passionate adult people with respect to the youngsters who lived instead of Virtual Experiences (MV, CVR, HVR). As an additional contribution, this framework could be adopted with any other kind of aesthetic experience.

3.1. Experimental session

In the experimental session, authors tested 70 participants, 10 for the live concert condition; 20 for the music video condition; 20 for the immersive condition with the Google Cardboard; and 20 for the immersive condition with the HTC-VIVE. We remind the original work for further details [40]. In the experimental session, the LC condition was executed at the Teatro Comunale "Pavarotti-Freni" in Modena, during the concert "Amarcord d'un Tango". The event took place outdoors, in the theatre courtyard. At the end of the concert, the authors asked volunteers to fill in a hard copy of both questionnaires. For the other three virtual conditions, they used a 360° video of the concerts, testing participants were tested at the Virtual and Augmented Reality Lab (VARLAB) of the University of Bologna. The Aesthemos questionnaire was furnished in the form of forty-two 5-point Likert scale scores referring to twenty-one emotion subscale[48]: *Feeling of beauty; Fascination; Being moved; Awe; Enchantment; Nostalgia; Joy; Humor; Vitality; Energy; Relaxation; Surprise; Interest; Intellectual challenge; Insight; Feeling of ugliness; Boredom; Confusion; Anger; Uneasiness; Sadness* as described in [48, 40].

3.2. Results and discussions

The author's collected data has undergone a reliability check to test for internal consistency and validate the research: they analyzed those constructs that exhibited a Cronbach's alpha index ≥ 0.70 [52]. All the constructs that passed this test, were subjected to statistical analyses to verify any significant differences among the four conditions, namely LC, MV, CVR, and HVR. Authors performed an Adjusted Wald-Confidence Interval test [53], which allows to check if the difference between two proportions is significant and how large the difference is. Authors selected this specific statistical test as they had four conditions and a low number of samples for each of them [54]. To adapt their data for the specific test, they binarized the Likert-scale

answers with a threshold of 4: ≥ 4 Likert scale answers were converted to 1, the lower to 0.

Question	Comparisons	(Wald) Inf. bound - Difference - Sup. bound
FB-E1	LC > CVR	0.18 - 0.47 - 0.76
FB-E1	LC > HVR	0.18 - 0.47 - 0.76
FB-E1	LC > MV	0.23 - 0.52 - 0.80
INT-E38	LC > CVR	0.07 - 0.39 - 0.71
INT-E38	LC > MV	0.22 - 0.53 - 0.83
INT-E38	**HVR > MV**	0.05 - 0.32 - 0.59
FA-E31	LC > CVR	0.02 - 0.34 - 0.66
FA-E31	LC > MV	0.22 - 0.53 - 0.83
FA-E31	**HVR > CVR**	0.04 - 0.32 - 0.60
FA-E31	**HVR > MV**	0.25 - 0.50 - 0.76

Table 1
The Table reports statistically significant comparisons for the Aesthemos, organized by construct-question *Feeling of Beauty*, *Interest* and *Fascination*, and reporting the Wald inferior bound, difference, and superior bound.

Table 1 reports some of the statistically significant results obtained for the *Feeling of Beauty* (FB-x items), *Interest* (INT-x items) and *Fascination* (FA-x items) constructs[1]. Each question item was codified to improve the readability of the table as follows: (FB-E1) *"I found it beautiful"*; (INT-E38) *"It piqued my interest"*; (FA-E31) *"I found it sublime"*.

The results highlighted how the Live Condition (LC) was in general the most effective condition in activating aesthetic emotions, in particular, related to *Feeling of Beauty* when compared with any of the virtual conditions. However, it can be also observed that, for the other two constructs, the superiority of the LC condition with respect to the HTC-Vive (HVR) one was not statistically significant. Moreover, considering the same constructs, HVR was superior to the Music Video (MV) and VR Google cardboard (CVR) conditions. Thus, even if the LC remains the best way to enjoy a music concert, the difference with the same experience lived through the HCT-Vive headset is not statistically significant for certain aesthetic and emotional constructs. This suggests that the fully immersive virtual reality is the "artificial experience" that can offer the spectator an experience more like the "real-live one" with respect to classical means such as 2D displays or mobile VR. To the best of our knowledge, this is one of the first empirical evidence of a major interest in musical aesthetic experiences when experienced in VR (HVR) than on a computer screen. This is also the first contribution that highlights how VR paradigms could be better than classical means to evaluate digital multimedia data such as 360° videos.

4. Conclusion

In this work, we discussed the possibility of measuring aesthetic and emotional constructs to evaluate the generative DL models exploiting XR paradigms. In fact, nowadays, the goodness of such models is evaluated from a qualitative human perspective using simple constructs

[1]We remind to [40] for the complete analysis.

such as the MOS and the Turing Test [11, 15, 12, 16, 17, 18, 19, 20, 22]. However, aesthetic and emotional factors could be used to positively improve generative model performance [33], even if challenging considering human subjectivity and the lack of validated scales. Extending the perspective to encompass digital content that goes beyond traditional images, videos, and audio, such as 3D models/scenes, and 360° panorama images and videos, conventional display hardware may not be the most effective means for human evaluation. A solution could be resorting to XR paradigms, which could also provide novel insights from a human-in-the-loop perspective [37, 36, 38, 39]. To the best of our knowledge, no previous work considered such a specific research direction. However, the framework introduced din [40] focused on the aesthetic and emotional evaluation of 360° videos in VR comparing them with their realistic counterpart and classical displays, providing one of the first empirical evidence that XR paradigms could be a better means to furnish the mentioned type of multimedia content to judge its quality (the approach is extensible also to generative deep learning material). In future works, we intend to explore such an approach with generative models involving 3D models/scenes, and 360° panorama images and videos and measure how the aesthetic and emotional perception changes with respect to classical fruition devices and paradigms.

References

[1] Y. Wang, Z. Su, N. Zhang, R. Xing, D. Liu, T. H. Luan, X. Shen, A survey on metaverse: Fundamentals, security, and privacy, IEEE Communications Surveys & Tutorials (2022).

[2] T. Huynh-The, Q.-V. Pham, X.-Q. Pham, T. T. Nguyen, Z. Han, D.-S. Kim, Artificial intelligence for the metaverse: A survey, Engineering Applications of Artificial Intelligence 117 (2023) 105581.

[3] L. Stacchio, M. Perlino, U. Vagnoni, F. Sasso, C. Scorolli, G. Marfia, Who will trust my digital twin? maybe a clerk in a brick and mortar fashion shop, in: 2022 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW), IEEE, 2022, pp. 814–815.

[4] L. Stacchio, A. Angeli, G. Marfia, Empowering digital twins with extended reality collaborations, Virtual Reality & Intelligent Hardware 4 (2022) 487–505.

[5] E. Morotti, L. Stacchio, L. Donatiello, M. Roccetti, J. Tarabelli, G. Marfia, Exploiting fashion x-commerce through the empowerment of voice in the fashion virtual reality arena: Integrating voice assistant and virtual reality technologies for fashion communication, Virtual Reality (2022) 1–14.

[6] Y. Y. Dyulicheva, A. O. Glazieva, Game based learning with artificial intelligence and immersive technologies: an overview, in: CEUR workshop proceedings, volume 3077, 2022, pp. 146–159.

[7] G.-J. Hwang, S.-Y. Chien, Definition, roles, and potential research issues of the metaverse in education: An artificial intelligence perspective, Computers and Education: Artificial Intelligence 3 (2022) 100082.

[8] S. Zhang, J. Li, L. Yang, Survey on controlable image synthesis with deep learning, arXiv preprint arXiv:2307.10275 (2023).

[9] J. P. Venugopal, A. A. V. Subramanian, J. Peatchimuthu, The realm of metaverse: A survey, Computer Animation and Virtual Worlds (2023) e2150.

[10] S. Mann, Y. Yuan, F. Lamberti, A. El Saddik, R. Thawonmas, F. G. Prattico, extended meta-uni-omni-verse (xv): Introduction, taxonomy, and state-of-the-art, IEEE Consumer Electronics Magazine (2023).

[11] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, B. Ommer, High-resolution image synthesis with latent diffusion models, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 10684–10695.

[12] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. L. Denton, K. Ghasemipour, R. Gontijo Lopes, B. Karagol Ayan, T. Salimans, et al., Photorealistic text-to-image diffusion models with deep language understanding, Advances in Neural Information Processing Systems 35 (2022) 36479–36494.

[13] K. Gao, Y. Gao, H. He, D. Lu, L. Xu, J. Li, Nerf: Neural radiance field in 3d vision, a comprehensive review, arXiv preprint arXiv:2210.00379 (2022).

[14] P. Cascarano, G. Franchini, F. Porta, A. Sebastiani, On the first-order optimization methods in deep image prior, Journal of Verification, Validation and Uncertainty Quantification 7 (2022) 041002.

[15] N. Deng, Z. He, J. Ye, B. Duinkharjav, P. Chakravarthula, X. Yang, Q. Sun, Fov-nerf: Foveated neural radiance fields for virtual reality, IEEE Transactions on Visualization and Computer Graphics 28 (2022) 3854–3864.

[16] M. Kwon, J. Jeong, Y. Uh, Diffusion models already have a semantic latent space, arXiv preprint arXiv:2210.10960 (2022).

[17] Z. Luo, D. Chen, Y. Zhang, Y. Huang, L. Wang, Y. Shen, D. Zhao, J. Zhou, T. Tan, Videofusion: Decomposed diffusion models for high-quality video generation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 10209–10218.

[18] P. Esser, J. Chiu, P. Atighehchian, J. Granskog, A. Germanidis, Structure and content-guided video synthesis with diffusion models, arXiv preprint arXiv:2302.03011 (2023).

[19] J. Ho, W. Chan, C. Saharia, J. Whang, R. Gao, A. Gritsenko, D. P. Kingma, B. Poole, M. Norouzi, D. J. Fleet, et al., Imagen video: High definition video generation with diffusion models, arXiv preprint arXiv:2210.02303 (2022).

[20] R. Gozalo-Brizuela, E. C. Garrido-Merchan, Chatgpt is not all you need. a state of the art review of large generative ai models, arXiv preprint arXiv:2301.04655 (2023).

[21] P. Cascarano, G. Franchini, E. Kobler, F. Porta, A. Sebastiani, Constrained and unconstrained deep image prior optimization models with automatic regularization, Computational Optimization and Applications 84 (2023) 125–149.

[22] L. Ruan, Y. Ma, H. Yang, H. He, B. Liu, J. Fu, N. J. Yuan, Q. Jin, B. Guo, Mm-diffusion: Learning multi-modal diffusion models for joint audio and video generation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 10219–10228.

[23] A. Wang, J. Dong, J. Shen, L.-H. Lee, P. Hui, Towards computational architecture of liberty: A comprehensive survey on deep learning for generating virtual architecture in the metaverse, arXiv preprint arXiv:2305.00510 (2023).

[24] G. Yariv, I. Gat, S. Benaim, L. Wolf, I. Schwartz, Y. Adi, Diverse and aligned audio-to-video generation via text-to-video model adaptation, 2023. arXiv:2309.16429.

[25] Y. Rubner, C. Tomasi, L. J. Guibas, The earth mover's distance as a metric for image retrieval, International journal of computer vision 40 (2000) 99–121.
[26] Z. Wang, A. C. Bovik, H. R. Sheikh, E. P. Simoncelli, Image quality assessment: from error visibility to structural similarity, IEEE transactions on image processing 13 (2004) 600–612.
[27] J. Shlens, Notes on kullback-leibler divergence and likelihood, arXiv preprint arXiv:1404.2000 (2014).
[28] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, S. Hochreiter, Gans trained by a two time-scale update rule converge to a local nash equilibrium, Advances in neural information processing systems 30 (2017).
[29] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, O. Wang, The unreasonable effectiveness of deep features as a perceptual metric, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 586–595.
[30] P. Achlioptas, O. Diamanti, I. Mitliagkas, L. Guibas, Learning representations and generative models for 3d point clouds, in: International conference on machine learning, PMLR, 2018, pp. 40–49.
[31] T. Wu, L. Pan, J. Zhang, T. Wang, Z. Liu, D. Lin, Density-aware chamfer distance as a comprehensive metric for point cloud completion, arXiv preprint arXiv:2111.12702 (2021).
[32] M. G. Keith, L. Tay, P. D. Harms, Systems perspective of amazon mechanical turk for organizational research: Review and recommendations, Frontiers in psychology 8 (2017) 1359.
[33] D. Joshi, R. Datta, E. Fedorovskaya, Q.-T. Luong, J. Z. Wang, J. Li, J. Luo, Aesthetics and emotions in images, IEEE Signal Processing Magazine 28 (2011) 94–115.
[34] NVIDIA, Understanding aesthetics in deep learning, https://developer.nvidia.com/blog/understanding-aesthetics-deep-learning/, 2016.
[35] H. L. O'Brien, E. G. Toms, The development and evaluation of a survey to measure user engagement, Journal of the American Society for Information Science and Technology 61 (2010) 50–69.
[36] S. Triberti, A. Chirico, G. La Rocca, G. Riva, Developing emotional design: Emotions as cognitive processes and their role in the design of interactive technologies, Frontiers in psychology 8 (2017) 1773.
[37] A. Mahdavi, H. Eissa, Subjective evaluation of architectural lighting via computationally rendered images, Journal of the illuminating Engineering Society 31 (2002) 11–20.
[38] A. Bellazzi, L. Bellia, G. Chinazzo, F. Corbisiero, P. D'Agostino, A. Devitofrancesco, F. Fragliasso, M. Ghellere, V. Megale, F. Salamone, Virtual reality for assessing visual quality and lighting perception: A systematic review, Building and Environment 209 (2022) 108674.
[39] X. Wu, L. Xiao, Y. Sun, J. Zhang, T. Ma, L. He, A survey of human-in-the-loop for machine learning, Future Generation Computer Systems 135 (2022) 364–381.
[40] C. Scorolli, E. N. Grasso, L. Stacchio, V. Armandi, G. Matteucci, G. Marfia, Would you rather come to a tango concert in theater or in vr? aesthetic emotions & social presence in musical experiences, either live, 2d or 3d, Computers in Human Behavior (2023) 107910.
[41] P. Milgram, F. Kishino, A taxonomy of mixed reality visual displays, IEICE TRANSACTIONS on Information and Systems 77 (1994) 1321–1329.
[42] M. Fiorentino, R. de Amicis, G. Monno, A. Stork, Spacedesign: A mixed reality workspace

for aesthetic industrial design, in: Proceedings. International Symposium on Mixed and Augmented Reality, IEEE, 2002, pp. 86–318.
[43] M. F. Shiratuddin, W. Thabet, D. Bowman, Evaluating the effectiveness of virtual environment displays for reviewing construction 3d models, CONVR 2004 (2004) 87–98.
[44] M. Leap, Magic Leap 2, The most immersive AR platform for enterprise, https://www.magicleap.com/en-us/magic-leap-2-video, 2023.
[45] A. Jimeno-Morenilla, J. L. Sánchez-Romero, H. Mora-Mora, R. Coll-Miralles, Using virtual reality for industrial design learning: a methodological proposal, Behaviour & Information Technology 35 (2016) 897–906.
[46] A. Chirico, P. Cipresso, D. B. Yaden, F. Biassoni, G. Riva, A. Gaggioli, Effectiveness of immersive videos in inducing awe: an experimental study, Scientific reports 7 (2017) 1218.
[47] Q. Zhang, J. Song, X. Huang, Y. Chen, M.-Y. Liu, Diffcollage: Parallel generation of large content with diffusion models, arXiv preprint arXiv:2303.17076 (2023).
[48] I. Schindler, G. Hosoya, W. Menninghaus, U. Beermann, V. Wagner, M. Eid, K. R. Scherer, Measuring aesthetic emotions: A review of the literature and a new assessment tool, PloS one 12 (2017) e0178899.
[49] J. de la Fuente Prieto, P. Lacasa, R. Martínez-Borda, Approaching metaverses: Mixed reality interfaces in youth media platforms, New Techno Humanities 2 (2022) 136–145.
[50] L. Geng, Y. Li, Y. Xue, Will the interest triggered by virtual reality (vr) turn into intention to travel (vr vs. corporeal)? the moderating effects of customer segmentation, Sustainability 14 (2022) 7010.
[51] S. Meeks, S. K. Shryock, R. J. Vandenbroucke, Theatre involvement and well-being, age differences, and lessons from long-time subscribers, The Gerontologist 58 (2018) 278–289.
[52] K. S. Taber, The use of cronbach's alpha when developing and reporting research instruments in science education, Research in science education 48 (2018) 1273–1296.
[53] A. Agresti, B. Caffo, Simple and effective confidence intervals for proportions and differences of proportions result from adding two successes and two failures, The American Statistician 54 (2000) 280–288.
[54] J. Sauro, J. R. Lewis, Quantifying the user experience: Practical statistics for user research, Morgan Kaufmann, 2016.