Virtual Reality (2023) 27:573–589
https://doi.org/10.1007/s10055-022-00671-1

ORIGINAL ARTICLE

Aesthetic judgments of 3D arts in virtual reality and online settings

Doga Gulhan[1] · Szonya Durant[1] · Johannes M. Zanker[1]

Received: 11 March 2021 / Accepted: 22 June 2022 / Published online: 26 July 2022

Abstract

Empirical aesthetics is beginning to branch off from conventional laboratory-based studies, leading to in-situ, immersive, often more accessible experiments. Here, we explored different types of aesthetic judgments of three-dimensional artworks in two contexts: virtual reality (VR), aiming for an immersive experience, and online settings aiming for an accessible setup for a remote audience. Following the pilot experiment conducted to select a set of 3D artworks, in the first experiment, participants freely engaged with virtual artworks via an eye-tracking-enabled VR headset and provided evaluations based on subjective measures of aesthetic experience such as ratings on liking, novelty, complexity, perceived viewing duration; and the objective viewing duration was also recorded. Results showed positive, linear, and mostly moderate correlations between liking and the other perceived judgment attributes. Supplementary eye-tracking data showed a range of viewing strategies and variation in viewing durations between participants and artworks. Results of the second experiment, adapted as a short online follow-up, showed converging evidence on correlations between the different aspects contributing to aesthetic judgments and suggested similarity of judgment strategies across contexts. In both settings, participants provided further insights via exit questionnaires. We speculate that both VR and online settings offer ecologically valid experimental contexts, create immersive visual arts experience, and enhance accessibility to cultural heritage.

Keywords Virtual reality (VR) · Empirical aesthetics · Art appreciation · Aesthetic judgment · Eye-tracking · Online experiment

1 Introduction

Evaluating visual artworks can be described as a partially overlapping extension of aesthetic experience, and as a complex cognitive-emotional process. Engaging with arts often involves both general emotions such as surprise, joy, or disgust, and emotions associated with arts such as sublime or aesthetic pleasure. On the other hand, when observers are asked to evaluate an artwork whilst they are interacting with art, they tend to assign and update a set of values towards the artwork. The assigned values can be related to any potential aspect of the artwork, for example, beauty, compositional properties, or monetary worth. These highly subjective assigned values are thought to depend on visual properties such as contrast and colour (Mallon et al. 2014), could be affected by contextual information (Grüner et al. 2019), artwork title (Turpin et al. 2019) or artists' names (Cleeremans et al. 2016) and can change over time (Isik and Vessel 2019). Taking on board these complexities from empirical studies, a set of design guidelines for using artworks as stimuli has recently been proposed (Hayn-Leichsenring 2017). The authors of these guidelines highlighted the vagueness of aspects of the research in this area. An additional conceptual challenge is that an observer can assign a value to an artwork either as an absolute judgment or a relative judgment. Nevertheless, previous work tends to propose that the evaluative aspect of aesthetic experience can be operationalized and thus at least partially measurable.

Evaluation of an artwork can be as simple as a single binary judgment of like or dislike, or a long interpretive narrative from an observer. On the theoretical level, aesthetic and non-aesthetic-based values can be assigned to evaluate arts, and they have the potential to influence each other (Aumann 2014). The conceptual richness of visual arts leads to the possibility of using many adjectives, adverbs, or metaphors to evaluate an artwork; and thus aesthetics becomes a challenging research topic for the philosophy of language

✉ Doga Gulhan
doga.gulhan@rhul.ac.uk

[1] Department of Psychology, Royal Holloway, University of London, Surrey, UK

and semantics as well (Young 2017). On the empirical level, researchers have previously investigated themes related to aesthetic values (often by incorporating relative judgments) in varied contexts, such as (i) Perceptual and representational attributes describing paintings as a basis to form an assessment tool (Chatterjee et al. 2010), (ii) Use of highly specific modes of expression such as "feeling like crying" in relation to aesthetic experience (Pelowski 2015), (iii) Predicting aesthetic preference by other perceived attributes such as meaning(fulness), and whether these attributions are robust to image manipulation such as blurring (Moore and West 2012), (iv) Category-dependent generality and specificity of word usage describing subsets of artworks, and aiming to form a language of aesthetics for the visual modality (Augustin et al. 2012), (v) Extent of choice reversal following a type of experimental biasing by pairing "average-beauty" paintings with either relatively more or relatively less beautiful paintings, where observer makes a binary preference choice between two abstract paintings (Belchev et al. 2018), among many others.

As a common framework implied in many studies and based on recent models and discussions (Leder et al. 2004; Locher 2011; Consoli 2015; Wassiliwizky and Menninghaus 2021), a general form of positive aesthetic judgment (such as finding an artwork "good") is often linked to other positive emotional, cognitive, or moral judgments (such as finding an artwork "pleasurable", "meaningful", or "beneficial",). Furthermore, these correlations between judgments are sometimes further broken down to investigate other correlates or mediating factors, for example, in the case of types of preference and meaningfulness judgments; artwork titles (Russell and Milne 1997; Turpin et al. 2019), individual traits (Landau et al. 2006), or viewing duration (Leder et al. 2006). A general interpretation regarding such associations between judgments is that most types of aesthetic judgments can be conceptually aligned along a single negative–positive judgment axis. Some research on the other hand has explicitly investigated the counter-intuitive associations between judgments, as well as specifically negative aesthetic emotions and judgments (Landau et al. 2006; Silvia and Brown 2007; Cooper and Silvia 2009; Wagner et al. 2014).

Since empirical research on aesthetic judgments is often conducted in laboratory settings, a common limitation is using reduced artworks such as 2D snapshots of paintings or a manipulated visual stimulus as a substitute for artworks. The tendency to favour well-controlled stimulus presentation in laboratory settings often results in a diminished resemblance between experimental paradigm and genuine aesthetic experience. The generalizability of findings outside the lab settings to a real world has been described as a common weakness of these studies (Locher et al. 1999; Brieber et al. 2015). Recent developments such as ease of implementing virtual reality (VR) environments, using 3D modeling software as an artistic tool, and more specifically photogrammetry methods to translate physical objects and environments into 3D models have come to offer, in some aspects, ecologically valid alternatives to real-world scenarios and useful tools for cultural heritage (Clini et al. 2018; Liarokapis et al. 2020). As immersive environments aim to enhance user experience in gallery and museum settings, many exploratory studies have started to investigate visitors' experience and the feasibility of these VR applications (Hoang and Cox 2017; Petrelli 2019; Parker and Saker 2020; Gulhan et al. 2021). Experiments have mostly focussed so far on the general cognitive implications of using these environments, for example, crowd movement on navigation decisions in VR (Zhao et al. 2020), mental imagery and eye movements in VR (Chiquet et al. 2020), visual search in 3D scenes (Helbing et al. 2020), replication of findings from a lab-based inattentional blindness paradigm in VR (Schöne et al. 2021), or episodic memory in virtual museum rooms (van Helvoort et al. 2020). Experimental aesthetics research in VR remains to be explored. Apart from screen-based and VR-based studies often conducted in a lab setting, internet-mediated research can be seen as a distinct research setting, and its validity can be linked to the increased viewing of visual arts by a remote audience, away from physical museums and galleries. Additionally, online research in general offers the possibility of a diverse and large sample and minimizing some biases, such as the observer-expectancy effect where researchers unintentionally influence the behavior of participants (Palan and Schitter 2018; Peer et al. 2017; Woods et al. 2015). In this sense, immersive experiments utilizing VR can be framed as a relatively better proxy for the real-world art experience compared to the restrictive lab-based experiments utilizing regular 2D monitors, whereas web experiments can be seen as a proxy for the online art experience.

The present research aimed to incorporate various commonly used judgment types for visual arts from previous studies including liking or novelty (Faerber and Carbon 2012; Fayn et al. 2015; Graf and Landwehr 2017; Song et al. 2021). Additionally, we aimed to include relatively "unusual" and infrequently investigated types of judgments such as liking from a third-person point-of-view (as a proxy for assessing normativity, i.e., whether participants' personal liking judgments align with their expected judgments from others' perspective) and perceived viewing duration (as a not directly aesthetic, artwork-based judgment). Although concepts related to these judgments (for example, normativity) were investigated in philosophical works (Ginsborg 2006), including experimental philosophy (Cova et al. 2012); to the best of the knowledge of the authors, the empirical aesthetics literature rarely goes beyond basic or "canonical" judgment types. The pilot experiment aimed to select a set of 3D models as artworks, the first experiment using VR was designed to measure observers' conscious decisions about a set of artworks as indexed by rating scales, and the second experiment as an online follow-up was a shortened version of the VR experiment. As a supplementary (and more implicit) measure, eye-tracking data were collected whilst participants were engaging with artworks in VR, to inspect visual exploration patterns of observers. Note, the eye-tracking as a supplementary proof-of-concept measure was not incorporated in quantitative analysis, as it was beyond the scope of the main research questions. In both the first and second experiments, exit questionnaires were included to provide additional insights into participants' attitudes towards visual arts, art-related arguments, and the experiments. The experiments' main aim was to investigate the strength of correlations between liking ratings and all other rating types in VR and online settings, and we expected significant correlations between aesthetic ratings in both settings. In this sense, the main alternative hypothesis can be formulated such that there are statistically significant positive linear relationships between liking ratings and other ratings, whereas the null hypothesis as the default state can be formulated that there are no relationships between liking and other ratings.

2 Pilot experiment: selection of 3D artworks

2.1 Methods

2.1.1 Participants

All three authors participated in the pilot experiment.

2.1.2 Stimuli and materials

The stimulus was a set of 2D snapshot images of 3D models from SketchFab (sketchfab.com), an online platform for publishing 3D content. The 3D models were a small subset of the collection, selected with the following criteria: the top hundred, most viewed, downloadable models (according to all-time website usage metrics provided by SketchFab) were listed, in line with the four suitable categories available on the website: architecture, art and abstract, cultural heritage, places, and travel. Since a digital model might belong to more than one category, eliminating duplicate models in this subset resulted in a total selection of 336 models instead of 400. After downloading a batch of 2D snapshot images of these 3D models (i.e., a single snapshot image for each 3D model), all 336 images were cropped proportionally and equated in size to 720 by 400 pixels using Adobe Photoshop (adobe.com). Each snapshot image has the same view that the model uploader selected as a "thumbnail view" of a particular model on the SketchFab website. Stimuli were presented using MatLab (mathworks.com) with Psychtoolbox (psychtoolbox.org), viewed on a screen of personal monitors of varying display resolution (where the resolutions of the two monitors were 1920 by 1080 pixels, and one monitor was 1440 by 900 pixels). Responses were recorded via the keyboard of personal computers.

2.1.3 Design

A single, binary variable of judgment (i.e., "interesting" or "not interesting") was present in the 2-alternative-forced choice (2-AFC) design, for each of the individually displayed artworks.

2.1.4 Procedure

Participants performed a judgment task to decide whether the presented model is "interesting enough to be included in the upcoming VR-based experiment". Each trial consisted of displaying a single 2D snapshot image of a 3D model at the center of the screen, and participants categorised them either as "interesting" or as "not interesting" by pressing rightward or leftward arrows on the keyboard. Each snapshot image, corresponding to a single model, was presented only once. A total of 336 artworks were presented in three blocks, with an estimated time of completion of 30 min. The order of presented images was randomized for each participant. (See Supplementary Fig. S1 for an overview of the general flowchart.)

2.1.5 Data analysis

Since the aim of this experiment was to choose models by unanimous agreement of all researchers, agreement percentages per artwork were calculated, followed by the determination of artworks which all participants unanimously agreed upon.

2.2 Results

Data revealed that 78 models out of 336 were found to be interesting by all participants, which was a 23.21% unanimous agreement, where the mean duration of decision per artwork was 1631 ms. Following on from that, the models were further categorised into two-by-two binary categories, based on "physicality" and "spatiality": physicality was operationalized as whether the 3D model was a recreation of a physical artwork (and labeled as "physical"), or completely created as a digital artwork (and labeled as "digital"). Spatiality was operationalized as whether the 3D model was a small-scale artwork (and labeled as "object"), or a large-scale artwork (and labeled as "space"). As a result, the two-by-two clustering resulted in four labels for four sets: physical object, physical space, digital object, and digital space.

To further clarify the concept of "physicality": Here, a "physical" model refers to the 3D recreation of an "existing", "real" object or space, and is often modeled using a photogrammetric workflow, such as the Nefertiti Bust currently displayed in the Neues Museum in Germany, the marble statue of Aphrodite currently displayed in the British Museum in England, or the archaeological ruins of the St. Olav's Church in Norway. On the other hand, a "digital" model refers to the 3D representation of a "new", "invented", or "designed" object or space, is often modeled from scratch, and does not have a direct reference point in real world, such as a 3D interpretation of a floating trefoil knot, or an interior space based on the video game Mirror's Edge.

Note that this clustering was performed by the lead author, and might be prone to miscategorization: for example, it is debatable whether a 3D model of a set of trees created via photogrammetric reconstruction from a public park is a physical object or physical space. To increase the diversity of types of selected artworks for the next experiment, from 78 models, four models were randomly selected per category, resulting in a total of sixteen artworks, including, for example, a digitally created room model and a photogrammetric model of a sculpture. Whenever required, a selected model might be disregarded due to the lack of feasibility of implementing it into a walkable virtual gallery space, therefore a new random selection was performed (see Fig. 1 for snapshot images of the final set of selected 16 artworks, including one artwork that was animated).

3 Experiment 1: aesthetic judgments in a virtual reality setting

3.1 Methods

3.1.1 Participants

Participants were students or members of staff from Royal Holloway, University of London, and they were recruited using convenience sampling. They were compensated monetarily (£5). A total of 31 participants (17 females, 11 males, $M_{\mathrm{Age}} = 22.74$ years, $SD_{\mathrm{Age}} = 4.83$ years, $R_{\mathrm{Age}} = 18–38$ years) were recruited for the experiment, and all were naïve to the hypotheses of experiments. All participants reported having

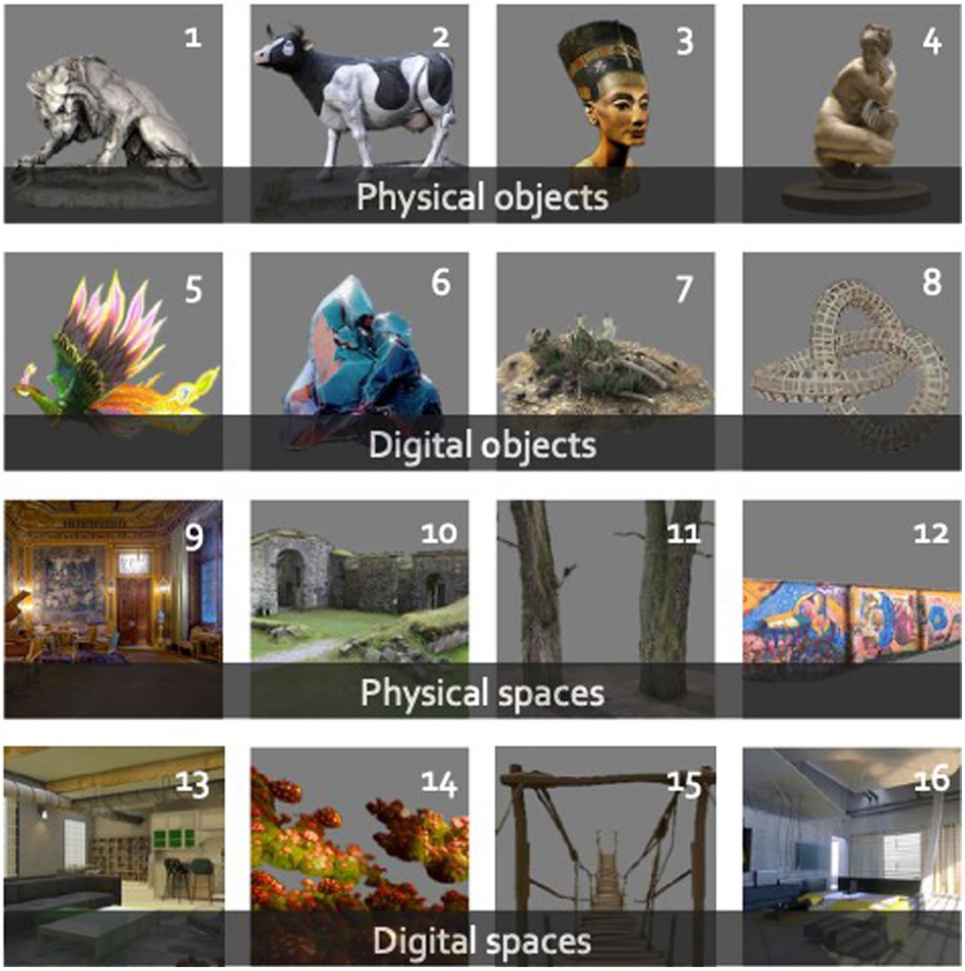


Fig. 1 Set of selected artworks after the Pilot Experiment. These artworks were a subset of 78 3D models which were found to be interesting by participants unanimously. For each of the 78 models, two descriptive tags were attached: whether the model was physical or digital (generated from an existing physical object/space using photogrammetry or not), and whether the model was an object or a space (depending on the model's relative size). Four models were randomly assigned to each category, resulting in a total set of 16 artworks to be used in Experiments 1 and 2. Also see Supplementary Fig. S2 listing attributions of individual 3D models including their titles, uploaders, and hyperlinks

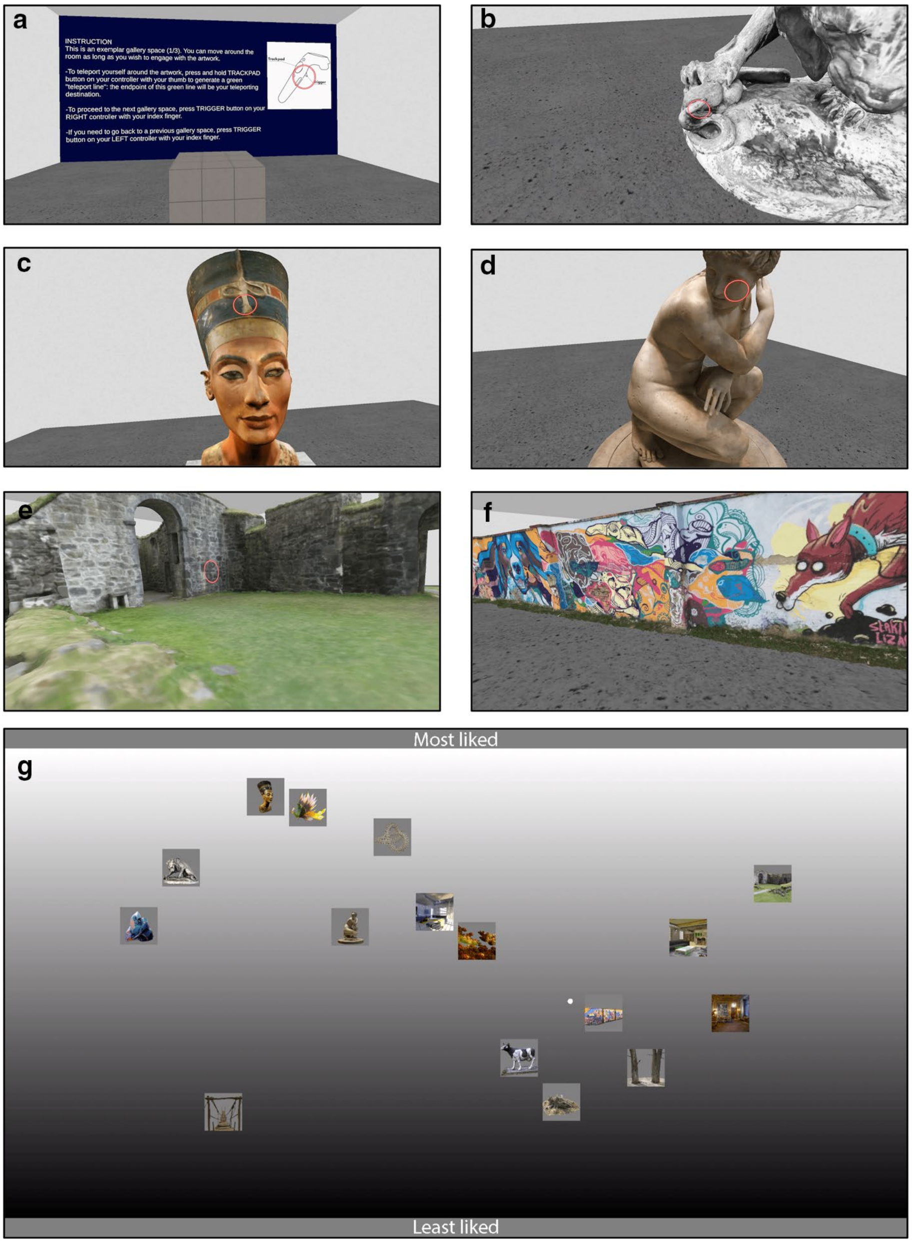
normal or corrected to normal vision. Participants could wear their glasses or contact lenses in the VR headset. All participants provided written informed consent prior to the experiment. All experimental protocols were approved by the Royal Holloway, University of London Research Ethics Committee. All methods were performed in accordance with the ethical guidelines and regulations of the Declaration of Helsinki.

A sensitivity power analysis using G*Power (gpower.hhu.de) for a sample of 31 participants in a one-tailed correlational design (as the main analysis) with a significance level of $\alpha = 0.05$, with an assumed power of 80% as a power level of $1 - \beta = 0.80$ was performed. The assumption about the null hypothesis was that there is no correlation in the population distribution, such that the correlation coefficient $\rho_0(H_0) = 0$. The resulted estimates of $r_{\text{critical}} = 0.30$ and $\rho_1(H_1) = 0.43$ entail the ability to detect positive relations with medium-to-large effect sizes (i.e. $r \geq 0.30$ or $r \geq 0.50$, following the conventional guideline values (Cohen 1992)), in line with our range of interest. Nevertheless, as a general disclaimer, this research had relatively low power to detect true low-to-medium effects, which in turn might minimise the likelihood of reproducibility of results presented.

3.1.2 Stimuli and materials

Stimuli of the first phase (referred to as VR pre-screening) were sixteen artworks in the form of 3D digital models, where participants were expected to engage with these artworks one by one in context, walking around on a 1:1 scale gallery space. All models used in the experiment were available to be used under Creative Commons licenses (refer from snapshot images illustrated in Fig. 1 to Supplementary Fig. S2 listing attributions of individual models including their titles, uploaders, and hyperlinks). Note, one randomly chosen 3D model (artwork #5) was originally animated, and we aimed to preserve this model's intended way of representation (as animated and not as static), and therefore used it as is, despite being a categorically different, potentially more salient stimulus. Models were digitally revised whenever needed for inter-stimulus consistency, using a set of modeling software such as Trimble SketchUp Autodesk Maya, Mudbox, and Blender (sketchup.com, autodesk.com, blender.org). These models were in different file formats and they were initially created and uploaded by different accounts using various software which is often not directly compatible with Unity: Therefore, these minor revisions were performed, such as resizing the 3D model by a factor of 10 or 100 (by scaling down or up the model to match it to "true" size), or correcting the orientation (by rotating the model along X, Y, or Z axis), or fixing the texturing issues (by remapping a texture to preserve transparency or alpha channel). Using Unity game engine (unity.com), models were placed individually in virtual gallery spaces, where all environments had roughly equal illumination levels. A generic Unity scene lighting setup was used for the scene, affecting each gallery space within the scene similarly (see the data repository for the details). The exact luminance level of an individual "frame" during the VR viewing was different from one another, and it depended on where in the virtual room and in what direction the participant was looking. Participants could freely move around by using physical space at the Psychology VR Lab, a walkable area of approximately 280 × 360 cm due to the limits of trackable area for the VR headset. This spatial limit on walking was explained to the participants, and participants were reminded if they got close to the boundaries of the trackable area. Participants could also use the teleport function via hand-held VR controllers to instantaneously shift to a further position (see Fig. 2a for an exemplar 1st person point-of-view of the instructions, and Fig. 2b–f for an exemplar 1st person point-of-view of the artwork viewing). Note that, to overcome any temporal order effect, conditions (as presentation order of gallery spaces) were randomized for each participant. To collect eye-tracking data, a software plug-in called Tobii Pro VR Analytics (tobiipro.com) was also implemented in the environment. Stimuli of the second phase (referred to as the judgment task) were static snapshot images of these digital models. All images of artworks were placed individually on a mid-grey, square background of 80 × 80 pixels ($\approx 1.5 \times 1.5°$ of visual angle), scalable to 240 × 240 pixels ($\approx 4.5 \times 4.5°$ of visual angle) by right mouse click (see Fig. 2g for a view of the judgment task, performed on a regular 2D monitor). No visual stimulus was present in the exit questionnaire, which was designed to collect demographic data, five-point Likert-scale rating questions, and some open-ended questions as feedback. Note that, one set of questionnaire items was particularly aimed to ascertain the task difficulty in making each judgment, phrased to participants as "how challenging was each of your judgments during the experiment", and we refer to these items as the level of difficulty.

Stimuli were displayed using an HTC Vive VR Headset (vive.com) with an embedded eye tracker. The headset had a display resolution of 1080 X 1200 pixels per eye at 90 Hz, with a gaze data output frequency of 120 Hz, weighing approximately 500 g. The field of view (FoV) was around 100° in both horizontal and vertical directions, where the actual FoV varies by the distance between the lenses which was adjusted for each participant based on their interpupillary distance. The headset was wirelessly connected to a PC (Lenovo ThinkStation, with Xeon E5-1630 @ 3.70 GHz CPU, Nvidia Quadro M5000 GPU, 40 GB of RAM, running on Windows 10 Pro). HTC Vive controllers were also used to teleport within the environment, and to proceed between gallery spaces. Software used to present VR stimulus and record gaze data was an executable file built using Unity.



INSTRUCTION
This is an exemplar gallery space (1/3). You can move around the room as long as you wish to engage with the artwork.

-To teleport yourself around the artwork, press and hold TRACKPAD button on your controller with your thumb to generate a green "teleport line": the endpoint of this green line will be your teleporting destination.

-To proceed to the next gallery space, press TRIGGER button on your RIGHT controller with your index finger.

-If you need to go back to a previous gallery space, press TRIGGER button on your LEFT controller with your index finger.

Fig. 2 **a**–**f** Exemplar virtual gallery space from the 1st person point-of-view, snapshots were taken from instructional gallery space (**a**), from artwork #1 (**b**), #3 (**c**), #4 (**d**), #10 (**e**), and #12 (**f**) during the VR pre-screening phase of Experiment 1, where participants engaged with individual artworks one by one in a virtual white-box gallery space. Here, the red circles on the artworks indicate fixations to illustrate the heatmaps' input as 3D fixation coordinates, and these fixational circles were not visible to the participants during the experiment. **g**. Exemplar view from the judgment task phase of Experiment 1. Here, participants were able to see and enlarge all the artworks simultaneously and they provided ratings on a "most *p* / least *p*" scale (such as most liked vs. least liked) by dragging and dropping the snapshot images of the artworks. Also see Supplementary Fig. S2 listing attributions of individual 3D models including their titles, uploaders, and hyperlinks

The judgment task was coded using MatLab with Psychtoolbox and was displayed via a separate PC (Dell Alienware 17, with Intel i7-8759H @ 2.20 GHz CPU, Nvidia GeForce RTX 2060 GPU, 16 GB of RAM, running on Windows 10 Home, with a display resolution of 1920 by 1080 pixels at 60 Hz refresh rate). The exit questionnaire was a simple form created using Google Forms (google.com/forms) and displayed again using the same PC.

3.1.3 Design

Designed mainly as a correlational experiment, nine variables from the task were the ratings provided by the participants for each artwork: liking, liking from the third-person point of view (POV), emotional valence, meaningfulness, novelty, artfulness, complexity, colourfulness, and perceived viewing duration. Liking from third-person POV was explicitly described to the participant such that this judgment should be based on the expectancy of how other participants would rate the images. During the task, these ratings as binary labels indicated minimum and maximum ends of the rating scale, as in "most-*p*/least-*p*", where *p* denotes a given variable of a judgment type (see Supplementary Fig. S3 for an overview of all judgment questions, and their labels as presented to the participants). Viewing duration was the tenth variable. The eye-tracking data was only used for visual inspection of the exploration patterns of observers. The main hypothesis was solely based on correlations between participant ratings: it was expected that participants' assigned liking ratings are positively correlated with other ratings towards artworks. Additionally, using a linear regression model, liking rating was defined as an outcome variable, and all other ratings as predictor variables, to check whether participants' ratings on these dimensions can predict liking rating. Lastly, mean liking ratings of individual artworks per category were compared using a 2 × 2 RM-ANOVA in line with 2 × 2 stimuli categories (namely, physical vs. digital arts, and objects vs. spaces), and post hoc comparisons using t-tests were Bonferroni corrected. The first factor was defined as physicality (namely, whether the displayed artwork was a representation of a physical artwork, or completely created as a digital artwork), where two levels were physical and digital artworks. The second factor was defined as spatiality (whether the displayed artwork was small-scale and defined as an "artwork-object" or is large-scale and defined as an "artwork-space"), where two levels were object and space. The reasoning was to check whether category-specificity affected liking ratings. Here, the dependent variable (DV) was the liking rating in percentage, and the two independent variables (IVs) were two artwork categories (labeled as physicality and spatiality). Note, the proof-of-concept eye-tracking was expected to provide supplementary data, but a detailed, quantitative analysis was beyond the scope of the research aims.

3.1.4 Procedure

Following the briefing and receiving consent from participants, the experiment consisted of three phases. In the first phase, referred to as the VR pre-screening, information on hardware, software, and the user interface of the VR headset were provided to the participant. Participants were in standing position in the Psychology VR Lab. After putting the head-mounted display (HMD) on, an eye-tracking calibration was executed to ensure the reliability of gaze data to be collected using the 5-point default in-built calibration. Following the calibration, participants firstly visited three gallery spaces. Using VR controllers, participants had a chance to practice traveling between these gallery spaces by the trigger button press on the VR controllers, as well as teleporting themselves around the artworks by pressing the trackpad-button. Participants then engaged with sixteen artworks one by one in randomized order, without any time constraints. Participants could revisit a specific artwork if they wished to do so, but none of the participants revisited a previously seen artwork again during the experiment. 3D objects and spaces were presented in an otherwise empty digital room, similar to a white-box gallery space. During this period, gaze data were recorded (see data repository for an exemplar movie recording of the VR gallery experience).

In the second phase, referred to as the judgment task, participants sat in front of a computer screen, approximately 57 cm away. For each question, participants could see and judge all artworks at once. Participants thus could drag and drop thumbnail images of artworks on screen, in relation to a "most-*p*/least-*p*" scale for the given property *p* such as most liked and least liked, visualized by a background gradient from black to white (see Fig. 2g for exemplar view from the judgment task). This way of sorting the stimuli aimed to allow for more precise and relative judgments from participants: Compared to sequential number-based ratings, spatially aligning the artwork images based on a rating scale (e.g., from most liked to least liked) while seeing all the stimulus set in one screen all the time during the judgment enables the participants to have an additional visual relation between artworks' ratings. Similarly, the visual grouping during the judgment task offers extra insight, for example, about the portion of the artworks rated as highly liked. Responses were recorded using a keyboard and mouse. Lastly, a brief exit questionnaire was also presented on-screen. During the VR phase and in-between sessions, participants were reminded that they may pause or stop the experiment whenever they feel discomfort or motion sickness (although none of the participants reported such issues, paused or ended the experiment). The experiment was conducted without a time limitation, but on average the time of completion was around 30 min. (See Supplementary Fig. S1 for an overview of the general flowchart, and see the data repository for an exemplar movie recording from the first-person point-of-view during the VR viewing phase.)

3.1.5 Data analysis

Three main data streams were formed: data from the VR eye-tracking, the rating judgments, and the exit questionnaire. Software used for the data analysis and visualisation were R and jamovi (r-project.org, jamovi.org). Apart from descriptive statistics, the eye-tracking data were only visualized in terms of fixation duration (referred to as heatmaps, associated with the amount of visual attention). In terms of the rating tasks, as a participant dragged and dropped images relative to background gradient, the image coordinates in pixels corresponded to the rating scores from 0 (on the lowest end of the scale, referring to the "least-*p*") to 100 (on the highest end of the scale, referring to the "most-*p*"). On screen, the difference between the lowest and highest possible y-coordinate of the center-point of the snapshot image that the participant can place was 920 pixels, and this pixel distance corresponded to the rating scale ranging from 0 to 100. The real viewing duration for individual artworks was measured by calculating the difference in the timestamps of entering and exiting (i.e., "teleporting in" and "teleporting out" using the trigger button on the VR controller) each individual gallery space in which only a single 3D model was exhibited. Linear correlations between liking ratings and all other ratings were indexed using the Pearson correlation coefficient since the rating scores were treated as continuous variables. Note, although the written and verbal instructions were aimed to indicate that the task was to provide judgment ratings on a continuous scale, due to the possibility that some participants might have treated it as an ordering task, a supplementary correlational analysis using Spearman correlation coefficient was conducted, treating the rating data as ordinal instead of interval. Additionally, mean liking ratings of individual artworks per category were compared using a 2 × 2 RM-ANOVA. A complementary set of analyses for the exit questionnaire included graphs indicating frequency distribution of rating responses on Likert-scale. Main reported descriptive values were mean (M) and standard error of the mean (± SEM) for any given analysis, unless stated otherwise.

3.2 Results

Participants spent more than half a minute viewing each artwork on average ($M_{\text{Duration}} = 44.35 \pm 4.12$ s). A 2 × 2 repeated-measures analysis of variance (RM-ANOVA) to check whether spatiality and physicality (as artwork categories) affected viewing duration yielded a significant difference in mean viewing duration for spatiality ($F_{(1,30)} = 29.574$, $p < 0.001$, ${\eta_p}^2 = 0.496$), but not for physicality ($F_{(1,30)} = 0.485$, $p = 0.492$, ${\eta_p}^2 = 0.016$) or interaction ($F_{(1,30)} = 0.617$, $p = 0.438$, ${\eta_p}^2 = 0.020$). The post hoc comparison ($t_{(20)} = 5.437$, $p < 0.001$) showed that participant spent more time engaging with artwork-spaces ($M_{\text{Space}} = 55.50$ s ± 6.24 s) compared to artwork-objects ($M_{\text{Object}} = 33.19 \pm 2.95$ s), and illustrated in Fig. 3a and in Supplementary Fig. S7*a*.

To measure the relation between liking rating and other ratings, a set of correlational analyses were calculated using Pearson correlation. For the analysis, continuous rating scores from thirty-one participants on each individual artwork out of sixteen were treated as a single data point, resulting in $N = 496$. Results showed that liking positively and significantly correlated with all measures, namely with liking from the third-person POV ($r_p = 0.538$), emotional valence ($r_p = 0.571$), meaningfulness ($r_p = 0.381$), novelty ($r_p = 0.360$), artfulness ($r_p = 0.441$), complexity ($r_p = 0.502$), colourfulness ($r_p = 0.443$), perceived viewing duration ($r_p = 0.680$), and real viewing duration ($r_p = 0.183$), where all $p < 0.001$. Perceived viewing duration also positively correlated with real viewing duration ($r_p = 0.214$, $p < 0.001$). Cross-correlation between all ratings showed mostly positive and moderate significant correlations as indexed by $0.3 < r_p < 0.7$, but some weak correlations as indexed by $0 < |r_p| < 0.3$ were present, especially with real viewing duration (see Fig. 4 for the overall cross-correlation matrix, Supplementary Fig. S4 for individual correlation plots, and Supplementary Fig. S5 for rating scores illustrated as boxplots and density curves drawn from the individual data points). Note, when the rating judgments were treated as ordinal data instead of interval data and the relations were calculated using Spearman correlation instead, highly comparable results were obtained in terms of the overall correlations (see

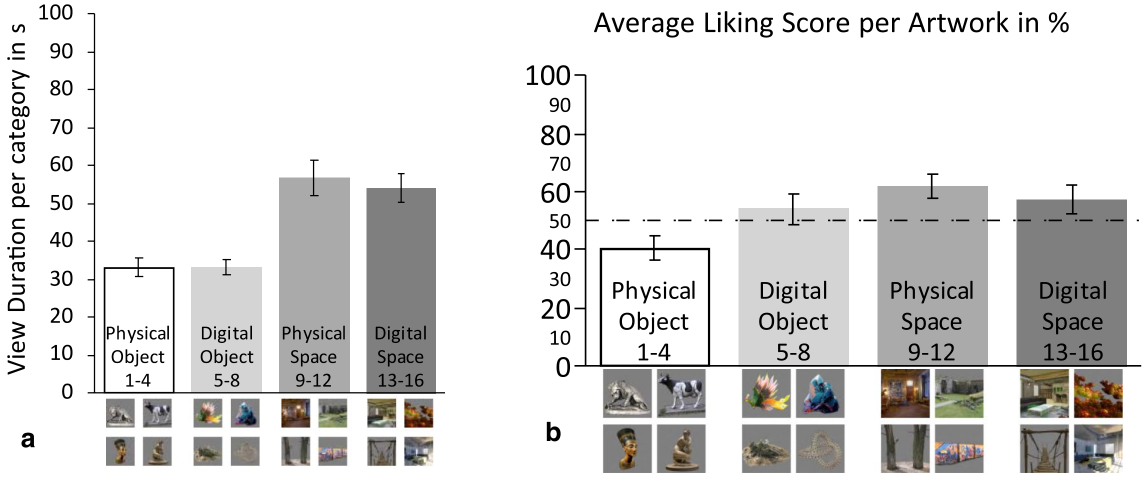


Fig. 3 Results of Experiment 1. **a** Average view duration per artwork category in seconds and **b** average liking score binned into artwork categories were illustrated as bar graphs, indicating means with error bars showing ± 1 standard error of the mean (SEM). In the second panel, a dashed line parallel to the *x*-axis denotes the mid-value of liking ratings. If the mid-value can be defined as a threshold point of a "neither liked nor disliked" artwork, then an artwork with a mean rating score (including variance) above the threshold can be defined as an overall liked artwork, and vice versa. The sample size was $N_{\text{Participant}} = 31$, and the number of observations for each category (out of four) was $N_{\text{Observation}} = 4$

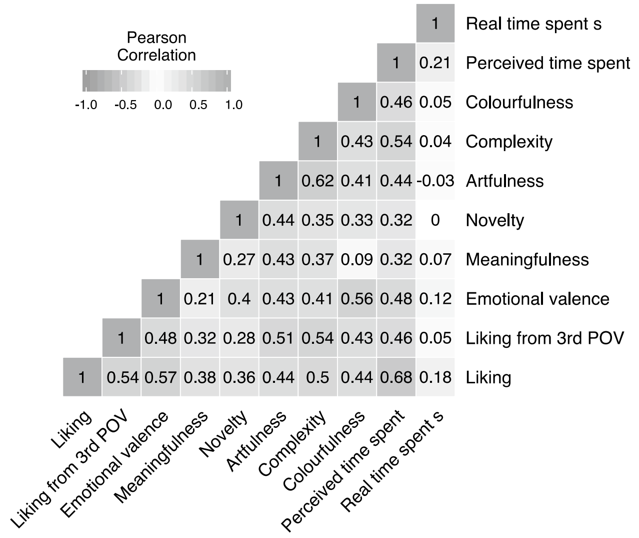


Fig. 4 Correlation matrix for Experiment 1 indicating positive relation as indexed by Pearson correlation, between all nine ratings that participants provided and one measured variable as real view time. Most of the individual responses showed positive and moderate cross-correlation to one another as indexed by $0.3 < r_p < 0.7$, except for real viewing time, which only showed either weak or no correlation to other variables as indexed by $0 < |r_p| < 0.3$. The sample size was $N_{\text{Participant}} = 31$, and the number of observations was $N_{\text{Observation}} = 16$. Note, for the correlation matrix using Spearman correlation, see Supplementary Fig S6

Supplementary Fig. S6 for the correlational matrix using Spearman correlation).

A multiple linear regression model was built to test whether liking can be predicted by other measured variables, following the previous findings on significant correlations. The analysis consisting of the liking rating as the dependent variable and all nine other metrics as the covariates resulted in a significant model: $F_{(9, 486)} = 77.11$, $p < 0.001$. The overall model explained 58.1% of the variance in liking ratings as indexed by adjusted R^2. Testing significance of individual predictors yielded four significant predictors as liking from the third-person POV ($\beta = 0.169$, $t = 4.467$, $p < 0.001$), positive emotional valence ($\beta = 0.237$, $t = 5.732$, $p < 0.001$), meaningfulness ($\beta = 0.123$, $t = 3.757$, $p < 0.001$), perceived viewing duration ($\beta = 0.389$, $t = 10.494$, $p < 0.001$); and five non-significant predictors as novelty ($\beta = 0.047$, $t = 1.504$, $p = 0.133$), artfulness ($\beta =$ -0.028, $t =$ -0.726, $p = 0.468$), complexity ($\beta = 0.034$, $t = 0.912$, $p = 0.362$), colourfulness ($\beta = 0.016$, $t = 0.508$, $p = 0.612$), real viewing duration ($\beta = 0.033$, $t = 1.543$, $p = 0.123$). To form the predictive model, all five non-significant predictors were removed from the model generation, and new constants were calculated based only on the significant predictors, therefore the parameter estimates became: $\hat{y} = 2.530 + 0.174x_1 + 0.261x_2 + 0.127x_3 + 0.416x_4$, where $\hat{y} =$ liking, $x_1 =$ liking from the third-person POV, $x_2 =$ positive emotional valence, $x_3 =$ meaningfulness, $x_4 =$ perceived viewing duration.

To check whether category-specificity affected liking ratings using a 2×2 RM-ANOVA, four ratings from a single participant corresponding to four artworks from the same category level was averaged (as a common practice to average data over participants,), thus resulting in $N_{\text{Participant}} = N_{\text{Observation}} = 31$, and $df = 30$. Physicality did not alter observer's liking ratings ($F_{(1, 30)} = 2.95$, $p = 0.096$, $\eta_p 2 = 0.090$), but spatiality significantly altered the liking ratings ($F_{(1, 30)} = 20.31$, $p < 0.001$, $\eta_p 2 = 0.404$), and observers liked spaces more compared to objects ($M_{\text{Space}} = 59.54 \pm 1.61\%$; $M_{\text{Object}} = 47.32 \pm 1.75\%$). A significant interaction effect was also present ($F_{(1, 30)} = 16.61$, $p < 0.001$, $\eta_p^2 = 0.356$), implying that liking ratings of two levels of physicality differed across the two levels of spatiality. The difference between Object$_{\text{Physical}}$ ($M = 40.70 \pm 2.11\%$) and Space$_{\text{Physical}}$ ($M = 61.57 \pm 2.09\%$) was pronounced in the post-hoc comparisons ($t_{(56.8)} =$ -6.06, $p < 0.001$), whereas no significant difference between Object$_{\text{Digital}}$ ($M = 53.98 \pm 2.67\%$) and Space$_{\text{Digital}}$ ($M = 57.52 \pm 2.43\%$) was observed ($t_{(56.8)} = -1.03$, $p = 0.999$). See Fig. 3b for average liking scores per artwork category, and Supplementary Fig. S7*b* for average liking scores per individual artwork.

Ratings from the exit questionnaire on the five-point Likert were used to calculate the most frequent responses and reported as a percentage for the highest frequency choice in brackets alongside the given questionnaire items. In terms of the amount of difficulty for each rating judgment (Supplementary Fig. S8*a*) based on the most frequent responses, overall, participants found that (i) Liking (45%), complexity (48%), and colourfulness (74%) were not at all difficult; (ii) Positive emotional valence (39%), meaningfulness (45%), novelty (32%), and perceived viewing duration (48%) were slightly difficult; (iii) Artfulness (35%) and liking from the third-person point of view (39%) was moderately difficult. In terms of general attitudes (Supplementary Fig. S8*b*), overall, participants (i) Strongly agreed that viewing experience in VR was enjoyable (81%); (ii) Strongly disagreed that the experiment was boring (48%); (iii) Disagreed that judgment tasks were challenging (32%); (iv) Strongly disagreed that video games are not art (45%); (v) Agreed that anything can be art (39%); (vi) Disagreed that all public art objects/spaces should be digitized and available online (29%); (vii) Disagreed that art museums and galleries are losing their significance (52%); (viii) Disagreed that aesthetic experience cannot be investigated empirically (35%). Note that, although part of these results indicated an overall positive response to the VR experience and the experiment, this finding might be prone to novelty effects of VR or experimenter bias. In terms of items related to the frequency of exposure to arts (Supplementary Fig. S8*c*), overall, participants (i) Sometimes visit art museum, art galleries, or art events (45%);

(ii) Never view art digitally (32%); (iii) Seldom pursue an artistic activity or a hobby, such as painting or participating workshops (48%); (iv) Never (26%) or sometimes (26%) play video games; and (v) Never use VR (55%). Open-ended feedback yielded an overall liking of the experiment and particularly diversity of artworks as commented by multiple participants. Various keywords from the feedback were extremely interesting, very enjoyable, interactive, amazed by the level of immersion, etc. Minor drawbacks such as rare connection issues and a potentially better user interface for the on-screen interactive questionnaire were also noted by some participants. Lastly, individual fixation duration plots as heatmaps showed viewing strategy differences among participants for each artwork. To briefly demonstrate the individual differences, exemplar heatmaps as a visualisation of fixation duration for two artworks from two randomly selected participants were plotted (Fig. 5), and individual heatmaps were uploaded to the data repository.

In summary, the main findings of Experiment 1 indicated that (i) Participants spent more time viewing spatial artworks compared to objects, (ii) Liking rating showed a linear and positive relation with all other judgment types, (iii) A linear model to predict liking ratings can be based on four judgments (namely, liking from the third-person POV, emotional valence, meaningfulness, perceived viewing duration), (iv) Although a relatively high variance was present for liking ratings per artwork subsets, an interaction between physicality and spatiality was observed, where the liking rating difference between spatiality depends on whether the artwork is physical or digital (such that spatial artworks were preferred more compared to artwork objects if they were physical artworks, but not if they were digital artworks), (v) The level of difficulty was reasonably low for judgment types, and (vi) Participants showed diverse opinions about presented arguments related to arts.

4 Experiment 2: aesthetic judgments in an online setting

4.1 Method

4.1.1 Participants

120 people (60 females, 60 males, $M_{\mathrm{Age}} = 33.28$ years, $SD_{\mathrm{Age}} = 11.51$ years, $R_{\mathrm{Age}} = 18–65$ years) were recruited for the online experiment via Prolific (prolific.co), an online participant recruitment tool. All participants were naïve to

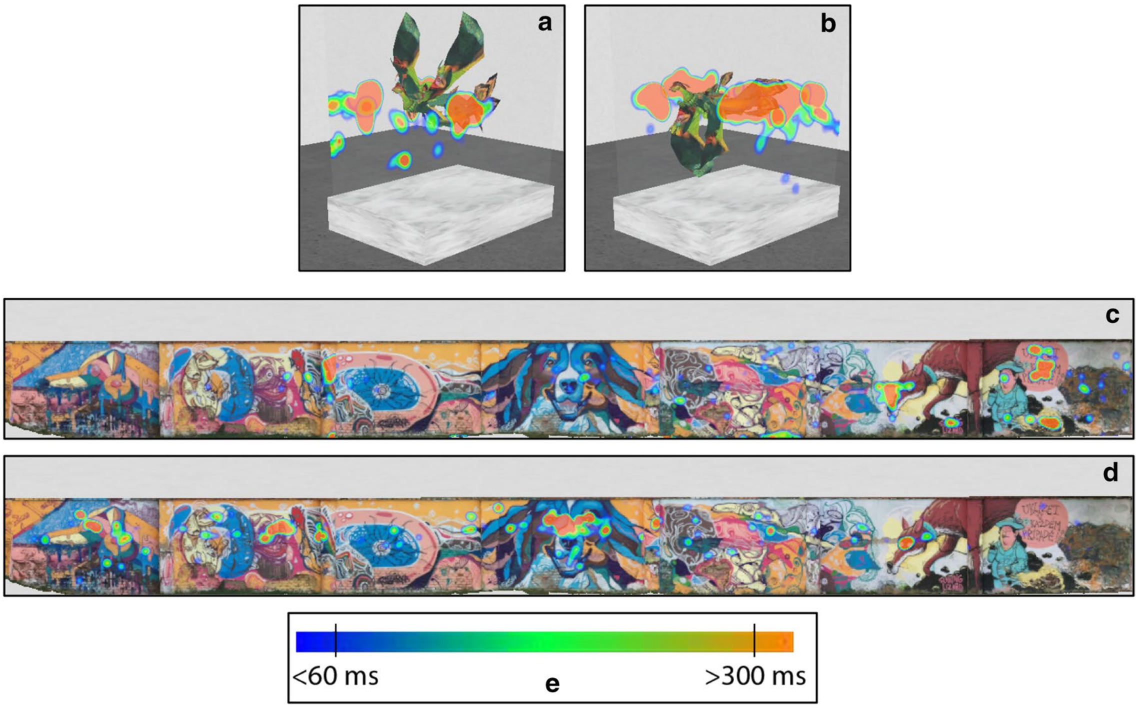


Fig. 5 Exemplar heatmaps from Experiment 1. **a**–**b** and **c**–**d** The visualisation of fixation duration for two artworks (as artwork #5 and #12) from two randomly selected participants with **e** a fixation duration scale ranging between 60 and 300 ms was generated to demonstrate various viewing strategies employed by participants. The heatmaps were created for participant's entire viewing time (here, 55 s and 105 s for a-b; 70 s and 71 s for **c**–**d**). Note that, artwork #5 was a moving digital artwork (with an animated wing motion) and displayed inside a glass box in VR, therefore the corresponding heatmaps were on the surface of this cuboid box, and not on top of the artwork surface. (See data repository for further heatmaps.)

the hypotheses of experiments and they were compensated monetarily (equivalent to £5/hour for a 5-to10-minute-long experiment). All were from the UK, as selected via Prolific pre-screening. All participants provided written informed consent prior to the experiment. All experimental protocols were approved by the Royal Holloway, University of London Research Ethics Committee. All methods were performed in accordance with the ethical guidelines and regulations of the Declaration of Helsinki.

Note, we only aimed to test whether the correlations between judgment ratings in this simplified online experiment were qualitatively similar to the results in the previous, more nuanced lab-based experiment. Therefore, here we (i) Reduced the quality of the presentation from 3D models in VR to 2D images on a monitor, (ii) Reduced the continuous rating scale to the Likert scale, and (iii) Reduced the number of artworks presented (whilst increasing the sample size to equate the number of observations for the correlational analysis). In this sense, this online experiment cannot be described as a direct, rigorous comparison to the VR experiment, but rather a derivative, exploratory, follow-up study.

The sample size for the online study is based on the assumption to equate the total number of observations between the VR and the online experiments (i.e., N_{Stimulus} X $N_{\text{Participant}}$): previously, 16 artworks were presented to 31 participants in the VR setting, here, for 4 artworks were presented to 120 participants in the online setting.

4.1.2 Stimuli and materials

One artwork from each category, aimed to represent each of the 2 × 2 categories, was selected from the sixteen artworks of the lab-based experiment, resulting in four artworks (number 1, 8, 12, and 14 from the lab experiment, see Fig. 1). Due to the nature of the online experiment, artworks were only displayed as 2D static images. A single snapshot image was displayed for each individual 3D model. Stimuli and the questionnaire were created on a simple online form using Google Forms and viewed on a screen of personal monitors. Responses were recorded via the keyboard of personal computers.

4.1.3 Design

In a correlational experimental design, as before, eight variables were ratings on aesthetic judgments: liking, liking from the third-person point of view, positive emotional valence, meaningfulness, novelty, artfulness, complexity, and colourfulness per artwork. Viewing duration could not be measured as the rating scales were simultaneously presented with the snapshot images of the artworks, to minimize the online experimental duration, thus the viewing time and decision time could not be separated.

4.1.4 Procedure

Following a written briefing and receiving a consent form from the participants online, the experimental workflow consisted of two brief phases: In the first phase, each artwork was presented as a 2D snapshot image of the 3D model, and participants viewed the snapshots without any time constraints. Rating questions per artwork were displayed beneath the images, and participants were asked to rate eight aesthetic judgments for each of the four artworks (see Supplementary Fig. S10 for a diagrammatic view of the judgment task). For the sake of simplicity, the ratings were on 5-point Likert scales, instead of a continuous interactive interface presented in Experiment 1. The second phase was a brief exit questionnaire, where participants were asked to rate the level of difficulty for each type of judgment, again on a 5-point Likert scale, and the same as in Experiment 1. This phase also contained two open-ended questions asking for (i) Any other terms, including adjectives or metaphors which might be useful in describing and judging artworks; and (ii) Any liked or disliked aspects of the experiment. The experiment was conducted without a time limitation, but on average the time of completion was expected to be approximately five minutes, based on previous piloting. (See Supplementary Fig. S1 for an overview of the general flowchart.)

4.1.5 Data analysis

Apart from descriptive statistics, for each judgment task, linear correlations between liking ratings and all other ratings were indexed using the Spearman correlation coefficient (as ratings were discrete variables). Software used for the data analysis was R and jamovi.

4.2 Results

The overall duration of the experiment was around five minutes with large variance ($M_{\text{Duration}} = 268.32$ s, $SD_{\text{Duration}} = 137.82$ s), where duration minima and maxima were approximately two and seventeen minutes.

For the correlation analysis, ratings of each individual artwork were treated as a single data point, resulting in $N = 480$ (see Fig. 6 for the overall cross-correlation matrix, and Supplementary Fig. S11 for individual correlation plots). Similar to the VR based experiment, results showed that liking positively and significantly correlated with all seven measures, namely with liking from the third-person POV ($r_s = 0.599$), positive emotional valence ($r_s = 0.442$), meaningfulness ($r_s = 0.503$), novelty ($r_s = 0.410$), artfulness ($r_s = 0.647$), complexity ($r_s = 0.382$), and colourfulness ($r_s = 0.262$), where all $p < 0.001$. Note that, although the correlations were significant both in Experiment 1 and

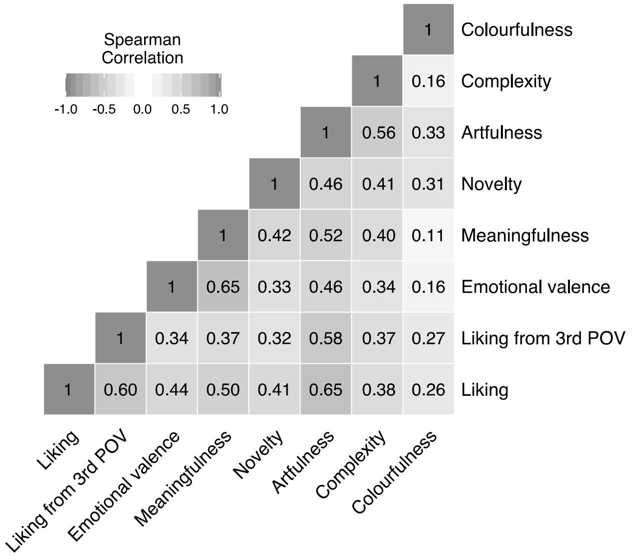


Fig. 6 Correlation matrix for Experiment 2 indicating positive relation as indexed by Spearman correlation, between all eight ratings that participants provided. Most of the individual responses showed positive and moderate cross-correlation to one another as indexed by $0.3 < r_s < 0.7$, with some exception of weak or no correlation as indexed by $0 < r_s < 0.3$. The sample size was $N_{\text{Participant}} = 120$, and the number of observations was $N_{\text{Observation}} = 4$

here in Experiment 2, the relation seemed to be relatively less pronounced here mainly due to higher variance present in the ordinal data (compare Supplementary Fig. S4 of Experiment 1 and Supplementary Fig. S11 of Experiment 2). Overall, the correlational patterns were similar (also see Supplementary Fig. S12 for a similarity assessment of the correlations between VR and online settings).

Building a prediction model was slightly different compared to the first experiment in VR since here the ratings were on a 5-point Likert scale, the ratings cannot be treated as continuous variables for a linear regression. Instead, following the findings on significant correlations, an ordinal logistic regression can be used. A model can be built to test the probability of a liking rating occurring given the known values of the other ratings. The analysis consisting of the liking rating as the dependent variable and all seven other metrics as covariates resulted in a significant model: $\chi^2_{(7)} = 366.006$, $p < 0.001$, $R^2_{\text{McF}} = 0.245$, where R^2_{McF} refers to McFadden R^2 and is not analogous to the R^2 in multiple linear regression. Testing significance of individual predictors yielded four significant predictors: liking from the third-person POV ($\beta = 0.122$, $z = 7.244$, $p < 0.001$), meaningfulness ($\beta = 0.121$, $z = 3.318$, $p < 0.001$), novelty ($\beta = 0.095$, $z = 2.300$, $p = 0.021$), artfulness ($\beta = 0.844$, $z = 6.724$, $p < 0.001$); and three non-significant predictors as positive emotional valence ($\beta = 0.084$, $z = 0.761$, $p = 0.447$), complexity ($\beta = -0.075$, $z = -0.845$, $p = 0.398$), colourfulness ($\beta = 0.060$, $z = 0.996$, $p = 0.319$). Note that, the two significant predictors as liking from the third-person POV and meaningfulness were the same as the previous model from the VR-based experiment, but here, the two other predictors were novelty and artfulness instead of emotional valence and perceived viewing duration (which was not a measured rating here).

In terms of the level of difficulty in making the judgements, based on the most frequently provided responses (where the percentage of responses reported in brackets), participants found (i) Colourfulness (62%), liking (49%), positive emotional valence (32%), and artfulness (30%) were not at all difficult; (ii) Complexity was slight to not-at-all difficult (30%:30%), (iii) Novelty was slightly difficult (34%), and (iv) Liking from the third-person point of view (32%) and meaningfulness were moderately difficult (35%), (see Supplementary Fig. S13). Open-ended feedback yielded an overall liking of the experiment and the simplicity of the design. Various keywords from feedback were enjoyable, fun, easy, etc. A minor suggestion was a potential addition of other viewing angles per artwork, and another participant suggested having more "traditional art" in the stimulus set. When asked to provide other terms for describing and judging the artworks, an extensive list of suggestions was produced, some of which were: provocative, inspirational, absorbing, deep, soothing, nostalgic, strange, etc. Some particular suggestions were disturbing, sad, deceptive, confusing, and dark, which may be inspirational for a relatively understudied research direction on aesthetic judgments associated with negative connotations (see Supplementary Fig. S9 for all suggested keywords visualised as a word cloud).

In summary, the main findings of Experiment 2 indicated that (i) Liking rating again showed a linear and positive relation with all other judgment types, similar to the lab-based VR experiment, (ii) A logistic model to predict liking ratings can be based on four judgments (namely, liking from the third-person, meaningfulness, novelty, artfulness), (iii) The level of difficulty was again reasonably low for judgment types, and (iv) Participants were able to suggest a diverse set of keywords which might be useful in aesthetic judgments.

5 Discussion

We began this work by demonstrating in the pilot experiment that adapting a 2AFC task on the judgment of interest could provide an alternative way of generating a stimulus set, whilst aiming to minimize stimulus selection bias. The first lab-based experiment in VR showed that the liking ratings significantly and moderately correlated with various other judgments with positive denotations such as positive emotional valence, meaningfulness, or novelty. A multiple linear regression model suggested that liking ratings can be predicted by some of those judgments (in this case, by liking from the third-person POV, emotional valence, meaningfulness, and perceived viewing duration ratings). Although the experimental design did not include an in-depth analysis of eye-tracking data, fixation duration visualised as 3D heatmaps, showed diverse viewing strategies of the artworks. Supplementary questionnaire results provided insights into the diverse opinions of participants towards arts, and an overall positive attitude towards the experiment itself, such that for example, the level of difficulty for the judgments was reasonable for all tasks, which implies the feasibility of the presented experimental method for future studies. The second experiment, in the form of an online follow-up, resulted in similar and comparable correlational trends between ratings. Although the participants' types of perceived immersion (such as spatial, cognitive, or emotional) and potential mediators for perceived immersion (such as engagement and realism) were not measured, the results suggested that reducing the intended spatial immersiveness of the artwork presentation medium (from the VR environment to the 2D image) did not radically change the relation between measured aesthetic ratings. An ordinal logistic regression model suggested that liking ratings can be predicted again by some judgments (in this case, by liking from the third-person POV, meaningfulness, novelty, artfulness ratings); a partially overlapping finding compared to the first VR-based experiment. Similar to the lab-based experiment, the level of difficulty for the judgments was again reasonable for all tasks in the online experiment. Additionally, participants were able to provide a diverse set of terms including adjectives or metaphors to be used in describing and judging artworks.

In terms of using artwork as stimuli, our approach had a relatively unconventional methodology compared to psychophysical tradition, where generally well-controlled or categorizable stimuli and high internal validity were often sought: for example, generating random dot textures with varying visual complexity levels (Friedenberg and Liby 2016) or comparing representational and abstract artworks (Schepman et al. 2015). Since the physical properties activating the senses can be generally described as an initial state of the aesthetic experience, these properties are often linked to the experience and the judgment of art. Following on from that, one common rationale behind well-controlled stimuli sometimes relies on the assumption that some intrinsic (physical) properties of artworks are the main (or only) factors that shape both the experience and the judgment of art. However, similar bold assumptions were recently started to be criticized, and for example, a conceptual dissociation between the evaluation of artworks (as a specific research case) and the aesthetic experience (as a more broad research area) is proposed (Skov and Nadal 2020). In this study, the stimuli selection process resulted in the inclusion of, for example, the bust of Nefertiti, a photogrammetric model of a graffiti wall, and an abstract digital sculpture resembling a trefoil knot. These items do not share many similarity or have an obvious common property. This intentional divergence across properties makes it even more interesting that we found relationships between aesthetic judgments themselves, irrespective of artwork properties or categories. Put differently, instead of parametrically modifying physical properties to see how judgements depend on these changes, we relied on the existing or "natural" variation of such properties to see how the different judgements are linked together. More broadly, investigating higher-level associations of aesthetic judgements separate from physical properties of visual arts can be indeed a meaningful empirical research context.

Following viewing artworks in VR or online, participants were able to provide elaborative judgments on rating scales for all given judgment items using a rating task, which aimed to eliminate serial dependence bias present in aesthetic judgments (Kim et al. 2019). In both experiments, participants were able to provide their judgments without much effort, as indexed by their responses indicating a low level of difficulty on the questionnaire. We compared perceptual judgements across a wide range of artworks, relying on the correlational analysis between judgment ratings, without directly incorporating "physical" properties of artworks or contextual effects in the experimental design. We did not aim to investigate, for example, whether a particular artwork set is liked more if presented in VR, or whether viewing 3D models in VR is a type of "genuine" art experience. Nevertheless, it is important to consider that previous research has raised questions about the contextual factors affecting aesthetic judgments, for example, authenticity and presentation context of artworks (Brieber et al. 2015) or originality (Locher 2016). Similarly, laboratory participants may not necessarily classify a stimulus as art in generic laboratory setups (Pelowski et al. 2017). Overall, these studies speak to the discrepancy between researchers' assumption about the art experience in labs and participants' attitudes towards it and underline the importance of conceptual soundness of experimental designs.

Regarding the conceptual link between physical properties of artworks and aesthetic judgments, one exception in line with this link in our analysis for the VR experiment was following the 2 × 2 categorization of presented artworks as spatiality (whether the artwork is large-scale or small-scale) and physicality (whether the artwork is a modelled version of a real-world object/space or a digital-only): the observed liking rating difference between spatiality depends on whether the artwork is physical or digital, and the difference between artwork-objects and artwork-spaces was pronounced only for physical artworks but not for digital artworks, but arguably, participants might not easily distinguish between physical and digital artworks, which were not introduced to participants. More specifically, irrespective of those categories formed by the researchers, in the VR experiment, all presented artworks might just be labelled as virtual by participants, whereas in the online experiment a common label might be digital. Additionally, the finding of a significant positive relationship between personal liking ratings and liking ratings from the third-person point of view suggests that people assume that other people think like them, and more speculatively, implies that a type of aesthetic normativity was present in the experiment. From this perspective, explicit measures as part of the experimental design can capture commonalities between types of judgments and may also contribute to evaluating conceptual frameworks related to aesthetic judgments.

On the other hand, in both experiments, although the regression models aimed to predict liking ratings did not depend on all the rating scores for the best fit, the existence of a significant, positive, and linear relationship between liking and all other ratings (albeit explaining only a medium amount of variance) does not entirely tally with existing research. For example, an inverted U-curve relation between complexity and preference has often been suggested (Berlyne 1958; Vitz 1966; Güçlütürk et al. 2016) and was not present in our results. Some potential explanations might be (i) The stimuli set did not cover the full range of (measurable) complexity levels, since the stimulus selection procedure did not specifically aim for it, (ii) Participants might have treated the concept of complexity not only as visual complexity, and therefore assigned varied meanings to it regarding other associated words such as complicated or hard to understand, among others.

Additionally, two main types of judgments can be formulated, either related to artworks' properties as perceived by the observer (such as pleasantness and interestingness) or related to properties as measurable features (such as colour and form). Some of those properties related to either of the judgment types can be treated as more elemental concepts and might be merged into relatively canonical factors such as arousal or regularity, respectively, for example using factor analysis on judgments (Marković and Radonjić 2008). Although building an overarching model is beyond the scope of this research, we can speculate that a potential inclusion of additional judgments with a positive connotation (such as interesting, successful, engaging, impressive; some of which were already suggested by the participants, see Supplementary Fig. S9) would still give similar results (namely, positive correlation with liking ratings). From this point of view, participants might have been using a common judgment strategy across all (or most of the) ratings, such as assigning an aesthetic value to an artwork as a common factor followed by providing isolated ratings. In this sense, targeting such potential judgment strategies, instead of isolated judgments, might be a promising research theme for future studies.

On considering viewing duration, a variable only present in the VR experiment, firstly, participants spent more time engaging with large-scale artworks (referred to as artwork spaces) compared to small-scale artworks (referred to as artwork objects), and the viewing durations were around 55 s and 33 s, respectively. Especially for the small-scale artworks, viewing duration in VR was similar to the real-world scenario, for example, museum-based research with a large sample size of 456 visitors found the average viewing duration as 28.63 s (Smith et al. 2017), where the researchers also underline the large variance between participants and between different artworks and attributed these arguably brief viewing durations to visitors' potential need for rapid art consumption. Here, one common-sense interpretation of duration increase for large-scale artworks might be that large-scale artworks simply provided more area to explore, in line with a previous finding from another museum-based study where larger viewing angles (as a derivative metric from painting size and viewing distance), a trend of longer viewing times were observed (Carbon 2017). Additionally, although the duration judgment (as the perceived time spent viewing artworks) was one of the moderately difficult ratings according to the exit questionnaire, it was a strong predictor of liking in the regression model. However, the weak correlation between real duration and perceived duration suggests that there were some misestimations of time and interestingly this was related to how much the artwork was liked. A related study involving a temporal reproduction paradigm in a between-groups design found a trend of duration underestimation for one group where the visual stimuli were described as "artworks" compared to another group, where the same stimuli were described as "photographs used in psychological experiments" (Arai and Kawabata 2016). Other previous research found expert-novice differences, such that the trained participants underestimated and naive participants overestimated the viewing duration of paintings, and conceptually linked this to perceptual and cognitive effort (Cupchik and Gebotys 1988). Although previous research underlined the importance of temporal dynamics during aesthetic judgments (Cupchik and Gebotys 1988; Smith et al. 2006; Muth et al. 2015), types of temporal distortion during viewing artworks, and more specifically, their relations to aesthetic judgments seem to call for further controlled experiments.

In terms of limitations and future directions, firstly, although the pilot experiment aimed to minimize the selection bias, implementing auto-generative algorithms into the experimental procedure might eliminate a potential selection bias. Emerging machine learning methods specific to media, visual arts, and cultural heritage might be relevant in terms of a stimulus generation, such as creating 3D objects from a single image (Chen et al. 2019) and constructing complex real-world scenes in 3D from a photo sample (Mildenhall et al. 2020). Secondly, although the VR as a research tool enables controlled experiments to approximate the real-world experience compared to the 2D monitor setups, the current available hardware has many limitations- limited field of view, unnatural locomotion substitutions such as teleportation, potential discomfort and motion sickness, to name but a few. Thirdly, the animated model (artwork #5 in Experiment 1) can be described as an oddball during the VR viewing phase, and previous work underlines the importance of motion, for example, in terms of effects of different 3D video games on visual attention guidance (El-Nasr and Yan 2006), of effects of types of motion on attention capture in sub-second range (Pratt et al. 2010): Although this aspect of motion was relatively irrelevant to the relations between judgments (as in the main correlational results), it did affect the judgments of this particular artwork (as shown in the additional analysis). Even though the models in the stimulus set do not have an obvious common property, "being static" might have been imposed to alleviate this categorical difference. Moreover, the spatially immersive experience using VR (and the ability to collect eye-tracking data) was not easily feasible for online experiments. However, the recent developments in online VR tools such as WebXR API, in line with the adoption of personal VR-HMDs with built-in eye trackers might provide more compatible online experiments soon. An immersive online experiment to show artworks as 360-degree images/videos, or as 3D environments can be created. Following on from that, a potential use of eye-tracking comparing online-2D and online-VR setups might provide a more direct measure of similarity between the two contexts. In both cases, the effects of the novelty of using VR might be minimized with longer experiments involving training sessions. Additionally, the relatively small stimulus set size and sample size (i.e., 16 artworks judged by 31 participants in VR experiment, and only 4 artworks judged by 120 participants in online experiment) limits the significance of the results and reduces the generalizability of the findings presented. Lastly, although this research was primarily concerned with the descriptive aspects of aesthetic judgments, predictive aspects beyond regression models can be further explored, especially using machine-learning-based tools: if a research direction is to build for example personalized predictive models (with a potential implication for online consumerism in art), many forms of artificial neural networks can be adapted to make better predictions about the liking judgment from other judgment types, or from supplementary measures including eye-tracking. Following the observed benefits of the multisensory interactions in VR museums (Koutsabasis and Vosinakis 2018), immersive experimental paradigms and novel approach to behavioural data analysis call to be extended to the multisensory research context, such as implementing immersive art pieces, video games or gamified experiences, to test for example whether aesthetic judgments hold their relationships across the senses.

6 Conclusion

Experimental aesthetics research often relies on restrictive laboratory setups, aiming for highly controlled experiments and tasks with high internal validity, but inevitably digressing from the genuine experience. Recently, in-situ experiments conducted in museums and galleries aim to counter this issue of experience, at the expense of a diminished control on environmental factors. In this sense, experiences using virtual, augmented, or mixed reality offer new research directions, but arguably they are still overlooked in aesthetic research. Here, the relations between aesthetic judgments towards a variety of 3D models were investigated in a relatively novel VR setting and followed-up with an online experiment. The VR experiment resulted in mostly moderate correlation between various types of judgments, along with the proof-of-concept 3D eye-tracking data showing different viewing trends among participants, and supplementary insights from participants were obtained via the exit questionnaire. The simplified online follow up experiment showed similar correlations, suggesting the persistence of these relations in a reduced experimental setup. In empirical aesthetics, in addition to laboratory-based, online, or in-situ conditions; immersive environments offer emerging directions for future research: The ability to provide relatively high ecological validity and simultaneously offer controllable experimental setups can be described as an otherwise improbable and novel experimental scenarios.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10055-022-00671-1.

Acknowledgements The authors would like to thank Royal Holloway University of London (RHUL) for the doctoral scholarship of Doga Gulhan and access to the Psychology VR lab.

Author contributions Conceptualization, DG, JMZ; Methodology, DG, JMZ; Software, DG; Formal Analysis, DG, SD; Investigation, DG, SD, JMZ; Writing – original draft, DG; Writing – review & editing, DG, SD, JMZ; Visualization, DG, JMZ; Supervision, SD, JMZ; Project management, JMZ.

Funding We received no funding for this study.

Data accessibility All anonymised data is accessible via the Open Science Framework (osf.io/ec46q).

Declarations

Conflict of interest The authors declare no conflicting interests.

Ethical approval All experimental protocols were approved by the Royal Holloway, University of London Research Ethics Committee. All methods were performed in accordance with the ethical guidelines and regulations of the Declaration of Helsinki.

References

Arai S, Kawabata H (2016) Appreciation contexts modulate aesthetic evaluation and perceived duration of pictures. Art and Perception 4(3):225–239. https://doi.org/10.1163/22134913-00002052

Augustin MD, Wagemans J, Carbon C-C (2012) All is beautiful? Generality vs. specificity of word usage in visual aesthetics. Acta Psychologica 139(1):187–201. https://doi.org/10.1016/j.actpsy.2011.10.004

Aumann A (2014) The relationship between aesthetic value and cognitive value: the relationship between aesthetic value and cognitive value. J Aesthet Art Critic 72(2):117–127. https://doi.org/10.1111/jaac.12073

Belchev Z, Bodner GE, Fawcett JM (2018) Inducing preference reversals in aesthetic choices for paintings: introducing the contrast paradigm. PLoS ONE 13(4):e0196246. https://doi.org/10.1371/journal.pone.0196246

Berlyne DE (1958) The influence of complexity and novelty in visual figures on orienting responses. J Exp Psychol 55(3):289–296. https://doi.org/10.1037/h0043555

Brieber D, Leder H, Nadal M (2015) The experience of art in museums: an attempt to dissociate the role of physical context and genuineness. Empir Stud Arts 33(1):95–105. https://doi.org/10.1177/0276237415570000

Carbon C-C (2017) Art perception in the museum: how we spend time and space in art exhibitions. I-Perception 8(1):1–15. https://doi.org/10.1177/2041669517694184

Chatterjee A, Widick P, Sternschein R, Smith WB, Bromberger B (2010) The assessment of art attributes. Empir Stud Arts 28(2):207–222. https://doi.org/10.2190/EM.28.2.f

Chen W, Ling H, Gao J, Smith E, Lehtinen J, Jacobson A, Fidler S (2019) Learning to predict 3D objects with an interpolation-based differentiable renderer. In: Wallach H, Larochelle H, Beygelzimer A, Alché-Buc F, Fox E, Garnett R (eds) Advances in neural information processing systems. Curran Associates Inc, pp 9609–9619

Chiquet S, Martarelli CS, Mast FW (2020) Eye movements to absent objects during mental imagery and visual memory in immersive virtual reality. Virtual Reality. https://doi.org/10.1007/s10055-020-00478-y

Cleeremans A, Ginsburgh V, Klein O, Noury A (2016) What's in a name? The effect of an artist's name on aesthetic judgments. Empir Stud Arts 34(1):126–139. https://doi.org/10.1177/0276237415621197

Clini P, Ruggeri L, Angeloni R, Sasso M (2018) Interactive immersive virtual museum: Digital documentation for virtual interaction. ISPRS Int Arch Photogramm Remote Sens Spatial Inform Sci. https://doi.org/10.5194/isprs-archives-XLII-2-251-2018

Cohen J (1992) A power primer. Psychol Bull 112(1):155–159. https://doi.org/10.1037/0033-2909.112.1.155

Consoli G (2015) Creativity and aesthetic evaluation. Two proposals to improve the model of aesthetic dis/fluency. Front Psychol. https://doi.org/10.3389/fpsyg.2014.01520

Cooper JM, Silvia PJ (2009) Opposing art: rejection as an action tendency of hostile aesthetic emotions. Empir Stud Arts 27(1):109–126. https://doi.org/10.2190/EM.27.1.f

Cova F, Pain N, Institute TH (2012) Can folk aesthetics ground aesthetic realism? Monist 95(2):241–263. https://doi.org/10.5840/monist201295214

Cupchik GC, Gebotys R (1988) The experience of time, pleasure, and interest during aesthetic episodes. Empir Stud Arts 6(1):1–12. https://doi.org/10.2190/5YN3-J3P8-FWHY-UDB3

El-Nasr MS, Yan S (2006) Visual attention in 3D video games. In: Proceedings of the 2006 ACM SIGCHI international conference on advances in computer entertainment technology - ACE '06, 22. https://doi.org/10.1145/1178823.1178849

Faerber SJ, Carbon C-C (2012) The power of liking: highly sensitive aesthetic processing for guiding us through the world. I-Perception 3(8):553–561. https://doi.org/10.1068/i0506

Fayn K, MacCann C, Tiliopoulos N, Silvia PJ (2015) Aesthetic emotions and aesthetic people: openness predicts sensitivity to novelty in the experiences of interest and pleasure. Front Psychol. https://doi.org/10.3389/fpsyg.2015.01877

Friedenberg J, Liby B (2016) Perceived beauty of random texture patterns: a preference for complexity. Acta Physiol (oxf) 168:41–49. https://doi.org/10.1016/j.actpsy.2016.04.007

Ginsborg H (2006) Aesthetic judgment and perceptual normativity. Inquiry 49(5):403–437. https://doi.org/10.1080/00201740600937898

Graf LKM, Landwehr JR (2017) Aesthetic pleasure versus aesthetic interest: the two routes to aesthetic liking. Front Psychol. https://doi.org/10.3389/fpsyg.2017.00015

Grüner S, Specker E, Leder H (2019) Effects of context and genuineness in the experience of art. Empir Stud Arts 37(2):138–152. https://doi.org/10.1177/0276237418822896

Güçlütürk Y, Jacobs RHAH, van Lier R (2016) Liking versus complexity: decomposing the inverted U-curve. Front Hum Neurosci 10:112. https://doi.org/10.3389/fnhum.2016.00112

Gulhan D, Durant S, Zanker JM (2021) Similarity of gaze patterns across physical and virtual versions of an installation artwork. Sci Rep 11(1):18913. https://doi.org/10.1038/s41598-021-91904-x

Hayn-Leichsenring GU (2017) The ambiguity of artworks: a guideline for empirical aesthetics research with artworks as stimuli. Front Psychol 8:1857. https://doi.org/10.3389/fpsyg.2017.01857

Helbing J, Draschkow D, Võ ML-H (2020) Search superiority: goal-directed attentional allocation creates more reliable incidental identity and location memory than explicit encoding in naturalistic virtual environments. Cognition 196:104147. https://doi.org/10.1016/j.cognition.2019.104147

Hoang TN, Cox TN (2017) Alternating reality: an interweaving narrative of physical and virtual cultural exhibitions. Presence Teleoper Virtual Environ. https://doi.org/10.1162/PRES_a_00307

Isik AI, Vessel EA (2019) Continuous ratings of movie watching reveal idiosyncratic dynamics of aesthetic enjoyment. PLoS ONE 14(10):e0223896. https://doi.org/10.1371/journal.pone.0223896

Kim S, Burr D, Alais D (2019) Attraction to the recent past in aesthetic judgments: a positive serial dependence for rating artwork. J vis 19(12):19. https://doi.org/10.1167/19.12.19

Koutsabasis P, Vosinakis S (2018) Kinesthetic interactions in museums: conveying cultural heritage by making use of ancient tools and (re-) constructing artworks. Virtual Reality 22(2):103–118. https://doi.org/10.1007/s10055-017-0325-0

Landau MJ, Greenberg J, Solomon S, Pyszczynski T, Martens A (2006) Windows into nothingness: terror management, meaninglessness, and negative reactions to modern art. J Pers Soc Psychol 90(6):879–892. https://doi.org/10.1037/0022-3514.90.6.879

Leder H, Belke B, Oeberst A, Augustin D (2004) A model of aesthetic appreciation and aesthetic judgments. Br J Psychol 95(4):489–508. https://doi.org/10.1348/0007126042369811

Leder H, Carbon C-C, Ripsas A-L (2006) Entitling art: influence of title information on understanding and appreciation of paintings. Acta Physiol (oxf) 121(2):176–198. https://doi.org/10.1016/j.actpsy.2005.08.005

Liarokapis F, Voulodimos A, Doulamis N, Doulamis (eds) (2020) Visual computing for cultural heritage. Springer. https://doi.org/10.1007/978-3-030-37191-3

Locher P (2011) Contemporary experimental aesthetics: state of the art technology. I-Perception 2(7):697–707. https://doi.org/10.1068/i0449aap

Locher P, Smith L, Smith J (1999) Original paintings versus slide and computer reproductions: a comparison of viewer responses. Empir Stud Arts 17(2):121–129. https://doi.org/10.2190/R1WN-TAF2-376D-EFUH

Locher P (2016) Comparison of the perception and evaluation of paintings viewed in the original versus reproduction formats. In: Mastandrea S, Maricchiolo F (eds) The role of the museum in the education of young adults. Motivation, emotion and learning. Roma Tre-Press, pp 27–34

Mallon B, Redies C, Hayn-Leichsenring GU (2014) Beauty in abstract paintings: perceptual contrast and statistical properties. Front Hum Neurosci. https://doi.org/10.3389/fnhum.2014.00161

Marković S, Radonjić A (2008) Implicit and explicit features of paintings. Spat vis 21(3–5):229–259. https://doi.org/10.1163/156856808784532563

Mildenhall B, Srinivasan PP, Tancik M, Barron JT, Ramamoorthi R, Ng R (2020) NeRF: Representing scenes as neural radiance fields for view synthesis. ArXiv E-Prints, arXiv:2003.08934

Moore KM, West AN (2012) Global perception, meaning, and aesthetic appreciation. Empir Stud Arts 30(1):23–38. https://doi.org/10.2190/EM.30.1.d

Muth C, Raab MH, Carbon C-C (2015) The stream of experience when watching artistic movies: dynamic aesthetic effects revealed by the continuous evaluation procedure (CEP). Front Psychol. https://doi.org/10.3389/fpsyg.2015.00365

Palan S, Schitter C (2018) Prolific.ac: a subject pool for online experiments. J Behav Exp Financ 17:22–27. https://doi.org/10.1016/j.jbef.2017.12.004

Parker E, Saker M (2020) Art museums and the incorporation of virtual reality: examining the impact of VR on spatial and social norms. Converg Int J Res into New Media Technol. 26(5–6):1–15. https://doi.org/10.1177/1354856519897251

Peer E, Brandimarte L, Samat S, Acquisti A (2017) Beyond the Turk: alternative platforms for crowdsourcing behavioral research. J Exp Soc Psychol 70:153–163. https://doi.org/10.1016/j.jesp.2017.01.006

Pelowski M (2015) Tears and transformation: feeling like crying as an indicator of insightful or "aesthetic" experience with art. Front Psychol 6:1006. https://doi.org/10.3389/fpsyg.2015.01006

Pelowski M, Gerger G, Chetouani Y, Markey PS, Leder H (2017) But is it really art? The classification of images as "art"/"not art" and correlation with appraisal and viewer interpersonal differences. Front Psychol 8:1729. https://doi.org/10.3389/fpsyg.2017.01729

Petrelli D (2019) Making virtual reconstructions part of the visit: an exploratory study. Digit Appl Archaeol Cult Herit 15:e00123. https://doi.org/10.1016/j.daach.2019.e00123

Pratt J, Radulescu PV, Guo RM, Abrams RA (2010) It's alive!: animate motion captures visual attention. Psychol Sci 21(11):1724–1730. https://doi.org/10.1177/0956797610387440

Russell PA, Milne S (1997) Meaningfulness and hedonic value of paintings: effects of titles. Empir Stud Arts 15(1):61–73. https://doi.org/10.2190/EHT3-HWVM-52CB-8QHJ

Schepman A, Rodway P, Pullen SJ, Kirkham J (2015) Shared liking and association valence for representational art but not abstract art. J vis 15(5):11. https://doi.org/10.1167/15.5.11

Schöne B, Sylvester RS, Radtke EL, Gruber T (2021) Sustained inattentional blindness in virtual reality and under conventional laboratory conditions. Virtual Real 25(1):209–216. https://doi.org/10.1007/s10055-020-00450-w

Silvia PJ, Brown EM (2007) Anger, disgust, and the negative aesthetic emotions: expanding an appraisal model of aesthetic experience. Psychol Aesthet Creat Arts 1(2):100–106. https://doi.org/10.1037/1931-3896.1.2.100

Skov M, Nadal M (2020) A farewell to art: aesthetics as a topic in psychology and neuroscience. Perspect Psychol Sci 15(3):1–13. https://doi.org/10.1177/1745691619897963

Smith LF, Bousquet SG, Chang G, Smith JK (2006) Effects of time and information on perception of art. Empir Stud Arts 24(2):229–242. https://doi.org/10.2190/DJM0-QBDW-03V7-BLRM

Smith LF, Smith JK, Tinio PPL (2017) Time spent viewing art and reading labels. Psychol Aesthet Creat Arts 11(1):77–85. https://doi.org/10.1037/aca0000049

Song J, Kwak Y, Kim C-Y (2021) Familiarity and novelty in aesthetic preference: the effects of the properties of the artwork and the beholder. Front Psychol 12:694927. https://doi.org/10.3389/fpsyg.2021.694927

Turpin MH, Kara-Yakoubian M, Walker A, Gabert NN, Fugelsang J, Stolz JA (2019) Bullshit makes the art grow profounder. Judgm Decis Mak 14(6):658–670. https://doi.org/10.2139/ssrn.3410674

van Helvoort D, Stobbe E, Benning R, Otgaar H, van de Ven V (2020) Physical exploration of a virtual reality environment: effects on spatiotemporal associative recognition of episodic memory. Mem Cognit 48:691–703. https://doi.org/10.3758/s13421-020-01024-6

Vitz PC (1966) Preference for different amounts of visual complexity. Behav Sci 11(2):105–114. https://doi.org/10.1002/bs.3830110204

Wagner V, Menninghaus W, Hanich J, Jacobsen T (2014) Art schema effects on affective experience: the case of disgusting images. Psychol Aesthet Creat Arts 8(2):120–129. https://doi.org/10.1037/a0036126

Wassiliwizky E, Menninghaus W (2021) Why and how should cognitive science care about aesthetics? Trends Cogn Sci 25(6):437–449. https://doi.org/10.1016/j.tics.2021.03.008

Woods AT, Velasco C, Levitan CA, Wan X, Spence C (2015) Conducting perception research over the internet: a tutorial review. PeerJ 3:e1058. https://doi.org/10.7717/peerj.1058

Young JO (ed) (2017). Semantics of aesthetic judgements (First edition). Oxford University Press

Zhao H, Thrash T, Grossrieder A, Kapadia M, Moussaïd M, Hölscher C, Schinazi VR (2020) The interaction between map complexity and crowd movement on navigation decisions in virtual reality. Royal Soc Open Sci 7(3):191523. https://doi.org/10.1098/rsos.191523