MONET: Museum Observatory of Neuro-affective Eye-Tracking

Raphaëlle Lemaire*
raphaelle.lemaire@unicaen.fr
GREYC, UNICAEN, ENSICAEN, CNRS
caen, France

Azamat Kaibaldiyev*
azamat.kaibaldiyev@unicaen.fr
GREYC, UNICAEN, ENSICAEN, CNRS
caen, France

Eléonore Mariette
eleonore.mariette@unicaen.fr
phIND, UNICAEN, Inserm
caen, France

Débora Viglieri
debora.viglieri@unicaen.fr
NIMH, UNICAEN, Inserm
caen, France

Jérémie Pantin
jeremie.pantin@unicaen.fr
GREYC, UNICAEN, ENSICAEN, CNRS
caen, France

Alexis Lechervy
alexis.lechervy@unicaen.fr
GREYC, UNICAEN, ENSICAEN, CNRS
caen, France

Fabrice Maurel
fabrice.maurel@unicaen.fr
GREYC, UNICAEN, ENSICAEN, CNRS
caen, France

Gaël Dias
gael.dias@unicaen.fr
GREYC, UNICAEN, ENSICAEN, CNRS
caen, France

Gaëtane Blaizot
gaetane.blaizot@unicaen.fr
UNICAEN
caen, France

Véronique Agin
veronique.agin@unicaen.fr
phIND, UNICAEN, Inserm
caen, France

Nicolas Poirel
nicolas.poirel@u-paris.fr
LaPsyDÉ, Université Paris Cité, CNRS
paris, France

Eric Bui
bui-th@chu-caen.fr
phIND, UNICAEN, Inserm, CHU Caen
caen, France

Hervé Platel
herve.platel@unicaen.fr
NIMH, UNICAEN, Inserm
caen, France

Denis Vivien
vivien@cyceron.fr
phIND, UNICAEN, Inserm, CHU caen
caen, France

Youssef Chahir
youssef.chahir@unicaen.fr
GREYC, UNICAEN, ENSICAEN, CNRS
caen, France

Abstract

This paper introduces MONET (Museum Observatory of Neuro-affective Eye-Tracking), a novel multimodal dataset collected over a nine-month period in a naturalistic museum setting, built to observe the effect of artwork on well-being. The core contribution of this work lies in the simultaneous capture of visual attention and affective responses from 148 visitors viewing original artworks, thereby ensuring an ecological validity often missing from traditional laboratory screen-based studies. The experimental design follows a unique two-phase protocol comparing solo and paired (two visitors simultaneously) visits, under both mediated and unmediated conditions. For each of the twelve paintings, the dataset provides individual saliency maps derived from mobile eye-tracking, valence and arousal scores obtained via an affect grid, and comprehensive demographic metadata. Exploratory findings reveal that while painting content remains the primary driver of inter-individual gaze variability, paired viewing leads to a significant convergence in affective responses between partners. Furthermore, the use of the Kullback-Leibler Divergence identifies a statistical link between spatial gaze distribution and emotional states, offering new pathways for modeling attention and affect in unconstrained environments. The dataset and supplementary materials are available at https://git.unicaen.fr/raphaelle.lemaire/monet, and the MONET dataset is distributed under a Data Use Agreement (DUA) license.

Figure 1: In-situ experiment with paired participants (B').

CCS Concepts

• **Applied computing** → *Health informatics*; **Fine arts**; **Document scanning**; • **Social and professional topics** → *Medical information policy*.

Keywords

Eye-tracking, Visual attention, Affect Grid, Artworks, In-situ dataset.

*Raphaëlle Lemaire and Azamat Kaibaldiyev contributed equally to this work.

MM '26, Rio de Janeiro, Brazil

ACM ISBN 979-8-4007-2213-4/2026/11
https://doi.org/10.1145/3767308.3838659

ACM Reference Format:
Raphaëlle Lemaire*, Azamat Kaibaldiyev*, Eléonore Mariette, Débora Viglieri, Jérémie Pantin, Alexis Lechervy, Fabrice Maurel, Gaël Dias, Gaëtane Blaizot, Véronique Agin, Nicolas Poirel, Eric Bui, Hervé Platel, Denis Vivien, and Youssef Chahir. 2026. MONET: Museum Observatory of Neuro-affective Eye-Tracking. In *Proceedings of the 34th ACM International Conference on Multimedia (MM '26), November 10–14, 2026, Rio de Janeiro, Brazil.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3767308.3838659

1 Introduction

Understanding aesthetic experience in naturalistic settings remains a fundamental challenge, particularly its relationship to well-being. Existing eye-tracking datasets are predominantly collected in controlled, in-silico environments, limiting their ability to capture the complex interplay between perception, attention, and affect that emerges during real-world cultural experiences[11, 14, 20]. Such studies often overlook critical factors, including physical proximity to the artwork, social context, and curatorial mediation. Meanwhile, existing in-situ research remains fragmented, typically focusing either on individual visitor trajectories [5, 17, 25] or on isolated sentiment analysis [9, 12, 16], without capturing their joint dynamics.

The MONET (Museum Observatory of Neuro-affective Eye Tracking) dataset address this gap through a large-scale, in-the-wild multimodal dataset that combines mobile eye-tracking recordings with affective self-reports. Experiments are collected in situ, during naturalistic visits at the Museum of Fine Arts of Caen (Figure 1). Collected over nine months, this corpus documents 148 participants across solo (Phase 1) and paired visits (Phase 2), including both mediated and unmediated conditions. For twelve paintings, we provide saliency maps generated via *Pupil Labs Core* glasses, valence and arousal scores via an affect Grid, and demographic metadata. This dataset serves as a novel benchmark for analyzing the connections between visual attention and emotion in ecological environments.

2 Related Work

Experimental Settings in Art Eye-Tracking. Analyzing visual attention in museum contexts faces inherent in-situ challenges, such as calibration constraints, free movement, and high inter-individual variability. Yet it offers an ecological validity that screen-based experiments cannot replicate [22]. Several studies have been conducted directly within museums, such as at the Studiolo del Duca of Urbino (Italy) [17] and the National Museum of Cinema in Turin (Italy) [5]. These works found that human figures and dynamic elements attract significant attention, with consistent fixation patterns across participants. More recently, research at the Liang Zhu Museum of Hangzhou (China) [25] with 8 participants identified strong correlations in fixation durations while distinguishing between gaze directed at artworks versus text panels. While some analyses utilize virtual reality [6, 29], most art-related datasets are collected in laboratory settings, where participants view screen reproductions. These studies generally follow two strategies: maximizing painting variety, with sets ranging from 150 [14] to 340 artworks [11], or maximizing participant numbers, such as studies involving 105 observers facing five paintings [20, 21]. However, these controlled environments often rely on downscaled versions, potentially obscuring subtle details and eliminating the contextual impact of the artwork's original scale, as well as withdrawing the social context of in-situ experiments.

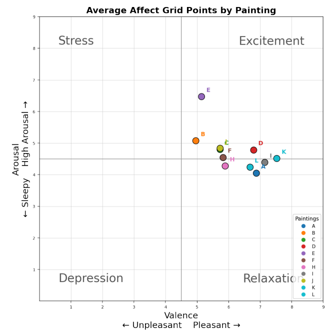


Figure 2: Mean coordinates per artwork on the affect grid.

Affective Response to Visual Art. The affective dimension of art viewing remains less explored than eye-tracking. Existing datasets are generally categorized into quantitative self-reports and textual descriptions. The former includes the use of an affect grid [23] to assess emotional responses to sculptures [16], and multidimensional questionnaires measuring aesthetic well-being [9]. Additionally, multi-stream CNNs have been developed to evaluate visitor satisfaction through facial and body expression recognition [12]. The latter category comprises text-based resources such as ArtEmis [2, 18], which provides emotion-grounded descriptions, and EmoArt [28], designed for emotion-aware artistic generation.

3 Experimental Protocol

To jointly study the visual attention and affective responses of museum visitors in an ecological setting, we deployed an experimental protocol involving 148 participants aged from 18 to 65 years old.

Participant Protocol. Each visitor, who either viewed the artworks of the experiments for the first time or had not seen them for over ten years, was equipped with *Pupil Labs Core* glasses [10] (eye cameras at 200 Hz, scene camera at 60 Hz), an Android smartphone running the *Pupil Mobile* application, and a lavalier microphone. The study received ethics committee approval and was conducted in two phases. In Phase 1 (solo visit), participants were randomly assigned to Group A (navigation only without mediation) or Group B (mediated condition with a museum guide). In Phase 2 (paired visit), pairs composed of one participant from Phase 1 (Group A) and one from Phase 1 (Group B) were formed, and assigned into Group A' (without mediation) or B' (with mediation).

The museum itinerary was fixed, covering 6 paintings (3 of which are in common in groups A and B) in Phase 1, and 8 paintings (2 seen by group A, 2 by group B, 2 by both groups, and 2 new ones) in Phase 2 across a broad range of artistic styles (see Table 1). A small number of participants completed Phase 2 individually (A1' and B1'), and 11 participants followed a different itinerary than their assigned group due to in-situ constraints. For each artwork, participants stood at a standardized distance to maximize painting coverage in the scene camera, observing it for 2 minutes and 30 seconds [17] (including a two-minute mediation for groups B/B'). After each painting, participants completed an affect grid to report their valence and arousal scores (Figure 2).

Eye-tracking Experimental Setup. Before each visit, the participant performed a 29-points calibration. It was performed to map

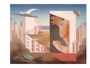
Paintings												
Short name	A	B	C	D	E	F	G	H	I	J	K	L
Distances	1.0m	2.0m	1.75m	1.5m	1.0m	1.5m	1.0m	1.0m	1.5m	1.75m	1.0m	1.0m
Groups	B	A’B’	BA’B’	AB	ABA’B’	A’B’	–	ABA’B’	AA’B’	AA’B’	BA’B’	A

Table 1: Paintings and viewing distances per groups A/B (Phase 1) and A’/B’ (Phase 2).

the full visual field covering all paintings, as to maximise real spacial accuracy. Calibration was conducted on a projected screen of dimensions $1.8 \times 1.0\ \text{m}^2$. In Phase 1, the participant was positioned facing the centre of the screen at a distance of 2.5 m. In Phase 2, the two members of each pair were positioned side by side at the centre of the screen. The relative positioning established during calibration (left-right assignment) was maintained throughout the museum visit (see Figure 1). During calibration, participants were instructed to keep their heads still and move only their eyes.

Because participants moved freely throughout the museum, gaze coordinates were recorded in the coordinate system of the scene camera. To enable cross-participant spatial analysis, these coordinates were projected onto high-resolution reference images of each painting. Scene camera frames were first preprocessed using Gamma adjustment and Contrast Limited Adaptive Histogram Equalization to compensate for the variable museum lighting conditions. Scale-Invariant Feature Transform keypoints were then extracted from both the video frames and the reference paintings. A homography matrix H was estimated to project egocentric gaze coordinates $g = (u, v)$ onto the 2D pixel space of the reference painting, yielding transformed coordinates $g' = (x, y)$. The Equation 1 formalise this, with $\sim$ denoting equality up to a scalar factor.

$$\begin{pmatrix} x \\ y \\ 1 \end{pmatrix} \sim H \begin{pmatrix} u \\ v \\ 1 \end{pmatrix}, \quad H \in R^{3\times 3} \tag{1}$$

Because recordings lasted up to one hour and covered multiple paintings, we manually placed marks in the *Pupil Player* timeline to isolate the exact temporal window corresponding to each painting.

Mobile eye-tracking in unconstrained environments is subject to gaze drift. To counter that, we applied a two stage drift correction. At first, we applied an initial global X/Y translation using the Offset function, based on visual inspection of the first painting, to compensate for the baseline drift of the 2D calibration. After projecting gaze onto the 2D reference paintings, for each painting, we visually inspected the projected gaze clusters overlaid on the reference image. We then applied a manual spatial translation to realign them with semantically salient anchors (e.g. faces, eyes) as in Equation 2,

$$\begin{pmatrix} x' \\ y' \end{pmatrix} = \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} \Delta x \\ \Delta y \end{pmatrix} \tag{2}$$

where (x, y) are the mapped gaze coordinates before correction, and $(\Delta x, \Delta y)$ is the manually estimated translation vector specific to each participant and each painting. This painting-level correction effectively reduced systematic spatial error, returning gaze accuracy to a level close to the initial calibration baseline [8].

Following drift correction, a continuous 2D grayscale saliency map was generated for each participant and each painting from the discrete fixation coordinates [14]. Each fixation location $l_k = (x_k, y_k)$ was modeled as a Dirac delta function, forming a discrete fixation map that was subsequently convolved with a 2D isotropic Gaussian kernel G_σ. The resulting map was then peak-normalized to the range $[0, 1]$, denoted by $\mathcal{N}$, as defined in Equation 3,

$$S_i(l) = \mathcal{N}\left(\sum_{k=1}^{K} \delta(l - l_k) * G_\sigma(l) \right) \tag{3}$$

where l is the coordinates in the 2D image of the painting and K is the number of fixations recorded for participant i on the painting. The standard deviation σ of the Gaussian kernel was set to represent exactly 1° of visual angle, following the convention for modelling the human foveal acuity region [8, 14]. For each painting n, σ was calculated dynamically from the physical canvas dimensions, the pixel dimensions of the reference image, and the average participant viewing distance D_n as in Equation 4,

$$\sigma_n = 2\tan\left(\frac{0.5\,\pi}{180}\right) \cdot D_n \cdot \frac{W_n^{\text{px}}}{W_n^{\text{phys}}} \tag{4}$$

where D_n is the average viewing distance to painting n (in metres), W_n^{px} is the width of the digital reference image in pixels, and W_n^{phys} is the physical width of the painting canvas in metres. The factor $2\tan(.) \approx 0.01745$ is the standard trigonometric constant for converting visual angle to a linear spatial extent [14].

Group	A	B	C	D	E	F	G	H	I	J	K	L	Total
A	1	0	1	65	70	0	0	61	61	63	2	57	381
B	65	0	69	66	67	0	1	61	0	0	63	0	392
A’	0	65	65	0	65	65	0	66	63	65	65	2	521
B’	0	67	67	0	67	67	0	68	62	68	68	6	540
A1’	0	3	3	0	3	3	0	3	3	3	3	0	24
B1’	0	2	2	0	2	2	0	2	2	2	2	0	16

Table 2: Affect grid and saliency map counts.

Group	#	Age mean	Age min-max	F (%)	M (%)	R (%)	L (%)
A	73	42.1±11.3	20-64	53 (73%)	20 (27%)	61 (84%)	12 (16%)
B	75	40.5±13.6	18.64	59 (79%)	16 (21%)	72 (96%)	3 (4%)
A’	71	40.7±13.5	18-64	51 (72%)	20 (28%)	62 (87%)	9 (13%)
B’	72	41.5±11.5	18-64	57 (79%)	15 (21%)	66 (92%)	6 (8%)
A1’	3	45.7±21.0	22-62	3 (100%)	0(0%)	3(100%)	0 (0%)
B1’	2	44.0±7.1	39-49	1 (50%)	1(50%)	2(100%)	0 (0%)
Total	296	41.24±12.5	18-64	224(76%)	72(24%)	266(90%)	30(10%)

Table 3: Participant distribution per group. F: female, M: male, R: right-handed, L: left-handed.

4 Dataset Description

Researchers can access the MONET dataset by completing a form available, along with access details, on Git. It contains attention maps and affect grid scores for each painting and each participant (see Table 2), as well as complementary demographic metadata. The number of recorded saliency maps and affect grid responses varies across paintings due to the exclusion of invalid gaze data resulting

Figure 3: Aggregated heat maps for Brueghel (D) on the left and Crespi (B) on the right. From left to right: the painting, all map, phase 1 map, group A map, group B map; the painting, all map, phase 2 map, group A' map and group B' map.

from hardware malfunctions, connectivity issues, or calibration failures.

Demographic Analysis. The dataset comprises 148 participants aged between 18 and 65 years. The sample is predominantly female (76%), exceeding typical museum attendance rates (55% [24], 63.3% [7]) but consistent with participation patterns reported in scientific studies (72.3% [15]). In addition, 90% of participants are right-handed, closely matching global population estimates (89.4% [19]). The presence of gaps in the 30–40 and 55–60 age ranges, combined with the non-normal distribution of ages, motivates the use of non-parametric statistical tests for age-related analyses.

To assess demographic balance (Table 3), we applied Mann-Whitney U tests for age and χ^2 tests for gender and handedness, excluding subgroups A1' and B1'. In Phase 1, groups A (73) and B (75) were balanced in age ($U = 2900.0, p = 0.53$) and gender ($\chi^2 = 0.45, p = 0.50$), though group A contained significantly more left-handed participants (16% vs. 4%; $\chi^2 = 4.99, p = 0.026$). In Phase 2, paired groups A' (71) and B' (72) were fully balanced across all variables: age ($U = 2492.5, p = 0.80$), gender ($\chi^2 = 0.68, p = 0.41$), and handedness ($\chi^2 = 0.33, p = 0.57$).

Furthermore, intra-pair age differences in Phase 2 showed no significant variation between A' and B' ($U = 2765.0, p = 0.33$). Despite assignment efforts aiming for gaps under 6 years, this right-skewed distribution reflects natural recruitment variability.

Calibration Quality Analysis. Eye-tracking data quality relies on two metrics provided by *Pupil Labs* [8, 10]: accuracy (defined as the mean angular error in degrees between the estimated and actual gaze position) and precision (the root mean square of the angular distance between successive gaze samples during fixation). Depending on participant eye characteristics, we employed either a 2D or a 3D calibration model [26]. Both share identical metric definitions but differ in their gaze estimation approach. The 2D calibration relies on a polynomial regression (with coefficients θ_{2D}) to map the 2D pupil position $p_i = (u_i, v_i)$ from the eye camera directly to normalised gaze coordinates g_i^{2D} in the scene camera as defined in Equation 5.

$$g_i^{2D} = f_{\text{poly}}(p_i\,;\,\theta_{2D}) \tag{5}$$

While highly accurate under ideal conditions, this mapping cannot compensate for glasses movement on the participant's head [1]. Conversely, the 3D pipeline dynamically fits a geometric eye model a sphere of radius r centred at c to successive pupil observations. Gaze is estimated as the ray g_i^{3D} cast from the eyeball centre c through the 3D pupil centre p_i^{3D} as defined in Equation 6.

$$g_i^{3D} = \frac{p_i^{3D} - c}{\|p_i^{3D} - c\|} \tag{6}$$

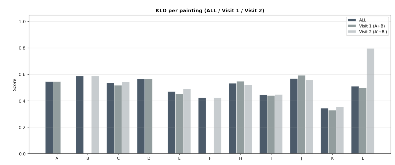


Figure 4: Histogram of KLD per painting against *all map* and *phase map*. The 1st column represents all visits, the 2nd column stands for Phase 1, and the 3rd column for Phase 2.

This continuous updating enables the 3D model to partially compensate for physical shifts of the glasses [10].

Regarding accuracy, the 2D calibration consistently outperformed the 3D model, producing lower mean errors and smaller interquartile ranges across both phases. In Phase 1, 2D calibration yielded a mean accuracy of $1.51 \pm 0.41°$, compared to $1.72 \pm 0.52°$ for 3D. In Phase 2, the paired setup slightly increased these errors to $1.85 \pm 0.54°$ for 2D and $1.96 \pm 0.59°$ for 3D. Conversely, precision remained remarkably stable across all groups and phases, averaging $0.10°$, with only rare outliers reaching $0.20°$ (2 participants in Phase 1, 4 in Phase 2). This consistently high precision ensures the stability and repeatability of the gaze measurements. Consequently, the saliency maps provided in this dataset reflect these post-recalibration adjustments, ensuring the gaze data can be reliably used for region-of-interest and saliency-based analyses.

5 Exploratory Analysis of MONET

Heat Map Analysis. We create mean saliency maps following the standard methodology of Le Meur et al. [14]. Since the dataset comprises two visit phases, we create the agregation map of all participants (*all map*), per group (*group map*) and per phase (*phase map*) for each painting (see Figure 3). To assess the variability of individual gaze behavior relative to group visual attention, we computed three metrics between each participant's saliency map and the three reference maps. The three metrics are the Pearson's Correlation Coefficient distance (CC-1), the symmetrised Kullback-Leibler Divergence (KLD), and the Inter-Observer Congruency (IOC) [27]. In particular, we will focus our attention on KLD but detailed results for CC-1 and IOC can be found in complementary materials.

The KLD values with respect to the three reference maps are low and consistent, with *all map* ranging from 1.546 (group B) to 1.865 (group A), and remain stable across the other two reference maps. This indicates that individual saliency maps are generally well-correlated with reference maps. At the painting level, KLD values vary across stimuli. The reduction in KLD when using the *group map* as compared to *all map* is modest across most paintings (on the order of 0.05 to 0.15), further indicating that *group maps* are close to the *all map* for most stimuli. An exception is the painting of Hennequin (L), which shows a discrepancy between Phase 1 (KLD = 1.423) and

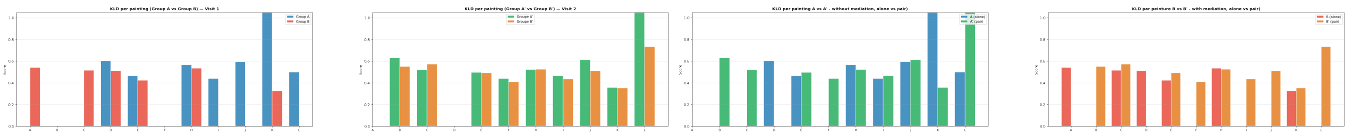


Figure 5: Histogram of KLD values per painting for the *all map*. Right to left: A/B, A'/B', A/A' and B/B' (first/second column).

Phase 2 (KDL = 1.694) (see Figure 4 and Figure 5). This difference is likely attributable to the small number of participants in Phase 2 (8) compared to Phase 1 (55). This painting was excluded from the t-test analysis comparing phases and groups, as well as paintings seen but not in regular visits. Taken together, the distance analyses indicate that individual saliency maps are well-aligned with all reference maps. The close similarity between distances computed against the reference maps seem to confirm that group membership and phase have limited influence on mean gaze patterns.

saliency distance	Demographic correlation	p-value
IOC	Sex (F/M)	0.051
	Handness (R/L)	0.591
	Age (-43y/+43y)	0.425
CC - 1	Sex (F/M)	0.051
	Handness (R/L)	0.704
	Age (-43y/+43y)	**0.046**
KLD	Sex (F/M)	0.159
	Handness (R/L)	**0.001**
	Age (-43y/+43y)	0.501

Table 4: T-test p-value for each demographic data for IOC, CC-1 and KLD. Bold indicates statistical significance ($p < 0.05$).

In addition in Table 4, we compute the t-test statistics between gender, laterality and age (below and above 43 years old) for IOC, CC-1 and KLD metrics. Only age (CC-1, $p = 0.046$) and handness (KLD, $p = 0.001$) show significative difference.

Affect-grid Analysis. The mean affect coordinates per painting are visualised on the Affect Grid space as illustrated in Figure 2. Mean coordinates cluster in the pleasant and moderately aroused quadrant of the grid, showing a generally positive affective tone of the museum experience. The mean valence scores per painting reveal substantial heterogeneity across stimuli in both phases. During Phase 1, valence scores range from 4.85 for Simon VOUET (E) to 7.81 for Maurice DENIS (K). During Phase 2, valence scores range from 4.97 for Bartolomeo MANFREDI (C) to 7.41 Maurice DENIS (K). Arousal scores generally show lower means than valence across both phases, with most paintings clustered in the moderate arousal (4.0 to 6.0). The mean arousal across all participants is 4.84 for group A, 4.81 for group B in Phase 1, and 4.65 for group A', 4.97 for group B' in Phase 2. This suggests that the museum visit induces a state of moderate emotional engagement. To quantify individual affective variability, we computed the Euclidean distance between each participant's affect coordinates and three reference mean points: the *global mean*, the *phase mean*, and the *group mean*. Mean distances to the *global mean* are consistent across groups, ranging from 1.998 (B') to 2.273 (A). Distances computed against the *phase mean* are nearly identical (differences < 0.05), confirming that the *global* and *phase mean* affect points are spatially very close in the valence-arousal space. With the Mann-Whitney test, no significant difference was found between groups A and B for any reference level (*global*: $p = 0.417$; *phase*: $p = 0.685$; *group*: $p = 0.810$), nor between A' and B' (*global*: $p = 0.205$; *phase*: $p = 0.332$; *group*: $p = 0.191$), indicating that mediation did not significantly modulate joint affective dispersion in either visit condition. However, significant differences emerged when comparing solo versus paired visit conditions (A/A': $p = 0.006$, B/B': $p = 0.004$). This consistent pattern across both valence and arousal simultaneously indicates that participants in the paired visit condition exhibited significantly lower affective dispersion than those in the solo visit.

Affect Grid	Demographic correlation	p-value
Distance	Gender (F/M)	**0.026**
	Handness (R/L)	0.398
	Age (-43y/+43y)	**0.022**
Valence	Gender (F/M)	0.243
	Handness (R/L)	0.092
	Age (-43y/+43y)	0.354
Arousal	Gender (F/M)	0.081
	Handness (R/L)	0.051
	Age (-43y/+43y)	**0.002**

Table 5: T-test p-value for each demographic data for euclidian distance. Bold indicates statistical significance ($p < 0.05$).

To show the impact of demographic data in the Affect Grid, we computed the t-test statistics for gender, laterality and age (below and above 43 years old) for the euclidian distance, arousal and valence (see Table 5). Two demographics variable show significant difference, i.e. age and gender.

saliency distance	ρ	p-value
CC-1	0.032	0.177
KLD	0.062	**0.009**
NSS	-0.005	0.824
SSIM	0.007	0.765

saliency distance	ρ	p-value
CC-1	0.055	0.223
KLD	0.116	**0.010**
NSS	0.067	0.138
SSIM	0.066	0.141

Figure 6: Spearman rank correlations (ρ) between saliency and Affect Grid. (Right) individual distance to global mean, (Left) distance between each pairs. Bold indicates statistical significance ($p < 0.05$).

Correlation Between Affect Grid and Heat Map. We investigate whether affective responses and visual attention are related. For that purpose, we computed for each participant and each painting two types of distances to the global mean: the Euclidean distance between the individual point and the *global mean* for the Affect Grid, and four saliency-based distances (Pearson correlation distance (CC-1), the symmetrised KL divergence (KLD), the Normalized Scanpath Saliency distance (NSS), and the Structural Similarity Index distance (SSIM) [3, 4, 13]) between the individual saliency map and the global mean saliency map (i.e. *all map*). Spearman rank correlations were then computed between these two types of distances, the null hypothesis stating that there is no significant correlation between saliency map distances and Affect Grid distances. Three out of four metrics yielded non-significant global correlations (see the right Table in Figure 6), CC-1 ($\rho = 0.032$, $p = 0.177$), NSS ($\rho = -0.005$, $p = 0.824$), and SSIM ($\rho = 0.007$,

$p = 0.765$). But, the KLD metric reached statistical significance at the global level ($\rho = 0.062$, $p = 0.009$). KLD differs from the other metrics in that it captures the full divergence between two probability distributions (here saliency maps) rather than a simple spatial or structural similarity. This may explain why it is sensitive to signals that CC-1, NSS and SSIM fail to detect.

Demographic and Visit correlations	ρ	p-value
Gender (F/M)	0.048/0.092	0.077/0.49
Handness (R/L)	0.065/0.053	**0.009**/0.459
Age (-43y/+43y)	0.053/0.069	0.112/**0.037**
Phase (V1/V2)	0.049/0.074	0.134/**0.027**
Group V1 (A/B)	0.049/0.074	0.134/0.27
Group V2 (A'/B')	0.082/0.043	**0.015**/0.189

Table 6: Spearman rank correlations (ρ) between KLD saliency distance and Euclidean Affect Grid distance, stratified by demographic and experimental groups. For each variable, results are reported for both subgroups (Female / Male). Bold indicates statistical significance ($p < 0.05$).

For KLD (see Table 6), significant correlations were found in male participants ($\rho = 0.092$, $p = 0.049$, $n = 454$ male among 1805 observation), right-handed participants ($\rho = 0.065$, $p = 0.009$, $n = 1604$), participant with an older age than the median (43 years) ($\rho = 0.070$, $p = 0.037$, $n = 894$), and in Phase 1 ($\rho = 0.072$, $p = 0.049$, $n = 743$). KLD appears to carry a weak but consistent signal linking heat maps distance to affective distance.

Pair Participant Correlation. To assess whether social interaction during the paired visit (Phase 2) is associated with convergence in visual attention and affective responses, we analysed the Spearman rank correlation between the saliency map distance and Affect Grid distance between the two members of each visitor pair. Analyses were conducted on 493 pair painting observations (241 in group A', 252 in group B'), covering 71 real pairs across the 9 paintings viewed in Phase 2, to determine whether any observed effects are specific to the actual pair composition. For KLD (see left Table of Figure 6), a significant positive correlation was found ($\rho = 0.116$, $p = 0.010$), indicating that pairs whose members showed more divergent gaze distributions also tended to report more divergent affective responses. This relationship was not significant for CC-1 ($\rho = 0.055$, $p = 0.223$) or SSIM ($\rho = 0.066$, $p = 0.141$).

saliency distance	pairs	ρ	p-value
CC-1	True pairs	0.055	0.223
	Fake pairs	0.009	0.570
	Constrained pairs	0.024	0.103
KLD	True pairs	0.116	**0.010**
	Faked pairs	-0.019	0.199
	Constrained pairs	0.017	0.245
NSS	True pairs	0.067	0.138
	Faked pairs	0.001	0.956
	Constrained pairs	0.022	0.134
SSIM	True pairs	0.066	0.141
	Faked pairs	0.014	0.347
	Constrained pairs	0.014	0.354

Table 7: Spearman rank correlations (ρ) for true pairs, faked pairs and constrained faked pairs. Bold indicates statistical significance ($p < 0.05$).

Results for real pairs were compared against two types of shuffled control pairs. The first shuffle is the constrained fake pairs, formed by randomly pairing one participant from group A with one from group B and did the visit 2 in the same group (A' or B') but in different pairs. The second shuffle is random fake pairs, formed without any group/phase constraint. We created 500 shuffled pairs per condition. When the same correlation was computed on constrained and random fake pairs (see Table 7), it became non-significant for all metrics (constrained fake pairs: $\rho = -0.019$, $p = 0.199$; random fake pairs: $\rho = 0.017$, $p = 0.245$). This specificity of the KLD correlation to real pairs suggests that the link between gaze distribution divergence and affective divergence is not a general property of any two observers viewing the same painting, but is specific to participants who actually visited together.

6 Applications and Downstream Task

The MONET dataset supports a broad range of research directions across computer vision, affective prediction and museum studies. Individual and aggregated saliency maps serve as a robust ground truth for evaluating and fine-tuning computational saliency models on in-situ museum data, complementing existing screen-based benchmarks [11, 14].

The synchronized availability of visual and emotional data should enable the development of models for automated emotion detection from eye-tracking data. We can design a pipeline that computes saliency evaluation metrics (KLD, CC-1, NSS and SSIM) to measure the distance between a visitor's actual saliency map and the map predicted by a model. These values can be utilized as input features for a machine learning trained to predict a participant's affective response. Because this approach relies on the mathematical comparison of gaze patterns, it solving the issue imposed by the 12 paintings present in the dataset.

Furthermore, the paired visit design, with 71 real pairs, provides a unique resource (as far as we know) for predicting gaze and affect distance between social partners. This supports the prediction of social influence in visual-affective perception, allowing for the prediction of how the partner's presence shifts individual gaze distributions or emotional scores. Comprehensive demographic information also allows for the prediction of affective scores or fixation patterns based on variables such as laterality, gender, age, and age gaps.

7 Conclusion

In this paper, we introduced MONET, the first in-situ dataset jointly providing saliency maps and affect grid scores for paintings viewed by visitors alone or in pairs. The dataset fills a critical gap between screen-based art eye-tracking datasets and the ecological reality of museum visits, offering a unique ressource for the prediction of visual attention, affective response, and social interaction in cultural heritage settings. Exploratory analyses establish two key findings. First, stimulus content is the dominant factor driving inter-individual gaze variability. Secondly, pair viewing is associated with significant affective convergence and with a correlation between saliency map and affect grid. The same observation holds true for individual participants. These findings highlight the complementary roles of visual attention and affective response in the museum experience. This also points to the social and contextual factors that selectively modulate each dimension. MONET together with the analysis provided, offers a foundation for future work on saliency prediction, affective computing and pairs behaviour modelling.

Acknowledgments

The ABC project was managed by the Centre Hospitalier Universitaire de Caen Normandie, 14000 CAEN, France (Ringgold ID: 26962). It was funded by the GIP Millénaire Caen 2025 as part of the ABC project (artbienetrecerveau.fr) and benefit from state aid managed by the National Research Agency under France 2030, CaeSAR (Caen, Stratégie pour l'Accélération en Recherche) (ANR-23-EXES-0001) and the Normandy region. The authors would like to thank Romain Goy, Alice Pelerin, Alice Poissonier, Elise Leroux, Juliette Tillet, and Dhia Merzougui for their contribution to the data collection process. We also like to thank the Museum of Fine Arts of Caen (France) which welcomed us for the study. This work was performed using the computing resources of the CRIANN.

References

[1] [n. d.]. Pupil Capture Documentation. https://docs.pupil-labs.com/core/software/pupil-capture/.

[2] Panos Achlioptas, Maks Ovsjanikov, Kilichbek Haydarov, Mohamed Elhoseiny, and Leonidas J Guibas. 2021. Artemis: Affective language for visual art. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 11569–11579.

[3] Zoya Bylinskii, Tilke Judd, Ali Borji, Laurent Itti, Frédo Durand, Aude Oliva, and Antonio Torralba. 2014. Mit saliency benchmark. 2015. *URL: http://saliency. mit. edu/results_mit300. html* 12 (2014), 13.

[4] Zoya Bylinskii, Tilke Judd, Aude Oliva, Antonio Torralba, and Frédo Durand. 2018. What do different evaluation metrics tell us about saliency models? *IEEE transactions on pattern analysis and machine intelligence* 41, 3 (2018), 740–757.

[5] Elena Di Giovanni. 2020. Eye tracking and the museum experience in Italy. *Altre Modernità: Rivista di studi letterari e culturali* 24 (2020), 10–24.

[6] Zihao Ding, Cheng-Tse Lee, Mufeng Zhu, Tao Guan, Yuan-Chun Sun, Cheng-Hsin Hsu, and Yao Liu. 2025. EyeNavGS: A 6-DoF navigation dataset and record-n-replay software for real-world 3DGS scenes in VR. In *Proceedings of the 33rd ACM International Conference on Multimedia*. 13126–13132.

[7] Scott Foster, Ian Fillis, Kim Lehman, and Mark Wickham. 2020. Investigating the relationship between visitor location and motivations to attend a museum. *Cultural Trends* 29, 3 (2020), 213–233.

[8] Kenneth Holmqvist, Marcus Nyström, Richard Andersson, Richard Dewhurst, Halszka Jarodzka, and Joost Van de Weijer. 2011. *Eye tracking: A comprehensive guide to methods and measures*. oup Oxford.

[9] Linhui Hu, Qian Shan, Lidan Chen, Siyin Liao, Jinxiao Li, and Guangpei Ren. 2024. Survey on the impact of historical museum exhibition forms on visitors' Perceptions based on eye-tracking. *Buildings* 14, 11 (2024), 3538.

[10] Moritz Kassner, William Patera, and Andreas Bulling. 2014. Pupil: an open source platform for pervasive eye tracking and mobile gaze-based interaction. In *Proceedings of the 2014 ACM international joint conference on pervasive and ubiquitous computing: Adjunct publication*. 1151–1160.

[11] Mohamed Amine Kerkouri, Marouane Tliba, Aladine Chetouani, and Alessandro Bruno. 2024. AVAtt: Art Visual Attention dataset for diverse painting styles. In *Proceedings of the 2024 Symposium on Eye Tracking Research and Applications*. 1–3.

[12] Do Hyung Kwon and Jeong Min Yu. 2024. Real-time Multi-CNN-based emotion recognition system for evaluating museum visitors' satisfaction. *ACM Journal on Computing and Cultural Heritage* 17, 1 (2024), 1–18.

[13] Olivier Le Meur and Thierry Baccino. 2013. Methods for comparing scanpaths and saliency maps: strengths and weaknesses. *Behavior research methods* 45, 1 (2013), 251–266.

[14] Olivier Le Meur, Tugdual Le Pen, and Rémi Cozot. 2020. Can we accurately predict where we look at paintings? *Plos one* 15, 10 (2020), e0239980.

[15] Lucas Lobato, Jeffrey Michael Bethony, Fernanda Bicalho Pereira, Shannon Lee Grahek, David Diemert, and Maria Flávia Gazzinelli. 2014. Impact of gender on the decision to participate in a clinical trial: a cross-sectional study. *BMC public health* 14, 1 (2014), 1156.

[16] Fahimeh Malekinezhad, Hassan Chizari, Hasanuddin bin Lamit, and Muhamad Solehin Fitry bin Rosley. 2013. A comparative study on designers and non-designers emotion of urban sculptures using affect grid. *Life Sci. J* 10, 3 (2013), 2056–2063.

[17] Serena Mandolesi, Danilo Gambelli, Simona Naspetti, and Raffaele Zanoli. 2022. Exploring visitors' visual behavior using eye-tracking: The case of the "Studiolo Del Duca". *Journal of Imaging* 8, 1 (2022), 8.

[18] Youssef Mohamed, Faizan Farooq Khan, Kilichbek Haydarov, and Mohamed Elhoseiny. 2022. It is okay to not be okay: Overcoming emotional bias in affective image captioning by contrastive data collection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 21263–21272.

[19] Marietta Papadatou-Pastou, Eleni Ntolka, Judith Schmitz, Maryanne Martin, Marcus R Munafò, Sebastian Ocklenburg, and Silvia Paracchini. 2020. Human handedness: A meta-analysis. *Psychological bulletin* 146, 6 (2020), 481.

[20] Roberto Pierdicca, Marina Paolanti, Simona Naspetti, Serena Mandolesi, Raffaele Zanoli, and Emanuele Frontoni. 2018. User-centered predictive model for improving cultural heritage augmented reality applications: An HMM-based approach for eye-tracking data. *Journal of imaging* 4, 8 (2018), 101.

[21] Roberto Pierdicca, Marina Paolanti, Ramona Quattrini, Marco Mameli, and Emanuele Frontoni. 2020. A visual attentive model for discovering patterns in eye-tracking data—a proposal in cultural heritage. *Sensors* 20, 7 (2020), 2101.

[22] Rodrigo Quian Quiroga and Carlos Pedreira. 2011. How do we see art: an eye-tracker study. *Frontiers in human neuroscience* 5 (2011), 98.

[23] James A Russell, Anna Weiss, and Gerald A Mendelsohn. 1989. Affect grid: a single-item scale of pleasure and arousal. *Journal of personality and social psychology* 57, 3 (1989), 493.

[24] J Schuster. 1991. *The Audience for American Art Museums. Research Division Report# 23.* ERIC.

[25] Wenjia Shi, Kenta Ono, and Liang Li. 2025. Cognitive insights into museum engagement: A mobile eye-tracking study on visual attention distribution and learning experience. *Electronics* 14, 11 (2025), 2208.

[26] Anca Velisar and Natela M Shanidze. 2024. Noise estimation for head-mounted 3D binocular eye tracking using Pupil Core eye-tracking goggles. *Behavior research methods* 56, 1 (2024), 53–79.

[27] Scott R. Walter, William T.M. Dunsmuir, and Johanna I. Westbrook. 2019. Inter-observer agreement and reliability assessment for observational studies of clinical work. *Journal of Biomedical Informatics* 100 (2019), 103317.

[28] Cheng Zhang, Hongxia Xie, Bin Wen, Songhan Zuo, Ruoxuan Zhang, and Wen-Huang Cheng. 2025. Emoart: A multidimensional dataset for emotion-aware artistic generation. In *Proceedings of the 33rd ACM International Conference on Multimedia*. 12644–12650.

[29] Yunzhan Zhou, Tian Feng, Shihui Shuai, Xiangdong Li, Lingyun Sun, and Henry Been-Lirn Duh. 2022. EDVAM: a 3D eye-tracking dataset for visual attention modeling in a virtual museum. *Frontiers of Information Technology & Electronic Engineering* 23, 1 (2022), 101–112.