scientific reports

Check for updates

OPEN

Understanding phototexts through an eye-tracking study on visual and emotional interactions between text and photo

Nunzio Langiulli[1]✉, Michele Cometa[2,3], Roberta Coglitore[2] & Vittorio Gallese[1,3]

Visual perception is a complex cognitive process that allows humans to extract meaningful information from their environment. Phototexts, artistic and literary compositions that integrate photographic and textual elements, offer a multimodal experience that requires readers to navigate between visual and linguistic modes of interpretation. This study investigates, in naïve participants, the influence of spatial layout configuration, emotional content, enjoyment, and attentional focus on the visual exploration of phototexts. Using eye-tracking technology, we examined participants' gaze patterns, focusing on metrics such as first fixation latency and fixation distribution across areas of interest (AOIs). Additionally, we explored the presence of left-gaze bias in text-photo stimuli systematically counterbalanced to mitigate positional effects. Our findings reveal key aspects of oculomotor behavior, as well as explicit preferences and judgments regarding phototext perception. Eye-tracking data indicated a natural reading bias, with shorter fixation latencies for text regardless of position, alongside a left-gaze bias that emerged specifically for photographs. Interestingly, attention patterns evolved dynamically over time, with an initial focus on text shifting toward the photographic component as viewing progressed. By analyzing phototexts as complex multimodal stimuli, this study provides novel insights into their synergistic effects on attention distribution and perceptual engagement.

Keywords Phototexts, eye-tracking, Left-gaze bias, Photo and text layout, Aesthetic enjoyment

Visual perception is a complex cognitive process that enables humans to interact effectively with their environment by extracting meaningful information from visual stimuli. Phototexts—artistic and literary compositions that seamlessly integrate photographic and textual components—invite readers into a multimodal experience, compelling them to oscillate between visual and linguistic modes of interpretation. This complex interaction challenges conventional distinctions between seeing and reading, forging new paths for understanding how humans process integrated sensory information. When presented with multimodal stimuli, such as phototexts, the visual system employs intricate mechanisms to process and integrate these inputs efficiently. This process involves not only the interpretation of distinct visual elements but also the synthesis of their combined semantic and emotional content. Eye-tracking has emerged as a powerful tool for investigating how individuals prioritize and process such multimodal information, offering key insights through metrics like the latency of the first fixation and the distribution of fixations within specific areas of interest (AOIs). The core principle behind eye-tracking is the eye-mind hypothesis, which posits that "the eye remains fixated on a word as long as the word is being processed"[1,2]. While this assumption has been challenged and discussed by findings on covert attention, the ability to direct attention to locations other than the fixation point [3–5], the principle that gaze location and duration provide a reliable proxy for cognitive engagement remains a cornerstone of the field[6] . Consequently, eye movements are understood to reflect a dynamic interplay between bottom-up (stimulus-driven) processes, such as attentional capture by salient visual features, and top-down (goal-driven) control, guided by the viewer's intentions and knowledge[1,7] .

Eye movement patterns are highly task-dependent. Reading, a specialized form of visual processing, is characterized by a remarkably consistent oculomotor pattern of sequential saccades and fixations: fixations

[1]Department of Medicine and Surgery, University of Parma, Unit of Neuroscience, Parma, Italy. [2]University of Palermo, Palermo, Italy. [3]Italian Academy for Advanced Studies in America at Columbia University, New York, USA. ✉email: nunzio.langiulli@unipr.it

typically last 200–300 ms, and saccades span approximately 7–9 letter spaces[7] . These parameters are modulated by linguistic properties of the text, such as word frequency and predictability, with more complex or unfamiliar text eliciting longer fixations and more frequent regressions[7,8]. Conversely, scene perception involves different oculomotor strategies. Here, fixations are more variable in duration (ranging from < 100 to > 500 ms), and saccades are generally longer, guided initially by low-level visual saliency[9] and subsequently by the viewer's semantic and emotional interpretation of the scene[10] . Furthermore, these general patterns can be modulated by individual and cultural factors. For instance, reading directionality inherent in a specific language shapes scanning habits that can transfer to other visual tasks[11] . By understanding these distinct oculomotor signatures for text and image processing, we can better investigate how they are integrated or compete when presented simultaneously within a single stimulus like a phototext.

The oculomotor strategies employed for image perception differ markedly from those used in reading. Unlike the systematic, sequential pattern of reading, image viewing is characterized by longer fixation durations and saccadic amplitudes, a strategy that facilitates the extraction of information from spatially complex stimuli[12] . This exploratory process is governed by a dynamic interplay between bottom-up factors, where visually salient properties of the stimulus (e.g., contrast, edges) automatically capture attention, and top-down control, where the viewer's goals, knowledge, and expectations guide the gaze[13] . For instance, visually complex scenes elicit more numerous fixations and broader scanning patterns than simpler compositions, while emotionally evocative images capture and hold attention, prompting longer and more thorough exploration compared to neutral stimuli[12,14] . This oculomotor behavior is further modulated by a range of observers' characteristics, including expertise, age, and cultural background. Art experts, for example, demonstrate unique scanning strategies sensitive to high-level features like composition and texture[15] , while age-related differences are evident in gaze patterns, with older adults sometimes adopting more focused strategies compared to the more exploratory patterns of younger adults[16] . Perhaps one of the most well-documented influences is cultural background. Research consistently shows that individuals from Western cultures tend to adopt an analytic viewing style, focusing on focal objects, whereas individuals from East Asian cultures employ a holistic style, allocating greater attention to contextual information and object relationships[17,18]. These distinct cognitive styles are directly reflected in eye movements: North American participants typically fixate on focal objects more quickly and for longer durations, whereas Chinese participants make more frequent saccades to the background[17] .

The spatial configuration of visual stimuli also plays a critical role in shaping perceptual and attentional outcomes. One well-documented phenomenon, particularly in Western cultures, is the leftward bias, or left-gaze bias[19,20]. This tendency to initially fixate or allocate more fixations to the left side of a visual stimulus has been observed across diverse contexts, including visual scenes[21], artworks[22] , faces [23–25], and bodies[26]. Additionally, some authors found that this bias can be modulated by emotional content and task demands[27,28]. Although the precise mechanisms underlying left-gaze bias remain debated, some researchers attribute it to an interaction between hemispheric lateralization and the directional scanning patterns established during evolution[29,30].

The theoretical tension between visual and textual elements has long been a topic of debate. Classical semiotics, as articulated by scholars like Roland Barthes, highlighted the complementary yet distinct ways in which images and text function to convey meaning. Barthes introduced the concepts of "anchorage" and "relay" to describe how text can ground the interpretation of an image or extend its narrative possibilities, suggesting a hierarchical relationship where text often guides the meaning of visual content[31] . In contrast, contemporary theories often advocate for a more symbiotic relationship, where visual and textual elements engage in a dynamic interplay, with each modality reciprocally influencing the interpretive process. This modern perspective has spurred empirical interest in how factors such as spatial layout, emotional content, and cultural context shape the perception of integrated multimodal stimuli like phototexts.

While the integration of text and image is now ubiquitous in digital media, artistic phototexts present a unique challenge to the visual system by deliberately manipulating the relationship between these two modalities. Lalla Romano's phototexts, particularly those from *Lettura di un'immagine* and *Romanzo di figure*, exemplify this intricate interplay Her works not only alternate textual and photographic elements spatially but also subvert traditional roles by positioning images as narrative agents and text as a visual accompaniment. This inverted hierarchy challenges readers to approach phototexts as integrated stimuli rather than as discrete components. Romano's assertion that "*In this book, the images are the text, and the writing is an illustration. The brief notes are meant to capture the reader's attention for a moment (as photos are usually viewed fleetingly) by suggesting perspectives of interpretation. There is no intention of providing information or telling stories: in the presence of photos, that would be indiscreet and misleading. Reading should not take place on that level*"[32] underscores her experimental approach, which aligns with broader trends in literary and visual theory that seek to blur the boundaries between modalities.

Consistent with the functional primacy of text for conveying explicit information, eye-tracking studies in educational contexts often reveal a text-driven processing bias. Research consistently shows that learners dedicate significantly more time to reading text than to viewing accompanying images, suggesting that comprehension is largely guided by the textual channel[33,34] . This text-centric pattern is particularly pronounced in young readers[35] . However, this bias is significantly modulated by spatial layout. When text and corresponding images are spatially integrated, viewers make more frequent "integrative transitions" (Spatial Contiguity Principle), saccades between the two sources, which is indicative of cognitive effort to build a coherent mental model and is often correlated with better learning outcomes[36] .

A critical limitation of the existing literature, however, is its predominant focus on didactic materials, where the primary goal is information acquisition. Far less is understood how viewers process artistic phototexts, where the relationship between text and image is often ambiguous, evocative, and intended to elicit an aesthetic or emotional response rather than convey factual knowledge. Moreover, the influence of spatial layout and emotional content on attentional allocation in these contexts is not yet fully understood. Addressing these gaps,

the present study investigates how spatial layout and emotional content of phototexts affect attentional patterns and users' engagement. Using eye-tracking, we analyzed early eye movement measures, such as first fixation latency and fixation counts within text and image components, under counterbalanced spatial arrangements. By investigating how participants navigate between textual and photographic elements, our research contributes to a new understanding of the cognitive and affective processes underpinning phototext engagement. The findings promise to illuminate broader implications for the design and interpretation of multimodal media, particularly in an era where digital platforms increasingly integrate text and image to create immersive storytelling experiences.

Materials and methods

Participants

A total of 48 participants (N = 48, 19 males and 29 females, with a mean age of 28.7 years and standard deviation of ± 6.3, within a range of 22 to 42 years) were selected to participate in the experiment. Participants were divided into two groups, each of which performed the same experimental task but with slight variations in the instructions provided regarding how to interact with the phototexts. The first group, consisting of 27 participants (N = 27, 10 males and 17 females, with a mean age of 26.5 ± 4, within a range of 21 to 35 years), was instructed to "read and freely explore the phototext in its entirety." The second group, consisting of 21 participants (N = 21, 9 males and 12 females, with a mean age of 26.2 ± 3.3, within a range of 22 to 33 years), was instructed to "freely explore and read the phototext in its entirety" to control for potential effects related to the order of instructions.

All participants were right-handed, as determined by the Italian version of the Edinburgh Handedness Inventory[37]. Written informed consent was obtained from all participants before the experiment commenced. The study received approval from the local ethics committee, "Comitato Etico Area Vasta Emilia Nord" (85/2019/DISP/UNIPR, 12/03/2019), and was conducted in compliance with the 1964 Declaration of Helsinki, including its later amendments and comparable ethical standards[38].

Stimuli

The experimental stimuli consisted of a set of phototexts by the Italian author Lalla Romano. Specifically, these phototexts were selected from the corpus of her work *Romanzo di figure* (Einaudi, 1986), which itself was a revised republication of her first phototext, *Lettura di un'immagine* (Einaudi, 1975). In *Lettura di un'immagine*, the author arranged her comments on the left-hand page, with corresponding photos on the right-hand page. In contrast, *Romanzo di figure* and *Nuovo romanzo di figure* featured an inverted layout, with texts appearing on the right-hand page and images on the left-hand page. While the textual content of 17 phototexts was modified lexically, including occasional additions, removals of individual words, or entire sentences, these modifications did not alter the core meaning of the texts.

The photos presented in the selected phototexts were originally taken by the author's father, Roberto Romano, mostly between 1904 and 1914 in Demonte, the author's hometown. All photographs were digitized from the original edition of *Romanzo di figure* (Einaudi, 1986) in high resolution (3600 × 2700 px). To maintain viewing times below 20 s per phototext, we selected 15 phototexts whose textual portions were sufficiently brief (mean character count without spaces = 172.6 ± 41.5). All photos were black-and-white, shared the same aspect ratio (3:4), and were counterbalanced based on their orientation (landscape: n = 8, portrait: n = 7; $\chi^2_{(1)} = 0$, $p > 0.05$). The images were also counterbalanced in terms of their content ($\chi^2_{(1)} = 0.27$, $p > 0.05$): some contained only landscapes (n = 6), while others depicted portraits of people (n = 9).

The phototexts were presented digitally across two Lateralization condition levels: one where the text was placed on the right side and the image on the left (Left Photo, LP), and the other where the image appeared on the right and the text on the left (Right Photo, RP). This manipulation was designed to investigate potential differences in attentional deployment depending on the spatial arrangement of text and image.

The textual component was formatted such that it covered an area of the screen similar to that covered by the corresponding image resulting in a balanced layout with the photo and the text juxtaposed side by side against a neutral grey (RGB: 128, 128, 128) background. The full set of experimental stimuli is publicly available via the Open Science Framework (OSF) repository for this project (please see the Data Availability Statement for access details). This formatting was implemented to ensure visual balance between the image and text components, preventing any disproportionate allocation of screen space that might inadvertently influence the participants' gaze behavior. By maintaining a consistent and proportional visual presentation, we aimed to isolate the effects of layout positioning without introducing confounds related to the relative size or prominence of the stimuli on the screen. We thus obtained 30 stimuli, with 15 in each experimental condition.

Apparatus

The experimental setup utilized a 4 K Ultra HD screen (28 inches; 39.3 cm × 65.7 cm) with a resolution of 3840 × 2160 pixels and a luminance of 30 lumens. The screen was positioned 60 cm from the participants to ensure an ideal viewing distance, and a chin rest was used to maintain consistent positioning and distance of participants relative to the screen.

Eye-tracking data was recorded using a Tobii Pro Eye-Tracker X3-120, which was mounted on the display and operated at a sampling frequency of 120 Hz. A dual-monitor configuration was employed: the primary monitor presented the experimental stimuli to participants, while a secondary monitor allowed the experimenter to observe participants' eye movements in real-time. The eye-tracking device was positioned beneath the primary screen to capture accurate gaze data without obstructing the visual field.

The stimuli were presented in full screen using Psychtoolbox-3, a toolbox implemented in MATLAB[39], which provided precise control over stimulus timing and presentation parameters. Eye movement recording and preprocessing were carried out using the MATLAB Titta toolbox[40], which provides a wrapper around the Tobii Pro SDK. Titta was specifically used for recording, as well as calibrating and validating participant gaze data.

Procedure

The experimental procedure began with an assessment session conducted one day before the experimental task. During this session, participants were asked to complete a series of questionnaires via Google Forms. The assessment included the Interpersonal Reactivity Index (IRI) to measure empathic traits[41], the Immersive Tendencies Questionnaire (ITQ) to assess participants' tendency to become immersed in experiences[42], the State-Trait Anxiety Inventory (STAI-Y) to evaluate trait anxiety levels[43], and the Edinburgh Handedness Inventory to determine hand dominance[44].

Before the experimental task, participants underwent a detailed calibration process to ensure reliable eye-tracking data. Participants were guided to align their head position using a chin rest and a reference target displayed on the screen as a blue circle. Calibration involved presenting a moving dot, which participants were instructed to follow with their eyes. This dot stopped at five predefined locations on the screen in a randomized sequence, where participants were required to fixate precisely. The validation process followed a similar protocol, with fixation points presented in a four-point diamond-shaped pattern to verify data accuracy across the viewing area. Calibration quality was assessed based on accuracy and precision, with acceptable calibration requiring an average accuracy deviation of less than 1° of visual angle and a precision (standard deviation) of less than 0.1° of visual angle. This setup allowed for the assessment of data quality by measuring accuracy and precision across the entire calibration space, thereby ensuring the reliability of the gaze tracking data collected.

As detailed in the Participants section, instructions were provided differently for each group. The phototexts were introduced as "literary works consisting of both a textual and photographic component." Participants in the first group were instructed to "read and freely explore the phototext in its entirety," while participants in the second group were instructed to "freely explore and read the phototext in its entirety." This difference was designed to eliminate any potential effects arising from the order of instructions.

During the experimental task, participants viewed 30 phototext stimuli in a randomized order. The experiment was divided into two blocks of 15 trials each, with a short break between blocks, and had a total duration of approximately 20 min. Each trial began with the presentation of a black fixation cross for 1.5 s on a gray background (RGB: 128, 128, 128), followed by the phototext. The presentation duration for each stimulus was held constant at 20 s to ensure that all eye-tracking metrics were collected under uniform temporal conditions, thereby eliminating viewing time as a potential confounding variable in the analysis. The fixation cross was presented randomly on either the left or right side within the half of the screen corresponding to the area where the photo or text of the stimulus would appear. This design choice was implemented to control participants' initial gaze location and prevent artifacts that could arise from a consistently fixed gaze starting location, thereby ensuring a balanced and unbiased exploration of the phototexts. After each stimulus, participants answered three questions presented in random order: (a) Valence: "How would you rate the emotional valence of the phototext?" (b) Arousal: "How would you rate the emotional intensity of the phototext?" (c) Self-Reported Attentional Focus: "What attracted your attention the most?" Participants responded using a Visual Analog Scale (VAS). For Valence, the scale ranged from -50 (negative) to 50 (positive). For Arousal, it ranged from 1 (low intensity) to 100 (high intensity). For Self-Reported Attentional Focus, it ranged from -50 (indicating focus on the photo) to 50 (indicating focus on the text). Participants were given 5 s to respond to each question.

Prior to the experimental task, participants completed a training session involving two trials, one for each experimental condition, with images not included in the experimental stimuli set. This training was designed to familiarize them with the experimental procedure and the response process. The experiment was conducted in a dark room to maintain controlled lighting conditions and minimize external distractions.

At the end of the experimental task, participants were asked a series of questions about their experience with the phototexts. These included: "Have you ever read a phototext before?", "Did you feel that the time provided was sufficient to complete the required tasks during the experiment?", and "In general, did you prefer the textual or photographic component?" Additionally, participants were shown the phototexts again and asked to rate their Enjoyment of each one by answering the question: "How much did you enjoy this phototext?" on a Likert scale from 1 (not at all) to 10 (very much).

Behavioral analysis

Behavioral analyses were conducted to determine how participants explicitly evaluated the phototexts. Behavioral responses to the question regarding what aspect of the phototext attracted participants' attention the most, whether the photo or the text, were analyzed using linear mixed-effects models. This approach allowed us to account for both fixed and random effects. Model fitting followed a hierarchical approach, beginning with a simple model and adding parameters sequentially to improve model fit. Likelihood ratio tests, along with Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC), were used to assess whether the inclusion of fixed effects, random effects, and interaction terms significantly improved the model. In the statistical models, Self-Reported Attentional Focus scores (i.e., whether participants focused more on the photo or the text) were entered as the dependent variable. Negative values indicate an attentional focus towards the Photo, whereas positive values indicate an attentional focus towards the Text. Scores for Valence and Arousal were transformed into categorical variables using the median as a cut-off (Valence MED = 52.3, Arousal Median = 1.5), resulting in two levels for each variable: High and Low. Specifically, 47.47% of the scores were categorized as High scores in Valence, while 48.75% were categorized as High scores in Arousal. Thus, the fixed independent variables included Valence (two levels: High, Low) and Arousal (two levels: High, Low). The random effect structure included intercepts by participant, with Valence and Arousal also included as random slope.

To assess the presence of influential outliers, the best fitting model obtained with the hierarchical approach was used to compute Cook's distances for each observation, and outliers were identified as those exceeding four times the mean Cook's distance. Approximately 5.13% of the data were identified as influential outliers and subsequently removed from the dataset for further analysis. A plot of Cook's distances was also generated

to visually inspect the distribution of influential observations. Model assumptions, such as normality and homoscedasticity, were evaluated by inspecting residual plots to ensure the validity of the inferences drawn from the models. Tukey's test was used for post hoc comparisons among means when significant main effects or interactions were identified.

To ensure that the data were not biased by the emotional content of the stimuli, which was not initially accounted for during stimulus selection, we verified that the distribution of participant scores for Valence and Arousal did not exhibit skewness. D'Agostino's K-squared Test was applied to evaluate whether the skewness of these scores significantly deviated from a normal distribution[45].

Additionally, we examined the responses given by participants to the question "In general, did you prefer the textual or photographic component?" and we performed a χ^2 test of proportions to compare the observed proportion to the expected proportion (H_0: 50%) and establish whether there was a clear preference towards a specific component of the phototext that can explain difference in attentional allocation.

Eye-tracking analysis

The eye-tracking analysis aimed to investigate potential lateralization biases and to determine whether visual exploration patterns were modulated by the different experimental conditions. To this end, each phototext was divided into two identical and symmetrical Areas of Interest (AOIs): one corresponding to the textual part and the other to the photographic part of the phototext. These AOIs were used to analyze the gaze data in a systematic manner. For each stimulus used in the experiment, a visualization with AOIs visually superimposed has been made publicly available via the Open Science Framework (OSF) repository for this project (please see the Data Availability Statement for access details).

For fixation classification, we used the Identification by Two-Means Clustering (I2MC) algorithm[46], which is well-suited for processing data across a range of noise levels, including when periods of data loss may occur. This algorithm provided robust fixation classification, even under challenging conditions.

We extracted the latency of the first fixation directed towards each AOI, as well as the mean number of fixations per AOI as parameters of interest. Latency of first fixation measures the time elapsed (in milliseconds) from stimulus onset to the initiation of the first fixation within the boundaries of a specific area of interest (AOI). This metric serves as a primary indicator of attentional capture and visual salience. Shorter latencies suggest that an AOI possesses high salience, capturing attention rapidly through predominantly automatic, bottom-up attentional processes. Conversely, longer latencies may indicate reduced salience of the target AOI[8]. Mean number of fixations per AOI represents the average count of fixations directed toward a given AOI during the viewing period, providing a measure of cognitive effort and processing depth. A higher number of fixations within an AOI suggests that the contained information is complex, ambiguous, or difficult to encode, thereby requiring more detailed processing and imposing greater cognitive load. In contrast, fewer fixations indicate that the information is processed more efficiently or is deemed less relevant to the experimental task[7]. These visual metrics were analyzed using linear mixed-effects models, allowing us to account for both fixed and random effects. For each parameter, model fitting followed a hierarchical approach, starting with a simple model and iteratively adding parameters to improve model fit. Likelihood ratio tests, Akaike Information Criterion (AIC), and Bayesian Information Criterion (BIC) were used to evaluate whether the inclusion of fixed effects, random effects, and interaction terms significantly improved model fit.

In the statistical models, each visual parameter was entered as the dependent variable. Scores for Self-Reported Attentional Focus were converted into categorical variables by dividing the original response scale (from -50, focus on the photo, to 50, focus on the text) into three symmetrical segments. This transformation yielded three distinct categories: Text, Photo, and Equal, where Text designates that participants focused more on textual content, Photo designates that participants focused more on photographic content, and Equal designates that participants focused approximately equally on both text and photo. Specifically, 32.84% of the scores were categorized as Equal, 43.6% as Photo, and 23.5% as Text.

In the latency of the first fixation model the fixed independent variables included AOI (two levels: Text, Photo), Lateralization (two levels: Right Image, Left Image) and Self-Reported Attentional Focus (three levels: Photo, Text, Equal). The random effect structure included intercepts by participant, with AOI also included as random slope. In the number of fixations model the fixed independent variables included AOI (two levels: Text, Photo) and Self-Reported Attentional Focus (three levels: Photo, Text, Equal). The random effect structure included intercepts by participant, with AOI also included as random slope.

To assess the presence of influential outliers, the best fitting models obtained with the hierarchical approach were used to compute Cook's distances for each observation, and outliers were identified as those exceeding four times the mean Cook's distance. Approximately 5.85% of the data in the latency of the first fixation model and 5.92% of the data in the number of fixations model were identified as influential outliers and subsequently removed from the dataset for further analysis. Plot of Cook's distances were also generated to visually inspect the distribution of influential observations. Model assumptions, such as normality and homoscedasticity, were evaluated by inspecting residual plots to ensure the validity of the inferences drawn from the models. Tukey's test was used for post hoc comparisons among means when significant main effects or interactions were identified.

To gain a deeper understanding of the temporal dynamics of visual exploration and examine how participants' visual attention evolved throughout the viewing of each phototext, we performed an additional analysis examining fixation density over time across both AOIs. This time-course analysis differs from spatial fixation density (which describes the concentration of fixations within an AOI) and instead focuses on when attention is directed to an AOI. This temporal analysis differs fundamentally from spatial fixation density, which quantifies the concentration of fixations within a given AOI, by instead focusing on the timing of attentional allocation to specific regions. Specifically, we divided the fixation density of each AOI into 20 time slices, each representing 1 s, to plot the fixation frequency in each AOI across the entire 20-s viewing duration[47], while also including the

variable Lateralization (two levels: Right Image, Left Image). This approach allowed us to capture fluctuations in attention allocation between the textual and photographic elements over time, highlighting both the temporal component and fixation density. These results are particularly relevant in relation to the two models examining fixation latency and the number of fixations, as they provide a complementary visualization of the dynamic nature of visual engagement and any emerging temporal patterns that could indicate shifts in attentional focus during the viewing period.

Correlation analysis

To better characterize the oculomotor behavior and the explicit judgments given by participants on the phototexts, the strength and direction of associations between variables of interest were assessed using Kendall's τ non-parametric correlation coefficient. This correlation method was used due to its robustness against outliers, its lack of assumptions regarding normality or equal variances, and its effectiveness in handling ties within the data[48] .

We specifically tested the correlation between the latency of first fixations directed towards each AOI and participants' scores on the STAI-Y2 questionnaire, which measures trait anxiety. This analysis aimed to determine the extent to which trait anxiety might have influenced oculomotor behavior, especially given that the viewing time for the phototexts was limited to 20 s. Additionally, we examined the relationship between Enjoyment scores attributed to the phototexts and participants' responses to the Self-Reported Attentional Focus question, which asked which aspect of the phototext (text or photo) attracted their attention the most. Both scores were further correlated with participants' results on the ITQ, measuring the capacity or tendency to become immersed in mediated environments, and the Fantasy Scale of the IRI (IRI-FS), assessing the propensity to imaginatively engage with fictional characters in books, movies, and plays. These analyses were conducted to explore potential associations between immersive tendencies and participants' engagement with the phototext components.

To identify multivariate outliers, we computed the Mahalanobis distance and then we used the alpha quantile of the χ^2 distribution set at $k = 2$ degrees of freedom and an alpha threshold of 0.025 to determine the cutoff value. Data points associated with a Mahalanobis distance above this cutoff value, which deviated significantly from the rest of the dataset, were excluded from the analysis.

All data analyses were conducted using R[49]. The lme4 package[50] was used for mixed-effects modeling and additional packages were used for model diagnostics[51], post hoc analysis[50,52] and data visualization[53,54].

Results

Behavioral results

Self-reported attentional focus model

The model explained 49% of the variance in dependent variable considering the random effects ($R^2_m = 0.09$; $R^2_c = 0.49$). The model revealed a significant main effect of Arousal ($F_{(1, 46.4)} = 12.7$, $p < 0.001$), indicating that participants reported that the Photo attracted more their attention when the stimuli were rated with High Arousal compared to those rated with Low Arousal ($t_{(46.3)} = -3.5$, $p < 0.001$) (High Arousal: M = -13, CIs = -19, -7; Low Arousal: M = -0.1, CIs = -6.5, 6.2). The model also revealed a significant main effect of Valence ($F_{(1, 47.1)} = 10.6$, $p < 0.001$) indicating that participants reported that the Photo attracted more their attention when the stimuli were rated with High Valence compared to those rated with Low Valence. ($t_{(46.8)} = -3.2$, $p < 0.001$) (High Valence: M = -10.1, CIs = -16, -5.5; Low Valence: M = -2.4, CIs = -8.4, 3.6). Additionally, the model revealed a significant Arousal by Valence interaction effect ($F_{(1, 2109)} = 19.1$, $p < 0.001$). Post-hoc comparisons (Fig. 1) revealed that when Arousal scores were High, Photo attracted more the attention of participants in High Valence compared to the Low Valence ($t_{(62.2)} = -4.7$, $p < 0.001$). In contrast, when Arousal scores were Low there was no difference in attention allocation between High and Low Valence. Furthermore, in both High ($t_{(55.5)} = -4.6$, $p < 0.001$) and Low ($t_{(55.5)} = -2.2$, $p < 0.05$) Valence, Photo attracted more the attention of participants in High Arousal than in Low Arousal (High Arousal – High Valence: M = -19.5, CIs = -25.8, -13.5; Low Arousal – High Valence: M = -2, CIs = -8.6, 4.6; High Arousal – Low Valence: M = -6.5, CIs = -13.4, 0.5; Low Arousal – Low Valence: M = 1.7, CIs = -5.5, 8.8).

The results of D'Agostino's K-squared Test indicated no significant deviation from normality, suggesting that the distribution of participant scores for Valence ($K^2 = 0.06$, $p > 0.05$) and Arousal ($K^2 = -0.08$, $p > 0.05$) did not exhibit skewness. Also, the results of the χ^2 proportion test indicated no difference ($\chi^2_{(1)} = 1$, $p > 0.05$,) between the observed proportion of phototext component preference (Photo: 41.7%) and the expected proportion (50%).

Eye-tracking analysis results

Latency of first fixations model

The model explained 59% of the variance in dependent variable taking into account the random effects ($R^2_m = 0.59$; $R^2_c = 0.18$). The model revealed a significant main effect of AOI ($F_{(1, 46.2)} = 32.9$, $p < 0.001$), indicating that latency of first fixations directed at Text AOI on average were shorter than mean latency of first fixations directed at Photo AOI ($t_{(47.1)} = 5.7$, $p < 0.001$) (Photo: M = 3.1, CIs = 2.4, 3.9; Text: M = 0.7, CIs = 0.6, 0.9). The model also revealed a significant main effect of Lateralization ($F_{(1, 2397.5)} = 23.4$, $p < 0.001$) indicating that latency of first fixations were shorter when the Photo were presented on the Left ($t_{(2399.5)} = -4.8$, $p < 0.001$) (LP: M = 1.8, CIs = 1.4, 2.1; RP: M = 2.1, CIs = 1.8, 2.5) compared to when the Photo were presented on the Right. Furthermore, the model revealed a significant main effect of Self-Reported Attentional Focus ($F_{(2, 1549.8)} = 15.7$, $p < 0.001$) indicating that the latency of first fixations were shorter when participants later reported that the textual component had attracted their attention the most , compared to when they later reported that the photographic content had attracted their attention the most ($t_{(2125.9)} = -5.4$, $p < 0.001$), or when they later reported that they distributed their attention approximately equally between text and photos ($t_{(1437.6)} = -4.1$, $p < 0.001$) (Equal: M = 1.8, CIs = 1.4, 2.2;

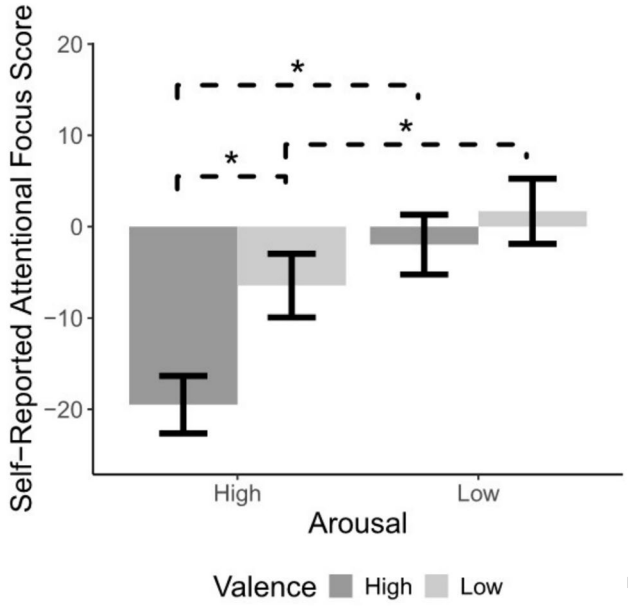

Fig. 1. Self-Reported Attentional Focus Model: Mean fitted effect values bar plots for Arousal by Valence interaction. Error bars indicate the standard errors (SE). *$p < 0.05$.

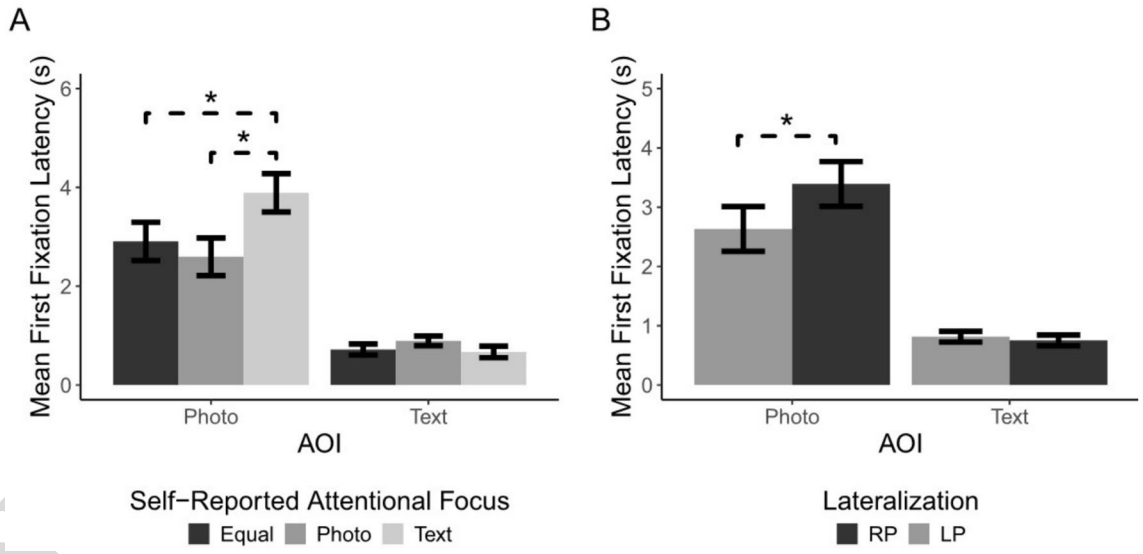

Fig. 2. Latency of first fixations Model: Mean fitted effect values bar plots for (**A**) AOI by Self-Reported Attentional Focus interaction and (**B**) AOI by Lateralization interaction. Error bars indicate the standard errors (SE). The y-axis represents the mean time in seconds from stimulus onset to the first fixation within each Area of Interest (AOI). Please note that all difference between Text AOI and Photo AOI are statistically significative ($p < .001$) and only additional significant difference among Lateralization and Self-Reported Attentional Focus levels are shown for clarity reasons. *$p < 0.05$.

Photo: M = 1.7, CIs = 1.4, 2.1; Text: M = 2.3, CIs = 1.9, 2.6). There was no difference between when participants focused primarily on photographic content compared to when they distributed their attention approximately equally between text and photos.

The model revealed a significant AOI by Self-Reported Attentional Focus interaction effect ($F_{(2, 2036.7)} = 28.9$, $p < 0.001$). Post Hoc comparisons (Fig. 2.A) revealed that independently from the Self-Reported Attentional Focus categories (Equal, Photo, Text) latency of first fixations directed at Text AOI on average were shorter than mean latency of first fixations directed at Photo AOI (Phot vs. Text Post Hoc Comparisons in Self-Reported Attentional Focus Equal: $t_{(57.1)} = 5$, $p < 0.001$; Self-Reported Attentional Focus Photo: $t_{(51.9)} = 4$, $p < 0.001$; Self-Reported Attentional Focus Text: $t_{(58.2)} = 7.4$, $p < 0.001$). Furthermore, only in Photo AOI latency of first fixations

were shorter when participants focused primarily on photographic content ($t_{(2423)} = -8.6$, $p < 0.001$) or distributed their attention approximately equally between text and photos ($t_{(2437.7)} = -5.5$, $p < 0.001$), compared to when they focused predominantly on text content (Photo AOI – Self-Reported Attentional Focus Equal: M = 2.9, CIs = 2.1, 3.7; Self-Reported Attentional Focus Photo: M = 2.6, CIs = 1.8, 3.6; Self-Reported Attentional Focus Text: M = 3.9, CIs = 3.1, 4.7). There was no difference between when participants focused primarily on photographic content compared to when they distributed their attention approximately equally between text and photos. Also, no specific differences emerged between Self-Reported Attentional Focus categories in the Text AOI (Text AOI – Self-Reported Attentional Focus Equal: M = 0.7, CIs = 0.5, 0.9; Self-Reported Attentional Focus Photo: M = 0.9, CIs = 0.7, 1.1; Self-Reported Attentional Focus Text: M = 0.7, CIs = 0.4, 0.9). Additionally, the model revealed a significant AOI by Lateralization interaction effect ($F_{(1, 2395.6)} = 32.3$, $p < 0.001$). Post Hoc comparisons (Fig. 2.B) revealed that independently from the Lateralization categories (LP, RP) latency of first fixations directed at Text AOI on average were shorter than mean latency of first fixations directed at Photo AOI (Phot vs. Text Post Hoc Comparisons in Lateralization LP: $t_{(50)} = 4.7$, $p < 0.001$; Lateralization RP: $t_{(50)} = 6.6$, $p < 0.001$). Furthermore, only in Photo AOI latency of first fixations were shorter when the Photo were presented on the Left ($t_{(2395.9)} = -7.2$, $p < 0.001$) (Photo AOI – LP: M = 2.7, CIs = 2, 3.5; RP: M = 3.5, CIs = 2.7, 4.3) compared to when the Photo were presented on the Right. No specific differences emerged between Self-Reported Attentional Focus categories in the Text AOI (Text AOI – LP: M = 0.8, CIs = 0.6, 1; RP: M = 0.7, CIs = 0.5, 0.9).

Number of fixations model

The model explained 79% of the variance in dependent variable considering the random effects ($R^2_m = 0.79$; $R^2_c = 0.68$). The model revealed a significant main effect of AOI ($F_{(1, 46.5)} = 423.7$, $p < 0.001$), indicating that number of fixations directed at Text AOI was higher than that directed at Photo AOI ($t_{(47.1)} = 20.6$, $p < 0.001$) (Photo: M = 18.6, CIs = 17.5, 19.6; Text: M = 45.9, CIs = 43.8, 47.9). The model also revealed a significant main effect of Self-Reported Attentional Focus ($F_{(2, 1902.2)} = 5.8$, $p < 0.01$) indicating that number of fixations were lower when participants later reported that the photographic content had attracted their attention the most, compared to when they later reported that the textual component had attracted their attention the most ($t_{(2345.8)} = -2.9$, $p < 0.01$), or when they later reported that they distributed their attention approximately equally between text and photos ($t_{(1749.6)} = -2.6$, $p < 0.05$) (Equal: M = 32.6, CIs = 31.5, 33.6; Photo: M = 31.4, CIs = 30.5, 32.4; Text: M = 32.7, CIs = 31.6, 33.7). There was no difference between when participants focused primarily on textual content compared to when they distributed their attention approximately equally between text and photos.

The model revealed a significant AOI by Self-Reported Attentional Focus interaction effect ($F_{(2, 2374.4)} = 139.4$, $p < 0.001$). Post Hoc comparisons (Fig. 3) revealed that independently from the Self-Reported Attentional Focus categories (Equal, Photo, Text) number of fixations directed at Text AOI was higher than that directed at Photo AOI (Phot vs. Text Post Hoc Comparisons in Self-Reported Attentional Focus Equal: $t_{(65.6)} = -19.4$, $p < 0.001$; Self-Reported Attentional Focus Photo: $t_{(55.9)} = -14.5$, $p < 0.001$; Self-Reported Attentional Focus Text: $t_{(66.4)} = -23.3$, $p < 0.001$). Furthermore, only in Photo AOI number of fixations were higher when participants focused primarily on photographic content ($t_{(1637.5)} = 4.2$, $p < 0.001$) or distributed their attention approximately equally between text and photos ($t_{(2301)} = 9.8$, $p < 0.001$), compared to when they focused predominantly on text content. Also in Photo AOI, number of fixations were higher when participants focused primarily on photographic content ($t_{(1578.5)} = -4.7$, $p < 0.001$), compared to when distributed their attention approximately equally between text and photos (Photo AOI – Self-Reported Attentional Focus Equal: M = 18.6, CIs = 17.3, 19.8; Self-Reported

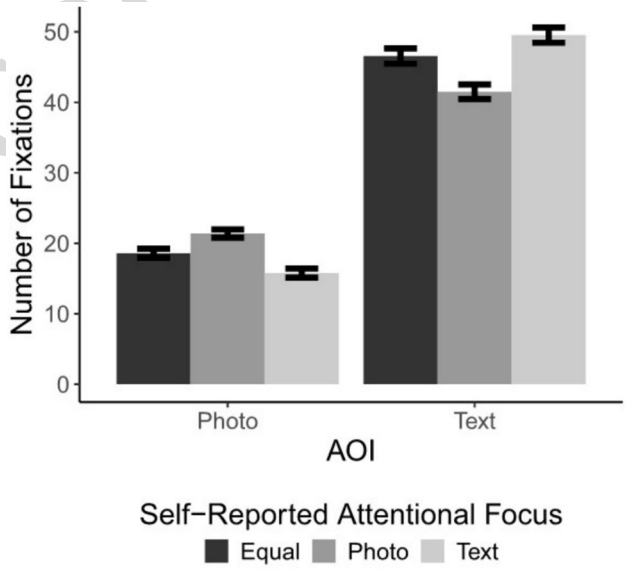

Fig. 3. Number of fixations Model: Mean fitted effect values bar plots for AOI by Self-Reported Attentional Focus interaction. Error bars indicate the standard errors (SE). Please note that all differences are statistically significative ($p < .05$).

Attentional Focus Photo: M = 21.4, CIs = 20.2, 22.5; Self-Reported Attentional Focus Text: M = 15.8, CIs = 14.5, 17.1). In contrast, in Text AOI number of fixations were higher when participants focused primarily on text content compared to when they focused predominantly on photographic content ($t_{(2439.9)} = -13.2$, $p < 0.001$) or distributed their attention approximately equally between text and photos ($t_{(2398.1)} = -4.1$, $p < 0.001$). Also in Text AOI, number of fixations were higher when participants distributed their attention approximately equally between text and photos ($t_{(2397.9)} = 7.7$, $p < 0.001$), compared to when they focused primarily on photographic content (Text AOI – Self-Reported Attentional Focus Equal: M = 46.6, CIs = 44.4, 48.7; Self-Reported Attentional Focus Photo: M = 41.5, CIs = 39.4, 43.6; Self-Reported Attentional Focus Text: M = 49.5, CIs = 47.4, 51.7).

Fixation density over time

In addition to the main latency of the first fixation and number of fixations models, we examined the fixation density across the 20-s viewing period by dividing each AOI into 20 time slices. This analysis revealed distinct temporal patterns in the distribution of visual attention between the Text and Photo AOIs. As illustrated in Fig. 4, the fixation frequency varied dynamically over time, suggesting notable shifts in attentional focus during the viewing process. It appears that fixation density in the initial time slices was considerably higher in the Text AOI, followed by a gradual increase in fixation density in the Photo AOI starting from the second half of the viewing period. The only deviation from this general time pattern occurred in the first time slice of the Photo AOI when the Photo was presented on the left side, where the fixation density was notably higher compared to when the Photo was presented on the right side. This pattern highlights an evolving interaction with the phototext components.

Correlation results

Kendall's τ values revealed no significant correlations between the latency of first fixations directed towards each AOI and STAI-Y2 scores (Photo AOI: $\tau = 0.02$, $p > 0.05$; Text AOI: $\tau = 0.03$, $p > 0.05$), indicating that trait anxiety did not influence participants' oculomotor behavior. Similarly, Kendall's τ values showed no significant correlations between Enjoyment and Self-Reported Attentional Focus scores ($\tau = -0.02$, $p > 0.05$), indicating that the degree of Enjoyment was not associated with whether participants focused on the text or the photo.

Furthermore, Kendall's τ values revealed a moderate to strong positive correlation between Enjoyment and ITQ scores (Fig. 5.A, $\tau = 0.16$, $p < 0.001$) and between Enjoyment and IRI-FS scores (Fig. 5.B, $\tau = 0.38$, $p < 0.001$), suggesting that higher levels of Enjoyment were associated with a greater capacity for participants to become immersed in mediated environments or imaginatively engage with fictional characters in books. In addition, Kendall's τ values revealed a small negative correlation between Self-Reported Attentional Focus and IRI-FS scores (Fig. 5.C, $\tau = -0.09$, $p < 0.001$), indicating that greater attentional focus on the photo (Self-Reported Attentional Focus negative scores) were associated with a greater capacity for participants to engage with fictional characters.

Discussion

The present study provides novel insights into the influence of spatial layout configuration, emotional content and aesthetic enjoyment on participants' visual exploration of phototexts. Additionally, we aimed to assess the presence of left-gaze bias in the context of stimuli composed of text and photo, which were systematically presented in a counterbalanced manner on the right and left sides to ensure balanced exposure and mitigate potential positional biases. To this end, eye movements were recorded while participants evaluated the Emotional Intensity, Valence, and Self-Reported Attentional Focus attentional focus arising from their engagement with the phototexts. It is worth noting that almost all participants who took part in the study had no prior knowledge of or experience with phototexts. Our findings reveal compelling aspects of oculomotor behavior, as well as self-reported explicit judgments and preferences regarding phototext perception.

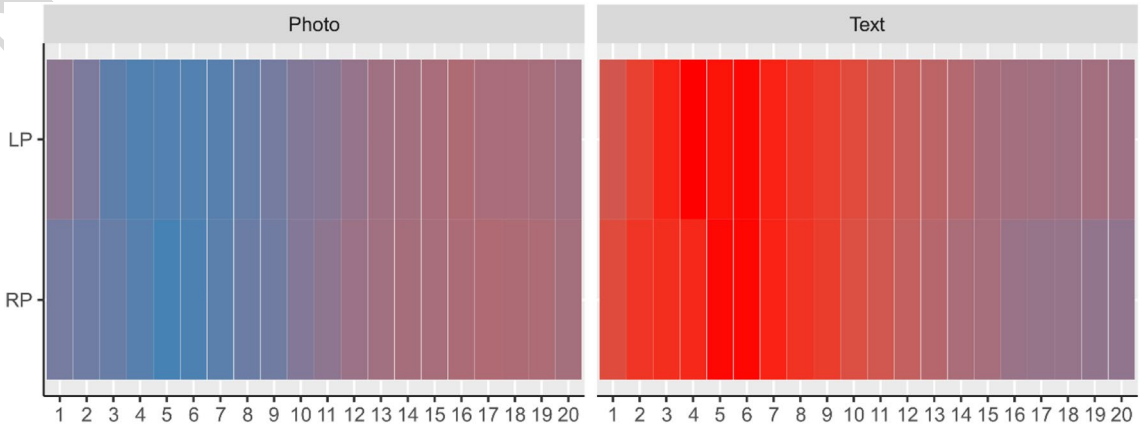

Fig. 4. Fixations' frequency in each AOI over time: fixations density was divided into 20-time slices (i.e. 1000 ms).

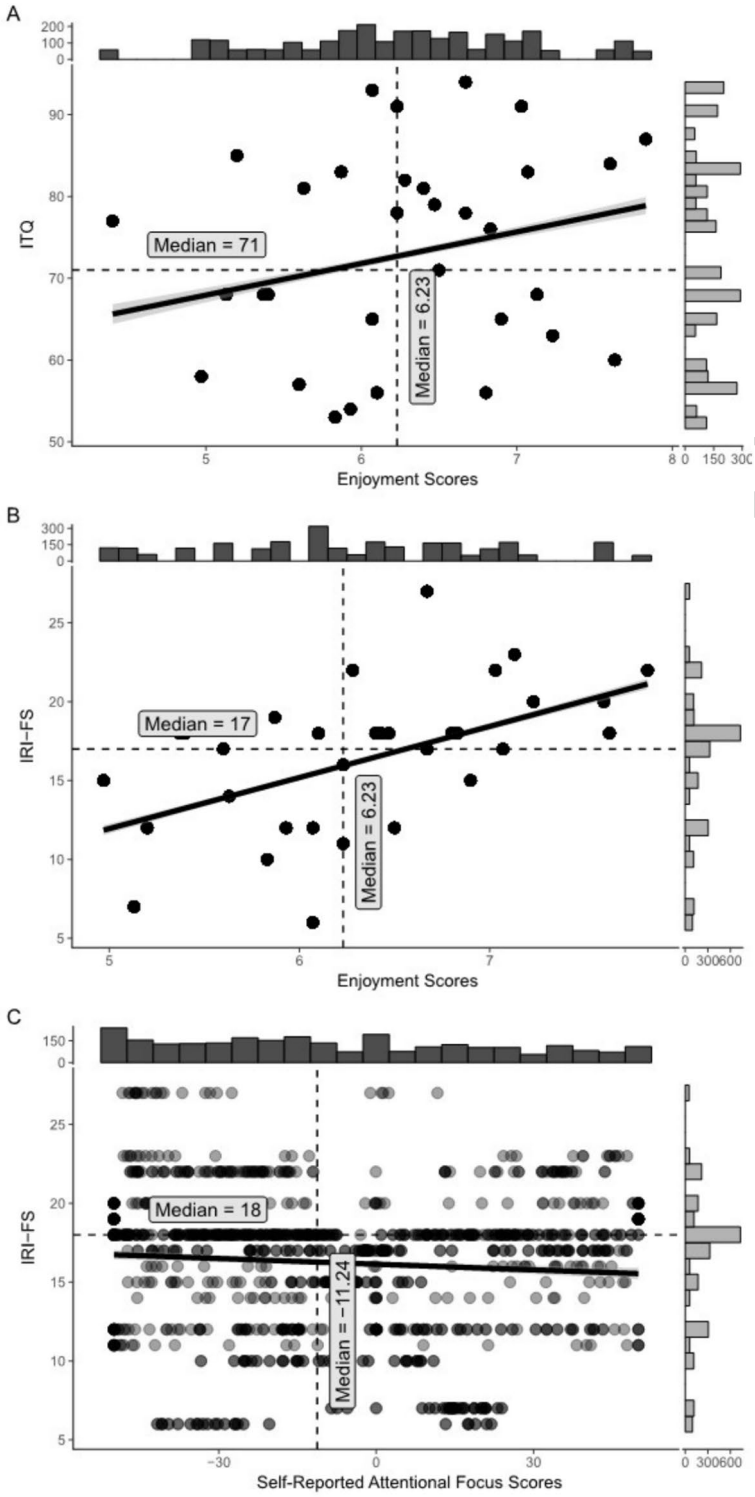

Fig. 5. Scatterplots of significant correlations between (**A**) Enjoyment and ITQ scores, (**B**) between Enjoyment and IRI-FS scores, and (**C**) between Self-Reported Attentional Focus and IRI-FS scores. A black linear regression line is overlaid to represent the trend, with 95% confidence intervals shaded in light gray. Dashed vertical and horizontal lines indicate the median values for each variable. The side panels display histograms representing the distribution of each variable.

In terms of Self-Reported Attentional Focus, the behavioral results highlighted a notable interaction between Arousal and Valence, suggesting that participants' attentional allocation to the photographic component was influenced by the emotional characteristics of the stimuli. Specifically, the photographic component of the phototext attracted more attention from participants when the phototexts were rated higher in arousal, regardless

of whether they were high or low in valence. This effect was particularly pronounced when valence ratings were also highly positive. Notably, this finding holds true even considering that the distribution of stimuli rated with negative and positive valence was balanced.

It is particularly noteworthy that no significant correlation was observed between participants' attentional focus on the text or photo and their explicit Enjoyment ratings and participants did not explicitly prefer either component of phototexts. This finding suggests that the subjective enjoyment of phototexts may not depend on which component dominates attentional focus, but rather on an integrated or holistic experience of the phototext as a unified stimulus. Additionally, our evaluation of immersive tendencies and empathy-related traits, such as the ability to imaginatively engage with fictional characters, provided further insight into individual differences that may shape phototext engagement. The significant positive correlations between Enjoyment and both ITQ and IRI-FS scores suggest that participants with a greater tendency to immerse themselves in different environments or project themselves into the experiences of others derived higher levels of enjoyment from the phototexts. Notably, participants with a greater capacity for immersion in mediated environments also exhibited a stronger attentional focus on the photographic component. This relationship highlights the potential for participants with higher immersive tendencies to engage more deeply with the visual elements of phototexts, which may evoke a richer and more emotionally engaging experience.

To investigate whether there was a left-gaze bias and a modulation of visual exploration patterns, we analyzed the latency of the first fixation, as well as the mean number of fixations, directed at Text and Photo AOIs. The eye-tracking data indicated a clear difference in attentional allocation depending on the components of the phototexts. The latency of the first fixation was significantly shorter in the Text AOI compared to the Photo AOI, suggesting a natural reading bias where participants are quicker to orient towards textual elements, regardless of their position on the left or right portion of the phototext. This finding aligns with a large body of research on multimodal learning, which often demonstrates a text-driven processing bias where viewers prioritize textual information to build an initial semantic framework[33,35]. This behavior is also consistent with the sequential pattern identified by Rayner and colleagues[8], where readers first engage with text, especially larger print, then smaller print, before examining accompanying images. However, our data also revealed that fixation latencies were modulated by lateralization, with shorter latencies observed when photos were presented on the left side compared to the right. This left-gaze bias emerged only when comparing the Photo components; otherwise, latencies at the Text AOI were consistently the shortest. This finding can be interpreted through the lens of the Spatial Contiguity Principle[36]. While text holds primary attentional priority, the principle suggests that viewers naturally attempt to integrate adjacent information. The shorter latency for left-side photos may indicate that this secondary integration process is facilitated when it aligns with the culturally ingrained left-to-right processing vector and left side bias, even when the primary target (text) is on the opposite side.

This apparent discrepancy between participants' explicit reports and their initial oculomotor behavior represents one of the most compelling findings of our study: a clear dissociation between implicit attentional orienting and explicit judgments of engagement. Our eye-tracking data consistently revealed a text-driven processing bias, where initial fixations were significantly faster to the textual AOI. This likely reflects a deeply rooted and automatic oculomotor pattern tied to reading habits, where the visual system prioritizes the decoding of linguistic information as an entry point to the stimulus. In stark contrast, participants' explicit judgments were heavily modulated by the affective properties of the phototext, with the photographic component being reported as the primary focus of attention when perceived as high in emotional arousal and valence. Taken together, these findings suggest a two-stage process of engagement with phototexts. The initial phase is characterized by a rapid, automatic orienting to text, driven by cognitive habits. This is followed by a more sustained, evaluative phase where the affective and semantic richness of the photograph dominates the enjoyment of the stimulus.

Interestingly, our correlation analysis did not reveal a significant relationship between trait anxiety and the latency of first fixations towards either AOI. This suggests that participants' trait anxiety did not modulate their oculomotor behavior in the context of limited viewing time. Unsurprisingly, the results for the mean number of fixations confirmed a prototypical pattern observed in reading[8]: participants made overall more fixations on the Text AOI than on the Photo AOI, considering all trials, even those where they stated that they focused more on the Photo AOI. These viewing patterns were consistent regardless of the order in which participants were instructed to "read" and "freely explore" the phototext.

The fixation density analysis over time revealed a dynamic evolution of attention between the text and photo components throughout the viewing period. Initially, participants directed more fixations towards the text, but from the mid-point onward, attention shifted more towards the photographic component. This temporal shift in attention suggests an evolving interaction between the text and image, where participants may initially seek contextual information from the text before engaging more deeply with the visual content.

In conclusion, our study provides novel insights into how spatial layout and emotional properties influence visual exploration and enjoyment of phototexts. The findings underscore the importance of considering both cognitive and affective factors when designing mixed-media content, as well as how individual differences in immersive tendencies can shape engagement.

Beyond contributing to the theoretical understanding of multimodal perception, our findings offer several practical applications. The clear link between emotional arousal and attentional focus on the photographic component is particularly relevant for designers of multimodal media. For instance, in advertising or UX design, where guiding a user's gaze to a key visual element is critical, employing images with high emotional arousal could serve as a powerful tool to capture and sustain attention. Similarly, in digital museum exhibits or online storytelling, pairing text with evocative, high-arousal images could enhance users' immersion and engagement with the visual narrative, a finding supported by the correlation we observed between enjoyment and immersive tendencies.

Furthermore, our results underscore that enjoyment is not tied to focusing on one component over the other but rather emerges from an integrated experience. This suggests that the most effective multimodal designs are not those where text and image compete, but where they function synergistically. Designers, authors and artists should therefore focus on the semantic and emotional congruence between text and image to create a unified and more engaging message.

Future research should expand on these findings also by incorporating more varied temporal exposures and examining additional individual-level variables and stimuli characteristics that might influence participants' engagement. For example, one limitation of our study was the relatively brief exposure time to each phototext, which may not fully reflect the naturalistic viewing conditions in which individuals interact with visual and textual materials over longer periods. However, this limitation was necessary for experimental design purposes to ensure consistent data collection across participants and all participants stated after the experiment that the time provided was sufficient to complete the required tasks. Future studies could explore how varying exposure durations, both shorter and longer, affect Self-Reported Attentional Focus and oculomotor behavior, particularly in relation to individual differences such as immersive tendencies.

Additionally, the stimuli used in this study did not have a clearly defined emotional connotation and were not explicitly validated for their emotional content. Nevertheless, participants' self-reported explicit valence ratings were balanced, which supports the reliability of the results. Future research could consider selecting and validating stimuli with distinct emotional characteristics to further investigate their impact on visual engagement and attentional patterns. Another potential avenue for exploration involves testing alternative layout configurations, such as an above-below arrangement, to uncover additional effects of spatial layout[55] on the processing of complex composite stimuli.

Furthermore, cultural influences on reading and visual exploration remain to be disentangled. This study was conducted with participants from Western cultures, characterized by a prototypical left-to-right reading direction. Future studies should aim to include participants from diverse cultural backgrounds to examine how reading direction and cultural context influence attention allocation. From a visual perspective, the phototexts in this study featured black-and-white photos in their original format. Future research could explore the use of phototexts with colored photos to examine whether color enhances participants' immersion and engagement with these types of stimuli. Similarly, our sample consisted of young adults; investigating these patterns in older adults, who may exhibit different oculomotor control, would be a valuable future direction.

Moreover, this study did not investigate the content relationship between text and image. In the phototexts analyzed, Lalla Romano's texts were often descriptive, but this is not always the case in the artistic and literary construction of phototexts by other authors. Exploring how variations in the textual content, descriptive, interpretative, or narrative, interact with photographic elements could offer further insights into the mechanisms driving engagement and interpretation of phototexts.

This study offers a significant contribution by investigating phototexts as complex multimodal stimuli that seamlessly combine textual and photographic elements, while uniquely examining their synergistic effects on attention distribution and perceptual experiences. By exploring these interactions, the study highlights previously unexplored dimensions of cognitive and emotional engagement with such integrated stimuli, setting the stage for future research into the design and interpretation of multimodal content.

Data availability

Stimuli and datasets generated and analyzed during the current study are available in the Open Science Framework (OSF) repository at the following link: https://osf.io/umk5q/?view_only=f6e2e3e3d71a480a812d935b318370bc. Further details regarding data processing and analysis can be found in the repository. Researchers interested in accessing specific data or additional materials are encouraged to consult the repository or contact the corresponding author.

Received: 18 February 2025; Accepted: 23 June 2026

References

1. Mayer, C. W., Rausch, A. & Seifried, J. Analysing domain-specific problem-solving processes within authentic computer-based learning and training environments by using eye-tracking: A scoping review. *Empir. Res. Vocat. Educ. Train.* **15**(1), 2. https://doi.org/10.1186/s40461-023-00140-2 (2023).
2. Just, M. A. & Carpenter, P. A. A theory of reading: From eye fixations to comprehension. *Psychol. Rev.* **87**(4), 329–354. https://doi.org/10.1037/0033-295x.87.4.329 (1980).
3. Ma, X., Liu, Y., Clariana, R., Gu, C. & Li, P. From eye movements to scanpath networks: A method for studying individual differences in expository text reading. *Behav. Res. Methods* **55**(2), 730–750. https://doi.org/10.3758/s13428-022-01842-3 (2023).
4. Kliegl, R., Nuthmann, A. & Engbert, R. Tracking the mind during reading: the influence of past, present, and future words on fixation durations. *J. Exp. Psychol. Gen.* **135**(1), 12–35. https://doi.org/10.1037/0096-3445.135.1.12 (2006).
5. Posner, M. I. Orienting of Attention*. *Q. J. Exper. Psychol.* **32**(1), 3–25. https://doi.org/10.1080/00335558008248231 (1980).
6. Kaakinen, J. K. What can eye movements tell us about visual perception processes in classroom contexts? Commentary on a special issue. *Educ. Psychol. Rev.* **33**(1), 169–179. https://doi.org/10.1007/s10648-020-09573-7 (2021).
7. Rayner, K. Eye movements in reading and information processing: 20 years of research. *Psychol. Bull.* **124**(3), 372–422. https://doi.org/10.1037/0033-2909.124.3.372 (1998).
8. Rayner, K. Eye movements and attention in reading, scene perception, and visual search. *Q. J. Exp. Psychol.* **62**(8), 1457–1506. https://doi.org/10.1080/17470210902816461 (2009).
9. Itti, L. & Koch, C. Computational modelling of visual attention. *Nat. Rev. Neurosci.* **2**(3), 194–203. https://doi.org/10.1038/35058500 (2001).
10. Lang, P. J., Bradley, M. M., & Cuthbert, B. N. (1997). Motivated attention: Affect, activation, and action. *Attention and Orienting: Sensory and Motivational Processes.*, 97–135.

11. Lee, S., Hwang, Y., Jin, Y., Ahn, S. & Park, J. Effects of individuality, education, and image on visual attention: Analyzing eye-tracking data using machine learning. *J. Eye Mov. Res.* **12**(2), 10.16910/jemr.12.2.4. https://doi.org/10.16910/jemr.12.2.4 (2019).
12. Olander, M. H., Brante, E. W. & Nyström, M. The Effect of Illustration on Improving Text Comprehension in Dyslexic Adults. *Dyslexia* **23**(1), 42–65. https://doi.org/10.1002/dys.1545 (2017).
13. Cisek, S. Z. et al. Narcissism and consumer behaviour: a review and preliminary findings. *Front. Psychol.* **5**, 232. https://doi.org/10.3389/fpsyg.2014.00232 (2014).
14. Bradley, M. M., Houbova, P., Miccoli, L., Costa, V. D. & Lang, P. J. Scan patterns when viewing natural scenes: Emotion, complexity, and repetition. *Psychophysiology* **48**(11), 1544–1553. https://doi.org/10.1111/j.1469-8986.2011.01223.x (2011).
15. Yang, H. Reading comics: The effect of expertise on eye movements. *J. Eye Mov. Res.* **17**(4), 10.16910/jemr.17.4.5. https://doi.org/10.16910/jemr.17.4.5 (2024).
16. Chaby, L., Hupont, I., Avril, M., Boullay, V. L. & Chetouani, M. Gaze behavior consistency among older and younger adults when looking at emotional faces. *Front. Psychol.* **8**, 548. https://doi.org/10.3389/fpsyg.2017.00548 (2017).
17. Chua, H. F., Boland, J. E. & Nisbett, R. E. Cultural variation in eye movements during scene perception. *Proc. Natl. Acad. Sci.* **102**(35), 12629–12633. https://doi.org/10.1073/pnas.0506162102 (2005).
18. Nisbett, R. E., & Masuda, T. (2006). Biological and cultural bases of human inference. 49–70. https://doi.org/10.4324/9780203774908-3
19. Ossandón, J. P., Onat, S. & Konig, P. Spatial biases in viewing behavior. *J. Vis.* **14**(2), 20–20. https://doi.org/10.1167/14.2.20 (2014).
20. Guo, K., Meints, K., Hall, C., Hall, S. & Mills, D. Left gaze bias in humans, rhesus monkeys and domestic dogs. *Anim. Cogn.* **12**(3), 409–418. https://doi.org/10.1007/s10071-008-0199-3 (2009).
21. Dickinson, C. A. & Intraub, H. Spatial asymmetries in viewing and remembering scenes: Consequences of an attentional bias?. *Atten. Percept. Psychophys.* **71**(6), 1251–1262. https://doi.org/10.3758/app.71.6.1251 (2009).
22. Calbi, M. et al. Haptic aesthetics and bodily properties of ori gersht's digital art: A behavioral and eye-tracking study. *Front. Psychol.* **10**, 2520. https://doi.org/10.3389/fpsyg.2019.02520 (2019).
23. Guo, K. Holistic gaze strategy to categorize facial expression of varying intensities. *PLoS ONE* **7**(8), e42585. https://doi.org/10.1371/journal.pone.0042585 (2012).
24. Calvo, M. G., Krumhuber, E. G. & Fernández-Martín, A. Visual attention mechanisms in happiness versus trustworthiness processing of facial expressions. *Q. J. Exper. Psychol.* **72**(4), 729–741. https://doi.org/10.1177/1747021818763747 (2018).
25. Butler, S. et al. Are the perceptual biases found in chimeric face processing reflected in eye-movement patterns?. *Neuropsychologia* **43**(1), 52–59. https://doi.org/10.1016/j.neuropsychologia.2004.06.005 (2005).
26. Calbi, M., Langiulli, N., Siri, F., Umiltà, M. A. & Gallese, V. Visual exploration of emotional body language: A behavioural and eye-tracking study. *Psychol. Res.* **85**(6), 2326–2339. https://doi.org/10.1007/s00426-020-01416-y (2020).
27. Thompson, L. A., Malloy, D. M. & LeBlanc, K. L. Lateralization of visuospatial attention across face regions varies with emotional prosody. *Brain Cogn.* **69**(1), 108–115. https://doi.org/10.1016/j.bandc.2008.06.002 (2009).
28. Calvo, M. G. & Lang, P. J. Gaze patterns when looking at emotional pictures: Motivationally biased attention. *Motiv. Emot.* **28**(3), 221–243. https://doi.org/10.1023/b:moem.0000040153.26156.ed (2004).
29. Vaid, J., & Singh, M. (1989). Asymmetries in the perception of facial affect: Is there an influence of reading habits? **27**(10), 1277–1287. https://doi.org/10.1016/0028-3932(89)90040-7
30. Chokron, S. On the origin of free-viewing perceptual asymmetries. *Cortex* **38**(2), 109–112. https://doi.org/10.1016/s0010-9452(08)70644-0 (2002).
31. Barthes, R. (1977). *Image-music-text* (H. and Wang, Ed.).
32. Romano, L. *Lettura di un'immagine* Vol. 553 (Einaudi, 1975).
33. Hu, J. & Zhang, J. The effect of cue labeling in multimedia learning: Evidence from eye tracking. *Front. Psychol.* **12**, 736922. https://doi.org/10.3389/fpsyg.2021.736922 (2021).
34. Jian, Y. Fourth graders' cognitive processes and learning strategies for reading illustrated biology texts: Eye movement measurements. *Read. Res. Q.* **51**(1), 93–109. https://doi.org/10.1002/rrq.125 (2016).
35. Jian, Y.-C. Reading instructions influence cognitive processes of illustrated text reading not subject perception: An eye-tracking study. *Front. Psychol.* **9**, 2263. https://doi.org/10.3389/fpsyg.2018.02263 (2018).
36. Johnson, C. I. & Mayer, R. E. An eye movement analysis of the spatial contiguity effect in multimedia learning. *J. Exp. Psychol. Appl.* **18**(2), 178–191. https://doi.org/10.1037/a0026923 (2012).
37. Oldfield, R. C. The assessment and analysis of handedness: The Edinburgh inventory. *Neuropsychologia* **9**(1), 97–113. https://doi.org/10.1016/0028-3932(71)90067-4 (1971).
38. World Medical Association. World Medical Association Declaration of Helsinki: Ethical Principles for Medical Research Involving Human Subjects. *JAMA* **310**(20), 2191–2194. https://doi.org/10.1001/jama.2013.281053 (2013).
39. Brainard, D. H. The Psychophysics Toolbox. *Spatial Vis.* **10**, 433–436 (1997).
40. Niehorster, D. C., Andersson, R. & Nyström, M. Titta: A toolbox for creating PsychToolbox and Psychopy experiments with Tobii eye trackers. *Behav. Res. Methods* **52**(5), 1970–1979. https://doi.org/10.3758/s13428-020-01358-8 (2020).
41. Albiero, P., Ingoglia, S. & Cocco, A. L. Contributo all'adattamento italiano dell'Interpersonal Reactivity Index. *TESTING PSICOMETRIA METODOLOGIA* **13**(2), 107–125 (2006).
42. Witmer, B. G. & Singer, M. J. Measuring presence in virtual environments: A presence questionnaire. *Presence Teleoperators Virtual Environ.* **7**(3), 225–240. https://doi.org/10.1162/105474698565686 (1998).
43. Spielberger, C. D., Gorsuch, R. L., Lushene, R. E., Vagg, P. R., & Jacobs, G. A. (n.d.). *State-trait anxiety inventory for adults: manual, instrument and scoring guide*. Mind Garden.
44. Dragovic, M. Towards an improved measure of the Edinburgh Handedness Inventory: A one-factor congeneric measurement model using confirmatory factor analysis. *Laterality* **9**(4), 411–419. https://doi.org/10.1080/13576500342000248 (2004).
45. D'Agostino, R. B. Transformation to normality of the null distribution of g_1. *Biometrika* **57**(3), 679. https://doi.org/10.2307/2334794 (1970).
46. Hessels, R. S., Niehorster, D. C., Kemner, C. & Hooge, I. T. C. Noise-robust fixation detection in eye movement data: Identification by two-means clustering (I2MC). *Behav. Res. Methods* **49**(5), 1802–1823. https://doi.org/10.3758/s13428-016-0822-1 (2017).
47. Coco, M. I. (2009). The statistical challenge of scan-path analysis. *2009 2nd Conference on Human System Interactions*, 372–375. https://doi.org/10.1109/hsi.2009.5091008
48. Kendall, M. G. A New Measure of Rank Correlation. *Biometrika* **30**(1–2), 81–93. https://doi.org/10.1093/biomet/30.1-2.81 (1938).
49. Team, R. C. (2022). *R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria*. https://www.R-project.org/
50. Bates, D., Mächler, M., Bolker, B. M., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, *67*(1). https://doi.org/10.18637/jss.v067.i01
51. Lüdecke, D., Ben-Shachar, M., Patil, I., Waggoner, P. & Makowski, D. Performance: An R package for assessment, comparison and testing of statistical models. *J. Open Source Softw.* **6**(60), 3139. https://doi.org/10.21105/joss.03139 (2021).
52. Lenth, R. V. (2022). *emmeans: Estimated Marginal Means, aka Least-Squares Means. R package version 1.8.3*. https://CRAN.R-project.org/package=emmeans
53. Wickham, H. *ggplot2 Elegant Graphics for Data Analysis* (Springer, 2009).
54. Fox, J. & Weisberg, S. *An R Companion to Applied Regression* 3rd Edition. (Sage Publications, 2019).

55. Cometa, M. Forme e retoriche del fototesto letterario. In *Fototesti. Letteratura e cultura visuale* (eds Cometa, M. & Coglitore, R.) 69–116 (Quodlibet, 2016).

Acknowledgements
Work supported by #NEXTGENERATIONEU (NGEU) and funded by the Ministry of University and Research (MUR), National Recovery and Resilience Plan (NRRP), project MNESYS (PE0000006) – A Multiscale integrated approach to the study of the nervous system in health and disease (DN. 1553 11.10.2022) awarded to V.G., and by the PRIN grant (2020YB7J25) "Phototexts: Rhetorics, Poetics, and Cognition" awarded to M.C. and V.G.

Author contributions
N.L. and V.S. conceptualized the idea and discussed it with all the authors. M.C. and R.C. digitalized the stimuli. N.L. edited the stimuli, performed data acquisition and analyses. N.L. and V.G. interpreted the results and wrote the manuscript. N.L. prepared all figures. All authors have contributed to, seen, reviewed and approved the manuscript.

Declarations

Competing interests
The authors declare no competing interests.

Additional information
Correspondence and requests for materials should be addressed to N.L.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.