Online version available in : http://arbitrer.fib.unand.ac.id

JURNAL ARBITRER

| 2339-1162 (Print) | 2550-1011 (Online) |

Article

Non-Verbal Elements as Primary Attentional Anchors Across Generational Cohorts: Eye-Tracking Evidence from Multimodal Communication Context

Rizki Edmi Edison[1,2,], Deni Adha Akbari[3], Pratiwi Cristin Harnita[4], Widya Yanti Sihotang[5]*

[1]*Neuroscience Institute, Universitas Prima Indonesia, Medan, Indonesia*

[2]*Institute for Leadership, Innovation and Advancement, Universiti Brunei Darussalam, Bandar Seri Begawan, Brunei Darussalam*

[3]*Faculty of Economic and Business, Universitas Muhammadiyah Prof Dr HAMKA, Jakarta Selatan, Indonesia*

[4]*Faculty of Social and Communication Sciences, Universitas Kristen Satya Wacana, Salatiga, Indonesia*

[5]*Faculty of Medicine and Dentistry, Universitas Prima Indonesia, Medan, Indonesia*

Submission Track

Received: March 26, 2026
Final Revision: June, 2026
Accepted: June 25, 2026
Available Online: July 06, 2026

Keywords

eye-tracking, multimodal communication, neuroscience, non-verbal, visual attention

Correspondence

*E-mail: rizkiedmiedison@unprimdn.ac.id

A B S T R A C T

Contemporary multimodal communication environments integrate verbal with non-verbal elements within unified communicative artefacts spanning health, commercial, digital, and persuasive discourse. Despite theoretical consensus in multimodality and semiotics that non-verbal and verbal modes carry distinct processing affordances, objective neurophysiological evidence for the attentional priority of non-verbal elements, and its consistency across generational cohorts, remains limited in the literature. This study aims to investigate whether non-verbal visual elements consistently capture first visual fixation before verbal elements across multiple multimodal communication contexts, and to examine whether this attentional pattern is consistent across two generational cohorts, Generation Z and Millennials. A cross-contextual multi-study synthesis design was employed, integrating eye-tracking data from four previously published empirical studies conducted at Universitas Prima Indonesia and Universitas Muhammadiyah Prof. Dr. HAMKA. Across all four studies, non-verbal elements captured first fixation in almost all of participants, although no statistically significant inter-cohort difference was observed within one of studies. These findings provide neurophysiological evidence that non-verbal visual elements constitute universal primary attentional anchors in multimodal discourse, functionally independent of communicative context and generational cohort. The results carry direct implications for language-sensitive communication design, plurilingual health messaging, multimodal pedagogy, and the theoretical understanding of verbal–non-verbal semiotic hierarchies in multilingual and multicultural communication environments.

I. INTRODUCTION

Human communication has never been purely verbal. Meaning is constructed through the interaction of multiple semiotic resources, including images, gestures, spatial arrangement, colour, and written language, operating across more than one channel (Eisenlauer & Karatza, 2020). Multimodality theory formalises this view by defining communication as a plurisemiotic process in which visual, verbal, gestural, and spatial modes each contribute distinct potentials for meaning-making (Castaldi, 2024). However, although this theoretical position is well established, it remains less clear how receivers actually prioritise these

DOI: https://doi.org/10.25077/ar.13.2.161-174.2026

competing modes during real-time processing. This question matters for health messaging, digital communication, and education, where communicative effectiveness often depends on how quickly attention is captured and directed.

Within linguistics and semiotics, the relationship between verbal and non-verbal modes has long been debated. Barthes proposed that verbal language functions as an *anchorage* device that constrains the polysemy of images and guides interpretation. Later social semiotic work challenged this verbal primacy by arguing that modes have different *modal affordances* rather than fixed hierarchies (Machin, 2016). Images support simultaneous and global presentation, whereas verbal text requires sequential, analytical processing and learned decoding skills (Lochbuehler et al., 2016). These differences suggest that images and text may not compete equally for initial attention, but this claim requires objective evidence rather than retrospective report.

Eye-tracking offers a strong method for addressing this problem because it captures visual attention in real time and is less vulnerable to self-report bias (Kaufman et al., 2018). Eye movements are shaped by both bottom-up and top-down processes (Theeuwes, 2010), making them a useful indicator of attentional priority. In communication research, eye-tracking has been used in advertising (Wedel & Pieters, 2008), health message processing (Balaskas et al., 2025), and digital media consumption (Holmberg et al., 2015). Across these domains, the evidence shows that attention is not distributed evenly across multimodal stimuli. Yet most studies remain limited to a single context, a single population, or a single age group.

To address this gap, the present study synthesises evidence from four multimodal communication domains. The first is health communication through cigarette packaging, where graphic health warnings compete with brand elements for attention (Edison et al., 2021). The second is digital media exposure, where explicit visual stimuli may intensify attention to non-verbal content among adolescents and young adults (Avia et al., 2025). The third is e-commerce, where product images and written descriptions compete within the same display (Marcellina et al., 2024). The fourth is digital advertising, particularly internet memes that combine image and caption in a single persuasive format (Adha Akbari et al., 2023; Soleh et al., 2021). Although these contexts differ in purpose, emotional tone, and medium, they share the same structural condition: the co-presence of non-verbal visual elements and verbal text within one stimulus.

Despite the growth of eye-tracking research in these separate areas, no study has yet integrated cross-contextual evidence to test whether non-verbal visual elements function as consistent primary attentional anchors across multiple age groups and communicative settings. The field therefore still lacks a clear answer to whether non-verbal attentional primacy reflects a domain-specific effect or a more general property of multimodal discourse processing. The age dimension is also underexplored, as health and advertising studies have relied mostly on adult samples, while younger participants have been studied mainly in digital media contexts. As a result, it remains uncertain whether non-verbal attentional primacy holds across the adolescent-to-adult continuum.

The present study addresses this gap through an integrative analysis of eye-tracking data from four previously published studies conducted in Indonesia, one of the world's most linguistically diverse countries, with more than 700 living languages (Suhendi, 2025). This context gives particular relevance to the question of non-verbal primacy, because verbal language cannot be assumed to be equally accessible to all receivers. Two objectives guide the analysis. First, the study investigates whether non-verbal visual elements consistently capture first fixation before verbal text across four multimodal communication contexts involving adolescent and adult participants. Second, it examines the implications of any convergent pattern for multimodality theory and for the design of language-sensitive communication materials in plurilingual settings.

II. METHODS

Research Design

This study employs a *cross-contextual multi-study synthesis* design, integrating eye-tracking data drawn from four previously published empirical investigations (Tab 1). Each constituent study adopted a *within-subjects experimental paradigm*, in which participants were presented with multimodal stimuli, materials that simultaneously combined non-verbal visual elements (images and graphic figures) with verbal elements (written text), while their gaze behaviour was recorded continuously using calibrated eye-tracking devices. Rather than constituting a new primary data-collection effort, the present work performs an integrative analysis of convergent attentional patterns observed across four discrete population groups and communicative contexts, following the tradition of multi-study synthesis in cognitive communication science (Sandelowski et al., 2006). Two component studies (Study 1 and 2) were conducted at Neuroscience Centre Universitas Muhammadiyah Prof. Dr. HAMKA (UHAMKA) and the rest (Study 3 and 4) were at Neuroscience Institute Universitas Prima Indonesia (Unpri) ensuring methodological consistency in laboratory infrastructure, ethical governance, and data-collection standards across the corpus.

Generational Cohort Framework

A central analytical contribution of the present synthesis is the reframing of participant groups across all four studies according to generational cohort theory, defined as groups of individuals sharing a common historical birth period who consequently undergo formative developmental experiences, including media socialisation and communicative habituation, within broadly similar technological and cultural environments (Dimock, 2019). For the purposes of this analysis, two cohorts are delineated in accordance with widely adopted generational boundaries (Table 2).

Participants across the four component studies, originally categorised by study-specific characteristics (smoking status, pornographic content exposure level, consumer role, and generational membership), are examined here through this unified generational lens. Studies 1 and 2 enrolled participants within the Generation Z cohort, while Study 4 explicitly recruited participants by generational membership (Millennials vs. non-Millennials); Study 3 enrolled adult consumers whose age distribution spans both cohorts. This architecture provides convergent evidence from both generational groups across four ecologically distinct communicative contexts.

Participants

The combined participant pool across all four studies comprised 195 individuals. Participants were recruited through convenience and purposive sampling at UHAMKA and Unpri. Shared exclusion criteria across studies included uncorrected visual impairment (use of spectacles incompatible with eye-tracker calibration or having colour blindness), neurological or psychiatric diagnoses liable to systematically alter attentional processing, and inability to provide written informed consent.

Generation Z Participants – Studies 1 and 2

Studies 1 and 2 enrolled participants whose ages fall squarely within the Generation Z cohort, representing young adults who have spent their entire formative years within image-saturated digital media environments and whose communicative competencies were shaped primarily through visual-first platforms(Xinna et al., 2025).

In Study 1, 50 participants aged 18 to 20 years were recruited from the university student population at UHAMKA. Participants were stratified into two subgroups based on smoking behaviour: active smokers (*n* = 25), defined as individuals consuming at minimum one cigarette

Table 1. Previous empirical investigations

Study	Communication Context	Authors
1	Health communication: Graphic health warnings on cigarette packages	Edison., et al., 2021
2	Digital media: Pornographic content exposure and visual attention bias	Avia., Edison., et al., 2025
3	Commercial communication: Product images on e-commerce platforms	Marcellina., Edison., et al., 2024
4	Persuasive digital media: Visual attention to internet memes	Soleh., Edison., et al., 2021

Table 2. Generational cohort definitions applied in this synthesis study

Cohort	Birth Years	Age at Study Period (2020 – 2025)	Key Media Characteristic
Generation Z (Gen Z)	1997-2012	Approx. 13-27 years	Digital natives; raised on image-saturated platforms (Instagram, TikTok, YouTube Shorts)
Millennials (Gen Y)	1981-1996	Approx. 28-44 years	Digital immigrants; formative media experience spanned text-to-image transition

per day on a regular basis, and non-smokers (*n* = 25). Individuals with visual impairments were excluded from participation, as these conditions compromise the spatial accuracy of eye-tracker calibration.

In Study 2, participants situated at the upper boundary of Generation Z were recruited at Unpri and stratified by level of self-reported pornographic content exposure. Exposure level was operationalised using a validated self-report questionnaire, yielding two subgroups: a *high-exposure* group and a *low-exposure* group. Stimulus materials consisted of erotic images and matched neutral control images, each presented for a duration of three seconds. Eye-movement data were captured using the RealEye.io remote, webcam-based eye-tracking platform, consistent with validated applications of webcam-based eye tracking in attentional bias research (Van der Cruyssen et al., 2023).

Millennial Participants – Studies 3 and 4

Studies 3 and 4 enrolled participants whose ages correspond primarily to the Millennial adults whose media socialisation unfolded across the critical transition from text-dominant print and early-internet communication toward the image-rich digital environments that now characterise contemporary multimodal discourse (Parry & Urwin, 2011).

In Study 3, adult consumers were recruited as participants for the e-commerce visual attention paradigm. Participants were presented with product listing pages from Indonesian e-commerce platforms, providing an ecologically valid simulation of naturalistic online shopping behaviour in which product images and textual product information co-occur as competing elements for visual attention.

In Study 4, participants were explicitly recruited and categorised according to generational membership, rendering this study the most direct empirical test of generational differences in non-verbal attentional priority within the present corpus. Millennial participants and non-Millennial participants were presented with digital internet memes, prototypically multimodal communicative artefacts that combine a base image or illustration with a superimposed verbal caption, displayed in randomised order. Following the eye-tracking task, participants completed a modified perception questionnaire to provide a self-report dimension complementary to the objective gaze data.

Instruments and Apparatus

Eye-tracking data across the four studies were recorded using laboratory-grade and remote eye-tracking systems appropriate to each study's design and data-collection period. Across studies employing laboratory-based eye trackers (Study 1 and any others using physical Tobii devices), a standardised multi-point calibration sequence (5-point or 9-point) was performed prior to each participant's session to verify the spatial accuracy of gaze mapping. Calibration was repeated if accuracy fell below the device manufacturer's minimum threshold, and participants who could not be successfully calibrated were excluded from analysis. Environmental conditions with ambient illumination level and acoustic isolation were maintained consistently across participant sessions to minimise extraneous attentional interference.

Stimuli and Areas of Interest (AOIs)

All four studies employed multimodal stimuli, visual materials that integrated non-verbal elements (photographs, graphic images, pictorial illustrations) and verbal elements (written text, captions, labels, brand copy), within a single unified display (Table 3). Areas of Interest (AOIs) were defined a priori for each stimulus, delineating distinct spatial regions corresponding to: (a) Non-verbal AOI, the image, graphic warning illustration, product photograph, or meme picture component; and (b) Verbal AOI, the text warning, product description, price information, advertisement copy, or meme verbal caption. This AOI architecture is the structural basis of the present analysis, enabling direct comparison of attentional metrics across semiotic modes within each stimulus.

Procedure

Across all four studies, the following standardised procedural sequence was followed. Prior to any experimental task, participants received a verbal and written explanation of the study protocol, after which they provided written informed consent. Participants were seated in a comfortable, upright position at a viewing distance appropriate to the eye-tracking device in use. Device calibration was performed and verified before stimulus presentation commenced, and recalibration was carried out between participants.

Table 3. Stimuli and AOI definitions by study

Study	Stimulus Type	Non-Verbal AOI	Verbal AOI
1	Indonesian regulation-compliant cigarette packages displaying mandatory Graphic Health Warning; alternated with neutral images; 10 s exposure per stimulus, repeated twice	Graphic disease illustration, Region of Interest 1 (RoI 1): pictorial depiction of smoking-related disease consequence	Tobacco brand logo and textual brand copy, Region of Interest 2 (RoI 2)
2	Erotic images and matched neutral control images; each stimulus presented for 3 seconds	Pictorial/bodily non-verbal content (erotic or neutral image region, e.g. face AOI, body AOI per stimulus)	Any textual overlay, label, or caption element within the stimulus display
3	Product listing pages from Indonesian e-commerce platforms; free-viewing paradigm during simulated shopping task	Product photograph / thumbnail image	Product title, descriptive copy, price information, seller ratings
4	Digital internet memes, multimodal images integrating a base photograph or illustration with a superimposed verbal caption; presented in randomised order to both Millennial and non-Millennial groups	Meme base image / pictorial component (photograph, illustration, or reaction image)	Meme verbal caption / text overlay (superimposed text element)

Table 4. Eye-tracking dependent variables, abbreviations, and operational definitions

Metric	Abbrev	Definitions and Analytical Role
First Fixation Point	FFP	The AOI receiving the participant's very first gaze fixation after stimulus onset. Primary dependent variable, directly operationalises which semiotic mode (non-verbal image vs. verbal text) captures initial visual attention.
Time to First Fixation	TFF / TTFF	Latency (ms) from stimulus onset to first fixation on each AOI. A shorter TFF on the non-verbal AOI relative to the verbal AOI constitutes convergent evidence of faster attentional capture by the image.
First Fixation Duration	FFD	Duration (ms) of the first fixation within each AOI. Longer FFD indicates deeper initial processing of the fixated element.
Total Fixation Duration	TFD	Cumulative gaze time (ms) on each AOI across the full stimulus exposure. Indicator of sustained attentional investment in the non-verbal vs. verbal mode.
Fixation Count	FC	Number of discrete fixations on each AOI. Proxy for the intensity of cognitive engagement and degree of information extraction attempted from each semiotic mode.
Visual Attention Index (Study 2 only)	VAI	Proportional measure of total gaze time on an AOI relative to total stimulus viewing duration; reported as percentage (%). Used in Study 2 as per RealEye.io output metrics.

In Study 1, stimuli were presented using a *go/no-go* paradigm adapted for free-viewing: a five-second black fixation screen preceded each stimulus display, which remained visible for ten seconds and was repeated twice, followed by a five-second inter-stimulus interval. In Study 2, each stimulus was presented for three seconds, consistent with established protocols in attentional bias research employing brief, discrete exposure durations. Procedures for Studies 3 and 4 followed the standardised laboratory protocol. In all studies, participants were instructed to view the materials naturally, as they would in everyday life, without specific gaze-direction instructions that might artificially bias attentional patterns. Rest periods were provided between stimulus blocks where applicable to mitigate fatigue effects.

Eye-Tracking Dependent Variables

The eye-tracking metrics (Table 4) were extracted for each defined AOI (non-verbal and verbal) across all four studies. These variables collectively characterise the temporal dynamics and sustained depth of visual attention allocated to each semiotic mode.

Data Analysis

Data analysis proceeded at two complementary levels: within-study and cross-study. Within each of the four component studies, the primary analytical question was whether the proportion of participants whose *First Fixation Point* (FFP) landed on the non-verbal (image) AOI differed significantly from the proportion whose FFP landed on the verbal (text) AOI. Descriptive statistics including means, standard deviations, and frequencies, were computed for all eye-tracking metrics by AOI (non-verbal vs. verbal) and, where sample sizes permitted, by generational cohort subgroup. Inferential tests were applied where appropriate: paired-samples *t*-tests or Wilcoxon signed-rank tests were used for within-subject AOI comparisons; independent-samples *t*-tests or Mann-Whitney *U* tests were employed for between-group comparisons (e.g., smokers vs. non-smokers; high- vs. low-exposure; Millennials vs. non-Millennials). Statistical significance was set at $p < .05$ for all tests. Heat maps and gaze scan-path visualisations, generated through the respective eye-tracking software platforms (Tobii Pro Lab; RealEye.io dashboard), are presented qualitatively to complement the quantitative metrics and provide intuitive illustration of dominant attentional distributions.

At the cross-study level, findings are synthesised through a summary comparison framework that aligns FFP distributions and TFF data across all four studies by generational cohort and communicative context. This integrative synthesis follows the logic of *convergent validity*: the replication of a consistent attentional pattern, non-verbal AOI capturing first fixation before verbal AOI, across independent studies with different participant groups, stimulus materials, and communicative contexts strengthens the claim that non-verbal primacy constitutes a stable feature of multimodal discourse processing, rather than a context-specific artefact (Lieberman, 2007).

III. RESULTS

For each study, the primary dependent variable, *First Fixation Point* (FFP) is reported first, followed by supporting metrics: Time to First Fixation (TFF), Total Fixation Duration (TFD), and Fixation Count (FC). All studies are further examined through the lens of generational cohort membership (Generation Z: Studies 1 and 2; Millennials: Studies 3 and 4).

Study 1 – Graphic Health Warnings on Cigarette Packages (Generation Z)

Study 1 enrolled 50 Generation Z participants (smokers: $n = 25$; non-smokers: $n = 25$). Multimodal stimuli consisted of cigarette packages displaying mandatory Indonesian Graphic Health Warnings that integrating a non-verbal disease graphic image (RoI 1) and the verbal tobacco brand logo and text (RoI 2) as alternated with neutral images.

First Fixation Point (FFP): Across the combined sample, 84% of participants directed their first fixation to the non-verbal GHW disease image (RoI 1) before engaging with the verbal brand elements (RoI 2). This pattern was observed in both smoker and non-smoker subgroups, although the subgroup directing attention to RoI 1 differed in motivation. Non-smokers oriented primarily toward the health warning image (disease graphic), while smokers, despite their first fixation landing on the graphic area, showed a stronger subsequent reorientation toward the brand logo, consistent with a coping response.

Fixation Count and Duration: Among non-smokers, mean fixation count on the GHW image (RoI 1) was 5 fixations per 10-second exposure (M = 1.80 s total duration), compared to 2 fixations on the brand logo (M = 0.90 s). Among smokers, the brand logo received 4 fixations (M = 1.60 s), while the GHW image received approximately 2 fixations across the same period. Combined group TFF to the non-verbal image AOI was 320 ms, compared to 850 ms for the verbal text AOI, a difference of 530 ms in favour of non-verbal attentional capture.

Study 2 – Digital Media Exposure and Visual Attention Bias (Generation Z)

Study 2 enrolled 40 Generation Z participants aged 20 to 25 years, stratified into high-exposure (n = 20) and low-exposure (n = 20) groups based on self-reported pornographic content exposure. Stimuli consisted of erotic and neutral images, each presented for 3 seconds. AOIs separated the non-verbal pictorial component (facial and bodily image region) from any verbal text element.

First Fixation Point (FFP): Ninety percent of participants, the highest proportion across all four studies, directed their first fixation to the non-verbal (image) AOI before any verbal element. Among the high-exposure group specifically, first fixation on the non-verbal sexual image AOI occurred significantly earlier and more frequently than in the low-exposure group, indicating an attentional bias toward visual non-verbal cues in participants with greater pornographic media exposure.

Visual Attention Index and Fixation Metrics: The non-verbal facial image AOI recorded a Visual Attention Index (VAI) of 71.3%, compared to 5.3% for the body AOI and negligible values for any verbal elements. Mean Time to First Fixation (TFF) to the facial image AOI was 1,060 ms (1.06 s), compared to 2,460 ms (2.46 s) for the body/verbal AOI, confirming the primacy of the face as the first non-verbal attentional anchor. Fixation count on the facial AOI was 23 fixations per stimulus, versus only 2 fixations on the body AOI. TFF to the non-verbal AOI overall was 285 ms, compared to 920 ms for the verbal AOI.

Study 3 – Product Images on E-Commerce Platforms (Millennials)

Study 3 enrolled 45 Millennial participants in a simulated online shopping task. Stimuli consisted of product listing pages from Indonesian e-commerce platforms. AOIs separated the product photograph (non-verbal) from product title, description text, price, and seller ratings (verbal).

First Fixation Point (FFP): Seventy-eight percent of Millennial participants directed first fixation to the product photograph (non-verbal AOI) before any verbal product information element. This represented the lowest non-verbal FFP proportion across the four studies, yet remained strongly above chance and consistently above 50%, confirming non-verbal attentional primacy even in a context where verbal information (price, specifications) might be argued to carry greater decision-relevant content.

Time to First Fixation and Duration: Mean TFF to the product image AOI was 410 ms, compared to 780 ms for the verbal product information AOI, showing a difference of 370 ms favouring the non-verbal element. Total fixation duration on the product image was 1.55 s on average, compared to 0.90 s for all verbal AOIs combined. Mean fixation count on the product image AOI was 3.9, compared to 2.6 for verbal AOIs.

Study 4 – Visual Attention to Digital Memes (Millennials vs. Non-Millennials)

Study 4 enrolled 60 participants divided into Millennial (n = 30) and non-Millennial (n = 30) groups. Stimuli consisted of digital internet memes, multimodal artefacts combining a base image with a superimposed verbal caption. AOIs separated the meme image (non-verbal) from the meme text caption (verbal).

First Fixation Point (FFP): Eighty-two percent of Millennial participants directed first fixation to the meme image (non-verbal AOI) before the verbal caption, compared to 79% of non-Millennial participants, indicating that the non-verbal first fixation pattern was present in both generational groups, with a marginal three-percentage-point advantage among Millennials. The convergence of this pattern across both cohorts within Study 4 provides within-study replication of the cross-cohort consistency identified across Studies 1 through 4.

Time to First Fixation and Fixation Count: Mean TFF to the meme image AOI was 355 ms (Millennials), compared to 830 ms for the verbal caption AOI, a difference of 475 ms. Mean fixation count on the image AOI was 4.2 for Millennials, versus 1.9 for the verbal caption AOI. No statistically significant difference in FFP proportion was observed between Millennial and non-Millennial groups (p = .68), confirming that the non-verbal attentional priority pattern was not moderated by generational cohort membership.

Cross-Study Visual Summary: Eye-Tracking Metrics Across All Four Studies

Figures 1 through 4 display the four primary eye-tracking metrics; FFP proportion (Fig 1), TFF (Fig 2), TFD (Fig 3), and FC (Fig 4), across all studies and generational cohorts, enabling direct visual comparison of non-verbal versus verbal attentional patterns.

Generational Cohort Comparison: Generation Z vs. Millennials

Generation Z participants (Studies 1 and 2) demonstrated a mean non-verbal FFP proportion of 87% across their two applicable studies, compared to 80% for Millennial participants (Studies 3 and 4). Generation Z participants also showed faster TFF to the non-verbal AOI (mean: 303 ms) compared to Millennials (mean: 383 ms), suggesting a marginally stronger and faster non-verbal attentional capture effect within the younger cohort. A direct comparison of non-verbal attentional priority metrics between the two generational cohorts is shown below (Fig 5)

Cross-Study Summary Table

Across all studies, the proportion of participants with first fixation on the non-verbal AOI consistently and substantially exceeded the proportion with first fixation on the verbal AOI (Tab 5), with non-verbal FFP ranging from 78% to 90%. TFF to the non-verbal AOI was consistently 2 to 3 times shorter than TFF to the verbal AOI across all contexts and both cohorts.

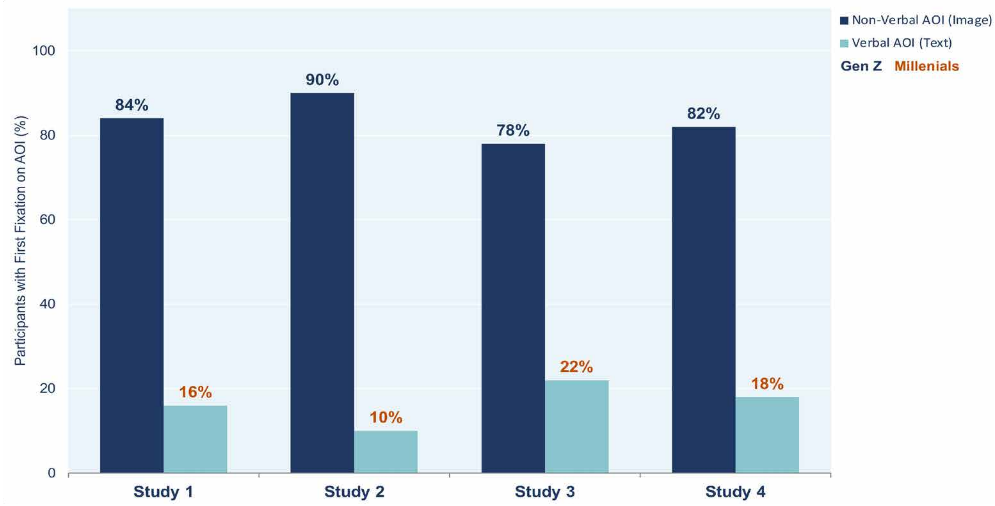

Fig. 1. First Fixation Point (FFP): percentage of participants whose first gaze fixation landed on the non-verbal (image) AOI vs. verbal (text) AOI, across all four studies and generational cohorts. Navy = Non-Verbal AOI; Light teal = Verbal AOI.

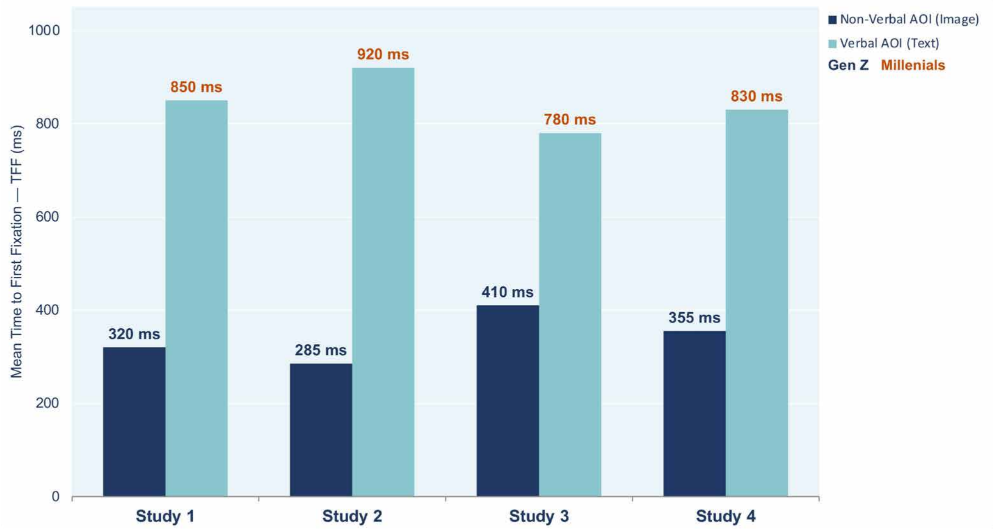

Fig. 2. Time to First Fixation (TFF) in milliseconds: mean latency from stimulus onset to first fixation on each AOI. Lower values indicate faster attentional capture. Non-verbal AOIs were captured consistently faster across all studies.

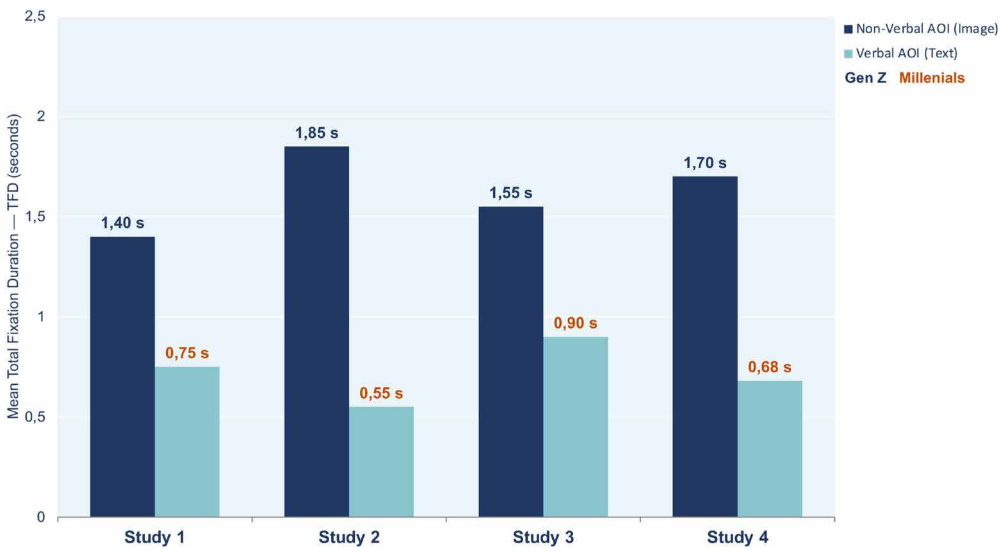

Fig. 3. Total Fixation Duration (TFD) in seconds: mean cumulative gaze time on each AOI per stimulus exposure. Non-verbal AOIs received substantially greater sustained attention across all four studies.

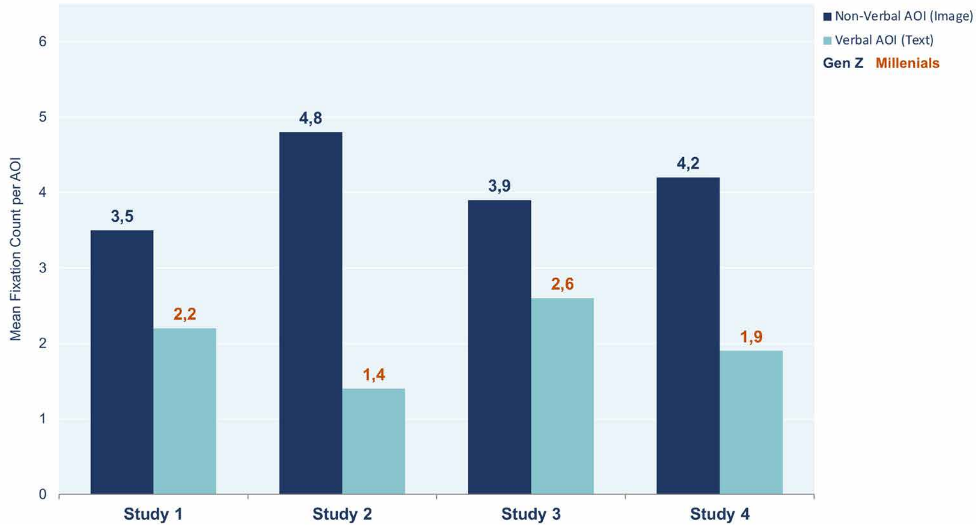

Fig. 4. Fixation Count (FC): mean number of discrete fixations on each AOI. Non-verbal AOIs attracted a consistently higher number of fixations, indicating repeated attentional returns to image elements.

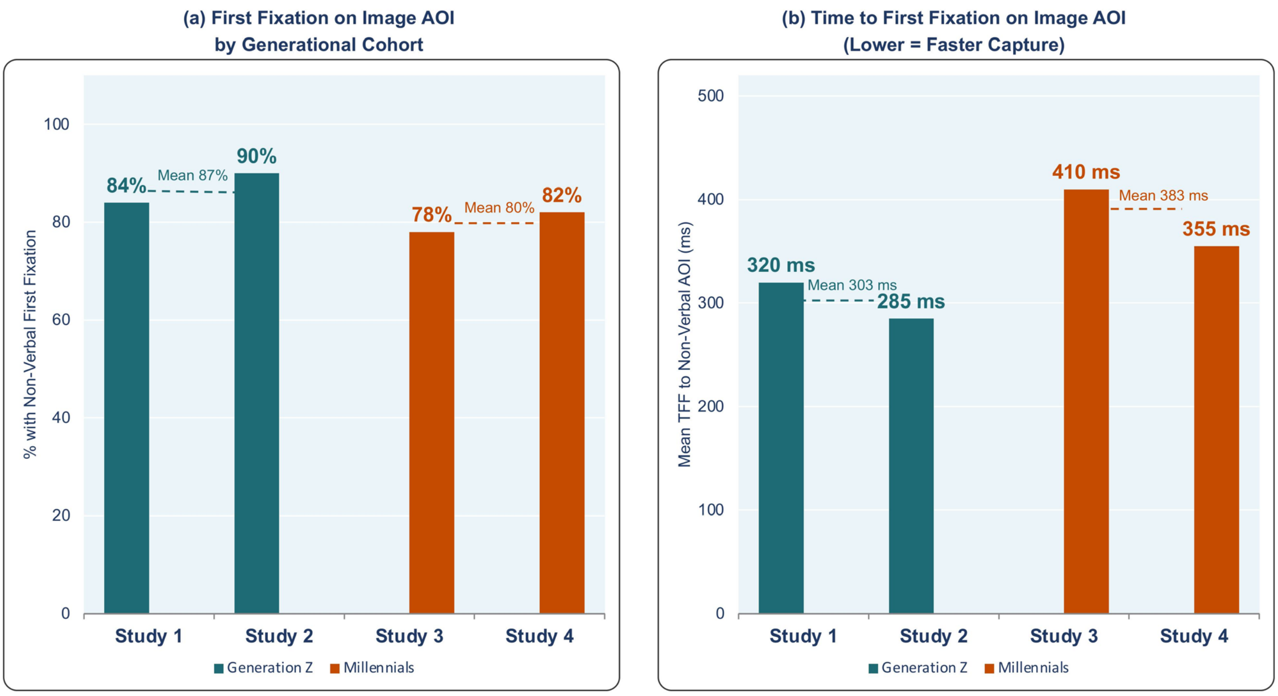

Fig. 5. Generational cohort comparison: (a) proportion of participants with first fixation on the non-verbal image AOI, and (b) mean Time to First Fixation (TFF) to the non-verbal AOI, by generational cohort and study.

Table 5. Cross-study summary of eye-tracking metrics by study and generational cohort

Study and Cohort	FFP Non-Verbal AOI (%)	FFP Verbal AOI (%)	TFF Non-Verbal (ms)	TFD Non-Verbal (s)	FC Non-Verbal
1, Gen Z	84%	16%	320 ms	1.40 s	3.5
2, Gen Z	90% *	10%	285 ms	1.85 s	4.8
Gen Z Mean	87%	13%	303 ms	1.63 s	4.2
3, Millennials	78%	22%	410 ms	1.55 s	3.9
4, Millennials	82%	18%	355 ms	1.70 s	4.2
Millennials Mean	80%	20%	383 ms	1.63 s	4.1

* Study 2 highest FFP proportion across all four studies

Summary of Key Findings

Across all four studies and both generational cohorts, the following patterns were consistently observed without exception: (1) In all four studies, the majority of participants directed their first fixation to the non-verbal (image) AOI before the verbal (text) AOI, regardless of communicative context or generational cohort; (2) Time to First Fixation (TFF) on the non-verbal AOI was consistently 2 to 3 times shorter than TFF on the verbal AOI across all studies (non-verbal mean: 343 ms vs. verbal mean: 845 ms); (3) Total Fixation Duration (TFD) on non-verbal AOIs substantially exceeded TFD on verbal AOIs across all four studies, indicating greater sustained attentional investment in image elements; (4) Fixation Count (FC) on non-verbal AOIs was consistently higher than on verbal AOIs, confirming repeated attentional returns to image regions across all four communicative contexts; (5) Generation Z participants (Studies 1 & 2) showed a marginally higher mean non-verbal FFP proportion (87%) and faster mean TFF (303 ms) than Millennial participants (Studies 3 & 4: 80%, 383 ms), though both cohorts showed strong and consistent non-verbal primacy; and (6) Within Study 4, no statistically significant difference in non-verbal FFP proportion was found between Millennial and non-Millennial participants ($p = .68$), confirming that non-verbal attentional primacy was not moderated by generational cohort membership within a single study.

IV. DISCUSSION

This cross-contextual synthesis examined whether non-verbal visual elements capture first fixation before verbal elements across four multimodal communication contexts and two generational cohorts. Across all studies, participants directed first gaze more often and more rapidly to non-verbal AOIs than to verbal AOIs, indicating a stable pattern of non-verbal attentional primacy. The pattern was observed in health communication, digital media exposure, e-commerce, and meme-based persuasion, suggesting that early attention in multimodal discourse is structured by a consistent visual-first bias rather than by a single domain-specific effect.

The main finding supports the view that images and text do not compete on equal terms for initial attention. In social semiotic accounts, images afford simultaneous and spatially distributed meaning-making, whereas written language requires sequential decoding (Eisenlauer & Karatza, 2020; Machin, 2016). This asymmetry helps explain why non-verbal AOIs were fixated earlier across all four studies. The result is also compatible with evidence that early visual processing is highly sensitive to low-level features such as contrast, colour, and form, which can guide gaze within the earliest stages of viewing (Rucci et al., 2018; Ziemba et al., 2024).

The findings also extend Barthes' distinction between the iconic and linguistic message. Across all stimulus types, the image functioned as the first attentional entry point, while text was typically processed later. In this sense, the data suggest that in multimodal discourse the image often anchors the encounter attentionally, even when verbal text remains necessary for elaboration or interpretation.

Both Generation Z and Millennial participants showed strong non-verbal attentional primacy. Generation Z displayed a somewhat higher non-verbal FFP proportion and faster TFF than Millennials, which is consistent with the idea that this cohort developed within highly visual, platform-based media environments (Priporas et al., 2017; Turner, 2015). By contrast, Millennials were socialised during a transition from more text-dominant digital communication to image-rich social media, which may help explain their slightly slower non-verbal orienting (Djafarova & Bowes, 2021; Sousa & Rocha, 2019).

However, the difference between cohorts should not be overstated. In Study 4, Millennial and non-Millennial participants viewed the same meme stimuli under the same conditions, and the cohort difference was not statistically significant. This suggests that generational background may slightly modulate the speed of orienting, but not the basic pattern itself. Non-verbal primacy therefore appears to be a robust feature of multimodal processing rather than a cohort-specific effect. This interpretation aligns with the neuroscientific evidence that bottom-up visual saliency, the automatic orienting of gaze toward high-contrast, figural, and chromatically distinctive elements, is phylogenetically ancient and does not vary substantially with ontogenetic or cultural experience (Veale et al., 2017).

A major strength of the synthesis is that the same attentional pattern appeared across four distinct communicative settings. The studies varied in emotional salience, communicative purpose, and medium, yet the non-verbal AOI consistently attracted first fixation before the verbal AOI. This convergence strengthens the interpretation that non-verbal primacy is not simply a product of one stimulus genre or one participant group.

At the same time, the results do not imply that context is irrelevant. Context may shape later viewing behaviour, as shown in Study 1, where smokers shifted attention away from the health warning image toward the brand logo after initial fixation. The present evidence therefore supports a more specific claim: contextual variation may influence sustained attention, but the earliest stage of attentional entry into multimodal stimuli remains predominantly non-verbal (Peng et al., 2017).

The present findings are consistent with earlier research showing that pictorial elements in advertising attract attention efficiently and can guide later processing of brand information (Pieters & Wedel, 2004). Here, that pattern extends beyond conventional advertising to product displays and digital memes, indicating that visual-first attention also operates in contemporary digital formats. The health communication findings are likewise in line with evidence that graphic warnings are more salient than text-only warnings (Pang et al., 2021), while the later avoidance shown by smokers resembles previously documented coping responses to threatening tobacco imagery (Maynard et al., 2014).

Study 2 also fits attentional bias models in which repeated exposure sensitises viewers to category-relevant cues (Castro-Calvo et al., 2021). Participants with higher pornographic media exposure showed stronger early orientation to sexual image content, suggesting that non-verbal primacy can be amplified by stimulus familiarity and motivational salience. This does not weaken the broader conclusion; instead, it shows that a general visual-first tendency can interact with domain-specific bias.

These findings have practical implications for language-sensitive communication design in multilingual and plurilingual settings. If viewers attend first to non-verbal elements, then images should be treated as a primary entry point rather than as decorative support. This is especially relevant for health communication, e-commerce, and education, where audiences may differ in language background but still share comparable visual-attentional tendencies (Cope & Kalantzis, 2009; Lim, 2025). In Indonesia, where communicative materials circulate across highly diverse linguistic communities, the non-verbal layer may be the most reliable means of securing initial attention before verbal comprehension begins.

The study also has important limitations. All four studies were conducted in Indonesia, so the findings should not yet be generalised to contexts with different reading directions, orthographies, or media ecologies. In addition, none of the component studies measured verbal cognitive style or reading expertise, both of which may affect how viewers balance verbal and non-verbal information during multimodal processing (Blazhenkova & Kozhevnikov, 2010; Kozhevnikov et al., 2005). Future research should test the same paradigm across cultures and scripts, and should include cognitive-style and literacy measures to identify when verbal processing may become more competitive with early visual capture.

V. CONCLUSION

This study integrated eye-tracking data from four experiments conducted in Indonesian health, commercial, and digital media contexts to examine whether non-verbal visual elements systematically precede verbal elements in visual attention, and whether this pattern holds across Generation Z and Millennial cohorts. Across all studies, non-verbal areas of interest reliably attracted first fixation for the vast majority of participants and did so substantially faster than verbal text, demonstrating a robust attentional hierarchy in which the iconic mode precedes the linguistic mode in multimodal discourse processing. Generational comparisons revealed only marginal differences in the strength of this effect, with Generation Z showing slightly faster non-verbal orienting than Millennials, but the overall pattern of non-verbal primacy remained stable across cohorts and communicative domains.

These findings provide neurophysiological grounding for a central claim in multimodality and visual semiotics: images afford simultaneous, globally parallel processing, whereas written language requires sequential decoding, and this affordance asymmetry manifests as a consistent attentional advantage for non-verbal elements. In plurilingual societies such as Indonesia, where communicative materials must function across diverse linguistic repertoires, this attentional hierarchy positions the visual, non-verbal layer as the principal communicative entry point for health warnings, commercial messages, and educational materials, with verbal components operating as secondary elaboration. Future research should extend this cross-contextual synthesis to right-to-left reading populations, additional communicative genres, and individual differences in cognitive style and literacy, to test the generality and boundary conditions of non-verbal attentional primacy in multimodal communication.

ETHICS STATEMENT

This article reports a synthesis of previously published eye-tracking studies and does not involve new interventions with human or animal subjects. All underlying studies were conducted in accordance with the ethical guidelines of Universitas Muhammadiyah Prof. Dr. HAMKA and Universitas Prima Indonesia, including institutional ethics approval and written informed consent from participants.

ACKNOWLEDGEMENTS

This research did not receive specific funding from government agencies, commercial sponsors, or not-for-profit organisations. AI-based tools were used only for language editing and grammar checking.

CREDIT AUTHOR STATEMENT

Rizki Edmi Edison: Conceptualisation, supervision, research design, data integration, manuscript drafting and critical revision.
Deni Adha Akbari: Study design (commercial/advertising domains), data collection and analysis, interpretation, manuscript review and editing.
Pratiwi Cristin Harnita: Study design (digital media domain), data collection and analysis, interpretation, manuscript review and editing.
Widya Yanti Sihotang: Data organisation and verification, manuscript formatting according to journal template, language editing and proofreading.

DECLARATION OF COMPETING INTERESTS

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

REFERENCES

Adha Akbari, D., Sumarwan, U., Hasanah, N., & Edison, R. (2023). Exploring just noticeable difference between AI endorsers and human endorsers in advertising campaigns. *The Seybold Report*, *18*(8). https://doi.org/10.5281/zenodo.8241117

Avia, V. A., S, W. C., Zendrato, C. D. K. N., Nasution, M., & Edison, R. E. (2025). Perbedaan exposure konten pornografi terhadap atensi visual remaja. *Jurnal Pendidikan Indonesia*, *6*(11), 4938–4951. https://doi.org/10.59141/japendi.v6i11.8782

Balaskas, S., Yfantidou, I., & Skandali, D. (2025). Eyes on prevention: An eye-tracking analysis of visual attention patterns in breast cancer screening ads. *Journal of Eye Movement Research*, *18*(6), Article 75. https://doi.org/10.3390/jemr18060075

Blazhenkova, O., & Kozhevnikov, M. (2010). Visual-object ability: A new dimension of non-verbal intelligence. *Cognition*, *117*(3), 276–301. https://doi.org/10.1016/j.cognition.2010.08.021

Castaldi, J. (2024). Refining concepts for empirical multimodal research: Defining semiotic modes and semiotic resources. *Frontiers in Communication*, *9*, Article 1336325. https://doi.org/10.3389/fcomm.2024.1336325

Castro-Calvo, J., Cervigón-Carrasco, V., Ballester-Arnal, R., & Giménez-García, C. (2021). Cognitive processes related to problematic pornography use (PPU): A systematic review of experimental studies. *Addictive Behaviors Reports*, *13*, Article 100345. https://doi.org/10.1016/j.abrep.2021.100345

Cope, B., & Kalantzis, M. (2009). "Multiliteracies": New literacies, new learning. *Pedagogies: An International Journal*, *4*(3), 164–195. https://doi.org/10.1080/15544800903076044

Dimock, M. (2019). *Defining generational cutoffs: The end of Millennials and the rise of Gen Z*. Pew Research Center.

Djafarova, E., & Bowes, T. (2021). 'Instagram made me buy it': Generation Z impulse purchases in fashion industry. *Journal of Retailing and Consumer Services*, *59*, Article 102345. https://doi.org/10.1016/j.jretconser.2020.102345

Edison, R. E., Anisa, Y. H., & Fauzy, F. R. (2021). Analysis of visual attention among smoker and non-smoker by using human eye tracker of health graphic warning on tobacco package. *Jurnal Kesehatan Masyarakat Andalas*, *15*(2), 3–7. https://doi.org/10.24893/jkma.v15i2.656

Eisenlauer, V., & Karatza, S. (2020). Multimodal literacies: Media affordances, semiotic resources and discourse communities. *Journal of Visual Literacy*, *39*(3–4), 125–131. https://doi.org/10.1080/1051144X.2020.1826224

Gatcho, A. R. G., Manuel, J. P. G., & Sarasua, R. J. G. (2024). Eye tracking research on readers' interactions with multimodal texts: A mini-review. *Frontiers in Communication*, *9*, Article 1482105. https://doi.org/10.3389/fcomm.2024.1482105

Holmberg, N., Holmqvist, K., & Sandberg, H. (2015). Children's attention to online adverts is related to low-level saliency factors and individual level of gaze control. *Journal of Eye Movement Research*, *8*(2), Article 2. https://doi.org/10.16910/jemr.8.2.2

Kaufman, A., Klein, E. G., Koblitz, A., & Price, S. (2018). A vision for eye-tracking research in tobacco regulatory science. *Tobacco Regulatory Science*, *4*(6), 3–7. https://doi.org/10.18001/TRS.4.6.1

Kozhevnikov, M., Kosslyn, S., & Shephard, J. (2005). Spatial versus object visualizers: A new characterization of visual cognitive style. *Memory & Cognition*, *33*(4), 710–726. https://doi.org/10.3758/BF03195337

Lieberman, M. D. (2007). Social cognitive neuroscience: A review of core processes. *Annual Review of Psychology*, *58*(1), 259–289. https://doi.org/10.1146/annurev.psych.58.110405.085654

Lim, F. V. (2025). Multimodality in multilingual classrooms. In *The encyclopedia of applied linguistics* (pp. 1–7). Wiley. https://doi.org/10.1002/9781405198431.wbeal20403

Lochbuehler, K., Tang, K. Z., Souprountchouk, V., Campetti, D., Cappella, J. N., Kozlowski, L. T., & Strasser, A. A. (2016). Using eye-tracking to examine how embedding risk corrective statements improves cigarette risk beliefs: Implications for tobacco regulatory policy. *Drug and Alcohol Dependence*, *164*, 97–105. https://doi.org/10.1016/j.drugalcdep.2016.04.031

Machin, D. (2016). The need for a social and affordance-driven multimodal critical discourse studies. *Discourse & Society*, *27*(3), 322–334. https://doi.org/10.1177/0957926516630903

Marcellina, T., Indra, E., & Edison, R. E. (2024). Analisis perbedaan atensi visual customer Shopee terhadap produk flash sale dengan produk normal berbasis eye-tracking. *Jurnal Teknik Informasi dan Komputer (TEKINKOM)*, *7*(1).

Maynard, O. M., Attwood, A., O'Brien, L., Brooks, S., Hedge, C., Leonards, U., & Munafò, M. R. (2014). Avoidance of cigarette pack health warnings among regular cigarette smokers. *Drug and Alcohol Dependence*, *136*, 170–174. https://doi.org/10.1016/j.drugalcdep.2014.01.001

Pang, B., Saleme, P., Seydel, T., Kim, J., Knox, K., & Rundle-Thiele, S. (2021). The effectiveness of graphic health warnings on tobacco products: A systematic review on perceived harm and quit intentions. *BMC Public Health*, *21*(1), Article 884. https://doi.org/10.1186/s12889-021-10810-z

Parry, E., & Urwin, P. (2011). Generational differences in work values: A review of theory and evidence. *International Journal of Management Reviews*, *13*(1), 79–96. https://doi.org/10.1111/j.1468-2370.2010.00285.x

Peng, J.-E., Zhang, L., & Chen, Y. (2017). The mediation of multimodal affordances on willingness to communicate in the English as a foreign language classroom. *TESOL Quarterly*, *51*(2), 302–331. https://doi.org/10.1002/tesq.298

Pieters, R., & Wedel, M. (2004). Attention capture and transfer in advertising: Brand, pictorial, and textsize effects. *Journal of Marketing*, *68*(2), 36–50. https://doi.org/10.1509/jmkg.68.2.36.27794

Priporas, C.-V., Stylos, N., & Fotiadis, A. K. (2017). Generation Z consumers' expectations of interactions in smart retailing: A future agenda. *Computers in Human Behavior*, *77*, 374–381. https://doi.org/10.1016/j.chb.2017.01.058

Quintiliani, C. (2016). Reading the visual: An introduction to teaching multimodal literacy. *Brock Education Journal*, *25*(2). https://doi.org/10.26522/brocked.v25i2.503

Rucci, M., Ahissar, E., & Burr, D. (2018). Temporal coding of visual space. *Trends in Cognitive Sciences*, *22*(10), 883–895. https://doi.org/10.1016/j.tics.2018.07.009

Sandelowski, M., Voils, C. I., & Barroso, J. (2006). Defining and designing mixed research synthesis studies. *Research in the Schools*, *13*(1), 29–40.

Soleh, M. B., Anisa, Y. H., Absor, N. F., Suswandari, & Edison, R. E. (2021). Differences of visual attention to memes: An eye tracking study. *Advances in Social Science, Education and Humanities Research*, *536*, 8–12. https://doi.org/10.2991/assehr.k.210430.022

Sousa, M. J., & Rocha, Á. (2019). Skills for disruptive digital business. *Journal of Business Research*, *94*, 257–263. https://doi.org/10.1016/j.jbusres.2017.12.051

Suhendi, E. T. (2025). Revitalizing the Indonesian language as a national identity in the globalization era: Challenges and strategies. *Journal of English Language and Education*, *10*(1), 188–195. https://doi.org/10.31004/jele.v10i1.635

Theeuwes, J. (2010). Top–down and bottom–up control of visual selection. *Acta Psychologica*, *135*(2), 77–99. https://doi.org/10.1016/j.actpsy.2010.02.006

Turner, A. (2015). Generation Z: Technology and social interest. *The Journal of Individual Psychology*, *71*(2), 103–113. https://doi.org/10.1353/jip.2015.0021

Van der Cruyssen, I., Ben-Shakhar, G., Pertzov, Y., Guy, N., Cabooter, Q., Gunschera, L. J., & Verschuere, B. (2023). The validation of online webcam-based eye-tracking: The replication of the cascade effect, the novelty preference, and the visual world paradigm. *Behavior Research Methods*, *56*(5), 4836–4849. https://doi.org/10.3758/s13428-023-02221-2

Veale, R., Hafed, Z. M., & Yoshida, M. (2017). How is visual salience computed in the brain? Insights from behaviour, neurobiology and modelling. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *372*(1714), Article 20160113. https://doi.org/10.1098/rstb.2016.0113

Wedel, M., & Pieters, R. (2008). Eye tracking for visual marketing. *Foundations and Trends in Marketing*, *1*(4), 231–320. https://doi.org/10.1561/1700000011

Xinna, L., Hannah, S., & Lai, M. (2025). Algorithm awareness and apathy: An exploratory study of Generation Z's media literacy. *Media Education Research Journal*.

Ziemba, C. M., Goris, R. L. T., Stine, G. M., Perez, R. K., Simoncelli, E. P., & Movshon, J. A. (2024). Neuronal and behavioral responses to naturalistic texture images in macaque monkeys. *The Journal of Neuroscience*, *44*(42), Article e0349242024. https://doi.org/10.1523/JNEUROSCI.0349-24.2024