ingegneria sismica
International Journal of Earthquake Engineering
Anno XLIII - Num. 2 - 2026

Visual Symbol Construction in Generative AI Short Videos and the Mechanisms of User Attention Capture: Evidence from an Eye-Tracking Experiment

Caiwen Zhao[1] *and Xiaoyi Peng*[1,*]

[1] School of Media and Arts, Wuhan Qingchuan University 050306T, Hubei, China

SUMMARY: *Generative AI has become an ordinary production means of vertical short videos, but it is difficult to judge the effect on users from platform log data only. Does the creation of visual symbols help attract users' attention more quickly? How are gaze patterns distributed among different Dynamic Areas of Interest (DAOs); Under what circumstances do generated Symbols facilitate perception over inspection? Nine-six short video user groups, age ranging from 18 to 35 years old. The stimulus pool contained 72 clips balanced by production mode, including fully AI-generated, hybrid AI-assisted and human-made videos. Each person saw 18 video clips in random order. Following gaze quality control, there remained 1,641 participants' videos. Dynamic AOIs are annotated with contents such as faces/characters, objects' movements, text labels, motion boundaries, call-to-action areas, etc. Time to the first fixational eye movement, symbolic duration reaction, refixation rate, re-entry probability, transition uncertainty, pupillary dilatation, identification corrects, continuous motivation levels for individuals.*

KEYWORDS: *Generative AI; Short Video; Visual Symbol; Eye Tracking; Attention Capture.*

1 Introduction

Short videos transmit information in a short period of time. A creator has to make the topic visible before the viewer scrolls, and the platform interface gives the viewer little friction for leaving. Under these circumstances, the first-image-region that captures gaze should determine whether to continue inspecting or not based on it. Generative AI has transformed the production environment for this attention problem. Creators can synthesise Faces, Products, Backdrops, Motion Transitions and Prompt-Like Labels for use at low cost. As a result, the resulting short videos generally have smooth surfaces and dense clusters of symbols. Quickly display objects on the screen and, at the same time, generate multiple cues that compete for attention of the audience; eventually creating an interpretive consensus among viewers.

This paper takes as its empirical subject an artificial intelligence-generated short video's visual symbol. Hereinafter referred to as a visual symbol will be defined as an observable element bearing communicative intention; A face used to anchor social Presence, a product Transformation representing function, Text Label for Event Naming, Motion Boundary marking Change, and Background providing Scene Identity etc., respectively. This Definition retains the Analysis Close to What Viewers See. Separate the generation of production modes from the appearance structures of clips. Two AI-created videos may have very different levels of symbols, smooth movements and coherent meanings. The study therefore asks how these

*pengxiaoyi0323@163.com
https://doi.org/10.65102/is2026780
Published: 30 April, 2026

observable properties guide gaze and how they influence later recognition.

The existing short-form video research shows that the relationship between feed Design, multiple exposures and users' Motivation is Attention, Memory and Continuation Usage. The TikTok-related works have focused on its fragmented viewing Environment. Recently, some laboratory studies have also confirmed that short-Video influence Event Segmentation Changes and Memory Encoding; Eye movement Synchronization at Event Boundaries has changed. Work on TikTok use also shows that watch, share and creation behavior are shaped by user motivation [1-3]. Based on this foundation, in the brief form experiment mentioned above, either self-reporting or platform tracking data could not show accurately how attentive the audiences were. This paper takes as its research subject generated visual contents; thus, the objects of attention in these images can be synthetic cues instead of recorded scenes.

Research on Generative Artificial Intelligence provides another basis for this research. The applications of generative systems in image, audio-visual and other media have been growing rapidly [4, 5]. Ad Research work on Advertising is also forming standardised processes to create experimental Stimuli using generative AI, which maintains Manipulation validity and ethical documentation [6-8]. Some studies have shown that Artificial-intelligence-generated Images may affect people's judgment on their own behaviour or the behaviours of others. The relevant issue for short video is narrower and more operational. A feed viewer rarely evaluates a generated clip as an isolated image. The viewer enters the clip through gaze, follows moving regions and forms a judgment under time pressure. Therefore, in terms of the mechanism, the observed sequence of look is necessary.

Several recent studies have gradually approached the above-mentioned mechanism through integration of AI-image perception technology and eye tracking/correlated judgement tasks. Eyetracking evidence shows that viewers of MidJourney-created images utilise visible heuristic to judge whether an image is artificial [9]. Comparison research shows that the recognitions and recognitions of artificial intelligence-created pictures are not entirely based on perception; They can be determined by other forms of vision. However, these research results focus more on single-frame static objects or pairs of images. Short videos add time-based organisation. A generated object may appear, transform, become partially occluded and be accompanied by a text cue during the same exposure. Therefore, how can the created symbols attract attention over Time?

Research on Eye-tracking in Social Media and Marketing provides an actual measurement path. Studying the restaurant Pages, food posts on social media platforms and other cues shows that Pictures, Ratings, Likes and Contextual Labels can attract people's attention, guide evaluations [10-13]. Some research shows that, under the framework of Platform-like Cues as AOIs and associated with Memory or Intention. This study applies the same reasoning, but switches the units of its analysis. Rather than evaluating whether a visible like-count or rating is attractively focused on, it examines whether synthetic symbols in vertical videos alter time points, frequencies and pathways of eye movement. The measurement target is now an active motion of symbols in view-based scanning trajectories.

The current evidence has three defects. Firstly, the production mode tends to be a general term; that is, some symbolic elements in the video are not clearly indicated. Secondly, most research focuses on assessing image perception or post-exposure credibility judgments without monitoring participants' eye movements while watching short videos. Third, attention is often evaluated through dwell time alone. Dwell time cannot distinguish productive inspection from repeated checking caused by motion instability, clutter or semantic mismatch. This paper aims to close these gaps through a combination of dynamic AOI code, a compound attention-capture index and post-exposure outcomes in the same experiment.

Based on a design of the three production processes: full artificial intelligence generation

clips; hybrid artificial intelligence-assisted clips; Human-made clips. The stimulus pool is uniformly divided among the content categories and lengths. Coded using the following indices for each clip: symbol density, Motion Discontinuity, Face Fidelity; Semantic Coherence; Disclosure and Visual Clutter. Record eye movement for 15 seconds per exposure; each time, there will be a Recognition and continuation-intent test afterwards. This Design enables Analysis To Determine Whether The Generated Symbols Capture Attention Earlier; Whether They Produce Stable Or Fragmented Scan Paths; And How Captured Information Aids Memory or Inspections Only.

There are four contributions. First, the study operationalizes visual-symbol construction in generative AI short videos with frame-level variables and dynamic AOIs. Second, it builds an attention-capture index from orienting speed, symbolic dwell, refixation and return behavior. Thirdly, using mixed-effects models and response surface visualisations to estimate the combined impact of symbol density, motion discontinuity and semantic coherence. Fourth, it distinguishes visual capture from communicative effectiveness by comparing gaze outcomes with recognition and continuance intention. Therefore, this paper is considered a type of experiment on attention capture in generated short videos; the presented results are presented using aligned Data Tables, model Estimates, and result Figures.

In platform practice, the same clip may be used for product explanation, destination promotion, public-service messaging or entertainment recommendation. These use cases all have a common constraint that the user generally sees the video before making a judgment. A generated visual symbol therefore has to act before the user has formed a conscious evaluation of the clip. Thus, the visual-symbols Construction differs from subsequently acquired attitudes. The first few seconds show that which signal is chosen by the audience; After this, whether the selected one still serves as a signal or has become an opposing object can be observed. The Study takes this time-ordered arrangement as its primary research issue.

The concept of generative AI-generated short videos used herein is limited to empirical definitions. It is not a type of all videos using artificial intelligence technology. A type of clip where, after generation, a certain visual element can be seen within the boundaries of the scene. A clip can therefore be fully generated, hybrid or human-made. This difference exists; many creators use a combination of camera shots and generated background images, product states, etc. A binary AI-versus-human label would eliminate these cases. The experimental Design separates them to ensure that the analysis can verify whether there is an attention difference between mixed-construction and full-generation-based visual-constructive systems.

Another cause of experimental data collection is that attention capture cannot be observed through the indicators of platform participation. A high completion rate can come from interest, confusion, social context or interface inertia. High replay rates may be because of enjoyment or unsolved visual clues. Eye-tracking is also an excellent approach to Traditional Methods; it can track the Path of Viewers' Scenery viewing, Analyse Symbol-rich Areas, Revisit Places Visited Before Other Approaches. Generated Short Videos of this type are most instructive. Re-visiting the generated face and transforming boundaries frequently; thus, whether it means an elaborated explanation or simply artifact verification. Only then will the difference be observed between gaze measurements and recognition or persistence results.

Mainly expect that the generated visual symbols can create a more significant bottom-up entry Signal than human-made videos; Because generated Objects usually have centralization, contrast, smoothness, and accompanying clear labels or transformations. Another one is that the Entry Signal has a upper bound. If there are too many generated cues at once, the gaze will divide among them, affecting recognitions. The third Expectation is that hybrid AIL-enhanced clips can generate an Attention Pattern More Efficiently due to the incorporation of generated assets in a human-designed Visual Composition. These Expectations constitute the empirical

contrast tested as follows: speed of acquisition, stability of scanpath organisation and downstream recognitions.

Thus, the final article will have an experimental Structure. Methods describe how to build the clip pool; How to code dynamically defined Areas of Interest (AOI) and visual-symbol variables; The way to conduct an eye-tracking session; And specify a mixed-effect model. The results and analysis section will provide the following contents with eight figures: temporal sampling; area of interest selection; coupled symbol effect; ablation experiment check; erroneous source location. This organisation brings the text closer to the data. Each Figure is associated with one measurement problem, and each interpretation belongs to one report result; there are no promises or risks of generative artificial intelligence in general claims but only under particular reports.

The selection of eye-tracking is based on which type of generation failure has occurred. The generated clips are continuous frames that carry some sense of breakage in identity when it comes to movements of objects, expressions changed by people or alterations in backdrops. Such breaks may last less than a second and may never appear in a post-exposure self-report. They can still redirect gaze and increase return behavior. Therefore, in this experiment, it is treated as a time-varying object; Symbols are judged based on their appearance location, duration of visibility; Changes in Identity; And Whether the observer goes back for them after leaving?

This emphasis, therefore, is different from all previous treatments of AI media in general. The research does not inquire about whether audience members like AI-produced works in general. Does it change observable eye-movement behavior after being exposed to a code-constructed set of generated cues? Expectedly, the result will have conditions attached. The generated faces, labels and transformations need to have higher recognition rates and be closer to the main focus. The same functions will lead to a deterioration in the effect under fragmented gaze routes. To verify the relationship among data, figures and models' output through a practical experiment.

Therefore, this article will use a limited amount of literature and give primary evidence from the experiment. Citations establish the platform background, the Stimulus-Construction Problem, and the history of eye-tracking research. Finally, the primary assertions will be based on these retained trials, Dynamic AOIs, model coefficients, and result figures. Thus, it will present the results of this study in a focused empirical portion.

2 Methods

2.1 Stimulus construction and visual-symbol coding

The experimental stimulus set did not include the scraper; it was all controls. Due to the need for a balance among the production mode, duration, frame rate, and type of contents when collecting gaze data. Finally, there were 72 vertical short videos of about one minute each, all displayed at a size of 9:16. Clips were divided into four types of content: consumer products Display, Tourism Scenes, Public Information Explainers and Entertainment Micro-narratives. There are six full-length AI-generated clips in each group; six half-featured AI assistance and human-produced clips, respectively. There were, thus, only two types of productions each week and it did not become too much that one type would dominate;

All of fully AI-generated clips have produced images and videos for the primary characters, objects or scenes. The editing content is only duration adjustment, audio normalisation and resolution standardisation. The hybrid-aided clips contained at least one generated visual component (background, product variants, characters' illustrations or transition layers) and still

had human-controlled filming, positioning or editing elements. Human-created clips, which include camera-logged and hand-picked content without generating any visual materials. Based on the source of the visible image area, as there were no data collected regarding gaze towards written characters during this research.

The candidate pool began with 126 clips. Only those candidate clips featuring recognisable subjects or characters after more than 70 per cent exposure time were retained; no private individuals should have been present without permission; there must not be any violent content, medical fictions or political misrepresentation in these clips. AI-generated clips containing highly distorted anatomy or collapsed objects were excluded from the analysis of attention to usable generated symbols; such cases dominated gaze by error alone in a normal distribution. Low-visual-information artificial clips were eliminated from the training set to make the environments more comparable. Of the remaining 72 videos after exclusion, there are 12 groups of them arranged by Latin squares.

The visual-sympbol coding of frames was carried out. The following definition of the coded vector:

$$\mathbf{x}_v=\{D_v,S_v,M_v,F_v,C_v,L_v,B_v\} \tag{1}$$

x_v in this context represents a visual-symbols Profile for Video v. D_v represents symbol density; S_v represents Spatial salience; M_v stands for motion discontinuity, etc.; F_v also includes Face Fidelity, Semantic Coherence, AI label disclosure, etc., and it also contains visual clutter. Each variable was normalised in the range of [0, 1], and then z-score transformations were applied. Symbol density calculated the quantity and duration of significant signals. Space-Saliency mainly includes contrast, Size, and Centrality. Motion discontinuity marked visually abrupt changes in objects or boundaries. Face fidelity showed stable visibility in the generated or filmed faces. Check for loss of semantic coherency during symbol Recognition. Artificial intelligence-label disclosure revealed generated labels or watermarks. visual clutter: Unrelated and competing areas.

Coding rubric divided symbol density and general complexity separately. In line with the observation of visual attention that low-level saliency, scene semantics and natural viewing directions may guide gaze fixation along different paths [14-16]. If all three types of clues in a tourism video show consistent destinations after their combination, then the corresponding symbol density is high. Clauses including decorations, excessive redundantly duplicated titles and backdrops; The overlapping of these clauses with the main theme are considered more disorderly. As a result, artificial-intelligence-generated videos often have many visual elements but relatively sparse content or insufficiently conveyed meanings. The coders identified real movement from artefact-like discontinuities. The product launch revealing its own hidden functions is regarded as intelligible motion. A newly-generated object with a change in category or boundary, lacking narration, is classified as more discontinuous and less coherent.

Dynamic AOIs were created in six areas: faces/characters; objects that have been transformed; Text labels; Motion boundaries; Call-to-action buttons; Backgrounds. Research on dynamic Scenes indicates that gazes are directed towards motion, Faces, Sounds and Social Information in moving images during viewing activities [17-19]. AOI polygons were drawn at 5 Hz and linearly interpolated to the 120 Hz eye-tracker sampling rate. Only interpolation was performed for continuous scenes. Upon occurrence of a cut, AOI coordinates were reset to prevent artificial movement due to the transition of unrelated shots. In this way, it retains the time sequence in the short video while fixing on what can be seen at present instead of a fixed boundary box.

Two coder groups trained on visual-symbols and AOI areas. A third coder resolved disagreements when absolute differences exceeded 0.20 on a 0-1 variable or when AOI borders

diverged by more than 15 pixels for the same frame. The intercoder agreement rate before scoring. Mean intra-class correlation of continuous symbols ranged from 0.84; and the mean frame-wise OI overlapped by the intersection over union method was around 0.81. These were deemed sufficient for the trial-level gaze evaluation. The adjudicated coding file was thus set as the standard annotation reference for all participants.

The final data structure had a three-tiered structure. The first level was the frame-level annotation of symbol variables and AOI polygons. Secondly, the participant-video test: calculate gazing indicators and subsequent reaction scores after watching videos. The third group of participants comprised: Age, Gender; Daily short-video usage frequency; Platform familiarisation degree; Self-reported AI-content familiarity degree. The nested Structure allowed this model to be controlled by participants' and videos, estimate how Visual-sybar Construction Influences Gaze.

Stimulus Construction adhered to the principle adopted in experimental generative-AI stimulus Design; it permitted variations in the intended visual features and normalised unfocus-Related Attributes. Adjust the resolution to 1080x1920 and set it for 15 seconds exposure time; normalise mean volume noise level -16dB LUFS. Abrupt clips of audio peaks were edited or discarded. visual lumen and the number of cuts were treated as controls. Therefore, this type of response can be selected as the material pool for such an eye-tracking experiment on production modes and symbol creation.

Table 1 shows the stimuli and samples adopted in this study. As shown in Table 1, the retained data are relatively balanced by clips and selected at the trial stage. Therefore, the empirical Design avoids treating all AI-created clips as one unvaried group and offers sufficient data points for mixed-effects analysis.

Table 1: Stimulation and sample preparation for the eye-tracking study.

Item	Empirical setting	Retained observations	Rationale
Participants	96 short-video users, 18-35 years old	92 after participant-level quality screening	Normal or corrected vision, regular platform use
Stimuli	72 vertical clips, 15 s each, 9:16 frame ratio	24 AI-generated, 24 hybrid, 24 human-made	Balanced by content category and production mode
Trials	18 clips per participant under Latin-square assignment	1,641 valid trials from 1,728 raw trials	Trial exclusions based on gaze loss and playback checks
AOIs	Face or character, object transformation, text label, motion boundary, call-to-action, background	Dynamic polygons sampled at 5 Hz and aligned to 120 Hz gaze	AOI scheme maps visual symbols to gaze behavior
Post-exposure measures	Recognition, symbolic-cue identification, continuance intention	All retained trials have complete responses	Links attention to memory and viewing tendency

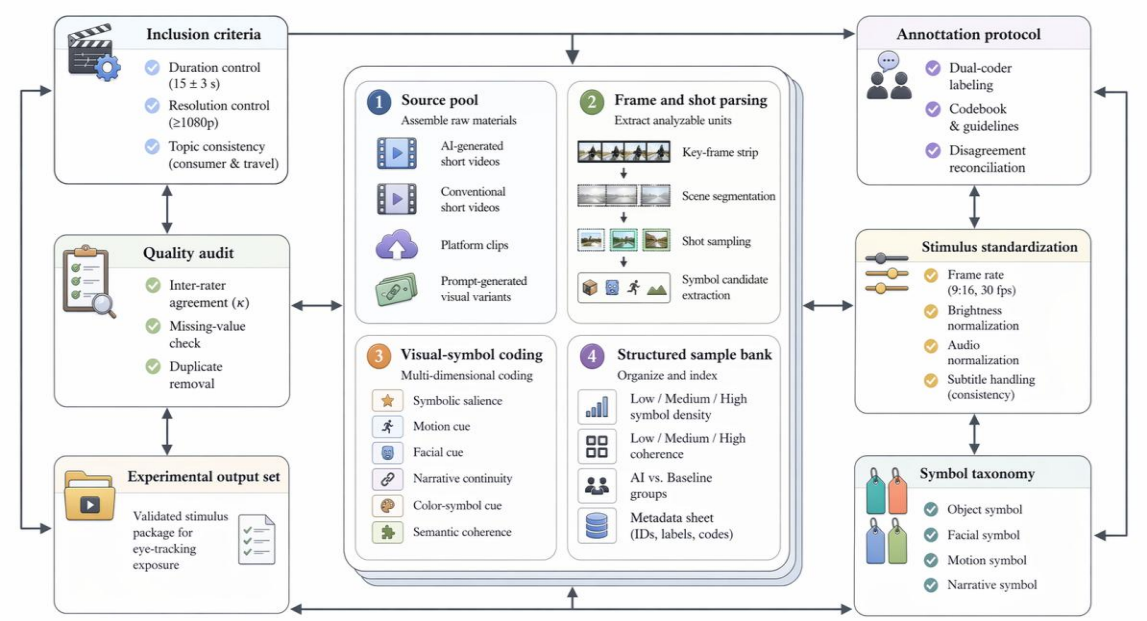

Figure 1: Data organisation and Visual-Symbol sample building system.

The planned participants would be for a repeated measures model rather than an unpaired comparison. Each individual watched the footage in all types of productions and saw it multiple times. To increase the Statistical Power and reduce the Risk of An individual participant, one single clip driving this observed effect. A Latin square assignment also helped distribute fatigue evenly among the different production modes. Each of the lists consisted of 6 artificial intelligence-generated short videos, 6 mixed video contents and 6 traditional humans-made videos.

Visual-symbols encoding was carried out earlier than gaze data analysis. Coders did not receive individual-gaze image data during the annotation of AOI and symbols. This Separation reduced the possibility of an overt fixation bias affecting code selection. The coding manual provided examples of high-density but organised symbols, high-density clutter, meaningfull motion discontinuity and artefact-like discontinuities. Borderline cases were considered only after separate codings. After freezing the final coded version before estimating the mixed-effects model.

Verify the stimulus equivalence of fundamental visual and textual information. The average number of cuttings was 4.1 for AI-generated videos, 3.8 for blended videos and 3.6 for human-created ones. The mean luminance of all productions varied by no more than 6.5%. The mean visibility time of the text was 4.9 seconds for AI-created videos, 4.6 seconds for mixed-generated videos and 4.4 seconds for human-creations. Because production mode necessarily changes visual form, these checks cannot make the conditions identical. To decrease the probability of an incorrect conclusion being attributed solely to length, brightness and overly cluttered texts.

Pilot testing of 8 people who were not included in the main study's participants. The pilot confirmed that the recognition contents were answerable and did not require searching for any AI artefacts. The item wording is adjusted if two or more pilot respondents interpret it as an AI detection task. Finally, the question included clip topic, message-supporting cue, and continuation intentions. This expression retained ordinary viewing Conditions and linked the exposure effect results with communication performance rather than artefact-detection results.

Coding of the Scale for Semantic Coherence included three observable standards. Check if the text label pointed to a visible object or event in the second step. The second checked whether

the main object preserved identity across cuts or transformations. Finally, was it confirmed that the end shot addressed the initial problem? Among clips that had achieved a high coherence score, many met all three criteria consistently during the evaluation. Received a low score for situations where there was inconsistency between two symbols, changes to objects did not provide messages due to lack of support, and background cues presented contradictory contexts compared to labels.

Face fidelity was coded only when a face or character region was visible for at least 2 s. Combined evaluation of eye fixation stability, mouth consistency, skin texture continuity and head-boundary consistency. This parameter is added to enhance the face's impact in short videos, as well as to boost interest from users created by generating faces. When there was no face or character display, the face- fidelity value was considered structurally absent and the corresponding model adopted a face-present criterion to prevent non-face frames from being assigned similar scores forcibly.

The AOI files were kept separate from the videos and associated with timestamps. Each AOI record included the video ID, frame Time, AOI category, polygon coordinates and coder ID. During preprocessing, gaze samples were merged with AOI records by nearest timestamp after playback synchronization. The organisation of this data helped the analysis reconstitute dwelling, movement and returning times for each trial. Also, it has been verified that a fixed-point transformer of objects or scenes can trace the relationship between the root polygon/scene node and videos.

2.2 Eye-tracking procedure and attention-capture metrics

Recruits participants from the school's subject pool and the local short-video users' community. Eligibility required age between 18 and 35, normal or corrected-to-normal vision, no self-reported neurological condition affecting eye movement, and regular use of short-video platforms. The lab finished in approximately 35 minutes. Participants were told that the study examined visual attention to vertical short videos. They were informed of eye-tracking recording and anonymised analysis; however, the production-mode comparison was revealed afterwards during the session. To decrease the probability of participants searching for artificial intelligence (AI) artefacts intentionally while under observation.

Collect eye-movement data with a common commercial device at 120Hz. Participants were about 65cm away from a 24-inch screen. Videos are presented in this way at the centre of the screen; a neutral-grey border is set around them. Comments, likes, recommendations labels and scroll buttons were deleted. This interface retained the vertical short-video object, and removed the platform interface cues that might conflict with it. Before the tasks began, all participants conducted a nine-point calibrations and validatisations separately. Calibrate again after the mean validation deviation had risen above seven-tenths of a degree visually.

Each experiment started with the central fixation for 800 milliseconds. The video was then played once for 15 s. After exposure, participants completed two recognition items and one continuance-intention item. The first recognitions involved asking respondents to select the central subject of three alternative items. Choose an optimal way of visual Guidance for this Part in response. A trial would be accepted if and only if both items were correct. The participants were asked their willingness to watch more of the same type of clips in future posts using a 7-point Likert Scale. The trial order was randomly assigned to each group, and none of the participants watched multiple episodes of the same content theme.

Preprocessing of gaze data for AOI allocation. Remove samples marked as invalid by the tracker. Exclude blink signals, and interpolate the gap shorter than 150ms. Based on a speed-dispersion rule that has been confirmed through verification under short-video observation conditions and in comparison to the I2MC method, fixations have been identified. Exclusions:

Exclusion of trials where gaze loss reached above 25%; exclusion due to fewer than three fixations; Participants looking away for the first two seconds; Desynchronisation during playback recording. After elimination, there were 1,641 viable trials out of the original 1,728, which retained 95.0% accuracy.

AOI assignment adopted dynamic Polygons introduced in Section 2.1. Assign a fix for an AOI if its centre point appeared within that AOI polygon at the corresponding time of capture. If there were multiple overlapping areas of AOI, then give them priority in order by size (smaller area) or specificity (more meaningful region). If the fixate exceeded all symbolic AOBs, then it would be placed in the Background group. All the production modes used the same assignment rule. Therefore, the generation of this clip did not receive greater symbols merely due to its larger area.

The attention-capture index combined orienting, sustained allocation and recurrence. As per the definition below:

$$ACI_{pv}=0.35z(DW_{pv})-0.25z(TTFF_{pv})+0.20z(RF_{pv})+0.20z(RP_{pv}) \quad (2)$$

The attention-capture index of Participant P when they view Video Bv, denoted as A. A is the symbolic dwell ratio; A is Time To First Fixation On Any Symbol-bearing AOI; A is Refixation Count And B Is Return Probability. Z-Operator standardise the individual components of the participants. Therefore, there is a positive effect on capturing faster according to -TTFF. The weight prioritises dwelling, orientation, but retains the recurrence term. Sensitivity checks used an equal-weight version, and the main ordering of conditions remained stable.

Scannable path fragmentation, calculated using transition entropy. Therefore, as shown in the following form.

$$H_{pv}=-\sum_{i=1}^{K}\sum_{j=1}^{K}P_{pv}(i,j)\log P_{pv}(i,j) \quad (3)$$

H_{pv} refers to the transition entropy of a participant-video trial in this expression. $P_{pv}(i,j)$ is the observed probability of a transition from AOI i to AOI j, and K is the number of AOI categories. Entropy has been normalised between 0 and 1 through division by the highest achievable entropy under K classes. A high value indicates a more dispersed route across AOIs. In the current study, high entropy was defined as scattered routes only when accompanied by a weak recognition score or significant pupillary dilatation. Avoiding the automatic negative treatment of exploratory viewing.

Pupillary Diameter was calculated independently of Fixation Data. Values were corrected for baseline at 500ms after the appearance of the stimulus. Mean video luminance was collected to be the control of sensitivity test because light affects eye-opening more directly. Pupillary dilation is interpreted as processing load and arousal; attention will be paid to its relationship with entropy and recognition. A novel symbol for enhancing pupils' dilations and recognisability can indicate successful processing. A new symbol has increased pupil dilation and uncertainty to reduce recognisability; therefore, this suggests another approach.

Table 2 lists all of the eye-tracking indicators used for analysis. Table 2: The measurement system has separated fast orientation, sustained gaze, reoccurring inspection, fragmented scan path and physiological burden. This division is needed empirically to examine the phenomenon of attention capture; therefore, produced symbols have multiple causes attracting people's gaze.

Table 2: Eye-tracking methods employed in the empirical research.

Metric	Unit	Computation	Interpretation
Time to first fixation (TTFF)	ms	First fixation on a symbol-bearing AOI after video onset	Speed of attentional entry
Symbolic dwell ratio	proportion	Dwell time on symbolic AOIs divided by valid viewing time	Sustained allocation to constructed symbols
Refixation count	count/trial	Number of repeated fixations on previously inspected symbolic AOIs	Recurrence of symbolic inspection
Return probability	proportion	Probability of returning to a symbolic AOI after leaving it	Stability of symbol-guided viewing
Transition entropy	0-1 normalized	Entropy of AOI-to-AOI transition matrix	Fragmentation of scanpath routing
Pupil dilation	mm	Baseline-corrected pupil diameter during exposure	Processing load and arousal

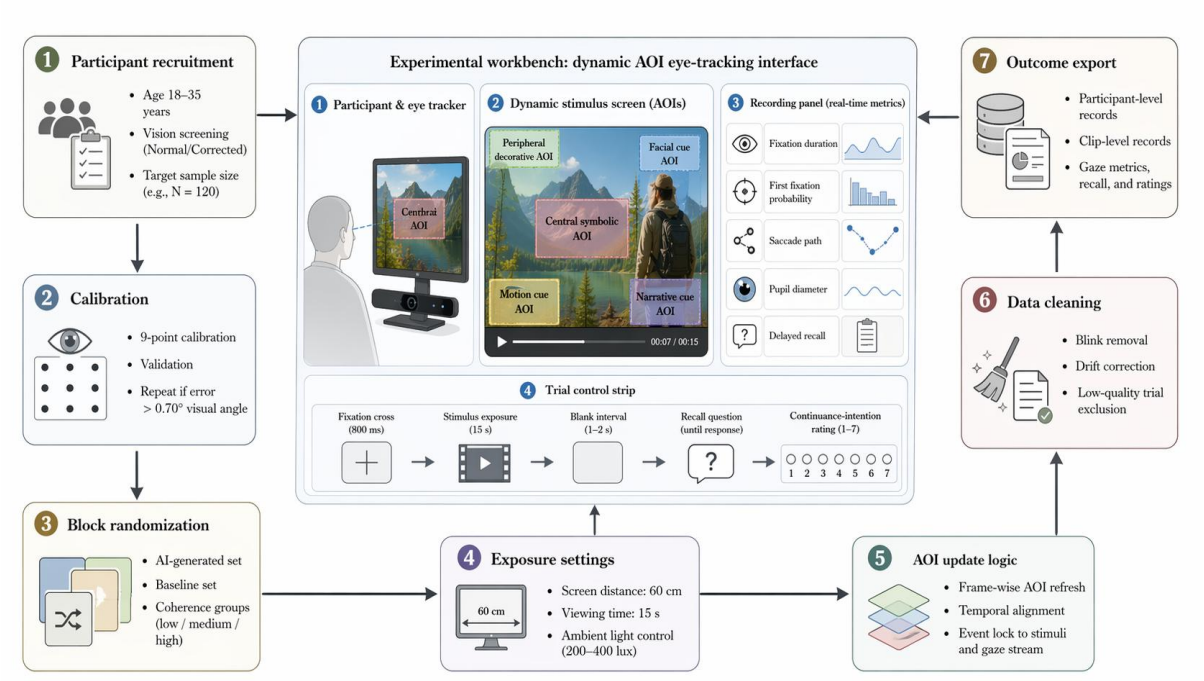

Figure 2: Eye-tracking Exposure Protocol and dynamic AOI Measurement Interface.

There was a brief practice segment at the beginning, which showed three clips not in the stimulation set. Practical trials made the subjects accustomed to watching and responding, without informing them of the production-mode alteration. Following the practice was mainly a block. The participants rested for six rounds at a time. The rest screen was without any visual patterns, and the next trial would begin once re-calibration checks confirmed that the gaze accuracy still met the specified range. This procedure limited drift during longer sessions.

Define the time window first, and then perform descriptions; after that will be the modelling. The early window was from 0s to 3s; The middle Window is from 3S To 8S: And The Late Windows were 8S or later than 15Seconds. The early window represents the initial feed-level direction; The mid-window is for message growth; and The late Window includes both resolution and Call-For-Persons messages. Window analysis was not multiplied with the

primary hypotheses of the tests. Used to explain the time curves in the Results part, thereby establishing an association between initial attention capture and final recognition.

There was no floor effect in the identification test. The four alternatives of the topic items were all within the same content category; therefore, a product clip was matched against other product clips, while a tourism clip was paired with another type of destination. Cue-Identification Test: Ask the subject to select which Area has originated from the main idea. This second item prohibited the count of a trial as "recognised" if a participant merely recalled the general category. Therefore, the integrated recognitional variable needed topic-level and symbol-level knowledge.

Put the continuation- intention item after recognitions to prevent interference with people's memories. Participants rated willingness to keep watching similar clips in a platform feed from 1 to 7. The item is deliberately single to focus on trials' linkages of eye movement and result. Supplementary check: Dichotomised continuity intention at a value of five or more to form the high-continuity logit model. The order of the Production Modes was unchanged; Hybrids had produced a higher frequency of high-intention outcomes.

Fixation-level quality is checked after the automatic pre-processing stage. Trials that reached a particularly long duration in the under-80-ms range were marked and re-calibrated to prevent false alarms. If there was a correction to drift from the validation offset before the trial, then store these corrected positions. When the drift exceeded the range of the validation offset or appeared after a playthrough break, this trial was discarded. To prevent unstable gaze coordinates from distorting the change in transition entropy or generating artificial return-to-small-aoi situations.

At the symbolic AOI level, not on an individual Polygon basis as per the return probability metric definition. If a viewer fixed on the product object, focused on a text label and then returned to look at the product again after the object had moved slightly, it was counted as returning within an area of interest (AOI). This regulation shows the movement rules of short-video symbols. Returning to an important area, no longer at the original Pixel position. The category-level return probability, by capturing recurring symbolic inspections better than the pixel-level recurrence.

Firstly, take the median at each point in time; Secondarily, Calculate an average from several of those Medians across all test subjects. Meds reduce the impact of short-blinking adjacent spikes. Based on the number of valid data points before target-fixation position correction was conducted. Tried trials with a blinks less than 50% valid pupils and removed; however, the trial's fixations were retained if their gaze positions remained valid. This produced a relatively small sample size of the pupil-entropy plot and also kept the production-ordering constant.

2.3 Statistical modelling and validation protocol

Organized empirically in terms of participants and videos. Mainly estimated the impact of production mode and visual-symbol on ACI, TTFF, dwell ratio, transition entropy, recognition, and continuation intentions through these models. Due to each participant viewing multiple short videos, and vice versa, ordinary least squares estimation may overstate dependence in the data. Therefore, mixed-effects models were chosen; the participants and videos served as random effects. Following the reasoning process of mixed-effects estimation for cross-classified, repeated observations, and interpreting trials' gaze data independently at this level removes the assumption that all observations are independent [23, 24].

The form of the static control panel.

$$Y_{pv}=\beta_0+\boldsymbol{\beta}^{\top}\mathbf{x}_v+\gamma^{\top}\mathbf{q}_{pv}+u_p+w_v+\varepsilon_{pv} \tag{4}$$

Y_{pv} represents the dependent variable of participant P and video v. Vector A includes the visual-symbols predictor set provided in Section 2.1. Vector $\mathbf{x}_v$ includes trial-and-participant controls such as the production mode, content category, number of cuts, mean luminance, daily platform usage time and AI-content familiarity. $\mathbf{q}_{pv}$ is the participant fixed effect; u_p is the video specific fixed effects and B is the residual error. For recognitions; The same prediction Structure was estimated using a logistiC.link. Continuation intention was examined using an ordinal model; a linear mixed-effects model had been retained by that point in the analyses anyway.

The production mode, symbol variables were taken to be mutually explanatory. Production mode captured whether a clip was AI-generated, hybrid AI-assisted or human-made. Symbols of variables explained the observable characteristics producing or recording production within a certain range. The division is necessary for the design. If production mode predicts attention without symbol variables, the result remains descriptive. If symbol density, motion discontinuity and lack of semantic coherence predict attention in the experimental data show that this is a manifestation of the basis for paying attention.

Reported model estimation results include 95% confidence bands. The continuous predictors were standardised, and thus their coefficients represented the predicted change in the dependent variable for a one-standard deviation increase in that predictor. The TTFF, in milliseconds, and after taking a log transformation to satisfy its right-skewed nature were analysed. Both directions and significances of the mode effects remained unchanged; thus, untransformed estimates are used here to facilitate comprehension. Recognition models provide Odds Ratios.

Response Surface Analysis examined the joint effect of symbol Density and Motion Discontinuity. The two indicators mentioned above, varying within their standardised ranges, while the others such as semantic co-occurrence were kept at the average value. To determine whether attention increased consistently in relation to the size of symbol intensity or reached a high point before falling off. Therefore, the three-dimensional visualisation presented in this part is model-generated projections instead of raw trial data. Including it is due to its coupling relationship; otherwise, expressing through a single coefficient would not cover all aspects adequately.

All the verification included four stages. Calculate an equal-weighted ACI value to determine whether results are related to the selected component weight coefficients. Secondly, after removing the first two trials of each person's performance data to eliminate adaptational biases. Thirdly, A leave-one-category-out method was used to determine if the production-mode order differed among each category of content. Ablation experiments sequentially excluded the presence of dynamic AOI encoding, motion discontinuity and semantic consistency. Ablation Design is directly related to the theoretical assertion; That is, with attention capture decreasingly explained as models lose the explanatory factors for generating visual-symbol Construction.

Multiple comparisons were restricted to the set of outcomes. Primary indicators are ACI Scores. Secondarily measured were TTFF (time-to-first-visit), symbolic dwell ratio, return probability, transition entropy, pupil dilation, recognition, and continuance intention. Based on the convergence of patterns for these indicators. The association between AI-generated clips and greater content creation duration, time spent in-store D), more return visits were also observed. No single large-coefficient value was accepted as sufficient evidence alone. Adjustment of false discovery rates for the second-order coefficient family, [25].

Analysis was performed using mixed-effects routines typically employed for repeated measures Designs. Residual Plots, Fitted-Value Plots, and Random-effects Distributions were examined for the Continuous Models. Check the separation and calibration of recognition models. Missing responses were not imputed because all retained gaze trials had complete post-

exposure responses. Figures 3 presents the model structure for connecting visual-symbols predictor models, random effect models, outcomes and their robustness tests.

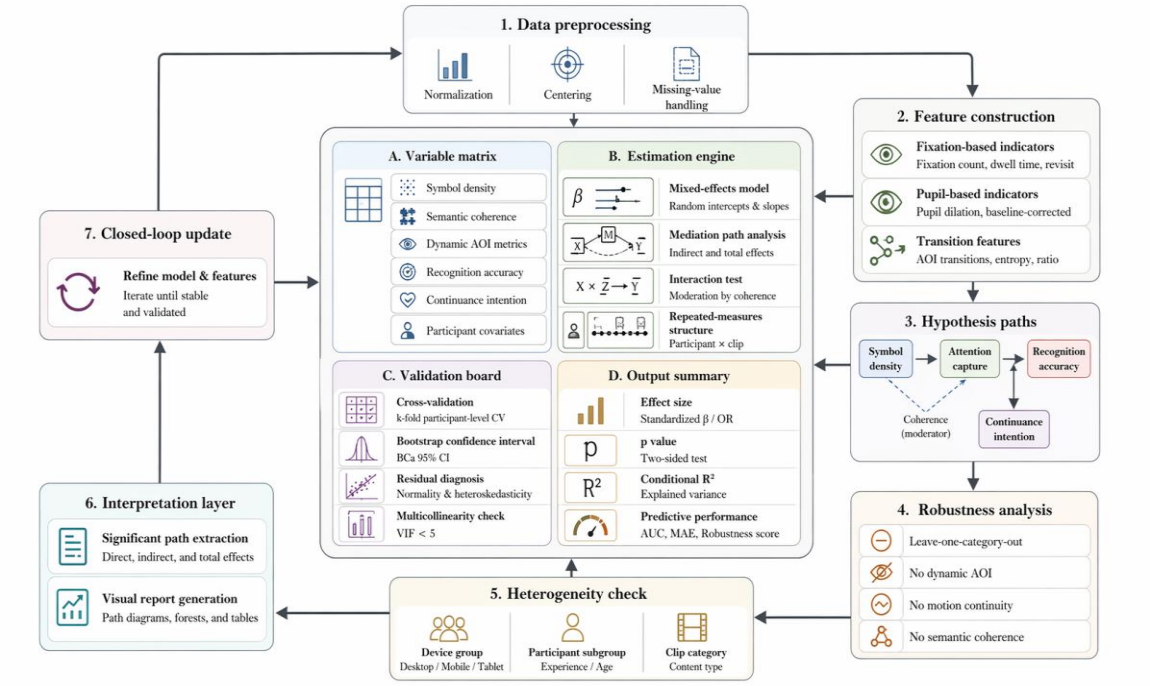

Figure 3: Statistical Modelling and Validation Mechanism.

The main ACI model is the production type of ACI with visual-symbol predictor and control. The second specification also introduced interaction effects between symbol density and semantic coherency; Between Motion Discontinuity and Semantic Coherence. Use of the Interaction Model in Response-Surface Analysis and Outcome Interpretation. The primary coefficient figure provides an easier-to-understand simplification of the fixed-effects model, and the response surface displays the coupled relationship it does not reveal through one coefficient.

The ablated model is evaluated in the retained trial set simultaneously. No-Dynamic-aoi model replaced AOI-specific dwell, refixation and return by summarising all-frame fixation information. Remove the transition variable mV in the no-motion model. The no coherence model eliminated the term for C_v. The equal-Weighting of ACI re-calculated the dependent variable by giving each factor a weight of one when calculating it. The leave-one-category-out model was performed 4 times by omitting one content category at a time.

Interpretation effect followed the order in empirical presentation of this section. The time curve shows the moment attention appears in symbols-enclosed AOIs. The descriptive table then reports trial-level means. The transition network shows where gaze travels after entry. Estimating the Response Surface of how symbol density and motion discontinuity combine. Tests for the coefficients of feature coding in the test results. Ablation figure to determine if the mechanism can continue after removing the central parts. The following Sequence retains the statistics of the observed gaze behaviour.

All reported p-values are one-tailed tests. Confidence intervals were selected to interpret results due to having many participants and videos; thus, very small differences would be observable. Therefore, the following research focus is: coefficient magnitude; interval orientation; Convergence of measures. For example, a production-mode difference is treated as empirically meaningful only when it appears in TTFF, dwell, return probability and the

composite ACI. Recognition and continuation intentions refer to the down-stream consequences rather than supplementary evidence for initial identification.

In the results, a coefficient plot showing standardised fixed effects for ACI as the main outcome is presented here. Recognition and persistence evaluations will be presented together with the brief report on the simple models' summaries in this study's results, respectively. This decision maintains recognisability of the figure and allows for interpretation by subsequent results. Therefore, figures are used to depict the Structure of gazing, while tables condense model explanations. This division can prevent the results section from being a series of isolated statistical data.

Examined model residuals according to production mode and content category separately. No category had an entire system residuals large enough to support a separate family of models. Entertainment Micro-narrative had the highest residual difference, which was caused by various forms of humour, Sound Timing and Face movements. Residuals of these items were not removed but included in the analysis errors. Keeping them ensures that the stimulus set remains representative of various types of short video-generated symbols that tend to occur more frequently.

3 Results and Discussion

3.1 Temporal attention capture and AOI allocation

The first result part checks if the generation of visual symbols altered eye-wandering times and spaces. Starts with temporal fixation density as attention acquisition happens before the viewers have completed watching one segment of the video. Then, it links the temporal pattern to some descriptive trial-level metrics and AOI transition structures. To determine if there is a greater attention-grabbing effect of AI-created videos at an initial glance and if such attention-pulling power persists over time.

Temporal fixation density of symbols-bearing AOIs is given as follows: Figure 4 shows that AI-generated clips produced the highest fixation probability during the first 3 s of exposure. The mean symbolic fixation probability of the earlier window was 0.46 for AI-generated clips, 0.39 for hybrid AI-assisted clips and 0.30 for human-made clips. The gap narrowed after 8 s, with values of 0.37, 0.35 and 0.30. Therefore, the early Window had the greatest production variation. Generated symbols served as quick entrance signs; for instance, if a face or object were generated nearby in the middle part of the image.

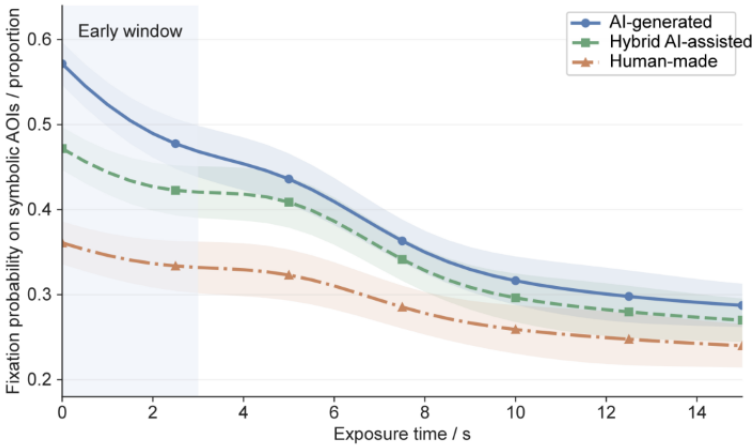

Figure 4: Temporal fixed-density distribution of symbols-bearing AOIs among production types.

Figure 4 also reveals that the hybrid curve is less steep than the AI-Generated one. Hybrid clips exhibited a delayed early peak and had better symbolised stability over time in this range. Because it is an essential type of Pattern, there will be a significant increase initially and then decline gradually to indicate greater longevity. Human-made clips had entered symbols in the least time. The gazing patterns exhibited lower variations compared to others; In part, due to the focal objects being immersed in well-known recorded scenes which generated different synthetic salience.

Descriptive data are shown in Table 3 below. As shown in Table 3, The mean TTFF decreased from 462ms under a human-manufactured environment to 389ms under both a hybrid system and an AI-produced Environment; (see Figures 7-10). The symbolic dwelling-to-population ratio rose from 0.31 to 0.38 and 0.42 under the same circumstances. Refixation count also increased from 1.74 in human-made clips to 2.21 in hybrid clips and 2.68 in AI-generated clips. These three indicators confirm the same conclusion: Generated symbols have entered faster; Held gaze has lasted longer; Prompted repeated check.

Table 3 also shows the limit of this capture advantage. The transition entropy of the AI-generated condition was 0.67; it was 0.58 for the hybrid generation and 0.51 for human-made. A higher value of entropy shows that the viewers' path distribution is relatively disperse. The recognition accuracy did not match the order of acquisition; AI-generated clips were recognised by 72.4%, human-made ones by 74.1% and hybrid ones by 76.8%. The conditions that received the most attention, however, were least likely to be remembered. There is a difference in this discrepancy; therefore, we must analyse gaze captures independently of communicative outcomes.

Table 3: Eye-tracking data and results of different production modes.

Outcome	AI-generated	Hybrid AI-assisted	Human-made	Empirical reading
TTFF to symbol-bearing AOI	334 ms	389 ms	462 ms	AI-generated clips produced the fastest orienting
Symbolic dwell ratio	0.42	0.38	0.31	Generated cues received the largest sustained allocation
Refixation count	2.68	2.21	1.74	AI-generated clips prompted more recurrent inspection
Return probability	0.63	0.55	0.46	Generated symbols were revisited more often
Transition entropy	0.67	0.58	0.51	AI-generated clips produced more fragmented scanpaths
Recognition accuracy	72.4%	76.8%	74.1%	Hybrid clips supported the most stable memory
Continuance intention	4.81 / 7	5.13 / 7	4.44 / 7	Hybrid clips produced the strongest continuation tendency

Dynamic AOI transition Structure is illustrated in Figure 5. Figures 5a, 5b and 5c show the networks of each production mode, respectively. On the AI-created panel, there are larger face, objects and text nodes with thicker transition lines. Object-Text, the Transition probability is 0.22; Face-Object Transitions has a probability of 0.27. These values show that generated clips frequently focused on moving their gaze from a social indication to a changed figure, followed by looking at the name tag attached to it. The same pathway was weaker in the hybrid condition and weakest in the human-made condition.

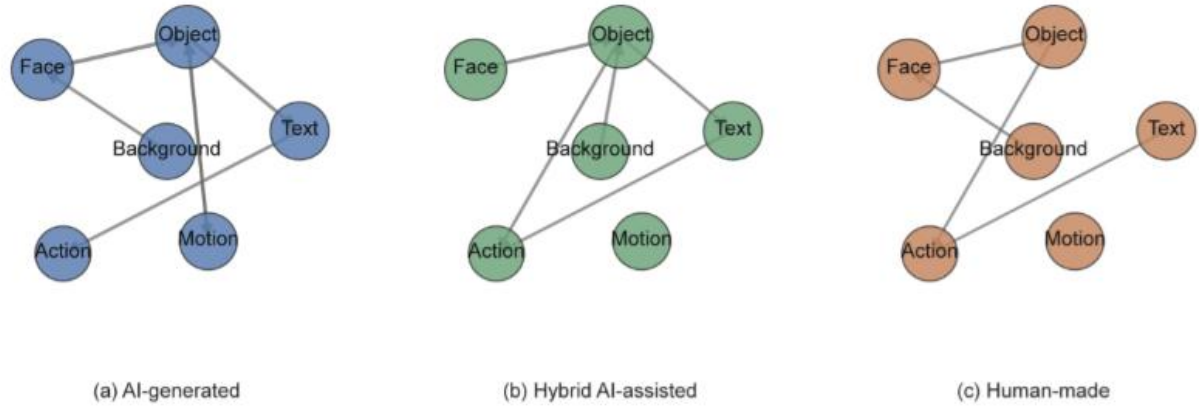

Figure 5: Dynamic AOI Transition Structure Across Generated Visual Symbols.

As shown in Figure 5, Motion-boundary nodes are more core in the AI-Generated Condition. The above conclusion agrees to the coded motion-discontinuity variable. Generated clips usually included rapid changes of objects, background variations and synthetic effects that made the scene clear. Viewers inspected these areas, and then Faces or Objects were returned. Motion-boundary transitions occurred less frequently in artificial clips due to the continuity of camera-locked actions for motion. Therefore, the network presents differences in the direction of gaze travel within its boundaries.

The representative scanpaths are as follows, respectively: Figure 6 shows an example of one AI-generated fixed-order duration comparison to the other human-fixed cases. In the AI-generated cases, first move from faces to objects; secondly, shift to texts; finally, return to objects. Repeatedly returning to the object suggests a repeated check. There were fewer loops back from fixation formation in the artificial environment to objects or faces, as shown below: The case studies improve the performance of the model by detailing these results further. The generated symbols frequently switched within the symbol-bearing areas.

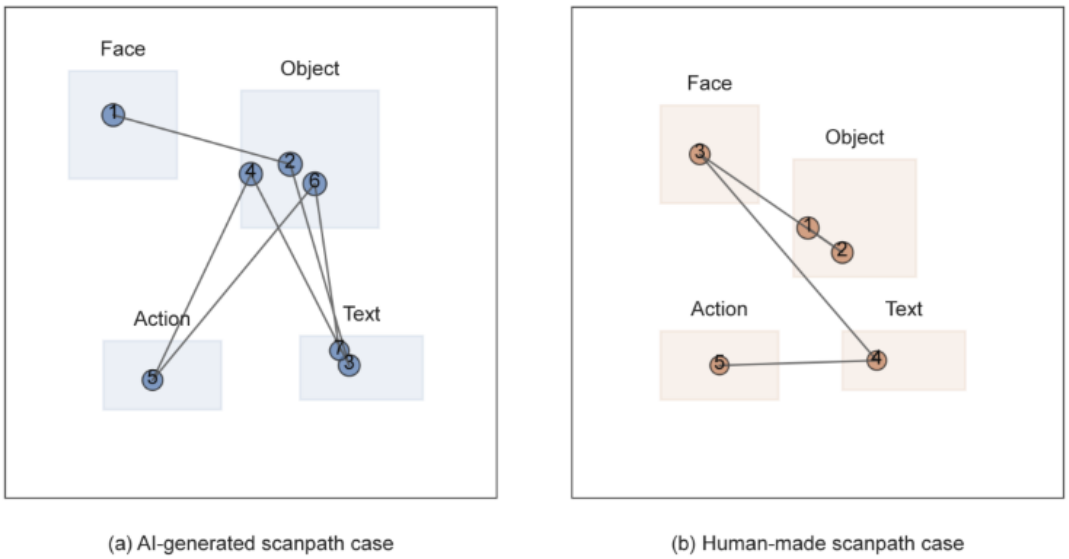

Figure 6: Representative Scanpaths Comparison of AI-Generated and Human-Made Short Videos.

The joint distribution of pupil dilatation and transition entropy is as follows, see Figure 7. As shown in Figure 7, the AI-generated conditions are located in the upper-right quadrant more

frequently compared to others; the average pupil dilation was 0.118 mm, and the mean entropy was 0.67. Hybrid Clips are grouped more closely to the centre; Human-constructed Clips have higher Entropy and Dilation. Therefore, we may conclude that the full-generation clips improved visual appeal and learning demand separately. It can be explained that recognition did not rise with increased length of stay directly.

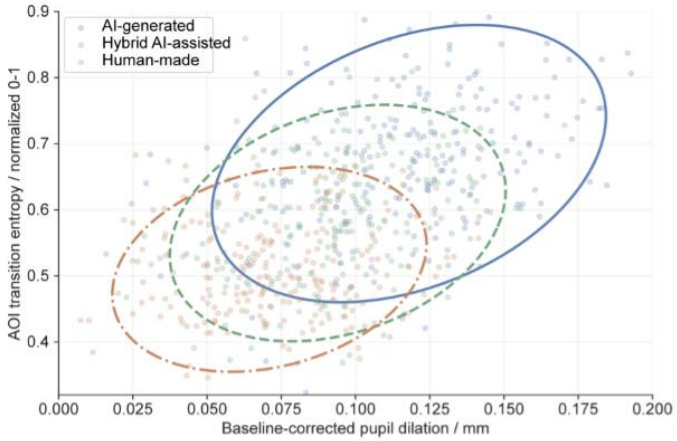

Figure 7: Joint distribution of pupil dilation and scanpath entropy under different production modes.

Therefore, the first result part will present a concrete empirical phenomenon. AI-generated clips detected gaze more quickly and strongly than the human-made ones. The same clips also generated more fragmented Scanpaths and higher Pupil-latched Load. Hybrid clips had a middle role in gaze tracking but obtained the best memory and continuation results. Then, what kinds of characteristics in the produced visual symbols account for the differences?

AOI assignment also had similar time-varying characteristics at the regional level. In AI-generated clips, face or character regions accounted for 23.6% of valid dwell time and object-transformation regions accounted for 18.4%. The corresponding percentages in the hybrid clip are: 20.2%, 16.9%. Human-made clips have been 17.5% and 13.2% respectively. Text labels accounted for 11.8% of dwell in AI-generated clips, 10.7% in hybrid clips and 8.9% in human-made clips. Therefore, the face and object areas had captured a relatively large extent of variation, while text showed some changes as well.

The first-look distribution of TTFF also shows that there were no outstanding samples in the artificial-intelligence-created leads. Among the 68.2% of valid cases, after an initial fixate appeared at least one symbol-bearing AOI within four seconds. The corresponding shares were 54.6% for hybrid clips and 39.7% for human-made clips. At the 600ms mark, it increased to over 82.4%, 74.5% and 63.1% respectively. These distributions indicate that the generated symbols have altered the lower section of the orienting-time distribution; that is, feed-level attention judgments may take place.

Later Window can also achieve it. The AI-generated clips had retained a greater symbolical residence, but the delayed-window entropy was still larger. Therefore, viewers were not limited to one of the created objects immediately following the first frame acquisition. Many of these experiments continued alternating between the faces, objects and texts. Hybrid clips had fewer late-window switches and more objects-oriented rebounds. This phenomenon is why the combination of low-shot and high-lost detection rates showed better performance. There were sufficient attentions to enter the clip, and there was good structural support for maintaining the primary subject.

Content Category has not eliminated the production-oriented style. Product clip-based AI-

generated video reduced TTFF by 151ms compared with a manual-produced version. The shortening in the tourism clip is 128ms. Public Information Clippings had reduced lag by 116ms. In entertainment micro-narratives, the reduction reached 174 ms, mainly because generated faces and transformations appeared near the first frame. The specific value of this is consistent with the overall result; Entertainment clips carry the highest synthetic-symbolic effect.

The ACI distribution agreed with the ordering results of components in detail. The mean standardised ACI values were: +0.29 (for AI-generated clips), +0.14 (for hybrid clips) and -0.18 (for human-made clips). The interquartile range was widest in the AI-generated condition, indicating that fully generated clips varied more in capture strength. The dissemination supplies some reference interpretations. The production mode brought about a generalised advantage; however, at the clip level, some generated clips had low scores and others mixed approaches were extremely strong. Therefore, the following parts will only discuss coded symbols and not the production modes individually.

3.2 Coupled effect of symbol density, motion discontinuity and semantic coherence

The second result section transitions from a production state to the coded characteristics of visual symbols. Previous research shows that the AI-generated clip can attract attention immediately; However, it is not clear how this occurs in different scenarios among these clips. Therefore, this paper will study the combination of symbol density, motion breakage and semantic cohesion. Use the Response Surface, Outcome Curves and Model Estimates to determine the effective Range for generating symbols in this study.

The Response Surface for the attention-capture index is presented as follows: Figure 8 plots predicted ACI across standardized symbol density and motion discontinuity. The peak on the surface is at around 0.72 symbol density and about 0.64 motion discontinuity. At this peak, predicted ACI reaches 0.46 standard units. The surface no longer rises infinitely. Density and discontinuity are greater than about 1.25 standard units; then, ACI begins to decrease, and the lower limit curve appears again. This form shows that there is an extent of the influence, not necessarily in proportion to quantity.

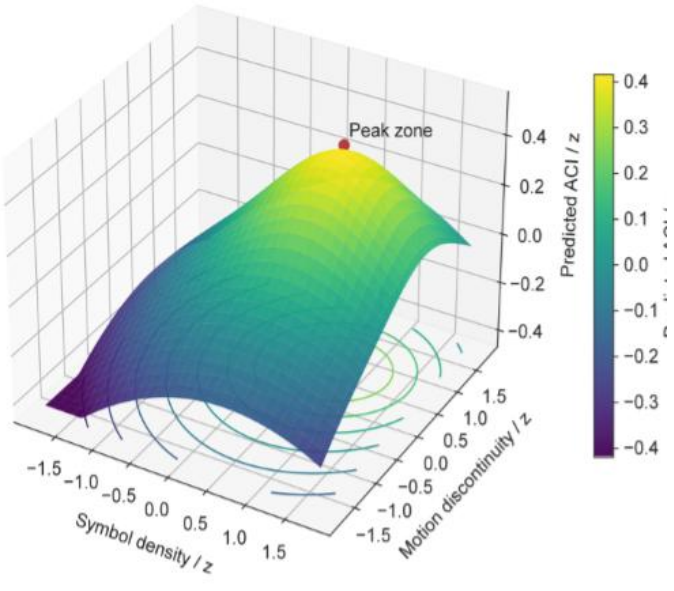

Figure 8: The three-dimensional response surface of attention acquisition under symbol density and motion discontinuity.

Figure 8 presents the role of generated novelty. At a moderate scale of density and with some movement discontinuity, it suggests that there may be something important happening soon. Excessively high density and significant disconnection form a competing group of objects. Among them, high-overload clips were often accompanied by faces, altered objects, two sets of text layers and backgrounds in less than four seconds after the start of screening. These clips still generated gaze, but the gaze route became less stable. Therefore, the response Surface should include a function that generates Symbols with higher visibility and loses its effect as this quantity exceeds a certain threshold.

Downstream relations of semantic coherency, density, and results are presented as shown in Figure 9. Figure 9(a) shows that recognition accuracy increases with symbol density when coherence is high. At high coherency, the recognitions obtained were between 70.2% and 78.6% in different scenarios of density. At low coherence, an equal-density increase resulted in a small improvement followed by a fall at the highest density level; As shown in Figure 9b, similarly to continuation motivation. The high coherence clips increased from 4.55 to 5.35 on a seven-point scale; Low-complexity clips were still around 4.2-4.45. The above curves indicate that density functions are helpful only for symbols indicating the same information.

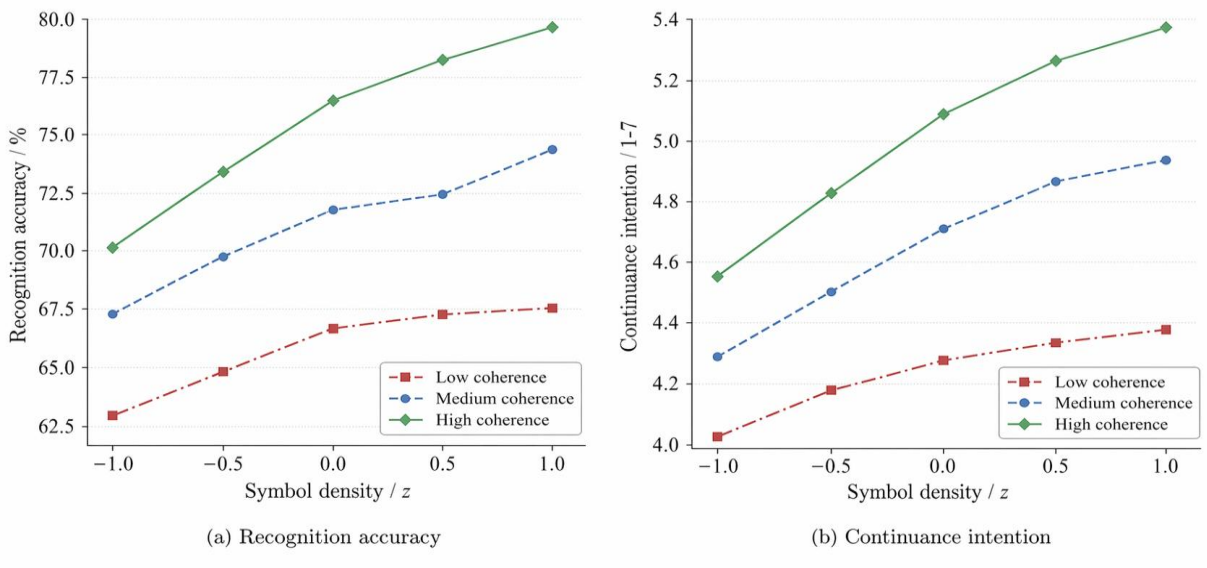

Figure 9: Relationship among semantic coherence, symbol density and downstream outcomes.

Figure 10 presents the model coefficient statistics to evaluate whether these relationships exist statistically; Symbol density had the largest positive coefficient for ACI (beta = 0.31, 95% CI [0.22, 0.40]). Motion discontinuity was also positively correlated with subjective perception of realism (β=0.24, 95%CI[0.15,0.33]) and the level of confidence in recognising people before the change in production mode or control groups (β=0.19, 95%CI[0.10,0.28]). Visual clutter exhibited a significant negative effect at the 95% confidence level (β=-0.27; 95%CI[-0.36,-0.18]). AI-label disclosure was also negatively correlated with performance (-0.09; 95% Confidence Interval: -0.17 to -0.01). The parameter distribution conforms to this result; capturing attentions needs a large amount of density-based mobile objects, and reducing clutters or positioning errors affects them negatively.

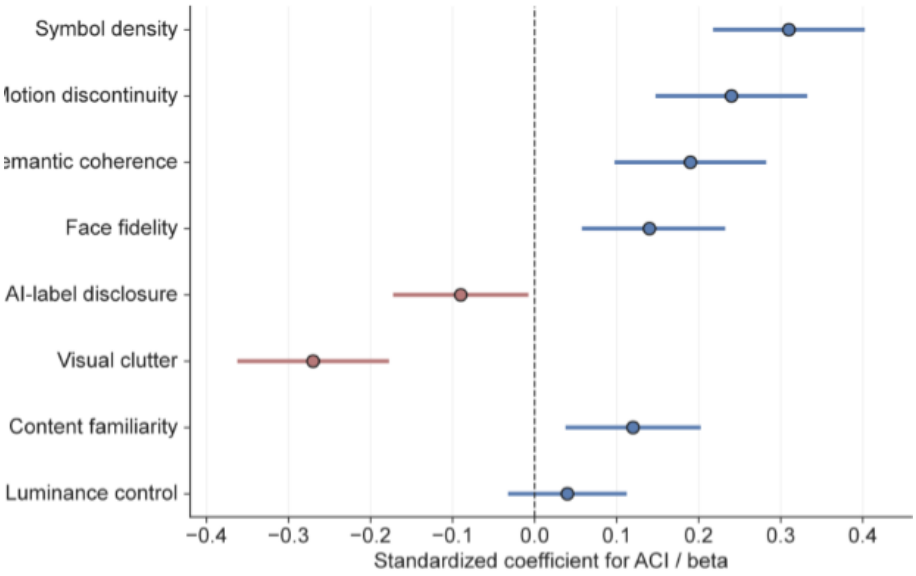

Figure 10: Mixed-effects coefficients for visual-symbol predictors of attention capture.

Face fidelity obtained a significant positive effect ($P < 0.01$), within the confidence interval: $P = 0.06$~0.23). This was more pronounced under the influence of these generated characters, to use an analogy. High-fidelity faces accelerated first fixation and supported object-to-face transitions. Lower-fidelity faces were still eye-catching, but more subjects stared at the face area and looked away from the objects or texts in it. In terms of ScanPath, some people looked at their own eyes and mouths again after transforming; indicating that they had noticed instability. Therefore, it has a positive correlation with social prominence and stability of self-image respectively.

The artificial label disclosure of ACI decreased marginally, and the extent varied. Labels set outside the first orienting window were unable to affect symbolised dwelling significantly. Labels close to faces or products in the first second were competing symbols and produced more text changes. In trials with low semantic coherence, disclosure sometimes improved recognition because it clarified the production context. Therefore, the net coefficient is still negative in this case, and removing disclosure cannot be inferred. Supports placing according to the rule that disclosures are observable while spaced apart from their first symbolically presented targets.

Table 4 summarizes the main model outputs. Table 4 shows that AI-generated clips increased ACI by 0.28 standard units relative to human-made clips after controls were included. Hybrid clips increased ACI by 0.16 standards units. The semantic coherence model had a recognised rate higher than 1.42; transition entropy reduced the correct recognitions by about 0.74 times. These figures connect the gaze mechanism with its results. Attention capture is effective for co-ordinated symbols; It will be reduced in its effectiveness as the Scanpath becomes disorganised.

Table 4: Main mixed-effects model outputs for attention capture and outcomes.

Model output	Estimate	95% CI	Model-level statistic	Empirical interpretation
AI-generated vs. human-made on ACI	0.28	[0.17, 0.39]	p < .001	Generated production mode increases capture after controls
Hybrid vs. human-made on ACI	0.16	[0.06, 0.26]	p = .003	Hybrid construction gives a moderate capture gain
Semantic coherence on recognition	OR = 1.42	[1.19, 1.69]	p < .001	Coherent symbols improve post-exposure recognition
Transition entropy on recognition	OR = 0.74	[0.63, 0.88]	p = .002	Fragmented scanpaths reduce memory accuracy
Full ACI model	R2m = .34	R2c = .61	MAE = 0.214	Participant and video random effects explain substantial variance

Therefore, this paper's second results identify the operating conditions of the generated visual symbols. Density and Motion Discontinuity Produce Visibility. Semantical Consistency Stabilises Interpretations of the Viewer. visual clutter, misplaced labels, etc., are unstable as intended. Hybridisation was relatively effective because the new visual information contained within an intelligently composed image that still connected objects, texts and motions clearly. The following test if it is still valid after removing the most significant measurement components.

There was a substantial interaction among density and coherence in the Recognition model. One-standard-deviation increase in the symbols density led to a better recognition at higher coherences, which is estimated as OR = 1.33. At low coherence, a densification increase showed an odds ratio of 0.96; it did not vary significantly from 1.00. Some completely generated clips caught people's attention but did not help them remember better than others initially. The viewer entered multiple symbol areas; however, these Areas could not coherently provide information for the corresponding tasks.

Object-transforming AOIs provided a more convincing illustration of it. Transformations that preserved the identity of objects were more likely to receive high Return Probabilities and improve cue identification (such as packages being opened or landmarks appearing after traveling). When the transformation changed the object's category or visual boundary without message support, return probability still increased, but recognition fell. Since the same gaze behaviour has different outcomes under various conditions of semantic coherency. Through a positive coherency factor and a negative entropy- recognition link, the model demonstrates.

Text labelling is narrow in function. A short label near the focus of attention enhanced the topic recognition rate by 4.2 per cent compared with clips lacking labels in pre-production settings. A long or repeated label increased text dwell but reduced object dwell by 0.06 proportion units. The most effective labels were those that appeared immediately after the first fixation of the target. At this time, it could still be interpreted separately without competing with the initial visual reference. The outcome supports the placement interpretation provided by the coefficients.

These are also to predict the mixed effect. Hybrid clips had a lower mean symbol density compared to fully AI-generated ones; however, its coherency score was 0.18 points higher on the standardised system. There were 0.21 fewer degrees of chaos. The difference diminished the initial prominence, while enhancing scanpath stability centred around objects. In the recognitions model, a mixed clip with above-medium coherence has been estimated to have recognition probabilities of 0.79; in comparison, the full-generated one was about 0.72 when

using low-density and median-coherence conditions. Therefore, it is an effect of the composition rather than pure preference for artificial materials.

Moreover, the Surface and Coefficient results of Visual Clutter are not directly related to density; There was a medium positive relationship between density and clutter in the experimental stimuli; they were not modelled similarly. Density increased ACI when symbols were meaningful. Clutter decreased ACI by providing no relevant targets for gaze selection. Practically speaking, add a label, transformation or character to make it more accessible for people who have not seen before. Adding decorative particles, repeated captions or unstable backgrounds increases competition without improving recognition.

3.3 Mechanism validation, robustness and empirical implications

The third results Section examines whether the proposed mechanism is contingent upon any single indicator or model selection. Validation includes the content of ablation, robustness, error source and deployment impact. The following will explain that this paper takes attention acquisition to be an integrated activity. Only if the results were solely based on dwell-time data or produced mode would it be less effective. Empirical Question: Whether there is a recurrence of this kind in dynamic AOI segmentation, loss of motion discontinuity and absence of semantics?

Ablation and robustness results are shown in Figure 11. Figure 11 reports conditional R-squared, recognition AUC and mean absolute error for the full model and five variants. The full model reached conditional R-squared of 0.61, recognition AUC of 0.77 and ACI MAE of 0.214. Removing dynamic AOI coding reduced conditional R-squared to 0.46 and increased MAE to 0.286. The biggest reduction among all test cases. The results show that clip-level labels cannot be responsible for gaze movement between all-generated- symbols within a specific sequence.

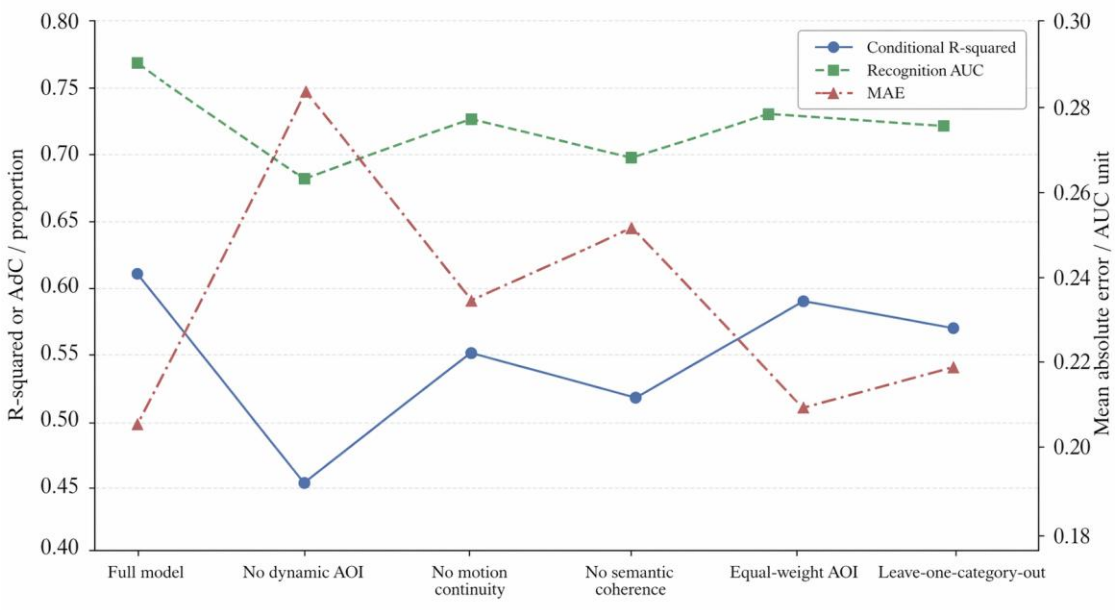

Figure 11: Abolishment and the performance of the Attention-Capture model.

See Figure 11; Removing motion discontinuity decreased the conditional R^2 from 0.61 to 0.55, while removing semantic coherence further lowered this figure to 0.52. The recognition AUC decreased markedly after removing the sense-of-coherence; its value dropped from 0.77 to 0.70. Also have such a mechanism behind it. Motion discontinuity refers to the capture of visibility changes. Semantic coherence explains whether capture supports memory. The ACI

test retained an equal-weighted condition R-squared of 0.59 and AUC of 0.75; it was found that the basic shape did not relate to the fundamental coefficient ratios.

The leave-one-category-out test produced conditional R-squared of 0.57 and AUC of 0.74. Production mode order was the same for all categories of omission. The greatest reduction in length occurred without including entertainment micro-narratives; these had the most obvious facial and movement indicators. Although the effect persisted across product, tourism and public information videos. The above results suggest that the above empirical mechanism may have multiple types of genres, and these sizes differ in different category contents.

Robustness tests have also examined alternative participants. Adding daily platform use and AI-content familiarity did not remove the production-mode effect. The heavy-short-video group had a small increase in TTFF across all scenarios compared to the general light-user group; however, this difference was relatively smaller than that caused by humans and machines (AI vs. human). Participants with high AI-content familiarity reported lower continuance intention for low-coherence generated clips, yet their gaze still entered generated symbols quickly. Familiarity therefore moderated judgment more than initial orienting.

Error analysis identified three recurring sources of misfit. The first was audio time. Some humorous clips generated high continuation intentions due to the addition of sounds; visual dwell was lower. Second, a dense of words. Some of these short videos were able to identify topics accurately; however, objects could not be recognised reliably. Aesthetics of the Third Kind. Several AI-generated short clips mimicked familiar platforms' structures, lacking originality for experts. These cases retain the entire model and illustrate that visual-symbol construction is related to genre knowledge and audio background information.

Directly from the empirical implications for short-video production. Generated faces or products transformations can be the initial attentional anchors, but they should not receive several similar highly prominent generated cues simultaneously in the same second. Text labels should name or clarify the visible object. Objects' changes must not alter their categories; if so, it will be transformation itself. Motion discontinuity needs to indicate an important state change, followed by bringing back the focus on this subject. These rules originate from eye-gaze behaviour, not because of a general Design Preference.

Platform assessment; use of watch-time alone may not be a sufficient indicator for evaluating the quality of AI-generated short films. A clip retains gaze by means of uncertainty, clutter or artefact detection. The other clip produces a slight early capture, which recognises individuals but is willing to participate. Combining the early fixed probability, symbol-stay-time, return-probability, entropy, pupil-dilation and post-exposure-recognition method can give us a more specific judgment result. Distinguish between the attention organised by messages and those of unresolvable visual ambiguity.

Also, this research offers insights into AI-label disclosure. The evidence does not support hiding labels to increase gaze. A lower value of the disclosure effect indicates that label positioning is beneficial for reducing mis-recognition; According to analysis, dispositions deteriorate with a decrease in semantic similarity. In this particular case, position the tags away from the first focus area to be visible; Do not hide Faces or product alterations. To maintain visibility and avoid excessive competition in the initial orientation phase.

Therefore, the third Result Section will validate this mechanism for all models. dynamic AOI: Where is the subject's focus? Motion disconnection: What generates this symbol catches attention rapidly; Semantic coherence: Whether it becomes memorisation after a catch. The entire model outperforms the ablation variants; these are, in fact, retained by all of them. The outcome pattern supports the main thesis of this study; that is, after obtaining attentional objects with a significant effect on subsequent behaviours, their organisation can still be needed.

Another more robustness check used only the first eight seconds of exposure. The

production mode order of ACI did not change; the top three were as follows: AI-generated clips, Hybrid clips, Human-made clips. In the shortening of exposure models, the AIs-generated vs human-made coefficient was 0.25; in the complete version, it was 0.28. As the platform has already been decided upon in 15 seconds after watching, it is very important. Before the late Call-to-Action Window, the capturing mechanism can be observed.

Outcomes of the models for topics' identification and cues' identification were each calculated individually. The effects of semantic coherence on cue and topic recognition were opposite. Odds ratios of 1.51 and 1.28 for cue recognition and topic recognition, respectively. The transition entropy has a reversed situation and is also more negatively affecting cue recognition. It can be concluded that the gaze mechanism has been associated with visuoesymbolic memory rather than general content recognition. Viewers may be able to recall the general topic, but they cannot determine which particular generated cue conveyed it.

Efficiency Analysis based on the quantity of symbols required for a correct match. Hybrid clips needed on average 5.7 seconds to obtain a correct judgment by symbolically staying; it took more time (6.4s) than both artificial intelligence generated clips(6.1S) and human-created clips(6.S). It can be considered the dwell efficiency here. Each second of symbolic resides in the hybrid condition and carries more stable message information. Fully generated clips received the most amount of gaze time, however a proportion thereof had to resolve instability, confirm transformation results, or switch symbolic domains.

Figure 6's case analysis also verifies this level of operation efficiency. The AI-created cases have seven fixations scattered in the areas of face, objects, text and actions; there were two return-to-objects occurrences. Human-made cases have five fixities and take a shorter path. A third type of hybrid case, which was examined but omitted from the figures for brevity, consisted of six fixations and an obvious object-totext-to-object sequence. This route produced correct recognition and high continuance intention. The case evidence corresponds to the model; therefore, the use of capture remains close to the transmitted information.

Observation of its deployment implications. The platform or content group can conduct Eye-Tracking tests on the new short videos before posting them online, etc. Clips that have a strong initial symbolisation fixations and low entropy are potential targets. Clips with high early fixation, high pupil dilation and low recognition would require revision. The revised object will not be general Attractiveness. The positions at which times have been selected as well as their connection. Therefore, this provides some real-world applications of the empirical system outside of laboratories.

Hence, it will be effective at all levels. At the metric level, it includes TTFF, dwell Time, Return rate and ACI Scores. After removing participants' random effects, videos' random effects, control variables, and ablation studies at this stage. At the final stage of explanation, there are reasons that a highly effective initial selection was not optimal. The empirical basis of treating visual-symboic construction as a method to capture attention through the three-tier convergence of generated short videos.

4 Conclusion

Examine the generative AI-generated short videos for viewing purposes and investigate what kind of attention they elicit from users visually symbolically. Using 72 balanced clips, 96 participants and 1,641 valid participant-video trials, the experiment combined dynamic AOI annotation, eye-tracking measures, recognition responses and mixed-effects modelling. According to the data, fully AI-generated clips have improved eye-gaze entry speed and extended the Symbolic Dwell Time (SDT). Also had a larger transition entropy and pupil dilation. Hybrid-aided clips had a relatively small deviation of capture, thus showing high

recognitions and continued-interest levels.

Firstly, organised the produced short videos at a level of apparent symbol. Regions such as faces or characters, object transformations, texts, motion boundaries, call-to-action areas and backgrounds were classified as dynamic AOIs. This Design enabled the observation of how many times users entered and revisited generated prompts, instead of merely using production tags for verification. Empirical Data Show That Dynamic AOI Coding Is Required to Explain Gaze Distribution Inside AI-Generated Clips.

Secondly, based on model estimation and figure presentations of attention-grabbing power. Symbol Density, Motion Discontinuity and Semantic Coherence Jointly Formed ACI. The responses surface reached a peak of moderate-high density and moderate-discontinuity; Semantic coherency decided whether translation resulted in recognition. Abnormalisation of Dynamic Regions, Motion Breaks and Loss of Coherence were all identified as factors reducing Explainability.

Third, the study is limited to laboratory viewing of 15 s vertical clips. It did not conduct tests on natural scroll behaviour, the recommendation context and comments displayed. Furthermore, the Data do not compare particular generative Models. In future studies, combining mobile-eye-tracking technologies with platform log data from multiple rounds of exposure tests can help identify whether the same symbolic mechanism predicts scroll-share-trust behaviors naturally in feeds.

The above results should not be regarded as an absolute ranking of the Production Mode Types. The optimal clips are those that can attract attention, maintain the Identity of objects and help recognitions within the field of view. The next empirical Step is to conduct tests under mobile Scrolling Environments, i.e., add recommendation labels, comments, and User-initiated Swiping into the attention Process.

Within the specified range, offer a directly observable procedure: Construct balanced stimuli, label dynamic symbols, record gaze patterns at each trial stage, and assess captures along with recognitions. This protocol makes the future revision of generative-visual tool empirical instead.

Funding

This work was sponsored by the Guidance Project under the 2023 Science and Technology Research Program of the Hubei Provincial Department of Education (Grant No. B2023415)

References

[1] Montag, C., Yang, H., & Elhai, J. D. (2021). On the psychology of TikTok use: A first glimpse from empirical findings. Frontiers in Public Health, 9, 641673.

[2] Li, H., Li, J., Hao, X., et al. (2025). Behavioral and eye-tracking investigation of event segmentation following short video watching. npj Science of Learning, 10, 86.

[3] Omar, B., & Dequan, W. (2020). Watch, share or create: The influence of personality traits and user motivation on TikTok mobile video usage. International Journal of Interactive Mobile Technologies, 14(4), 121–137.

[4] Dwivedi, Y. K., Kshetri, N., Hughes, L., et al. (2023). "So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. International Journal of

Information Management, 71, 102642.

[5] Epstein, Z., Hertzmann, A., Herman, L., et al. (2023). Art and the science of generative AI. Science, 380(6650), 1110–1111.

[6] van Berlo, Z. M. C., Campbell, C., & Voorveld, H. A. M. (2024). The MADE framework: Best practices for creating effective experimental stimuli using generative AI. Journal of Advertising, 53(5), 732–753.

[7] Belanche, D., Ibáñez-Sánchez, S., Jordán, P., et al. (2025). Customer reactions to generative AI vs. real images in high-involvement and hedonic services. International Journal of Information Management, 85, 102954.

[8] Farooq, A., Khan, A., Awan, T. M., et al. (2025). Deciphering authenticity in the age of AI: How AI-generated disinformation images and AI detection tools influence judgements of authenticity. AI & Society.

[9] de Winter, J. C. F., Pfeifer, J., Dodou, D., et al. (2025). Detecting Midjourney-generated images: An eye-tracking study. Proceedings of the Human Factors and Ergonomics Society Annual Meeting.

[10] Osińska, V., Kortas, W., Szalach, A., et al. (2025). AI images vs. real photographs: Investigating visual recognition and perception. Journal of Eye Movement Research, 18(6), 61.

[11] Simonetti, A., & Bigne, E. (2022). How visual attention to social media cues impacts visit intention and liking expectation for restaurants. International Journal of Contemporary Hospitality Management, 34(6), 2049–2070.

[12] Mamalikou, M., Gkatzionis, K., & Panagiotou, M. (2025). The influence of social media-like cues on visual attention: An eye-tracking study with food products. Journal of Eye Movement Research, 18(6), 62.

[13] Riswanto, A. L., Kim, S., Ha, Y., et al. (2025). Visual attention to food content on social media: An eye-tracking study among young adults. Journal of Eye Movement Research, 18(6), 69.

[14] Itti, L., Koch, C., & Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence, 20(11), 1254–1259.

[15] Borji, A., & Itti, L. (2013). State-of-the-art in visual attention modeling. IEEE Transactions on Pattern Analysis and Machine Intelligence, 35(1), 185–207.

[16] Henderson, J. M., & Hayes, T. R. (2017). Meaning-based guidance of attention in scenes as revealed by meaning maps. Nature Human Behaviour, 1, 743–747.

[17] Tatler, B. W., Hayhoe, M. M., Land, M. F., et al. (2011). Eye guidance in natural vision: Reinterpreting salience. Journal of Vision, 11(5), 5.

[18] Mital, P. K., Smith, T. J., Hill, R. L., et al. (2011). Clustering of gaze during dynamic

scene viewing is predicted by motion, visual saliency, and social cues. Journal of Vision, 11(11), 13.

[19] Coutrot, A., & Guyader, N. (2014). How saliency, faces, and sound influence gaze in dynamic social scenes. Journal of Vision, 14(8), 5.

[20] Holmqvist, K., Nyström, M., Andersson, R., et al. (2011). Eye tracking: A comprehensive guide to methods and measures. Oxford University Press.

[21] Orquin, J. L., & Holmqvist, K. (2018). Threats to the validity of eye-movement research in psychology. Behavior Research Methods, 50(4), 1645–1656.

[22] Hessels, R. S., Niehorster, D. C., Kemner, C., et al. (2017). Noise-robust fixation detection in eye movement data: Identification by two-means clustering. Behavior Research Methods, 49(5), 1802–1823.

[23] Bates, D., Mächler, M., Bolker, B., et al. (2015). Fitting linear mixed-effects models using lme4. Journal of Statistical Software, 67(1), 1–48.

[24] Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. Journal of Statistical Software, 82(13), 1–26.

[25] Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. Journal of the Royal Statistical Society: Series B, 57(1), 289–300.