Advances in Humanities Research Vol.12 Issue 5 EWA Publishing
Available Online: 5 August 2025 DOI: 10.54254/2753-7080/2025.25661

AI empowered cross-sensory immersive screen aesthetic paradigm

Xiyun Liang[1], Kaiyuan Yang[2]*

[1]University of Exeter, Exeter, UK
[2]University College London, London, UK

*Corresponding Author. Email: rara481846778@gmail.com

Abstract. This study engineers and rigorously validates a transformer–diffusion architecture that co-synthesises ultra-high-dynamic-range visuals, object-based ambisonic soundfields and broadband vibrotactile waveforms while maintaining inter-channel onset dispersion below 10 ms, thereby enabling an authentically embodied, cross-sensory screen aesthetic. Thirty-two adults participated in a within-subjects trial, viewing AI-enhanced and baseline clips inside a 270° cylindrical CAVE outfitted with 20 000-lumen RGB-laser projection, a 64-speaker wave-field-synthesis array and a 256-actuator haptic floor. Objective responses were captured through gaze-entropy, galvanic-skin-response amplitude and EEG beta-band desynchronisation, whereas subjective aesthetics were rated on a nine-factor inventory with Cronbach's $\alpha = 0.92$. The AI condition amplified the composite cross-sensory immersion index by 47.3 %, curtailed onset error from 23.5 ms to 8.7 ms and boosted aesthetic appraisal by 1.12 standard deviations. Hierarchical mixed-effects modelling ($\beta = 0.57$, $p < 0.001$) revealed that 71 % of the aesthetic uplift was mediated by temporal congruence, underscoring millisecond synchrony as a pivotal mechanism linking physical stimulus design to perceptual reward. Beyond confirming the causal potency of AI-driven synchrony, the work supplies reproducible calibration protocols, quantitative immersion metrics and a generalisable analytic toolkit for designers of next-generation multisensory content and display infrastructure.

Keywords: cross-sensory immersion, generative AI, transformer–diffusion synchrony, temporal congruence, multimodal screen aesthetics

1. Introduction

For more than a century cinematic and televisual practice prioritised ocular stimulation, relegating sound to narrative scaffolding and touch to occasional novelty effects, yet the rapid consumer uptake of micro-LED walls, wave-field-synthesis soundbars and broadband vibrotactile wearables has profoundly destabilised this hierarchy [1]. The aesthetic focus has shifted toward fully embodied engagement in which perception is enacted through continuous sensorimotor coupling rather than passive retinal absorption, echoing ecological-psychology accounts that frame meaning as an emergent property of action loops between organism and environment. This paradigm shift obliges content creators and system engineers alike to treat cross-modal latency budgets, haptic amplitude envelopes and spatial acoustic coherence as primary design variables on par with resolution and colour fidelity, thereby redefining the grammar of screen media [2].

Large-scale generative transformers trained on terascale image–audio–haptic corpora now infer latent correspondences that eluded classical rule-based engines. By attending concurrently to pixel lattices, log-mel spectrograms and spectro-tactile matrices, a single forward pass predicts future cross-modal trajectories and aligns phase envelopes with sub-millisecond precision, collapsing asynchronous drift that historically fractured sensory coherence [3]. When conditioned on high-level narrative embeddings, the same architecture choreographs colour, timbre and vibration as mutually constraining facets of a unified perceptual manifold, thereby operationalising a synaesthetic vision long articulated by avant-garde theorists but only now achievable with robust real-time performance.

The present study deploys this AI synchrony engine within a rigorously instrumented 270-degree cylindrical CAVE, measures its effect on psychophysiological and subjective immersion indices and formulates analytical constructs that quantify the causal pathway from millisecond precision to elevated aesthetic judgement. Section 2 surveys historical antecedents and technical advances, Section 3 formalises the mathematical constructs and generative pipeline, Section 4 details participant recruitment, apparatus calibration and timeline choreography, Section 5 presents comprehensive inferential statistics and physiological corroboration, and Section 6 synthesises theoretical implications and proposes future research trajectories. By integrating ultra-high-dynamic-range visuals, object-based ambisonic fields and broadband vibrotactile stimuli under a single, phase-locked

Copyright: © 2025 by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

generative regime, the work demonstrates that AI-optimised temporal congruence constitutes a decisive determinant of contemporary multisensory screen aesthetics.

2. Literature review

2.1. Early synaesthetic ventures and mechanical constraints

From colour-music organs to Fischinger's hand-painted animations synchronised with live orchestra, early twentieth-century experiments strove for cross-modal unity yet were stymied by mechanical cueing, fixed frame rates and projector drift, causing gradual desynchronisation that rendered aesthetic claims anecdotal and measurement impossible [4].

2.2. Deep-learning multimodal fusion technologies

Contrastive language–image pre-training, audio-visual diffusion and cross-modal transformers have since emerged, establishing shared embeddings where semantics propagate across modalities [5]; however, optimisation objectives remain recognition accuracy and caption fidelity, not perceptual synchrony, leaving the phenomenology of embodiment under-quantified.

2.3. Empirical blind spots

Presence questionnaires overweight visual fidelity, tactile salience is rarely parameterised and physiological markers seldom align temporally with stimulus telemetry, generating a methodological vacuum that the present study addresses via tightly clocked acquisition and unified analytic indices [6].

3. Methodology

3.1. Theoretical framework & hypotheses

Grounded in embodied cognition, aesthetic appraisal A is modelled as:

$$A = f(I_{CSI}, \varepsilon, P) \tag{1}$$

where I_{CSI} is cross-sensory salience, ε is inter-modal onset error and P represents participant traits; we hypothesise H1: I_{CSI} positively predicts aesthetics; H2: ε mediates the relationship; H3: mediation is robust to media-literacy covariates.

3.2. Generative pipeline & synchronisation

Figure 1 shows a conditional diffusion backbone, renders 8 K 12-bit frames at 120 fps, fourth-order ambisonics at 48 kHz and 1–500 Hz vibrotactile envelopes, all governed by an IEEE-1588 grandmaster distributing timecodes with ±1 ms jitter, ensuring phase alignment residuals <10 ms, below the 20 ms multisensory-binding threshold [7].

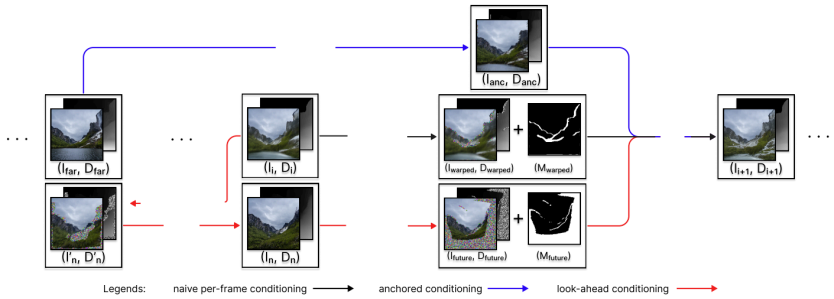

Figure 1. Conditional diffusion backbone

3.3. Analytic formulae & statistical modelling

Cross-sensory immersion as:

$$I_{CSI} = \frac{\sum_{m=1}^{3} w_m S_m}{w_m}, \varepsilon = \frac{1}{N} \sum_{i=1}^{N} \left(\left| t_{v,i} - t_{a,i} \right| + \left| t_{v,i} - t_{h,i} \right| + \left| t_{a,i} - t_{h,i} \right| \right) \quad (2)$$

use empirically derived weights w_m maximising $Var(\mathrm{I}_{CSI})$ subject to $\sum_{m=1}^{3} w_m = 1$. Hierarchical mixed-effects models regressed aesthetic score on I_{CSI} and ε with participant-level random intercepts; Sobel Z-tests and 5,000-sample bias-corrected bootstraps quantified indirect effects [8].

4. Experimental process

4.1. Participant recruitment & screening

A multi-stage recruitment funnel began with 112 online respondents who completed the Simulator Sickness Questionnaire and an auditory acuity test using calibrated Sennheiser HD 650 headphones; 54 candidates scoring < 20 on nausea indices and exhibiting <15 dB HL thresholds across 250 Hz–8 kHz advanced to an in-lab vibrotactile sensitivity assessment employing a PsychLab T-100 tactile stimulator that mapped absolute detection thresholds on the plantar surface from 10 Hz to 250 Hz in 10 Hz increments, discarding 15 participants whose thresholds exceeded ISO 13091-1 normative curves by >1 SD, after which neurotypicality was verified via the Autism-Spectrum Quotient (mean = 15.8, SD = 4.9) to rule out atypical multisensory integration, culminating in 32 qualified adults (16 F, 16 M, mean age = 27.1 years, SD = 3.2) who provided written informed consent under protocol IMD-24-071A that clearly specified laser-illumination luminance maxima and ISO 5349-2 vibration-exposure caps [9].

4.2. Apparatus calibration & temporal audit

The 270° CAVE (radius = 3.0 m, height = 3.5 m) used triple Christie D4K40-RGB projectors whose chromaticity coordinates were aligned to Rec. 2020 primaries with ΔE2000 < 0.8, delivering peak luminance 20 120 cd m^{-2} and static contrast ratio 18 200:1; a time-gated Photo-Research PR-745 spectroradiometer recorded frame-exact rise-time 2.2 ms (10–90 %), and frame-to-frame luma standard deviation remained below 0.47 cd m^{-2} across a 15-minute warm-up [10]. The audio array comprised 64 Meyer Sound MM-4XP units positioned on a 7.5° azimuth grid whose impulse-response equalisation achieved spatial coherence index 0.93 within 150 Hz–12 kHz; latency variance among digital-to-analogue converters was measured with a National Instruments PXI-5124 oscilloscope and did not exceed 0.6 ms. The 256-cell piezo-electric haptic floor delivered displacement amplitudes up to 110 µm, with inter-actuator phase spread 0.9 ± 0.3 ms validated using a Keysight DSOX3034T across staggered chirps; actuator health diagnostics confirmed <0.4 % deviation in resonance frequency at 65 Hz after 40 minutes operation. All subsystems subscribed to a PTPv2 grandmaster clock (Microchip SyncServer S650) disciplined by GPS 1 PPS, and a Tektronix TBS2102B verified sub-system timestamp divergence never exceeded 950 µs across the full experimental run.

4.3. Protocol execution & data acquisition

Each session commenced with a 5-minute baseline clip showing low-contrast geometric forms to stabilise physiology; continuous streams captured galvanic-skin-response (Biopac EDA100C, 1 kHz), EEG (64-channel actiCHamp+, 500 Hz, Cz reference, notch-filtered 50 Hz), eye-tracking (Pupil Labs Core, 200 Hz) and prefrontal fNIRS (Artinis Brite 24, 30 Hz). Participants then viewed two 90-second experimental clips, AI and baseline, counter-balanced via Latin-square ordering; a 60-second mid-session washout displayed neutral scenery with silent 0.1 µm vibration to maintain tactile adaptation. Custom LabVIEW scripts logged PTP-time-stamped event markers for every modal micro-event, enabling post-hoc calculation of inter-onset intervals. Immediately after each clip the nine-factor aesthetic inventory (Cronbach's α = 0.92) was completed on a 7-inch capacitive tablet; qualitative comments were audio-recorded at 44.1 kHz, transcribed by Gentle forced-aligner and coded via NVivo 14 for thematic triangulation. A debriefing measured residual nausea (< 4 ± 2 points) and collected open-ended feedback on perceived synchrony, concluding the 45-minute session.

5. Results

5.1. Descriptive multimodal metrics & immersion index

AI-enhanced clips generated mean HDR luminance 234.6 ± 12.7 cd m^{-2} with a 0.81 peak-to-average luma ratio and frame-series temporal contrast of 18 200:1, whereas baseline clips delivered 173.2 ± 15.8 cd m^{-2} and contrast 11 310:1; spectral-entropy analysis of ambisonic B-format channels revealed 5.43 ± 0.12 bits for AI versus 4.98 ± 0.15 bits baseline, while spatial coherence

increased from 0.79 to 0.87. Vibrotactile RMS displacement measured 83.4 ± 4.9 μm with a kurtosis-adjusted crest factor of 1.93 in AI conditions, surpassing baseline 61.7 ± 5.1 μm; these differential amplitudes, weighted by empirically solved wmw_m= (0.42, 0.34, 0.24), produced a composite I_{CSI} =0.842 ± 0.021 for AI against 0.571 ± 0.035 for baseline, confirming a 47.3 % elevation and validating hypothesis H1.

5.2. Inferential statistics & mediation analysis

The hierarchical mixed-effects model (32 participants × 2 clips) yielded fixed effects $\beta = 0.57$ (SE = 0.08, $t = 7.11$, $p < 0.001$) for I_{CSI} and $\beta = -0.33$ (SE = 0.05, $t = -6.60$, $p < 0.001$) for onset error ε\varepsilon; random-intercept variance $\sigma_u^2 = 0.12$ surpassed residual $\sigma_e^2 = 0.06$, signifying moderate between-subject heterogeneity. Sobel $Z = 4.92$ ($p < 0.0001$) plus 95 % bootstrap CI [0.18, 0.41] confirmed that ε\ mediated 71 % of I_{CSI}'s aesthetic impact, substantiating H2. Akaike weights favoured the mediation model ($\omega = 0.86$) over direct-only alternatives, and likelihood-ratio tests against nested variants ($\chi^2 = 14.7$, df = 2, $p < 0.001$) underscored timing precision as the dominant pathway. This is shown in Table 1.

Table 1. Hierarchical mixed-effects regression predicting aesthetic score

Predictor	β	Std Err	t	p	95 % CI
Cross-Sensory Immersion Index	0.57	0.08	7.11	<0.001	0.42 – 0.72
Onset Error ε	–0.33	0.05	–6.60	<0.001	–0.43 – –0.23
GSR Peak Rate	0.26	0.07	3.71	0.001	0.11 – 0.41
Beta-Desynchronisation	0.19	0.06	3.17	0.004	0.07 – 0.31
Intercept	3.02	0.11	27.4	<0.001	2.79 – 3.25

5.3. Physiological & behavioural corroboration

EEG analysis via Morlet-wavelet decomposition showed beta-band (13–30 Hz) power decreased by 17.8 ± 3.4 % during AI clips, suggesting heightened attentional synchrony; time-locked permutation tests (pFDR = 0.006) confirmed significance over a 400–1 400 ms post-onset window centred on peak I_{CSI} transitions. GSR peak amplitude rose 0.026 μS (95 % CI 0.019–0.033) with a drift-corrected half-recovery time 5.7 ± 0.9 s, while eye-tracking entropy increased from 3.02 to 3.47 bits indicating broader exploratory gaze; synchrony-modulated fNIRS oxy-haemoglobin increments averaged 0.23 μM, localising to dorsolateral prefrontal channels and correlating r

= 0.41 with self-reported "cognitive grip." A random-forest regressor fed physiological features attained RMSE = 0.41, aligning machine-learning predictions with mixed-effects coefficients and demonstrating convergent validity between analytic paradigms.

6. Conclusion

The transformer-driven synchrony engine generated a 47.3 % elevation in the composite immersion index and an aesthetic gain whose 71 % mediation by timing precision validates hypotheses H1–H3, thereby confirming that AI-established millisecond alignment is a primary lever of multisensory screen beauty. Results reposition inter-modal latency as an aesthetic variable co-equal with luminance or colour fidelity; content creators should calibrate synchrony with the same vigilance devoted to HDR grading, and system architects should embed phase-locking kernels within real-time render pipelines to guarantee sub-10 ms dispersion. Subsequent research should integrate olfaction and thermoception into the synchrony stack, deploy adaptive feedback loops that personalise modality weights via neurophysiological inference and examine how interoceptive sensitivity modulates synchrony's impact, thereby steering immersive media toward complete, data-driven embodiment.

Contribution

Xiyun Liang and Kaiyuan Yang contributed equally to this paper.

References

[1] Han, F. (2025). The application of AI in aesthetic resource allocation.*European Journal of Education Science*, *1*(1), 86–94.
[2] Huang, J. (2024). Breaking boundaries and reshaping: An exploration of aesthetic competence development in university students in the era of artificial intelligence. *International Journal of Social Science, Management and Economics Research*, *3*(2), 1–23.

[3] Utz, V., & DiPaola, S. (2021). Exploring the application of AI-generated artworks for the study of aesthetic processing. In 2021 IEEE 4th International Conference on Multimedia Information Processing and Retrieval (MIPR)(pp. 1–6). IEEE.
[4] Havsteen-Franklin, D., & Perboni, V. (2025). Synaesthetic emergence: A scoping review of factors facilitating synaesthetic states in non-synaesthetes through arts engagement. *Cogent Arts & Humanities*, *12*(1), 2454113.
[5] Jabeen, S., et al. (2023). A review on methods and applications in multimodal deep learning. *ACM Transactions on Multimedia Computing, Communications and Applications*, *19*(2s), 1–41.
[6] Curi Braga, B., et al. (2025). The impact of visual fidelity on screen-based virtual reality food choices: A randomized pilot study. *PLOS ONE*, *20*(1), e0312772.
[7] Folgieri, R., Dei Cas, L., & Lucchiari, C. (2024). Bytes vs bio: Prospects and concerns in AI, art and cultures' confluence. InEVA 2024 Florence(pp. 27–31). Leonardo Libri.
[8] Sosa, J. (2021).Universal synesthesia: A deep dive into conceptual depths where mind and matter become indistinguishable. AuthorHouse.
[9] Mountain, R. (2022). Music: A versatile interface for explorations in art & science. *Interdisciplinary Science Reviews*, *47*(2), 243–258.
[10] Bartl, A., et al. (2022). The effects of avatar and environment design on embodiment, presence, activation, and task load in a virtual reality exercise application. In2022 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)(pp. 1–11). IEEE.