Journal Publication of International Research for Engineering and Management (JOIREM)
Volume: 04 Issue: 09 | Sep-2026
ISSN (O) 3107-6696

Cross-Modal Design Systems: Translating Synesthetic Hallmarks into Multi-Sensory UI Tokens to Reduce Cognitive Load and Perceptual Error in Digital Interfaces

Pankaj Jangir, Ruchi Gaur

Student, Department of UX & Digital product design, World University of Design (WUD), Sonipat, India
Assistant Professor, Department of UX & Digital product design, World University of Design (WUD), Sonipat, India
pankajjangir60511@gmail.com

Abstract - Contemporary digital interfaces are overwhelmingly biased toward the visual modality, demanding sustained visual attention that rapidly exhausts human working memory. While multi-modal feedback systems have been introduced to mitigate this visual fatigue, these systems frequently rely on arbitrary, non-ecological mappings between visual, auditory, and haptic cues.

The arbitrary pairing of cross-modal stimuli inadvertently triggers sensory Stroop interference and exacerbates cognitive load. Under conditions of high visual demand, the Colavita visual dominance effect dictates that the human nervous system will often attenuate or completely ignore incongruent auditory and haptic feedback, so secondary sensory channels function as extraneous noise rather than meaningful data conduits.

Grounded in synesthetic neuroscience and crossmodal correspondence theory, this study introduces the Synesthetic Design Token Model — mathematically rigorous, isomorphic mappings between visual properties (luminance, geometry, scale) and acoustic (frequency, envelope) and haptic (vibration frequency, sharpness) parameters. A within-subjects design (N = 30) evaluated the model across three conditions: Visual-Only, Incongruent Multi-Modal, and Congruent Synesthetic, measured via reaction time, error rate, and NASA-TLX cognitive load.

The Congruent Synesthetic interface produced significantly faster reaction times and lower error rates than both other conditions. The Incongruent condition produced the highest cognitive load and slowest reaction times, confirming cross-modal Stroop interference; the Congruent condition significantly reduced mental demand, effort, and frustration.

Extending W3C design tokens to biologically congruent audio-haptic values yields a measurable "diversity gain" in information processing, letting DesignOps teams mitigate the Colavita effect and build higher-throughput, more accessible interfaces.

Keywords: Human-Computer Interaction, Synesthetic Design, Design Tokens, Cognitive Load, Multisensory Integration.

1. INTRODUCTION

The evolution of Human-Computer Interaction (HCI) has proceeded under an enduring paradigm of visual dominance. From complex aviation dashboards and medical informatics displays to everyday mobile applications, modern digital environments demand unprecedented levels of sustained visual attention. This profound reliance on a single sensory channel routinely exceeds the intrinsic processing capacity of human working memory. Sweller's (1988) cognitive load theory, first articulated in the context of instructional design and since widely adapted to HCI, holds that the human brain allocates a finite pool of mental resources to process information; when a UI requires users to track dense, rapidly changing visual variables without support from other modalities, extraneous cognitive load scales rapidly, precipitating decision fatigue and perceptual error.

To alleviate visual saturation, designers have increasingly adopted multi-modal feedback, but this is often executed arbitrarily. Feedback lacking a natural, structural relationship with the visual event induces a cross-modal variant of the classic Stroop interference phenomenon first documented by Stroop (1935): the brain is forced to resolve response competition between incongruent stimuli across different modalities, and processing latency rises accordingly.

This is compounded by the Colavita visual dominance effect. Colavita (1974) demonstrated that naive observers frequently fail to respond to a clearly audible tone when it is presented simultaneously with a visual flash, establishing that simultaneously presented auditory and visual stimuli are not processed with equal weight — the auditory channel is disproportionately neglected. Sinnett, Spence, and Soto-Faraco (2007) later confirmed the robustness of this effect under more complex, ecologically valid stimulation conditions and showed that selective attention can modulate, but not eliminate, the bias. In the context of HCI, the Colavita effect implies that simply appending non-congruent sounds or vibrations to visual UI components does not guarantee multi-sensory integration; under high cognitive load, the central nervous system prioritizes the visual channel and treats unaligned acoustic or haptic signals as extraneous noise.

To bypass the Colavita effect and suppress cross-modal Stroop interference, HCI must pivot toward interfaces that leverage naturally occurring, biologically grounded associations between the senses. Synesthesia — a condition in

which stimulation of one sensory pathway leads to automatic experiences in a secondary pathway — provides a theoretical template for this endeavor. By abstracting synesthetic hallmarks and universal crossmodal correspondences into a scalable design system, practitioners can act as "synesthetic orchestrators," aligning visual, acoustic, and haptic design tokens to facilitate pre-attentive multisensory integration.

This study formalizes these principles into a structural UI architecture termed the Synesthetic Design Token Model, leveraging Shannon's Information Theory and Tononi's Integrated Information Theory (IIT) to investigate how congruent, synesthetically mapped multi-sensory tokens generate a perceptual "diversity gain" that reduces computational demand on the brain.

1.1 AIMS AND OBJECTIVES

The core objectives of this study are as follows:

1. To determine whether a congruent, synesthetically mapped multi-sensory design token system reduces extraneous cognitive load relative to visual-only and incongruent multi-modal interfaces.

2. To measure the impact of synesthetic design tokens on perceptual error rates and reaction times in a high-density data verification task.

3. To formalize a mathematically rigorous, machine-readable mapping between visual, acoustic, and haptic design-token parameters that can be embedded in existing design-system infrastructure.

4. To evaluate whether the framework measurably mitigates the Colavita visual dominance effect and cross-modal Stroop interference under controlled experimental conditions.

5. To situate the proposed token architecture relative to prior ecological-acoustics and synesthetic-orchestration frameworks, identifying what a formal, machine-readable specification adds beyond qualitative design guidance.

6. To translate the empirical findings into actionable DesignOps guidelines for engineering teams building high-density digital interfaces.

1.2 RESEARCH PROBLEM

Existing multi-modal feedback implementations in HCI remain largely qualitative and ad hoc, relying on designer intuition rather than isomorphic, quantifiable mapping rules linking visual primitives to acoustic and haptic parameters. Foundational neuroscience and psychology — cross-activation theory, crossmodal correspondences, the principle of inverse effectiveness, and the "synesthetic orchestrator" model — establish that congruent sensory mappings are neurologically real and statistically robust, yet none of this literature translates into a mathematically rigorous, machine-readable specification suitable for modern design-token pipelines. Consequently, while the Colavita effect and cross-modal Stroop interference are well documented as barriers to effective multi-sensory design, no empirical study has tested whether a formally specified, token-based synesthetic mapping system can measurably overcome these effects in a controlled, task-based digital interface. This project addresses that problem directly.

2. LITERATURE REVIEW

The theoretical foundation of this research synthesizes synesthetic neuroscience, ecological acoustics, mathematical models of communication, and the psychology of crossmodal and multisensory integration.

2.1 Synesthetic Neuroscience and Cross-Activation Theory

Synesthesia provides a distinct neurological framework for understanding effortless, pre-attentive multisensory binding. Ramachandran and Hubbard (2001) introduced the cross-activation theory, proposing that grapheme-color synesthesia arises from hyperconnectivity between contiguous cortical regions during early neural development — specifically between the number/grapheme area and the color-processing center (V4) in the fusiform gyrus, alongside hyperconnectivity involving the angular gyrus.

While clinical synesthesia occurs in only two to four percent of the population, the underlying architecture of cross-activation suggests a continuum of sensory connectivity present in all human brains. Corciulo and Bochicchio (2024) proposed elevating synesthesia from a localized neuropsychological anomaly to a foundational model for HCI, arguing that encoding synesthetic pathways into digital systems allows machines to transmit information aligned with the brain's latent structural expectations.

2.2 Crossmodal Correspondences and Ecological Acoustics

Crossmodal correspondences are universal, statistical associations shared across the general population, distinct from idiosyncratic clinical synesthesia. Spence (2011) mapped robust, innate linkages between sensory dimensions — most prominently between auditory pitch and visual size or elevation, where high-frequency sounds map to small, bright, elevated objects and low-frequency sounds map to large, dark, lower objects. Spence categorizes these correspondences as structural, statistical, or semantic, aligning with Gaver's (1986) work on ecological "auditory icons" that leverage everyday sound physics for intuitive information conveyance.

2.3 The Designer as a Synesthetic Orchestrator

Riccò, Belluscio, and Guerini (2003) positioned the communication designer as a "synesthetic orchestrator" who deliberately coordinates multisensory elements for total

intersensory coherence, distinguishing intersubjective (quantitative) dimensions — pitch, scale, luminance, vibration amplitude — that can be mapped isomorphically, from subjective (qualitative) dimensions such as hue or timbre. Failing to respect quantitative mappings, they argue, produces severe cross-modal interference.

2.4 Information Theory, Diversity Gain, and Integrated Information (Φ)

This framework applies Shannon's Information Theory and Tononi's Integrated Information Theory (IIT). Corciulo and Bochicchio (2024) treat human perception as a communication channel susceptible to noise and entropy; congruent acoustic and haptic signals achieve a sensory "diversity gain" analogous to redundant transmission across independent fading channels in telecommunications. Tononi's Φ measures the degree to which a system's visual, acoustic, and haptic tokens are perceived as a single, irreducible event rather than competing, incongruent stimuli.

2.5 Multisensory Integration and the Principle of Inverse Effectiveness

A further physiological basis for the Synesthetic Design Token Model comes from single-neuron work on multisensory integration in the superior colliculus. Meredith and Stein (1983) and, more extensively, Stein and Meredith (1993) documented that neurons receiving convergent input from two sensory modalities can produce a response exceeding the arithmetic sum of the two unisensory responses — a superadditive enhancement. Critically, this enhancement follows the principle of inverse effectiveness: the proportional benefit of combining two weak, ambiguous, or degraded unisensory signals is far greater than the benefit of combining two already-strong signals. Applied to interface design, this principle predicts that congruent multi-sensory tokens should yield their largest performance gains precisely under the high-density, high-uncertainty conditions rapid visual streams, subtle distractors, degraded attention where the visual channel alone is least reliable, which is exactly the task context evaluated in the present study.

2.6 Sound Symbolism and the Bouba-Kiki Effect

A second, closely related line of evidence for non-arbitrary sound-shape mapping comes from sound-symbolism research. Köhler (1929) first observed that naive participants overwhelmingly matched the nonsense word "takete" to a jagged, angular figure and "baluba" to a rounded, amoeboid figure; Ramachandran and Hubbard (2001) later replicated and extended this using the now-canonical labels "kiki" and "bouba," finding agreement rates above 95% across widely different linguistic populations. This effect is normally discussed as a word-shape phenomenon, but the underlying mechanism — an association between acoustic sharpness (rapid formant transitions, high-frequency plosive consonants) and visual angularity — is the same structural correspondence the Synesthetic Design Token Model formalizes at the level of continuous acoustic parameters (waveform, frequency) rather than discrete phonemes. Framed this way, corner-radius-to-timbre mapping (Section 3.1, Rule 3) is a direct extension of the Bouba-Kiki effect from language into interface geometry.

2.7 Cognitive Load Theory in Interface Design

Sweller's (1988) cognitive load theory distinguishes intrinsic load (inherent to the task), extraneous load (imposed by how information is presented), and germane load (devoted to schema construction). Multi-modal HCI research generally treats additional sensory channels as a way to reduce extraneous load by offloading part of the informational burden from vision onto audition or touch — but only if the offloaded signal does not itself require additional interpretive effort. An incongruent or arbitrary auditory or haptic cue does not reduce extraneous load; it adds a second, competing source of extraneous load, which is precisely the mechanism the present study's Incongruent condition is designed to expose. The Synesthetic Design Token Model can therefore be read as an attempt to guarantee that added sensory channels remain extraneous-load-neutral or extraneous-load-reducing by construction, rather than leaving that outcome to individual designer judgment.

2.8 Commercial and Industry Precedents

Elements of crossmodal congruence already appear piecemeal in commercial software, even without a formal theoretical grounding. Apple's Human Interface Guidelines recommend pairing specific Taptic Engine haptic patterns with specific system sounds for standard interactions (e.g., a lighter "peek" haptic with a higher-pitched confirmation tone, a heavier "impact" haptic with a lower-pitched error tone), which is directionally consistent with the luminance/frequency and mass/envelope correspondences formalized in Section 3.1, though without an explicit, publicly specified mathematical mapping. Game engines similarly couple visual "juice" (screen shake, particle scale, hit-flash brightness) with audio mixing parameters (pitch, transient sharpness) and controller rumble, but this coupling is typically tuned by ear on a per-interaction basis by sound designers rather than derived algorithmically from a shared token source. The Synesthetic Design Token Model's contribution relative to this existing practice is not the intuition that congruent multi-sensory feedback feels better — which experienced designers already exploit informally — but the formalization of that intuition into a deterministic, machine-readable specification that can be authored once and propagated automatically across an entire design system, and that can be empirically validated rather than relying solely on designer judgment.

2.9 Research Gap

None of the reviewed literature translates crossmodal correspondence theory into a mathematically rigorous,

machine-readable specification embeddable in modern design-system infrastructure, and no empirical study has tested whether a formally specified, token-based synesthetic mapping system measurably overcomes the Colavita effect and cross-modal Stroop interference in a controlled digital interface. Table 1 summarizes this gap relative to the closest prior frameworks: ecological auditory icons and synesthetic-orchestration theory offer qualitative design guidance without a machine-readable specification; the W3C Design Tokens Community Group (2025) format standardizes visual-only tokens without any crossmodal dimension; and Corciulo and Bochicchio (2024) call for synesthesia to be adopted as a foundational HCI model but stop short of proposing or validating an operational framework. This study addresses that gap by introducing and empirically testing the Synesthetic Design Token Model.

2.10 RESEARCH QUESTIONS AND HYPOTHESES

RQ1: To what extent does a congruent, synesthetically mapped multi-sensory design token system reduce extraneous cognitive load compared to visual-only and incongruent multi-modal interfaces?

RQ2: How does the application of synesthetic design tokens impact perceptual error rates and reaction times in high-density data verification tasks?

Table 1

Positioning of the Synesthetic Design Token Model Relative to Prior Frameworks

Prior Framework	Core Contribution	Machine-Readable?	Empirically Tested?
Gaver (1986) – Auditory Icons	Ecological sound-event mapping	No	Partial
Riccò et al. (2003) – Synesthetic Orchestration	Designer-led intersensory coherence	No	No
W3C DTCG – Design Tokens	Visual-only token interoperability	Yes	N/A
Corciulo & Bochicchio (2024)	Synesthesia as HCI model (conceptual)	No	No
Present Study	Cross-modal isomorphic token layer	Yes	Yes (N=30)

3. CONCEPTUAL FRAMEWORK: THE SYNESTHETIC DESIGN TOKEN MODEL

Design tokens are the atomic, platform-agnostic building blocks of a design system — colors, spacing, typography, and similar values — typically serialized in JSON for machine consumption. The Design Tokens Community Group (DTCG), operating under the W3C Community process, standardizes how such tokens are formatted and exchanged across design and engineering tooling, reaching its first stable specification (2025.10) after several years of collaborative development. However, this standard, like the token architectures it formalizes, is confined to a two-layer model of primitive tokens (raw values) and semantic tokens (contextual usage) that addresses only the visual modality.

The Synesthetic Design Token Model introduces a third layer to this architecture: Cross-Modal Isomorphic Tokens. This layer establishes strict mathematical mapping rules linking visual rendering properties directly to real-time acoustic generation parameters (via the Web Audio API) and haptic playback parameters (via Apple CoreHaptics or Android-equivalent APIs), so that a single semantic token — "Success," "Warning," "Error" — deterministically generates a congruent value in all three modalities rather than requiring a designer to separately author and reconcile them.

H1: Participants interacting with the Congruent Synesthetic interface will demonstrate significantly lower overall cognitive load (NASA-TLX) than those interacting with Visual-Only or Incongruent Multi-Modal interfaces.

H2: The Congruent Synesthetic condition will yield significantly faster reaction times and lower error rates, demonstrating measurable mitigation of both the Colavita effect and cross-modal Stroop interference.

3.1 Mathematical Mapping Rules

1. Luminance → Frequency & Sharpness: Background lightness L ∈ [0,100] maps logarithmically to oscillator frequency $f_{hm} = f_{min} \cdot (f_{max}/f_{min})^{(L/100)}$, mirroring the auditory-pitch/visual-brightness correspondence documented by Spence (2011), and haptic sharpness S = L/100.

2. Scale → Envelope & Duration: Normalized bounding-box area A_{norm} scales ADSR decay/release time and haptic pulse duration P_{ms} proportionally, reflecting the ecological-acoustics principle that larger physical objects produce longer, more resonant sounds.

3. Corner Radius → Timbre & Intensity: Corner radius R_{px} interpolates waveform ω between sine (ω=0) and sawtooth (ω=1), directly operationalizing the Bouba-Kiki-style association between rounded geometry and smooth timbre, and inversely scales haptic intensity I ∈ [0,1].

3.2 Structured Token Specification

These mappings are serialized into an extended JSON schema, structurally compatible with the DTCG Format Module, that unifies the three modalities at the architectural level rather than treating acoustic and haptic feedback as an external, hand-tuned layer bolted onto a purely visual design system.

Table 2

Synesthetic Design Token Specification

Token Level	Visual (CSS)	Acoustic (Web Audio)	Haptic (AHAP)	Semantic Meaning
Success	#1D9E75; radius 8px	880 Hz sine, short attack	S=0.8, I=0.6	Task completion, low friction
Warning	#F59E0B; radius 0px	440 Hz sawtooth	S=0.9, I=0.9, 100ms	Attention required
Neutral	#F9FAFB; large shadow	220 Hz triangle, long decay	S=0.2, I=0.4, 200ms	Background elevation
Error	#DC2626; radius 0px	110 Hz square, distortion	S=1.0, I=1.0	System error

3.3 Illustrative Token Payload

The following abbreviated payload illustrates how a single semantic token resolves to all three modalities at build time, consistent with the DTCG's JSON-based exchange format:

{ "token": "success.primary", "visual": { "color": "#1D9E75", "radius": "8px" }, "acoustic": { "freq": 880, "wave": "sine", "attack": "short" }, "haptic": { "sharpness": 0.8, "intensity": 0.6 } }

4. EMPIRICAL METHODOLOGY

To empirically validate the Synesthetic Design Token Model, a controlled, within-subjects experimental design measured the impact of crossmodal congruence on cognitive load, reaction time, and error rate.

4.1 Participants

A priori power analysis (G*Power) indicated N = 30 was required for 0.80 power at medium-to-large effect sizes (η^{p2} = .15). Participants (14 male, 16 female, aged 22–38, M = 26.4, SD = 3.8) reported normal vision, hearing, and tactile sensation, and were screened using the Eagleman Battery for Synesthesia to exclude clinical synesthetes, ensuring the sample reflected the general population's reliance on universal statistical crossmodal correspondences rather than idiosyncratic synesthetic wiring.

4.2 Apparatus

Stimuli were presented on a 144Hz monitor to eliminate display-lag confounds. Audio was synthesized via the Web Audio API through Sennheiser HD600 headphones, with envelopes managed via the exponentialRampToValueAtTime() method for precise ADSR control. Haptics used Wideband Linear Resonant Actuators (LRAs) rather than Eccentric Rotating Mass (ERM) motors: ERM motors suffer from 50–100 ms spin-up/spin-down latency and inextricably couple frequency to amplitude, whereas LRAs offer ~5–20 ms start/stop latency and independent frequency control — a hardware requirement for rendering the precise haptic sharpness and transient pulses specified by the AHAP payloads used in the study.

4.3 Conditions

Condition A (Visual-Only): standard visual feedback only, with zero acoustic or haptic output.

Condition B (Incongruent): tri-modal feedback with mappings that deliberately violate crossmodal correspondence — e.g., a small, bright, rounded success notification paired with a low-frequency, heavily distorted square wave and a long, low-frequency haptic rumble.

Condition C (Congruent Synesthetic): tri-modal feedback aligned per the Synesthetic Design Token Model — small, bright, rounded elements trigger high-frequency sine-wave transients and sharp, brief haptic pulses.

4.4 Task and Procedure

Participants completed 150 trials per condition (450 total) in a high-density data-verification task, categorizing cards as "Valid" or "Invalid" against a target alphanumeric sequence while ignoring visually similar semantic distractors. Condition order was counterbalanced via Latin Square to mitigate order, learning, and fatigue effects. NASA-TLX (Hart & Staveland, 1988) was administered after each block, assessing Mental Demand, Physical Demand, Temporal Demand, Performance, Effort, and Frustration.

4.5 Statistical Analysis Plan

All primary outcome variables (reaction time, error rate, and aggregate NASA-TLX) were analyzed using one-way repeated-measures ANOVA with condition (A, B, C) as the within-subjects factor. Mauchly's test of sphericity was applied to each omnibus test prior to interpretation; where violated, degrees of freedom were adjusted using the Greenhouse-Geisser correction rather than assuming compound symmetry. Significant omnibus effects were followed by pairwise post-hoc comparisons with Bonferroni correction to control the family-wise error rate across the three pairwise contrasts (A vs. B, A vs. C, B vs. C) per outcome measure. Effect sizes are reported as partial eta-squared (η^{p2}) for omnibus tests and Cohen's d for pairwise comparisons, following conventional small/medium/large thresholds of 0.2/0.5/0.8 for d and 0.01/0.06/0.14 for η^{p2}.

5. RESULTS AND QUANTITATIVE ANALYSIS

Data were analyzed via Repeated-Measures ANOVA (α = .05), with Greenhouse-Geisser correction applied where Mauchly's test indicated sphericity violation.

5.1 Reaction Times and Error Rates

Feedback condition significantly affected reaction time, $F(2,58) = 45.32$, $p < .001$, $\eta^{p2} = .61$. Condition C was fastest ($M = 312$ ms, $SD = 34$) versus Condition A ($M = 358$, $SD = 41$), $p < .01$, $d = 1.22$, and Condition B was slowest ($M = 415$, $SD = 48$), significantly slower than both A ($p < .01$, $d = 1.28$) and C ($p < .001$, $d = 2.48$) — a textbook cross-modal Stroop signature.

Error rate followed the same pattern, $F(2,58) = 31.18$, $p < .001$, $\eta^{p2} = .52$: Condition C lowest (2.4%), Condition A moderate (5.8%), Condition B highest (11.2%), supporting the sensory "diversity gain" hypothesis and consistent with the principle of inverse effectiveness — the congruent multisensory condition delivered its largest relative benefit precisely on the trials where visual discrimination was most ambiguous.

5.2 Cognitive Load (NASA-TLX)

Condition significantly affected aggregate workload, $F(2,58) = 28.74$, $p < .001$, $\eta^{p2} = .50$. Condition C lowered Mental Demand and Frustration versus both other conditions despite presenting more raw sensory data — consistent with high Integrated Information (Φ) fusing the tri-modal input into a single perceptual gestalt. Condition B produced a marked spike in Effort and Frustration, confirming that arbitrary multi-sensory feedback actively harms the user experience.

5.3 Summary Results

Table 3

Reaction Time, Error Rate, and NASA-TLX Outcomes by Condition

Metric	Cond. A	Cond. B	Cond. C	Stats
RT (ms)	358±41	415±48	312±34	p<.001
Error (%)	5.8±1.2	11.2±2.1	2.4±0.8	p<.001
NASA-TLX	62.4±8.5	78.1±9.2	45.3±6.7	p<.001
Frustration	65.2±10.1	85.4±8.8	32.1±7.5	p<.001
Mental Demand	68.0±9.5	82.3±10.2	48.5±8.1	p<.001

Note. Values are M ± SD; RT = reaction time; TLX = Task Load Index.

These results support both H1 and H2: synesthetic design tokens function as cognitive prosthetics, reducing latency,

error, and fatigue by harmonizing data transmission across parallel sensory pathways.

6. DISCUSSION AND DESIGN FRAMEWORK

6.1 Defeating the Colavita Effect

Condition B illustrates how sensory mismatch forces the brain to exert executive effort suppressing non-visual anomalies, driving up RT, error, and frustration. Condition C shows the Colavita effect is not insurmountable: congruent tokens achieve true pre-attentive binding, fusing modalities into a maximally irreducible conceptual structure (MICS) before executive processing, so the user reaps Shannon's diversity gain and cognitive load falls even as more sensory data is presented.

6.2 Actionable Guidelines for DesignOps

Unified Token Authoring: Design-system tooling (Figma variables, Style Dictionary) should algorithmically derive acoustic/haptic tokens from visual primitives via CI/CD, using a DTCG-compatible schema extended with the acoustic and haptic attributes defined in Section 3.

Sensory Diversity by Default: High-stakes environments (dashboards, flight decks, medical terminals) should offload status updates to haptic/acoustic channels, freeing the visual channel for priority tasks — precisely the condition under which the principle of inverse effectiveness predicts the largest multisensory benefit.

Accessibility: Congruent multisensory tokens reduce cognitive friction for neurodivergent users (e.g., ADHD, ASD) without requiring a separate accessibility mode. They also directly serve WCAG Success Criterion 1.3.3 (Sensory Characteristics), which requires that instructions for operating content not rely solely on a single sensory characteristic such as shape, color, or sound; a token architecture that guarantees every visual state has a congruent, non-arbitrary acoustic and haptic counterpart provides this redundancy by construction rather than as an afterthought layered on for compliance.

6.3 Domain-Specific Implications

Aviation and Safety-Critical Displays: Cockpit and control-room interfaces already use auditory alerts, but rarely with a formalized crossmodal grammar; a token layer would let alert severity be expressed consistently and predictably across sight, sound, and touch, potentially reducing the time to correctly identify a warning under high workload.

Gaming and Immersive Media: Game engines already couple visual and acoustic feedback tightly for “juice” and feel; a formal token layer would let that congruence be authored systematically rather than tuned by ear per interaction, and would extend naturally to haptic controllers.

Enterprise and Productivity Software: Notification-heavy tools (email, chat, project management) could use the model to differentiate urgency levels across modalities without adding visual clutter, addressing a common complaint about alert fatigue in such tools.

6.4 Comparison to Post-Hoc Sonification Approaches

It is worth distinguishing the present approach from post-hoc sonification, in which an existing visual interface is separately augmented with sound after the fact — the dominant practice criticized in Section 2.9's research gap. Post-hoc sonification necessarily treats the acoustic channel as secondary and derivative, authored by a different team (often audio engineers rather than UI designers) working from a finished visual spec rather than a shared token source. This separation is precisely what allows incongruent mappings like those tested in Condition B to enter production software: nothing in the post-hoc workflow structurally prevents a sound designer from choosing a sharp, high-arousal chime for a soft, rounded success state, because the two decisions are made independently, often at different points in the development timeline, by people with no shared specification connecting them. The Synesthetic Design Token Model's single-source-of-truth architecture (Section 3) removes this failure mode by construction rather than by process discipline or design review, which is a meaningfully different — and more scalable — guarantee for large design systems maintained by dozens of contributors over multiple years.

7. THREATS TO VALIDITY

7.1 Internal Validity

Condition order was counterbalanced via Latin Square specifically to control for practice and fatigue effects, which otherwise threaten internal validity in any within-subjects design spanning 450 trials. A further threat is experimenter or stimulus-construction bias in Condition B: because the incongruent mappings were deliberately authored to violate crossmodal correspondence, it is possible that they were, on average, more jarring or aversive than a naturalistic incongruent mapping a real design team might accidentally ship. This was mitigated by matching all three conditions for stimulus loudness (LUFS-normalized) and haptic amplitude range, so that Condition B's cost reflects incongruence specifically rather than simple intensity.

7.2 External Validity

The high-density data-verification task was chosen to maximize visual working-memory load and thereby give the Colavita effect and cross-modal Stroop interference the best possible opportunity to manifest, consistent with the study's confirmatory aim. This same choice, however, limits generalization to lower-density, more typical consumer interfaces, where the baseline visual-only

condition may already be well within users' processing capacity and the relative benefit of congruent tokens correspondingly smaller — a direct prediction of the principle of inverse effectiveness (Section 2.5).

7.3 Construct Validity

NASA-TLX is a well-validated general workload instrument, but it was not designed specifically to isolate crossmodal integration cost from general task difficulty. Reaction time and error rate provide convergent, more directly interpretable behavioral evidence for H2, and the consistency of all three measures (RT, error rate, and NASA-TLX) pointing in the same direction across conditions strengthens confidence that the underlying construct — crossmodal congruence — rather than a nuisance variable, is driving the observed effects.

8. LIMITATIONS AND FUTURE DIRECTIONS

8.1 Hardware Constraints

Most consumer hardware still uses ERM motors with 50–100 ms latency and coupled frequency-amplitude response, which would likely recreate Condition B's interference. Piezoelectric actuators (1–400 Hz, near-instantaneous latency) are recommended for future deployments.

8.2 Longitudinal Adaptation

This cross-sectional design cannot show whether cognitive-load reductions persist over months of daily use or diminish through neural habituation, an open question given that repeated exposure to any consistent sensory pairing can, in principle, shift from effortful to automatic processing.

8.3 Sample and Ecological Validity

The sample (N = 30, university-recruited) and the simulated data-verification task, while methodologically controlled, may not generalize to the full diversity of real-world interface contexts, professional user populations, or task types; replication with domain-specific tasks (e.g., actual dashboard monitoring, actual cockpit simulation) is needed before the guidelines in Section 6 can be treated as validated for those settings specifically.

8.4 Spatial Computing and XR

Future work should extend the model into XR, mapping depth (Z-axis) and spatial audio (HRTFs) to spatial haptic arrays to prevent disorientation and cybersickness in immersive environments.

8.5 Engineering Cost of Adoption

A reasonable objection to the model is implementation cost: does adding a third token layer meaningfully increase design-system maintenance burden? In the baseline (visual-only) workflow, a designer authors one value per state; under the Synesthetic Design Token Model, the designer still authors only the visual value, since the acoustic and haptic values are algorithmically derived rather than separately hand-tuned (Section 3.1). The marginal engineering cost is therefore concentrated in a one-time implementation of the mapping functions themselves — the logarithmic frequency mapping, the ADSR scaling, and the waveform interpolation — rather than in ongoing per-token authoring effort. This is analogous to how a color-contrast-checking rule, once implemented in a design-system's tooling, does not require each individual color token to be manually re-validated by a human reviewer.

9. IMPLEMENTATION CASE STUDY: A WORKED EXAMPLE

To make the abstract mapping rules in Section 3 concrete, this section walks through applying the Synesthetic Design Token Model to a single, common interface component: a form-submission confirmation toast.

9.1 Baseline (Visual-Only) Implementation

In a conventional design system, a designer specifies only the visual token: a rounded (8px radius), high-luminance green surface (#1D9E75) that fades in over 200ms. No acoustic or haptic behavior is defined; if a developer later wants sound, they must separately choose a sound effect from a library, with no guarantee that its pitch, envelope, or timbre bears any principled relationship to the visual token it accompanies.

9.2 Synesthetic Token Resolution

Under the proposed model, the same semantic token ("success.primary") resolves deterministically to all three modalities via the mapping rules in Section 3.1: the 8px corner radius and #1D9E75 luminance jointly determine a sine-wave oscillator near 880 Hz with a short attack and medium release, and a haptic sharpness of 0.8 with intensity 0.6. A developer changing only the visual radius — for instance, softening it to 16px for a friendlier brand refresh — automatically shifts the acoustic waveform further toward pure sine and reduces haptic intensity, without any additional design or engineering decision. This is the practical payoff of Section 3.3's schema: the crossmodal relationship is enforced structurally, not by convention or code review.

9.3 Failure Mode Illustrated

The Incongruent condition tested in this study corresponds to a common real-world failure mode: a designer reuses a generic "notification" sound effect (often a sharp, high-arousal chime chosen for attention-grabbing rather than semantic fit) across both success and error states. Under the token model, this failure is structurally prevented, because the acoustic and haptic values are derived from the same visual parameters that

already distinguish success from error, rather than authored independently for each state.

10. ETHICAL CONSIDERATIONS

All participants provided informed consent prior to the experimental session and were informed that they could withdraw at any time without penalty. Because the study involved sustained exposure to auditory and haptic stimuli, including deliberately incongruent and mildly aversive combinations in Condition B, participants were screened for any reported sensory sensitivity or discomfort with vibrotactile stimulation prior to enrollment, and sessions were capped in length to limit fatigue. The Eagleman Battery for Synesthesia screening, while primarily methodological, also served to ensure that participants who might experience atypical or more intense involuntary cross-sensory percepts were not inadvertently placed in a condition designed to induce sensory conflict. Given the accessibility motivation underlying this work, any future field deployment of the model in production software should similarly include an opt-out or intensity-adjustment mechanism for users with auditory or tactile sensitivities, consistent with the accessibility guidance in Section 6.2.

11. TOWARD A STANDARDIZED CROSSMODAL SPECIFICATION

A natural extension of this work is formal standardization. The Design Tokens Community Group's Format Module reached its first stable specification only recently, after several years of iterative, consensus-driven development among design-tooling vendors; its scope remains deliberately limited to visual and structural token types (color, dimension, typography, and similar categories). Extending that specification to a formally ratified acoustic and haptic module would require the same kind of multi-stakeholder process: tool vendors (Figma, Style Dictionary, and comparable systems) would need to agree on a shared vocabulary for acoustic parameters (oscillator type, frequency, envelope stage durations) and haptic parameters (sharpness, intensity, duration) analogous to the CSS-derived vocabulary the visual module already uses.

Two practical obstacles stand between the present study and that standardization goal. First, haptic APIs remain fragmented across platforms — Apple's CoreHaptics AHAP format, Android's VibrationEffect API, and various third-party game-controller SDKs each expose different parameter sets and units, so a cross-platform token specification would need an abstraction layer analogous to how CSS colors abstract over different underlying color-management pipelines. Second, and more fundamentally, the mapping rules formalized in Section 3.1 were derived from, and validated against, a single experimental task and a single participant population; before they could be proposed as a default behavior for a public specification, they would need replication across a broader range of tasks, hardware configurations, and user populations, consistent with the external-validity concerns raised in Section 7.2. The present study should therefore be read as an empirical proof of concept for the mapping principle, not as a finished specification ready for standardization.

12. IMPLICATIONS FOR DESIGN EDUCATION

Beyond its engineering implications, the Synesthetic Design Token Model has implications for how multi-sensory design is taught. Riccò et al.'s (2003) "synesthetic orchestrator" framing already implies that designers need some grounding in cross-sensory perception, but design curricula have historically treated visual design, sound design, and haptic/product design as largely separate disciplines with little shared vocabulary. A token-based framework offers a concrete, teachable artifact — a small set of parametrized mapping rules — that could be introduced alongside typography and color theory in interaction-design coursework, giving students a structural, testable starting point for multi-sensory decisions rather than relying purely on trial-and-error intuition built up over a career. This pedagogical angle is speculative and was not tested in the present study, but it follows naturally from treating crossmodal correspondence as a formal design constraint rather than a matter of taste.

APPENDIX A: EXTENDED MAPPING WORKED EXAMPLES

The four semantic tokens specified in Table 2 (Section 3.2) represent the minimal viable token set for a status-communication system. A production deployment would extend this set to cover additional common UI states — for example, a "Loading / In-Progress" token (mid-luminance, ascending pitch glide, low-intensity continuous haptic pulse to signal an ongoing, not-yet-resolved process) and a "Disabled" token (minimum luminance, no acoustic output, zero haptic intensity, signaling the deliberate absence of an interactive affordance). Extending the token set in this way follows directly from the mapping rules in Section 3.1 rather than requiring new design decisions for each additional state, which is the central engineering advantage the model claims over ad hoc multi-sensory design.

A second practical extension concerns internationalization and cultural variation. Spence's (2011) tripartite classification of correspondences (structural, statistical, semantic) implies that the luminance-to-frequency and scale-to-envelope mappings used here, being structural and statistical, should generalize broadly across cultural contexts, whereas any qualitative (hue- or timbre-based) token extensions would need local validation, consistent with Riccò et al.'s (2003) distinction between intersubjective quantitative dimensions and more culturally contingent qualitative ones.

APPENDIX B: GLOSSARY OF SYMBOLS AND TERMS

L: Normalized luminance/lightness of a visual token, L ∈ [0, 100].

f_{hm}: Oscillator fundamental frequency (Hz) assigned to a visual token.

S: Haptic sharpness parameter, S ∈ [0, 1] (CoreHaptics AHAP).

I: Haptic intensity parameter, I ∈ [0, 1] (CoreHaptics AHAP).

A_{norm}: Bounding-box area of a UI element, normalized to viewport area.

R_{px}: Corner radius of a UI element, in pixels.

ω: Waveform interpolation parameter between sine (0) and sawtooth (1).

Φ (Phi): Tononi's Integrated Information measure; degree to which a system's tri-modal tokens are perceived as a single irreducible event.

MICS: Maximally Irreducible Conceptual Structure (IIT terminology).

CONCLUSION

This research demonstrates that the solution to visual sensory overload is embedded in the brain's own cross-modal architecture. By acting as synesthetic orchestrators, HCI practitioners can extract isomorphic mapping rules from cross-activation theory, ecological acoustics, and the principle of inverse effectiveness. The Synesthetic Design Token Model provides a rigorous framework for extending design tokens into the acoustic and haptic domains, and the empirical evidence is unequivocal: congruent multi-sensory binding drastically reduces reaction time, mitigates error, and slashes cognitive load, pointing toward interfaces that communicate with the human nervous system in its own natural language.

REFERENCES

[1] F. B. Colavita, "Human sensory dominance," Perception & Psychophysics, vol. 16, no. 2, pp. 409–412, 1974.

[2] S. Corciulo and M. A. Bochicchio, "Synesthesia as a model for HCI: A systematic review," arXiv preprint arXiv:2404.09303, 2024.

[3] Design Tokens Community Group, "Design Tokens Format Module," W3C Community Group Report, 2025.

[4] W3C Web Accessibility Initiative, "Understanding Success Criterion 1.3.3: Sensory Characteristics," Web Content Accessibility Guidelines (WCAG) 2.1, 2018.

[5] W. W. Gaver, "Auditory icons: Using sound in computer interfaces," Human-Computer Interaction, vol. 2, no. 2, pp. 167–177, 1986.

[6] S. G. Hart and L. E. Staveland, "Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research," Advances in Psychology, vol. 52, pp. 139–183, 1988.

[7] W. Köhler, Gestalt Psychology. New York, NY: Liveright, 1929.

[8] M. A. Meredith and B. E. Stein, "Interactions among converging sensory inputs in the superior colliculus," Science, vol. 221, no. 4608, pp. 389–391, 1983.

[9] V. S. Ramachandran and E. M. Hubbard, "Synaesthesia—A window into perception, thought and language," Journal of Consciousness Studies, vol. 8, no. 12, pp. 3–34, 2001.

[10] D. Riccò, A. Belluscio, and S. Guerini, "Design for the synesthesia: Audio, visual and haptic correspondences experimentation," in Proc. 1st Int. Meeting Sci. Technol. Design, Lisbon, Portugal, 2003.

[11] D. A. Sinnett, C. Spence, and S. Soto-Faraco, "Visual dominance and attention: The Colavita effect revisited," Attention, Perception, & Psychophysics, vol. 69, no. 5, pp. 673–686, 2007.

[12] C. Spence, "Crossmodal correspondences: A tutorial review," Attention, Perception, & Psychophysics, vol. 73, no. 4, pp. 971–995, 2011.

[13] B. E. Stein and M. A. Meredith, The Merging of the Senses. Cambridge, MA: MIT Press, 1993.

[14] J. R. Stroop, "Studies of interference in serial verbal reactions," Journal of Experimental Psychology, vol. 18, no. 6, pp. 643–662, 1935.

[15] J. Sweller, "Cognitive load during problem solving: Effects on learning," Cognitive Science, vol. 12, no. 2, pp. 257–285, 1988.