MAV-C: A Framework for the Joint Objective Estimation of Audio-Visual Complexity in Immersive Virtual Environments

Luca Resti*[1,2], Amelia Gully[3], Michael McLoughlin[2], Gavin Kearney[2], and Alena Denisova[1]

[1] *Department of Computer Science, University of York, UK*
[2] *AudioLab, School of Physics, Engineering and Technology, University of York, UK*
[3] *Department of Language and Linguistic Science, University of York, UK*

Abstract

We present the Motion-Aware Audio-Visual Complexity Metric (MAV-C), a reference-free framework for the joint objective estimation of audio-visual complexity. The metric combines entropy-based audio features (temporal, spectral, and spatial) with visual features (Sobel gradient magnitude, chromatic uniqueness, and optical flow) via a parametric fusion stage, producing a continuous joint complexity score $C_{AV}(t) \in [0,1]$. We validate MAV-C on two datasets: a controlled synthetic corpus (SYN) of stimuli with known signal characteristics and a naturalistic gameplay corpus (GAM) of 60 clips drawn from the SAFEPLAY-X dataset. On SYN, the metric exhibits strong validity: the audio score C_A and visual score C_V are each insensitive to changes in the opposite modality (CoV < 0.003), the joint score C_{AV} spans $[0.00, 0.90]$ across all parameter combinations, and single-axis feature sweeps produce monotone trajectories (Spearman ρ up to 0.995). On GAM, C_V differs significantly across content categories (Kruskal-Wallis $p = 0.021$) while C_A does not, and the two sub-scores are uncorrelated ($r = 0.03$), confirming they operate on independent signal dimensions. OFAT sensitivity analysis identifies a two-tier parameter hierarchy, with modality balance (w_a) and visual regularization (ρ_v) as most significant tunable parameters. Full subjective calibration is planned as future work.

*Corresponding author: luca.resti@york.ac.uk

1 Introduction

Many multimedia applications are inherently audio-visual: their perceptual impact is determined by the joint interaction of both modalities, not from either in isolation. Evaluating such content therefore requires assessment frameworks that jointly process audio and visual signals rather than as independent streams. Furthermore, many existing objective quality metrics often focus on signal-level distortions introduced by compression or transmission, but are blind to the informational and semantic demand of the content itself. In locally-rendered environments such as video games, encoding artefacts are absent or minimal, and existing metrics are therefore not well-suited to help practitioners assess the quality of audio-visual scenes. Dense visual signals and multi-source acoustic environments can saturate cognitive capacity and produce cross-modal interference [12], with consequences for user experience that distortion metrics cannot capture.

The relationship between the structural complexity of audio and visual signals and human perception has been a growing area of research in recent years. Acoustic Scene Complexity (ASC) [13] was found to be largely shaped by the spectro-temporal and spatial distribution of sound events. Signal-based features such as energy and spectral entropy have recently been successfully used to characterize ASC objectively [6], and extended to account for the spatial distribution of sound sources through Spherical Acoustic Spatial Entropy (SASE) [8]. In the visual domain, perceived complexity is influenced by structural and chromatic signal variety [7], and

Multi-Scale Sobel Gradients (MSG) and Multi-Scale Unique Color (MUC) have been identified as interpretable features for predicting complexity ratings [9]. Both the subjective datasets and the objective models derived from them [13, 7, 9] address a single modality in isolation. Visual complexity research has focused on static images: to our knowledge, no annotated dataset of perceived complexity exists for video (with audio), and no framework jointly models audio and visual complexity over time on video signals. There is thus a significant gap for applications such as video games, where ASC and visual scene content are known to jointly influence perception [4].

We propose Motion-Aware Audio-Visual Complexity Metric (MAV-C), a reference-free framework that estimates Audio-Visual Complexity (AVC) by combining entropy-based audio features with visual complexity features mediated by a temporal motion component. The metric is parametric, with component weights designed to be calibrated against human complexity ratings; no such joint audio-visual dataset currently exists, so subjective validation is deferred to future work.

2 The MAV-C Architecture

MAV-C estimates joint audio-visual complexity as a continuous score $C_{AV}(t) \in [0, 1]$ from synchronized audio-visual stimuli. Each signal feature is normalized against its physical upper bound (e.g. maximum possible entropy given the number of bins), rather than against statistics of a particular dataset. This makes the scale comparable across stimuli and resolutions without recalibration. The architecture has three sequential stages: feature extraction, logarithmic range compression, and late fusion.

2.1 Feature Extraction

Audio and visual features are extracted and aligned via a shared time grid.

Audio Modality: The audio sequence is analyzed at 100 Hz with a 30 ms analysis window and 10 ms hop length, yielding feature vector $\mathbf{x}_A(t) \in \mathbb{R}^K$. All components are Shannon entropies across different physical domains: time (Energy Entropy (EE)), frequency (Spectral Entropy (SE)), and space (SASE), giving $\mathbf{x}_A(t) = [EE(t),\, SE(t),\, SASE(t)]^\top$. SASE requires audio source trajectory metadata and is optional: $K = 3$ in spatial mode and $K = 2$ in stereo mode when such metadata is unavailable.

EE quantifies the uniformity of energy distribution within an analysis frame [6]. Each 30 ms frame is divided into $N_E = 10$ equal sub-frames. Let $p_t(k)$ be the fractional signal power of sub-frame k, the temporal entropy is:

$$EE(t) = -\sum_{k=1}^{N_E} p_t(k) \log_2 p_t(k). \tag{1}$$

Impulsive transients concentrate energy in few subframes, yielding low EE; sustained broadband signals distribute energy uniformly, approaching the analytical bound of $\log_2(10) \approx 3.32$.

SE quantifies the uniformity of the acoustic power distribution across the frequency spectrum [6]. From the Hamming-windowed STFT, letting $P_t(m)$ be the normalized power at bin m over the perceptually relevant band [16 Hz, 10 kHz] ($M = 300$ bins at 48 kHz):

$$SE(t) = -\sum_{m=1}^{M} P_t(m) \log_2 P_t(m) \tag{2}$$

The theoretical maximum is $\log_2(300) \approx 8.23$.

Finally, SASE is an optional feature that evaluates the spatial dispersion of acoustic energy using source trajectory metadata [8]. The spherical acoustic map is partitioned into $N = 64$ equal-area discrete solid angles Ω_i; letting $I_t(\Omega_i)$ denote the normalized active intensity from sector i:

$$SASE(t) = -\sum_{i=1}^{N} I_t(\Omega_i) \log_2 I_t(\Omega_i) \tag{3}$$

Spatial complexity reaches its maximum of $\log_2(64) = 6.0$ when energy is equally distributed across all sectors (perfectly diffuse field) and approaches its minimum when all energy arrives from a single direction.

Visual Modality: Video is processed at its native frame rate, with features merged onto the 100 Hz audio grid. The visual feature space is composed of structural, chromatic, and temporal components.

Let $\mathbf{I}(t) \in [0, 1]^{H \times W \times 3}$ denote a multi-channel video frame at time t. Structural spatial complexity

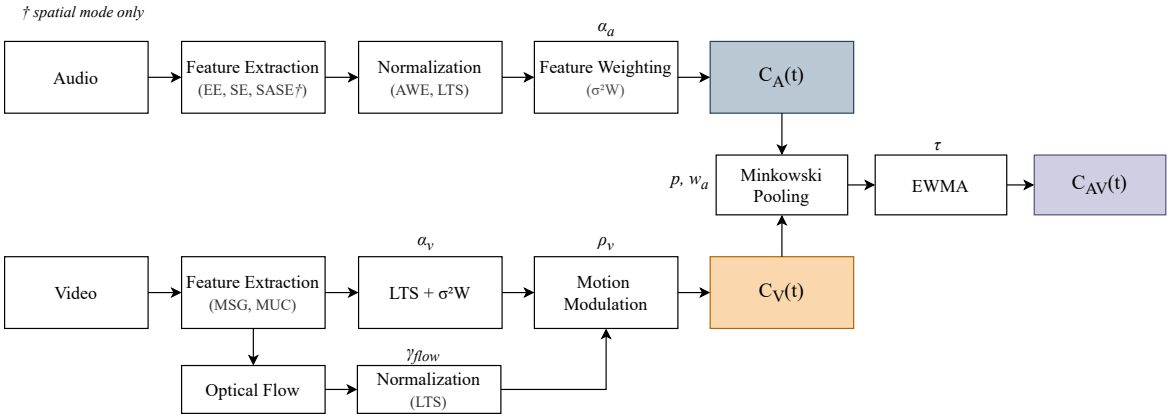


Figure 1: MAV-C block diagram. Parallel audio and visual pipelines produce sub-scores $C_A(t)$ and $C_V(t)$, fused into joint score $C_{AV}(t) \in [0, 1]$. Annotated parameters are tunable.

is quantified via MSG [9]. Applying horizontal and vertical Sobel convolution operators $\mathbf{D}_x$ and $\mathbf{D}_y$ to the color channels $c \in \{1, 2, 3\}$ at each of four spatial scales $s \in \{1, 2, 4, 8\}$, where s is a downsampling factor giving a frame of size $H/s \times W/s$, the per-pixel gradient magnitude matrix $\mathbf{G}_{c,s} \in \mathbb{R}^{H_s \times W_s}$ is:

$$\mathbf{G}_{c,s}(t) = \sqrt{(\mathbf{I}_{c,s}(t) * \mathbf{D}_x)^{\circ 2} + (\mathbf{I}_{c,s}(t) * \mathbf{D}_y)^{\circ 2}} \quad (4)$$

where $*$ denotes 2D discrete convolution, $(\cdot)^{\circ 2}$ denotes element-wise squaring (Hadamard power), and the $\sqrt{\cdot}$ is applied element-wise to yield a per-pixel gradient magnitude map. The global MSG score is the weighted summation of the spatial means of these pixel-wise gradient magnitudes across all color channels and spatial scales:

$$MSG(t) = \sum_s w_s \left(\frac{1}{3} \sum_{c=1}^{3} \frac{1}{H_s W_s} \sum_{x=1}^{W_s} \sum_{y=1}^{H_s} \mathbf{G}_{c,s}(x, y, t) \right). \quad (5)$$

The scale weights are $w_s \in \{0.4, 0.3, 0.2, 0.1\}$ for $s \in \{1, 2, 4, 8\}$, assigning greater importance to full-resolution gradients [9]. The analytical maximum, derived from the Sobel kernel norm applied to a unit-range signal, is $\sqrt{20} \approx 4.47$.

Chromatic spatial complexity is evaluated via MUC [9]. At each scale s, the color channels are uniformly quantized via bit-depth reduction, mapping the continuous color space to a discrete set of states. Let $\mathcal{U}(\cdot)$ denote an operator that returns the cardinality of the set of unique discrete vectors in a matrix. The MUC score is the weighted summation of unique quantized colors present at each scale, using the same weights as MSG ($w_s \in \{0.4, 0.3, 0.2, 0.1\}$):

$$MUC(t) = \sum_s w_s |\mathcal{U}(\mathbf{I}_{q,s}(t))|. \quad (6)$$

Because the number of unique colors is bounded by pixel count, the upper bound of MUC depends on native resolution and quantization depth. This maximum is computed per stimulus at runtime and used for normalization, preserving the $[0, 1]$ output range across any input resolution.

The temporal visual complexity of a video sequence is quantified using dense optical flow. Let $\mathbf{v}_{x,y}(t) \in \mathbb{R}^2$ represent the motion vector at pixel coordinates (x, y). The mean motion magnitude $OF(t)$ is the spatial mean of the L_2 norms of the flow vectors, normalized by the frame diagonal $D = \sqrt{H^2 + W^2}$:

$$OF(t) = \frac{1}{D} \left(\frac{1}{H \cdot W} \sum_{x=1}^{W} \sum_{y=1}^{H} \|\mathbf{v}_{x,y}(t)\|_2 \right). \quad (7)$$

This restricts the relative optical flow sequence to the interval $[0, 1]$.

2.2 Logarithmic Range Compression

To ensure the homogeneity of the parameter scales, MAV-C projects the raw feature vectors into a $[0, 1]$

subspace. Let $x_j(t)$ be the j-th raw feature extracted at time t, and $x_{j,max}$ be its corresponding theoretical maximum.

This is implemented via Log-Theoretical Scaling (LTS), a nonlinear mapping $\mathcal{L}(\cdot)$ defined elementwise:

$$\hat{x}_j(t) = \mathcal{L}(x_j(t)) = \frac{\log_2(1 + \gamma_j x_j(t))}{\log_2(1 + \gamma_j x_{j,max})} \quad (8)$$

where $\gamma_j > 0$ is a curvature parameter. The logarithmic form of $\mathcal{L}$ reflects the Weber-Fechner law of sensory perception. For all auditory and visual features except optical flow, a fixed value $\gamma_j = 10$ is used. This is not a tunable parameter in this study, as subjective complexity ratings are required to meaningfully anchor the perceptual scale, which is deferred to future work. Optical flow requires a separate parameter γ_{flow}. Normalization by the frame diagonal D makes $OF(t)$ resolution-independent, with the trade-off that even fast motion produces values well below 0.1, since the mean per-pixel displacement is always a small fraction of D. Applying $\gamma_j = 10$ to such a signal would utilize only the lowest portion of the $\mathcal{L}$ curve. γ_{flow} compensates by increasing the curvature to spread the active flow range across the full $[0, 1]$ output. The sensitivity of this parameter is quantified in Section 4.1.

2.3 Modality Sub-Scores

After log scaling, feature vectors are aggregated into independent auditory and visual complexity terms, $C_A(t)$ and $C_V(t)$.

Amplitude-Weighted Entropy: Raw entropy features are not sensitive to signal level, whereby a quiet hum and a loud impulsive sound (e.g., an explosion) may yield near-identical EE and SE values. To account for the perceptual salience of loud and complex sounds, both features are scaled by an amplitude weight $w(t) \in [0, 1]$, defined as the per-frame RMS mapped linearly between the ITU-R BS.1770 silence gate (lower bound, $w = 0$) and the EBU R128 true-peak ceiling (upper bound, $w = 1$) [5, 3]. This yields the weighted vector $\tilde{\mathbf{x}}_A(t)$, which guarantees $C_A(t) = 0$ for silent content.

Acoustic complexity C_A: The audio complexity score is the inner product of the LTS-processed and amplitude-weighted feature vector $\tilde{\mathbf{x}}_A(t)$ and a weight vector $\mathbf{w}_A \in \mathbb{R}^K$. $\mathbf{w}_A$ is computed via *sigma-squared weighting* (σ^2W), which assigns each feature a weight proportional to its within-clip variance, then regularizes toward a uniform prior via parameter $\alpha_a \in [0, 1]$:

$$w_j = \frac{\alpha_a}{K} + (1 - \alpha_a) \frac{\sigma_j^2}{\sum_{k=1}^{K} \sigma_k^2}, \quad \mathbf{1}^\top \mathbf{w}_A = 1 \quad (9)$$

where σ_j^2 is the within-clip variance of the j-th LTS-compressed feature. At $\alpha_a = 0$ weights are fully data-driven while at $\alpha_a = 1$ all features contribute equally. The acoustic complexity score is then:

$$C_A(t) = \mathbf{w}_A^\top \mathcal{L}(\tilde{\mathbf{x}}_A(t)). \quad (10)$$

MAV-C operates in two configurations: a stereo mode ($K = 2$, using EE and SE) applicable to any stereo audio track, and a spatial mode ($K = 3$, adding SASE) when directional source localization metadata is available.

For the visual domain, the static spatial complexity $S_V(t)$ is computed as the inner product of the LTS-processed MSG and MUC features using σ^2W weights with visual regularization parameter α_v:

$$S_V(t) = \mathbf{w}_V^\top \left[\mathcal{L}(MSG(t)),\ \mathcal{L}(MUC(t))\right].^\top \quad (11)$$

The final visual complexity score $C_V(t)$ scales $S_V(t)$ by a motion term, where $\rho_v \in [0, 1]$ controls how much spatial complexity is retained when there is no motion:

$$C_V(t) = S_V(t)\Big[\rho_v + (1 - \rho_v)\mathcal{L}(OF(t))\Big]. \quad (12)$$

2.4 Feature Fusion and Cognitive Integration

The modality sub-scores are fused into a joint complexity score $C_{AV_{raw}}(t)$. Let $\mathbf{c}(t) = [C_A(t), C_V(t)]^\top \in [0, 1]^2$ and let $\mathbf{w} = [w_a, w_v]^\top$ be the modality weight vector with $\|\mathbf{w}\|_1 = 1$. The fusion is a weighted Minkowski pooling via the L_p norm:

$$C_{AV_{raw}}(t) = (w_a C_A(t)^p + w_v C_V(t)^p)^{\frac{1}{p}} \quad (13)$$

where $p \geq 1$ is the pooling exponent. As $p \to \infty$, the L_p norm approaches the Chebyshev distance (L_∞), biasing the joint score in favor of the dominant modality. This behavior is motivated by attentional capture, as salient events in one modality are known to dominate attention [12].

Temporal integration is modeled as a causal first-order Exponentially Weighted Moving Average (EWMA) filter applied to $C_{AV_{raw}}$:

$$C_{AV}(t) = \beta_t C_{AV}(t-1) + (1-\beta_t) C_{AV_{raw}}(t). \quad (14)$$

Coefficient $\beta_t = \exp(-\Delta t/\tau)$ depends on the sampling interval Δt and a time constant τ reflecting the integration time of human perception.

The sensitivity of all fusion parameters is characterized via One-Factor-At-a-Time (OFAT) parameter sweeps in Section 4.1.

3 Stimuli

Two stimulus sets were used: a synthetic dataset (**SYN**) for controlled verification of expected metric behavior, and a naturalistic video game gameplay dataset (**GAM**) for real-world evaluation.

SYN: SYN comprises 18 audiovisual stimuli (15 s, 30 fps, 1280×720; 48 kHz audio), processed in spatial mode ($K = 3$, including SASE) with all frames sampled. Audiovisual complexity was varied independently across different axes. Audio spans three dimensions: spectral content (mix ratio of broadband noise to a pure sine tone), temporal density (rectangular on/off gating duty cycle), and spatial distribution (number of virtual sources $N \in [1, 64]$ and their angular velocity $R \in [0°, 90°/\text{s}]$); spatial audio was rendered binaurally via SADIE II KEMAR HRIRs [1]. Video spans texture density (uniform noise block size from 1 to 64 px) and motion speed (0-15 px/frame). The 18 conditions span floor and ceiling anchors, a midpoint baseline, independence and SASE boundary conditions, and single-feature continuous sweeps (see Table 1).

GAM: GAM comprises 60 gameplay clips (5 s each) drawn from SAFEPLAY-X [10], a dataset originally assembled for violence detection. Clips were stratified across three content categories (20 each): *threat* (suspense/stealth), *assault* (active combat), and *lethal* (culminating action). These labels reflect narrative content type and carry no implied perceptual complexity ordering. All clips were processed in stereo mode ($K = 2$, EE and SE only) due to the lack of trajectory metadata.

Table 1: SYN conditions referenced in the results.

Condition	Description
Floor	Silence with a static, low-texture canvas
Ceiling	All axes simultaneously at maximum complexity
Joint	Audio and visual complexity increase together across time
Audio-only	Audio complexity increases, visual content held at low fixed level
Independence	One modality sweeps its full range while the other is held fixed
SASE boundary	Single localized source (floor) vs. fully distributed 64-source field (ceiling); C_V held fixed
Audio	Full sweep of spectral, temporal, and spatial audio axes (EE, SE, SASE)
Texture	Texture density swept from coarse (64 px) to fine (1 px blocks)
Motion	Motion speed swept from static (0) to fast (15 px/frame)

4 Results

4.1 Default Parameters and Sensitivity

MAV-C has one *fixed design constant*, the LTS curvature $\gamma_j = 10$ (shared across all features), and seven *tunable fusion parameters*. Pending subjective calibration, these are initialized to midpoints: $p = 2$, $w_a = 0.5$, $\alpha_a = 0.5$, $\alpha_v = 1.0$, $\rho_v = 0.60$, $\gamma_{flow} = 500$, $\tau = 400$ ms; this combination is referred to as *"default parameters"* throughout. To characterize how each parameter independently affects the output, OFAT sweeps are run across all GAM and SYN items, varying one parameter at a time while holding the rest at default values.

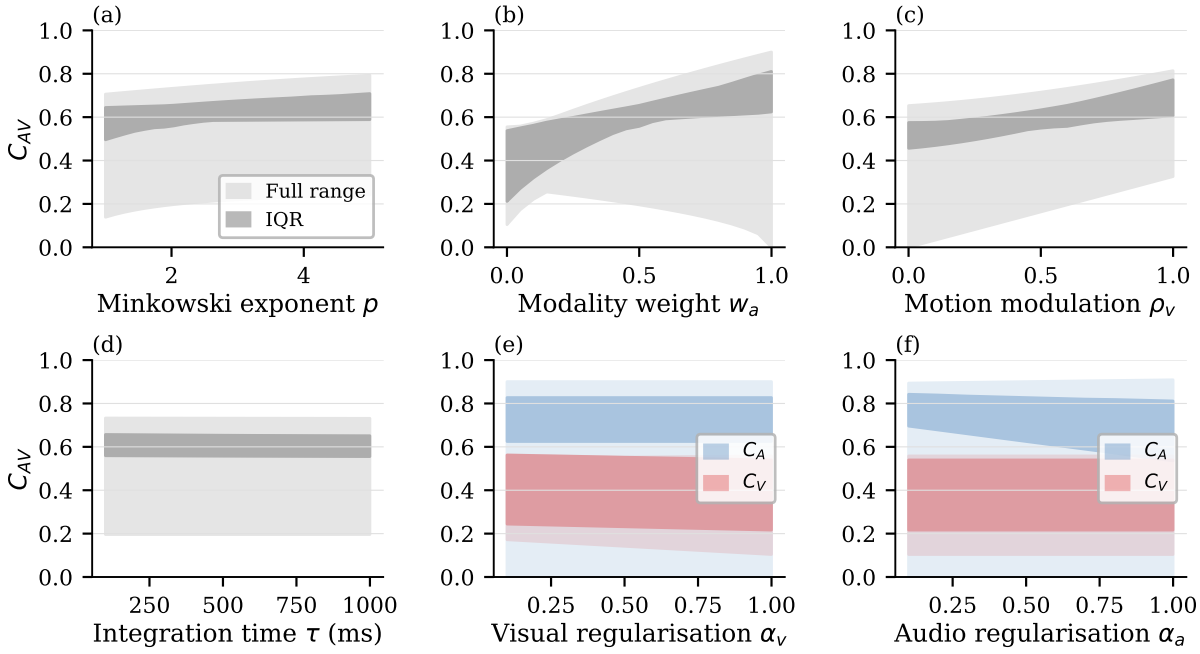


Figure 2: One-Factor-at-a-Time (OFAT) parameter sensitivity on SYN. Each panel varies one parameter while holding all others at their default values. Panels (a)-(d): C_{AV}; panels (e)-(f): C_A (blue) and C_V (red). Light bands: full range across 18 conditions. Dark bands: IQR. (a)-(c) primary fusion; (d) temporal integration; (e)-(f) adaptive weight regularization.

4.2 Synthetic Dataset

Figure 2 shows the OFAT sensitivity across all 18 SYN conditions. Across all parameter combinations, C_{AV} spans $[0.00, 0.90]$. The light band shows the full range and the dark IQR band shows that most conditions cluster in the mid-range, with the anchors driving the extremes. At default parameters the anchor conditions yield a dynamic range of 0.54 (floor: 0.19, ceiling: 0.73). Single-axis sweeps produce strongly monotone metric trajectories: Spearman $\rho = 0.995$ (p<0.001) for texture, $\rho = 0.899$ (p<0.001) for audio, and $\rho = 0.896$ (p<0.001) for motion.

Modality independence is confirmed by the independence conditions (Coefficient of Variation (CoV) < 0.003): when one modality sweeps its full range the other sub-score varies by less than 0.010, confirming each sub-score is driven solely by its own modality. SASE boundary conditions show that C_A rises from 0.624 (single localized source) to 0.786 (fully diffuse 64-source field), while C_V remains constant ($\Delta = 0.000$). Across all 110 parameter combinations, the joint condition reaches a higher C_{AV} ceiling than either unimodal condition alone, confirming that audio and visual complexity contribute independently and additively to C_{AV}.

4.3 Gameplay Dataset

Figure 3 shows the distribution of per-clip mean C_{AV}, C_A, and C_V across all 60 GAM clips and all 110 parameter combinations. Across all parameter combinations, C_{AV} spans $[0.00, 0.94]$. At default parameters the distribution narrows to $[0.35, 0.68]$ (median 0.60), situating naturalistic gameplay content in the upper-middle portion of the metric's range and well within the $[0.19, 0.73]$ interval established by the SYN anchors, confirming that the metric is neither saturated nor floored on naturalistic content.

Figure 4 shows the per-clip mean C_A plotted against C_V. The two sub-scores are statistically independent across the GAM corpus: Pearson $r = 0.03$ (p=0.824; bootstrap 95 % CI $[-0.20, 0.33]$),

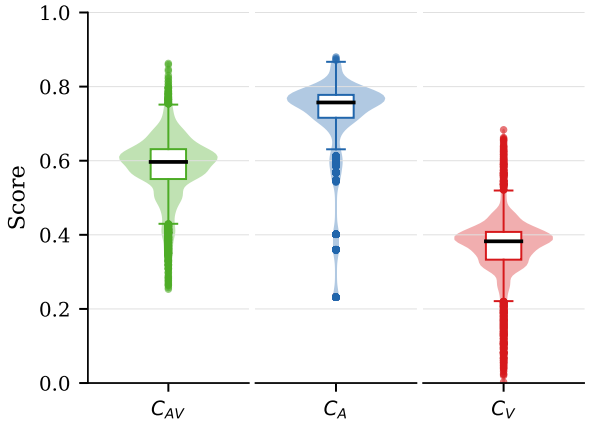


Figure 3: Distribution of per-clip mean C_{AV}, C_A, and C_V across all 60 GAM clips and all parameter combinations (6600 observations: 60 clips × 110 parameter combinations). Violins show full density; boxes show median and IQR; outliers beyond 1.5×IQR are plotted individually.

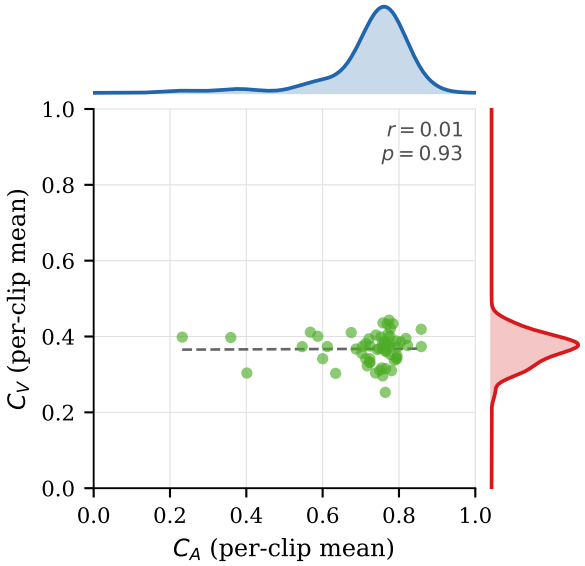


Figure 4: Per-clip mean C_A vs. C_V across 60 GAM clips, averaged over all parameter combinations. Dashed line: linear fit ($r = 0.03$, p=0.824). Marginal KDE curves for C_A (blue, top) and C_V (red, right). The near-zero correlation indicates that C_A and C_V vary independently on this corpus.

Spearman $\rho = 0.13$ (p=0.310). This indicates that C_A and C_V vary independently on this corpus. A Kruskal-Wallis test across the three signal categories (threat, assault, lethal) reveals significant separation in C_V ($H = 7.70$, $p = 0.021$) but not in C_A ($p = 0.437$), consistent with the visual-dominant escalation characteristic of action gameplay: optical flow and texture density increase with content intensity while the audio mix remains broadly constant.

5 Discussion

The SYN corpus confirms that the underlying architecture functions as intended: modality sub-scores are mathematically isolated (CoV < 0.003) and respond monotonically to targeted feature variations (Spearman $\rho \geq 0.88$). The GAM results support the argument for a joint metric: C_V tracks narrative escalation across content categories (KW $p = 0.021$) while C_A does not ($p = 0.437$), and the near-zero C_A/C_V correlation ($r = 0.03$) confirms the two sub-scores capture largely independent signal dimensions. A unimodal measure would therefore be blind to whichever modality is omitted, regardless of that modality's contribution to scene complexity.

The appropriate modality balance varies with content type, task, and observer. The visual dominance effect [2, 11] suggests that visual salience often captures attention preferentially, which may warrant adjusting w_a for visually rich content. MAV-C's parametric structure is designed to support perceptual calibration. The sensitivity analysis (Figure 2) reveals a two-tier hierarchy across both datasets with w_a and ρ_v acting as primary levers (OFAT spans: SYN 0.33 and 0.17; GAM 0.35 and 0.15). p, α_v, and γ_{flow} provide fine-tuning (GAM spans: 0.09, 0.03, and 0.01) and τ has negligible effect on both (<0.01). The exception is α_a, which spans 0.06 on SYN but only 0.01 on GAM, indicating its influence is content-dependent and strongest for stimuli with extreme spectral variation. With C_{AV} spanning $[0.00, 0.94]$ across all parameter combinations on GAM, the metric has the requisite flexibility to be tuned to specific subjective complexity ratings. The equal-weight default ($w_a = 0.5$) is a conservative starting point pending such data.

6 Limitations and Future Work

No subjective validation of MAV-C was performed, as no joint audiovisual complexity dataset (featuring video) with perceptual ratings currently exists. A subjective study to obtain such a dataset and tune metric parameters is planned. This will also enable empirical determination of γ_j, currently fixed at a perceptually motivated value. Corpus generalizability is also a constraint: GAM only features violent videogame scenes, and the parameter hierarchy observed here may thus shift for other content types. On the signal processing side, the spatial audio mode requires directional source metadata not always available in production pipelines. Extending the framework to approximate source trajectories would broaden its applicability. Dense optical flow is computationally intensive and may not capture high-level motion salience, motivating the evaluation of lightweight alternatives.

Acknowledgments

This work was supported by CoSTAR (Convergent Screen Technologies and Performance in Realtime) Live Lab, funded by the Arts and Humanities Research Council, grant reference AH/Y001079/1 and by the Engineering and Physical Sciences Research Council, grant reference EP/W524657/1. This work was conducted in collaboration with Electronic Arts SEED.

References

[1] Cal Armstrong, Lewis Thresh, Damian Murphy, and Gavin Kearney. A perceptual evaluation of individual and non-individual HRTFs: A case study of the SADIE II database. *Applied Sciences*, 8(11), 2018. ISSN 2076-3417. doi:10.3390/app8112029. URL `https://www.mdpi.com/2076-3417/8/11/2029`.

[2] Francis B. Colavita. Human sensory dominance. *Perception & Psychophysics*, 16(2):409–412, 1974. doi:10.3758/BF03203962.

[3] European Broadcasting Union. Loudness normalisation and permitted maximum level of audio signals. EBU Recommendation R 128, European Broadcasting Union, 2020. URL `https://tech.ebu.ch/publications/r128`.

[4] Stefan Fichna, Thomas Biberger, Bernhard U. Seeber, and Stephan D. Ewert. Effect of acoustic scene complexity and visual scene representation on auditory perception in virtual audio-visual environments. In *2021 Immersive and 3D Audio: from Architecture to Automotive (I3DA)*, pages 1–9. IEEE, 2021. doi:10.1109/I3DA48870.2021.9610916.

[5] International Telecommunication Union. Algorithms to measure audio programme loudness and true-peak audio level. ITU-R Recommendation BS.1770-4, International Telecommunication Union, 2015. URL `https://www.itu.int/rec/R-REC-BS.1770`.

[6] Erik Jorgen Jorgensen. *Acoustic complexity in real-world noise and effects on speech perception for listeners with normal hearing and hearing loss*. PhD thesis, The University of Iowa, 2022.

[7] Cameron Kyle-Davidson, Elizabeth Yue Zhou, Dirk B. Walther, Adrian G. Bors, and Karla K. Evans. Characterising and dissecting human perception of scene complexity. *Cognition*, 231, February 2023. doi:10.1016/j.cognition.2022.105319. URL `https://linkinghub.elsevier.com/retrieve/pii/S0010027722003080`.

[8] Luca Resti, Amelia Gully, Michael McLoughlin, Gavin Kearney, and Alena Denisova. Spherical acoustic spatial entropy: Predicting acoustic scene complexity in virtual environments. *IEEE Open Journal of Signal Processing*, 7:144–153, 2026. doi:10.1109/OJSP.2026.3657297.

[9] Karahan Saritas, Peter Dayan, Tingke Shen, and Surabhi S. Nath. Complexity in complexity: Understanding visual complexity through structure, color, and surprise. In *Annual Meeting of the Cognitive Science Society*, 2025. URL `https://escholarship.org/uc/item/2g6408z2`.

[10] Aakash Singh, Vinayak Bansal, Muskan Saini, Deepawali Sharma, and Vivek Kumar Singh. SAFEPLAY-X: A comprehensive gameplay

video dataset for violence detection with explainable deep learning applications. *Expert Systems with Applications*, 316:131724, 2026. doi:10.1016/j.eswa.2026.131724.

[11] Scott Sinnett, Charles Spence, and Salvador Soto-Faraco. Visual dominance and attention: The Colavita effect revisited. *Perception & Psychophysics*, 69(5):673–686, 2007. doi:10.3758/BF03193770.

[12] Durk Talsma, Daniel Senkowski, Salvador Soto-Faraco, and Marty G. Woldorff. The multifaceted interplay between attention and multisensory integration. *Trends in Cognitive Sciences*, 14(9):400–410, 2010. doi:10.1016/j.tics.2010.06.008.

[13] Adam Weisser, Jörg M Buchholz, and Gitte Keidser. Complex acoustic environments: Review, framework, and subjective model. *Trends in Hearing*, 23, 2019. doi:10.1177/2331216519881346.