SEND: A SPATIAL EVENT NEURAL DETECTOR FOR INTENTIONAL OBJECT MOTION IN IMMERSIVE MUSIC MIXING

Xu Gan Linhao Zhao Zhenhai Yan

Tencent Music Entertainment
Shenzhen, China
{limegan, linozhao, zhenhaiyan}@tencent.com

ABSTRACT

Deciding exactly when to move audio objects in immersive mixes is a labor-intensive artistic task. Current tools react strictly to instantaneous frequency overlaps, lacking the macroscopic awareness required for musically intentional spatial transitions. To model these decisions, we propose SEND (Spatial Event Neural Detector). Its dual-stream architecture analyzes the target track against its background context, combining a Spec-TNT backbone and a Temporal Convolutional Network (TCN) to capture hierarchical spectral features and precise rhythmic cues. Their dynamic interplay is modeled via a novel Cross-Track Gating Interaction (CTGI) mechanism. SEND significantly outperforms reactive baselines (Frame-F1: 0.7675, Event-IoU: 0.6305). Subjective tests and our Structural Boundary Alignment (SBA) metric confirm it effectively mimics human expert intentionality, ensuring superior spatial stability and musical appropriateness. Audio samples are available at: https://9ime.github.io/projects/send/

1. INTRODUCTION

Immersive audio has fundamentally shifted the mixing paradigm from channel-based routing to object-based spatial orchestration. In the context of this work, "immersive audio" specifically refers to object-based formats that combine static channel-based beds with dynamic, 3D-positioned audio objects. In this landscape, dynamic spatialization is not a random effect but a deliberate artistic decision used to enhance musical narrative. Professional mixing is characterized by intentional sparsity: spatial objects typically remain stable, transitioning only during specific musical cues to maintain image clarity. Despite the rapid advancement of intelligent mixing, current research primarily focuses on static positioning [1], upmixing [2], or reactive rule-based masking relief [3] [4]. While effective for corrective tasks, these heuristic approaches lack the macroscopic musicality and intentionality inherent in human mixing. Crucially, the automated generation of professional-grade spatial metadata—specifically, the decision-making process of when an object should move—remains largely unexplored.

We argue that the essence of professional spatialization lies in the articulation of intentional boundaries; therefore, we formally define Musical Spatial Event Detection (MSED). By modeling spatialization as a sequence of musically-justified spatial moving events, we focus on localizing the precise temporal boundaries where motion serves the artistic narrative.

To address this task, we propose SEND (Spatial Event Neural Detector), a dual-stream architecture that mimics an engineer's holistic listening by decoupling audio into a target stream (the dynamic object) and a background context stream (the static accompaniment mix). At its core, a Spectral-Temporal Transformer in Transformer (Spec-TNT) backbone and parallel Temporal Convolutional Network (TCN) blocks extract high-resolution spectral features and long-range temporal dependencies. A novel Cross-Track Gating Interaction (CTGI) mechanism explicitly intertwines these streams to learn complex spectral coordination and musical interplay. The primary contributions of this paper are summarized as follows:

1. We formalize the task of MSED for object-based immersive mixing, providing a neural framework to localize discrete spatial moving events within professional audio content.
2. We propose SEND, a dual-stream architecture that models the interplay between target objects and their musical context via a CTGI mechanism.
3. We conduct extensive objective and subjective evaluations across professional immersive content. The results validate SEND's ability to identify spatial event boundaries that align with the underlying musical context in object-based mixing.

The paper is organized as follows: Section 2 reviews related work and identifies gaps in modeling intentional motion. Section 3 details the dual-stream SEND architecture and the CTGI mechanism. Section 4 outlines the experimental setup and proposed domain-specific metrics. Section 5 analyzes the objective and subjective evaluation results, and Section 6 concludes the paper.

2. RELATED WORK

2.1. Rule-Based Spatial Automation

In Intelligent Music Production, spatial parameters are primarily treated as functional tools for technical clarity and unmasking. Foundational frameworks by Hafezi and Reiss [5] established autonomous optimization for minimizing spectral overlap, a cross-adaptive philosophy extended to the spatial domain by Pestana and Reiss [4] via frequency-band splitting. Feature-based automation, such as mapping spectral centroids to the stereo field [6], further ensures frequency-complementary positioning. However, these rule-based schemes rely on rigid, instantaneous mapping curves, lacking the capacity to model the long-term musical context and artistic intent essential for professional production.

2.2. Deep Learning for Immersive Audio Research

Deep learning has profoundly revolutionized immersive audio via two main paradigms: physical reconstruction and parameter es-

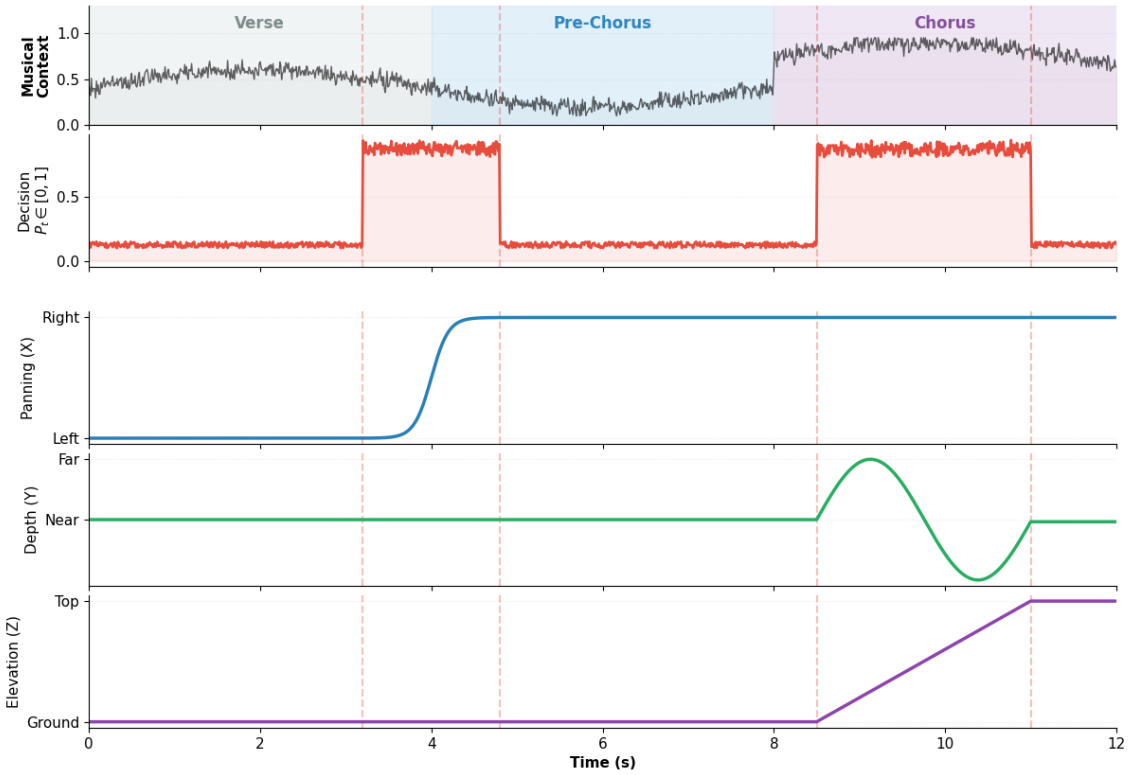

Figure 1: Conceptual framework of probability-driven spatial movement events. By analyzing the inherent musical context, the model infers a frame-wise probability P_t. Multidimensional trajectories (X, Y, Z) undergo synchronized transitions exclusively within the detected event windows (red dashed intervals). This gating ensures spatial modulations are both musically intentional and perceptually stable.

timation. Tasks such as source localization [7], self-supervised Ambisonics generation from mono [8], and spatial mixing [1] [9] focus on reconstructing physical reality or mapping formats rather than creative expression. Recently, research has shifted to learning mixing conventions via differentiable consoles [10] and style transfer [11]. However, these systems are largely designed for static placement, global spectral balancing, or vision-guided trajectories. In these contexts, spatialization is treated as a reactive response to acoustic or visual cues. Consequently, there remains a research gap in capturing autonomous, musically-intentional motion—where spatial transitions are driven not by physical necessity, but by the underlying narrative and structural evolution of the music itself.

2.3. Music-Driven Motion Synthesis and Frame-Level Event Prediction

While underexplored in IMP, mapping audio features to coherent physical motion is well-established in cross-modal fields like human skeletal dynamics [12] and image animation [13]. These systems prove that the temporal evolution of music contains sufficient cues to drive coherent, discrete physical actions. To capture the discrete nature of professional mixing, we frame spatial automation as a frame-level binary classification task, analogous to VAD [14] and Music Onset Detection [15].

3. METHODOLOGY

3.1. Formalization of Musical Spatial Event Detection (MSED)

As shown in Figure 1, we conceptualize spatial trajectories as discrete moving events rather than continuous drifts. We formally define Musical Spatial Event Detection (MSED) as the task of localizing the precise temporal boundaries of musically-intentional spatial transitions. Given a dynamic target object and its static background context, MSED predicts a sequence of frame-wise binary triggers dictating exactly when the object should move to serve the artistic narrative.

3.2. Model Architecture

The proposed SEND is an end-to-end dual-stream architecture designed to predict spatial motion triggers by analyzing the interplay between a target object and its musical context. As illustrated in Figure 2, the pipeline consists of four integrated stages:

1. **Dual-Stream Encoding:** Target and background audio are transformed into Mel-spectrograms and processed by parallel ResNet-based Encoders to extract hierarchical time-frequency features.
2. **Cross-Track Interaction (ST-CTGI-Unit):** Encoded features enter the ST-CTGI-Unit, where a Spec-TNT backbone performs spectral-temporal decoupling. The CTGI mech-

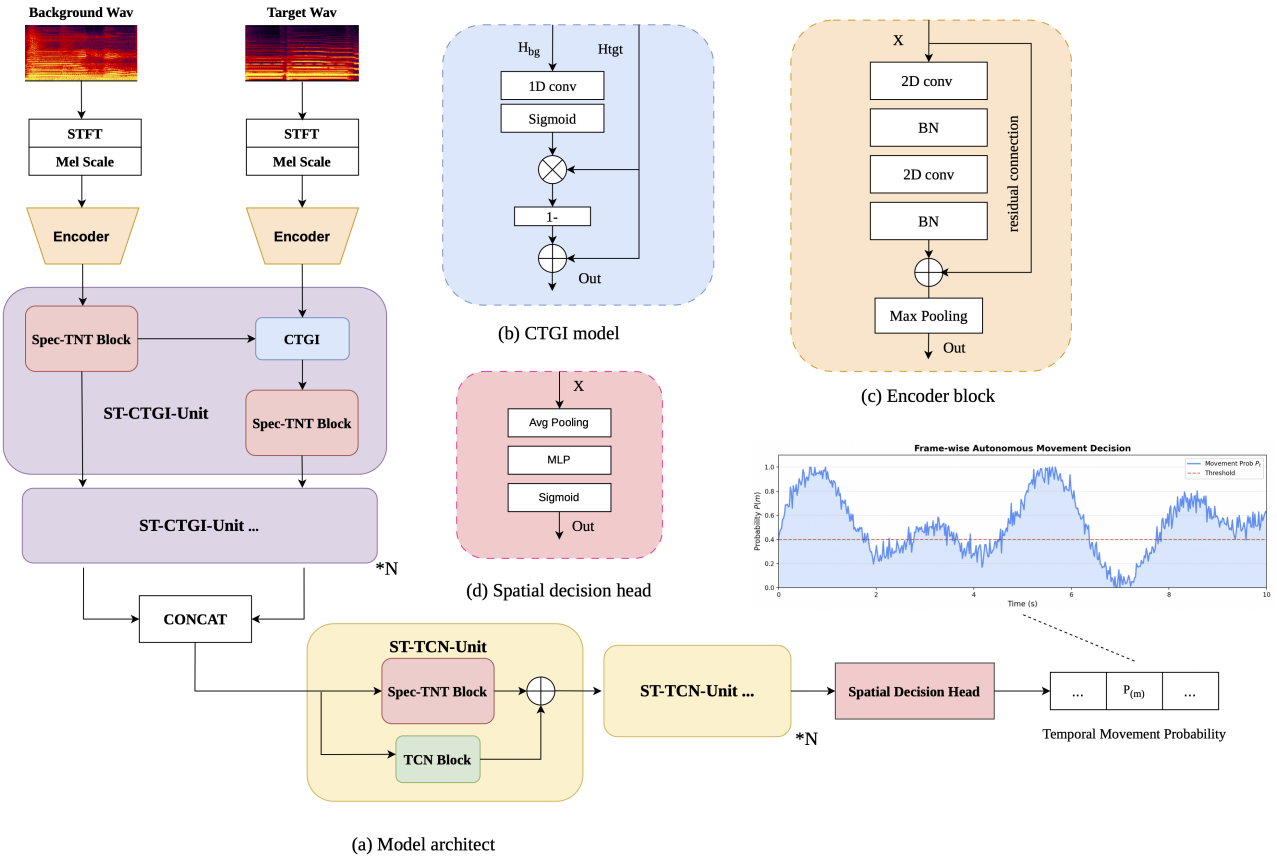

Figure 2: Schematic of the proposed SEND architecture. (a) Global Architecture: The dual-stream pipeline processes Target and Background Mel-spectrograms through ResNet-based Encoders. Feature interaction is facilitated by the ST-CTGI-Unit, followed by temporal refinement in the ST-TCN-Unit (a parallel configuration of Spec-TNT and TCN blocks) and a Spatial Decision Head to output frame-wise movement probabilities. (b) Detail of the CTGI. (c) Detail of ResNet-based Encoder. (d) Detail of Spatial Decision Head

anism intertwines and gates information flow to capture spectral collisions and rhythmic cues.

3. **Spatio-Temporal Refinement (ST-TCN-Unit):** The fused representations are processed by parallel Spec-TNT and TCN blocks, balancing global structural awareness with high-resolution temporal precision.
4. **Spatial Decision:** The Spatial Decision Head decodes refined features into a frame-wise probability sequence, identifying discrete spatial moving events for spatial transitions.

3.3. Spec-TNT and Hybrid Temporal Modeling

3.3.1. Spectral-Temporal Factorization

To capture both fine-grained harmonic content and macroscopic musical structures, we employ the Spec-TNT architecture [16] as a backbone for hierarchical representation learning. Proven effective in MIR tasks like beat tracking [17] and audio understanding [18], this module factorizes spectral-temporal dependencies into two interleaved stages: a spectral encoder that aggregates frequency-wise cues into latent tokens, and a temporal encoder that models long-range transitions across the time axis.

In our framework, Spec-TNT is utilized in two functional capacities. First, in the interaction stage, it independently extracts deep spectral-temporal features from the dual streams to facilitate cross-track gating. Second, in the refinement stage, it processes the integrated representations in parallel with a TCN branch, balancing global structural awareness with localized temporal precision. This decoupled design ensures a global receptive field while maintaining computational efficiency for multi-track audio analysis.

3.3.2. Auxiliary Temporal Refinement via TCN

To complement Spec-TNT's global dependencies, a lightweight TCN is integrated as a parallel residual branch. This TCN refiner utilizes dilated convolutions to capture local energy fluxes and onset nuances. By fusing these with Spec-TNT's semantic representations, the hybrid design ensures both macroscopic structural understanding and frame-level rhythmic sensitivity, stabilizing the predicted motion trajectories.

3.4. Cross-Track Gated Interaction

To capture dynamic dependencies between dual streams, we propose the CTGI mechanism. Unlike speech enhancement tasks [19]

that align corrupted signals with references, CTGI treats the background stream $H_{bg}^{(l)} \in \mathbb{C}^{C\times T\times F}$ as a contextual constraint. We derive a masking-aware gating vector $g^{(l)} \in [0,1]^{C\times T\times F}$ via a 1×1 convolution $\Phi^{(l)}$ and Sigmoid activation:

$$\mathbf{g}^{(l)} = \sigma\left(\Phi^{(l)}(\mathbf{H}_{bg}^{(l)})\right) \quad \text{where } \mathbf{g}^{(l)} \in [0,1]^{C\times T\times F} \tag{1}$$

The gate $\mathbf{g}^{(l)}$ identifies critical latent states where the background environment imposes structural constraints on the target track's spatial presence. Rather than merely reacting to spectral overlaps, this gate functions as a contextual priority map.

Inspired by the Gated Linear Units (GLU) for controlled information flow [20] and the masking-based feature extraction success in speech separation [21], we apply a complementary gating logic to the target features $\mathbf{H}_{tgt}^{(l)}$. The modulation is combined with a residual connection to ensure training stability and preserve the target's core harmonic integrity:

$$\mathbf{H}_{tgt}^{in(l)} = \mathbf{H}_{tgt}^{(l)} \odot (1-\mathbf{g}^{(l)}) + \mathbf{H}_{tgt}^{(l)} \tag{2}$$

$$\mathbf{H}_{tgt}^{(l+1)} = \text{Spec-TNT-Block}^{l}\left(\mathbf{H}_{tgt}^{in(l)}\right) \tag{3}$$

The $(1-g^{(l)})$ term implements a subtractive gain reduction, where the background stream "carves out" latent space for the target. Interleaving this gated interaction across layers ensures that predicted spatial events are both rhythmically precise and perceptually consistent with the evolving spectral context.

3.5. Data Representation and Training Objective

We compute Log-Mel Spectrograms using an STFT with a Hann window ($N = 2048$, $H = 512$) and $F = 128$ Mel bands. This yields a dual-stream input $X \in \mathbb{R}^{2\times T\times F}$ for the target and background tracks.

The model outputs frame-wise probabilities $\hat{y} \in [0,1]^T$. Binary triggers b_t are generated via a decision threshold τ^*:

$$b_t = \begin{cases} 1, & \text{if } \hat{y}_t \geq \tau^* \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

To ensure an unbiased evaluation, we determine the optimal decision threshold τ^* by maximizing the Frame-F1 score through an exhaustive grid search. To prevent data leakage and ensure generalizability, this threshold is optimized strictly on the validation set. The frozen τ^* is then applied uniformly to the test set to convert continuous probabilities into discrete binary triggers [22] [23].

We employ Weighted Binary Cross-Entropy (WBCE) loss to mitigate class imbalance between sparse "move" frames and "stay" frames:

$$\mathcal{L} = -\frac{1}{T}\sum_{t=1}^{T}[w\cdot y_t \log(\hat{y}_t) + (1-y_t)\log(1-\hat{y}_t)] \tag{5}$$

where w is a positive weight hyperparameter tuned to penalize false negatives. The proposed model is implemented using PyTorch and trained on a single NVIDIA L20 GPU. We use the AdamW optimizer with an initial learning rate of 10^{-4} and a batch size of 32. Training typically converges within 20 hours.

4. EXPERIMENTS

4.1. Dataset Curation and Preprocessing

To evaluate SEND, we curated a dataset from 1,000 professional multi-track object-based immersive audio projects encompassing diverse musical genres. These raw projects encapsulate a 7.1.2 channel-based bed governed by static spatial definitions, alongside independent audio objects capable of carrying dynamic, time-varying spatial metadata.

To prevent acoustic interference and spatial data leakage, each dynamic track is isolated as the target object, while its corresponding background context is rendered binaurally from the static 7.1.2 bed and all strictly static objects, with all other dynamic objects explicitly muted. This ensures the model learns to trigger motion based solely on the target's dedicated, unaltered background. By iterating through all dynamic objects within each project as potential targets, we generate a comprehensive dataset totaling 260 hours.

To convert continuous spatial trajectories into discrete labels ($y_t \in \{0,1\}$), we apply a frame-to-frame spatial deviation rule to the target's normalized 3D coordinates ($x, y \in [-1,1]$ and $z \in [0,1]$). A spatial moving event ($y_t = 1$) is triggered if the Euclidean distance between consecutive frames t and $t-1$ exceeds a sensitivity threshold ϵ. Under this strict criterion, active movement frames account for approximately 30% of the dataset.

All audio is processed at 48 kHz/24-bit and normalized to -20 LUFS. To strictly prevent data leakage, the dataset is partitioned by project into 930 training, 50 validation, and 20 testing sets. We use 20-second clips for training, with a 10-second overlapping sliding window during evaluation to ensure temporal continuity.

4.2. Baselines and Model Variants

4.2.1. Baselines

To rigorously evaluate the proposed framework, we establish two distinct baseline systems for comparison:

DSP Baseline [4]: This system triggers panning based on instantaneous spectral masking using a greedy algorithm. To bridge the dimensional mismatch between its continuous bin-wise output and our track-level binary triggers, we apply an Energy-Normalized Relative Shift $\Delta\tilde{R}_j(n)$:

$$\Delta\tilde{R}_j(n) = \frac{\sum_{k=0}^{N-1}\left|D_{j,bg}(n,k) - D_{j,bg}(n-1,k)\right|\cdot\left|X_j(n,k)\right|}{\sum_{k=0}^{N-1}\left|X_j(n,k)\right| + \epsilon} \tag{6}$$

where $D_{j,bg}(n,k) = |p_j(n,k) - p_{bg}(n,k)|$ represents the absolute distance between the target and background context at frame n and frequency bin k, $X_j(n,k)$ is the complex spectrum, and ϵ is a small constant to prevent division by zero during silent frames.

To ensure a fair comparison, we binarize the baseline's continuous output $\Delta\tilde{R}_j(n)$ using Sparsity Matching, selecting the top $x\%$ peak values (where $x\%$ is the percentage of active movement frames triggered by our proposed model). This ensures both systems operate under a unified movement budget. Crucially, rather than penalizing the DSP baseline, this Sparsity Matching strategy theoretically favors it by isolating only its most confident, highest-energy unmasking decisions. This ensures that any subsequent performance degradation of the baseline is inherently due to its lack of macroscopic musical context, rather than an unfair thresholding process.

CRNN Baseline: A classic CRNN (Convolutional Recurrent Neural Network [24] + Bidirectional LSTM [25]) paradigm, which has been widely adopted in audio event detection and music information retrieval tasks. The CRNN baseline shares the same input/output interface and comparable parameter count as our proposed model, ensuring a fair comparison.

4.2.2. Ablation Configurations

To validate the necessity of each architectural component, we further evaluate three internal variants:

- **Single-Stream (S-S):** Removes the background branch to test the impact of acoustic context.
- **Fusion-Add / Fusion-Concat:** Replace the CTGI mechanism with either element-wise addition (Fusion-Add) or channel-wise concatenation (Fusion-Concat) before the ST-TCN-Unit. These variants bypass the dynamic cross-track gating logic to evaluate the specific necessity of interactive feature exchange in capturing musical interplay.
- **w/o TCN:** Excludes the TCN module to investigate the model's reliance on local rhythmic anchoring.

4.3. Evaluation Framework

To evaluate the model's ability to capture both instantaneous and macroscopic spatial intents, we employ the following metrics:

4.3.1. Statistical Metrics

- **Frame-F1 & MCC:** Point-wise classification accuracy, and Matthews Correlation Coefficient (MCC) specifically accounts for the inherent class imbalance between motion and static states.
- **AUC-ROC:** Evaluates the model's discriminative power across all decision thresholds, representing its overall sensitivity to musical triggers.
- **Event-IoU:** To assess event-level integrity, we employ a matching-based mean IoU (mIoU). Each ground-truth window W_g is paired with its best-matching prediction W_p via temporal overlap. The score is defined as:

$$\text{mIoU} = \frac{\sum \text{IoU}(W_p, W_g)}{N_g + N_{fp}} \tag{7}$$

  where N_g is the total ground-truth count and N_{fp} is the number of unpaired (false) predictions. This formulation strictly penalizes both missed events and redundant triggers by assigning them an IoU of 0.
- **Beat Alignment Score:** To quantify the rhythmic synchronization of the predicted triggers, we propose the Beat Alignment Score (BAS). Unlike traditional discrete metrics, our BAS is calculated directly on the continuous probability map P_t to evaluate its alignment with the musical grid. We define two variants: $\textbf{BAS}_{\textbf{all}}$ (alignment with all beats) and $\textbf{BAS}_{\textbf{down}}$ (alignment specifically with downbeats). The score is formulated as the expectation of a Gaussian-decayed alignment over time:

$$\text{BAS} = \frac{\sum_t P_t \cdot \exp\left(-\frac{|t - t_{ref}|^2}{2\sigma^2}\right)}{\sum_t P_t} \tag{8}$$

  where t_{ref} represents the timestamp of the nearest (down) beat, and $\sigma = 50$ms is the tolerance constant. This probabilistic formulation ensures that the metric rewards the model for concentrating high-probability density near rhythmic anchors while penalizing "off-beat" spatial activations. Intuitively, a higher BAS physically and musically indicates that the model's predicted spatial moving events are more tightly 'locked' onto the strong beats or downbeats, reflecting a highly rhythmic and musically-aware spatialization strategy.

4.3.2. Domain-Specific Objective Metric

Through empirical observation of our dataset, we identified a consistent hallmark of human expert mixes: spatial transitions are deliberately anchored to structural musical shifts. To bridge the evaluation gap and quantify this human-like artistic intent, we introduce a novel domain-specific metric.

Structural Boundary Alignment: In professional music production, macroscopic structural transitions (e.g., shifting from a verse to a chorus) serve as primary artistic motivations for dynamic spatial panning. To evaluate whether the model's predictions align with this macro-level musicality—effectively measuring how closely the model mimics a professional engineer's structural awareness—we propose the Structural Boundary Alignment (SBA) metric.

We capture timbral and harmonic evolution by concatenating frame-wise MFCCs and Chromagrams into a feature vector $\mathbf{F}(n)$, constructing a Self-Similarity Matrix:

$$\text{SSM}(i, j) = \text{dist}(\mathbf{F}(i), \mathbf{F}(j)) \tag{9}$$

A structural boundary curve, $S_{str}(n)$, is derived by convolving a Gaussian checkerboard kernel along the SSM diagonal. Peak-picking is then applied to $S_{str}(n)$ to extract precise structural timestamps.

Crucially, professional immersive mixing adheres to *intentional sparsity*—objects need not move at every transition, but any triggered motion should heavily align with a structural boundary. To mathematically reflect this artistic prior without penalizing stable spatial states, we transform binary predictions into continuous movement intervals. SBA is then defined as the proportion of these predicted spatial moving events that successfully encompass a structural peak within a perceptual tolerance window of ± 50 ms. In essence, SBA functions as a boundary-aware precision metric, explicitly evaluating the structural intentionality behind every triggered motion.

4.3.3. Subjective Evaluation Setup

To perceptually validate the efficacy of our interval-based approach, we conducted a double-blind, multiple-stimulus listening test with 20 experienced listeners, including audio engineers and musicians. All trials were conducted using high-quality circumaural headphones in a controlled environment at a standardized 85 dB SPL. We selected five representative multi-track excerpts, each evaluated under five randomized conditions:

- **Hidden Reference (Human Expert Mix):** Custom mixes crafted by professional mixing engineers. This condition serves as the absolute acoustic and musical ground truth, representing unrestricted human artistic intent and optimal

Table 1: Comprehensive evaluation metrics for spatial motion prediction. Best results are highlighted in bold. (BAS denotes Beat Alignment Score).

Model / Configuration	Frame-F1	AUC-ROC	MCC	Event-IoU	BAS_{all}	BAS_{down}
DSP Baseline	0.4496	0.4988	-0.0192	0.0495	8.01%	1.28%
CRNN Baseline	0.5994	0.7572	0.2819	0.4519	10.43%	4.76%
S-S	0.1605	0.5330	0.0389	0.0126	6.85%	0.00%
Fusion-Add	0.5659	0.7032	0.4778	0.2706	11.36%	2.27%
Fusion-Concat	0.6645	0.8014	0.5710	0.4532	11.57%	1.85%
w/o TCN	0.7244	0.8764	0.6487	0.6027	8.33%	0.00%
Proposed (SEND)	**0.7675**	**0.8777**	**0.7013**	**0.6305**	**17.84%**	**4.85%**

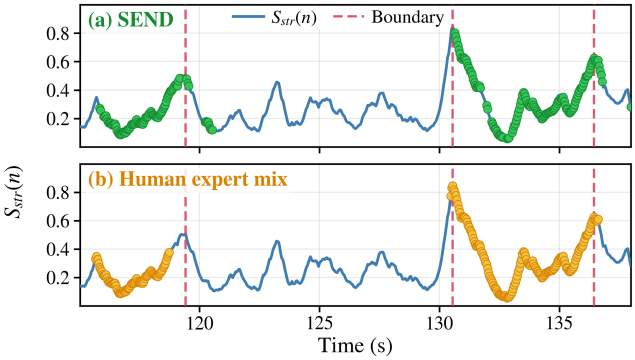

Figure 3: Structural boundary curve $S_{str}(n)$ with the predicted frame-wise trigger sequence from SEND (top, green) and the trigger sequence annotated by a human expert mix (bottom, orange). Red dashed lines indicate detected structural boundary peaks. The highlighted segment shows that both SEND and the human expert tend to initiate spatial movement near structural boundaries, with SEND exhibiting a denser activation pattern.

spatial unmasking decisions without any algorithmic constraints.

- **Hidden Anchor (Static Mix):** An unprocessed mix of the background music and target track, fixed at the center with zero spatial movement.
- **Proposed (SEND):** Spatial trajectories were authored by professional mixing engineers, with the movement onset and duration strictly constrained to the intervals predicted by our proposed SEND. Outside these intervals, the signal remained strictly centered. This condition evaluates whether the model-identified triggers align with professional artistic timing.
- **CRNN Baseline:** Prepared identically to the Proposed condition, but utilizing active movement intervals predicted by the CRNN baseline.
- **DSP Baseline:** The continuous, time-frequency bin-based greedy algorithm directly rendered the spatial output, utilizing the optimal parameters reported in the original paper. No human trajectory editing was applied.

Subjects rated the randomized stimuli on a scale of 1 to 5 across three distinct dimensions. To ensure consistent evaluation, the following criteria were explicitly defined for the participants: 1) **Spatial Stability**: Evaluates whether the target instrument's spatial position is natural, penalizing any distracting, dizzying, or unnatural spatial jitter. 2) **Musical Appropriateness**: Assesses whether the spatial movement aligns with the musical structure, emotional arc, and phrasing, resembling a deliberate artistic decision. 3) **Clarity**: Measures the distinctness between the target instrument and the background music, evaluating whether the spatial distribution effectively reduces auditory clutter and spectral collision.

5. RESULTS AND ANALYSIS

5.1. Statistical Performance and Ablation Analysis

Overall Performance: As shown in Table 1, the proposed SEND architecture outperforms all baselines across all evaluated metrics. Notably, SEND achieves a Frame-F1 of 0.7675 and an Event-IoU of 0.6305, representing a significant improvement over the CRNN baseline (0.4519 IoU) and the DSP-based rule mapping (0.0495 IoU). The lower MCC and IoU scores of the baselines highlight their susceptibility to temporal fragmentation and false triggers, whereas SEND maintains higher structural integrity.

Effect of Dual-Stream Context: The ablation study confirms the necessity of the dual-stream design. The S-S model, which lacks background context, fails to localize spatial events effectively, resulting in a near-zero Event-IoU (0.0126) and a low MCC (0.0389). This confirms that spatial intent in professional mixing is intrinsically linked to the relationship between the target object and the overall musical environment.

Integration Strategy: Regarding feature fusion, both Fusion-Concat (0.4532 IoU) and Fusion-Add (0.2706 IoU) underperform relative to the proposed CTGI mechanism. Specifically, the sharp decline in Fusion-Add performance suggests that simple element-wise aggregation is insufficient for modeling the complex spectral dependencies required for MSED.

Temporal Modeling: The removal of the TCN block (w/o TCN) leads to a complete collapse in BAS_{down} performance (0.00%). This result demonstrates that while spectral features can identify localized motion cues, long-term temporal modeling is mandatory for synchronizing spatial events with macroscopic musical structures such as downbeats.

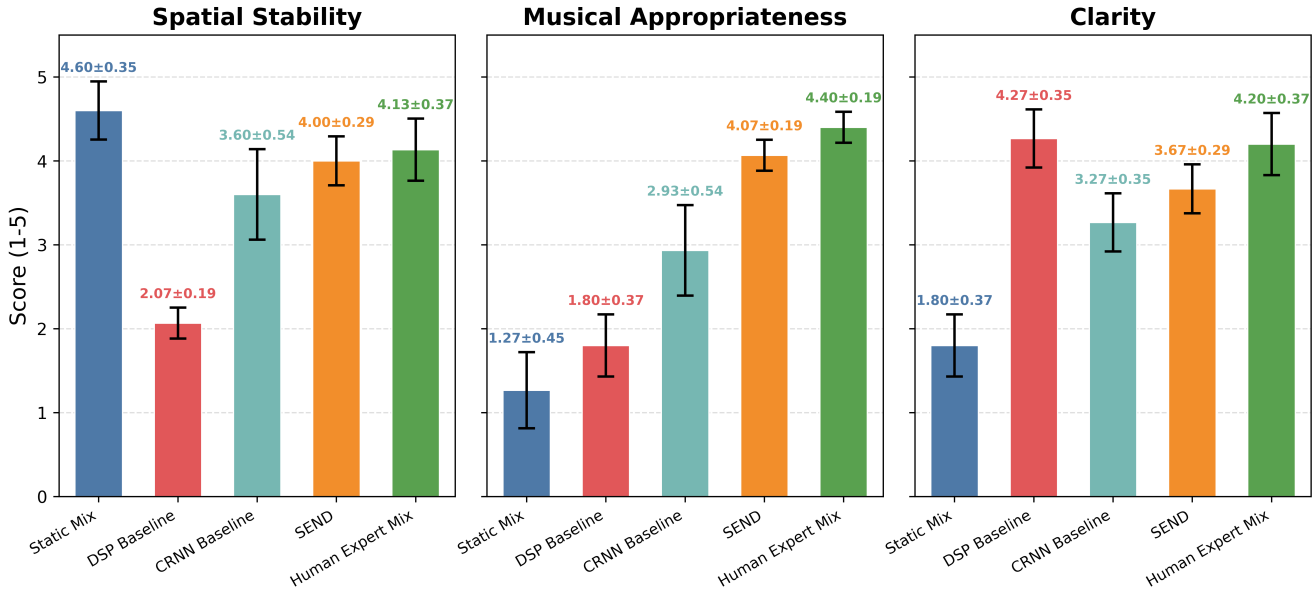

Figure 4: Results of the multiple-stimulus subjective evaluation across three perceptual dimensions: Spatial Stability, Musical Appropriateness, and Clarity. The bar plots display the mean scores evaluated on a 5-point scale for the Static Mix (anchor), DSP Baseline, CRNN Baseline, our proposed SEND model, and the Human Expert Mix (reference). Error bars denote the 95% confidence intervals. Numerical values indicate the mean ± confidence interval. As shown, the proposed SEND model significantly outperforms both baselines in preserving spatial stability and musicality, achieving performance comparable to the human expert upper bound while maintaining effective masking release.

5.2. Perceptual and Musical Analysis

5.2.1. Objective Domain-Specific Results

The SBA metric highlights the limitations of reactive signal processing and conventional learning paradigms. The DSP and CRNN baselines scored a mere 0.14 and 0.31, respectively, as they remain structurally "blind" to macroscopic musical chapters. In contrast, our proposed SEND achieved an SBA of 0.67—a substantial improvement over both baselines that closely approaches the human expert mix (0.72). This validates that SEND comprehends the overarching musical narrative, explicitly deploying spatial panning as an arrangement tool at pivotal structural boundaries.

To visually illustrate this structural awareness, Figure 3 compares the detection patterns of SEND and a human expert alongside the structural boundary curve $S_{str}(n)$. Both the human expert and the proposed model exhibit a strong tendency to initiate spatial transitions in close proximity to the structural peaks (indicated by red dashed lines). However, a closer inspection of the temporal trajectory also reveals minor limitations in SEND's current frame-level stability. For instance, at approximately 121s and 132s, the model produces brief, isolated spatial triggers ("jitters") that do not perfectly align with macroscopic structural shifts. While our temporal refinement modules mitigate most of these anomalies, such transient activations can occasionally lead to abrupt spatial jumps, highlighting a specific area for future optimization.

5.2.2. Listening Test Results

The subjective evaluation in Figure 4 confirms a statistically significant preference for our approach across all dimensions ($p < 0.01$).

Spatial Stability: The DSP Baseline scored the lowest (2.07), as instantaneous unmasking introduces distracting high-frequency drift. Our Proposed model (4.00) significantly outperformed both baselines, with a 95% confidence interval overlapping the Human Expert Mix (4.13), confirming the efficacy of our interval-based execution.

Musical Appropriateness: Our model (4.07) performed comparably to the human benchmark (4.40), decisively outperforming the DSP (1.80) and CRNN (2.93) baselines. This suggests the Fusion architecture effectively learns to trigger movements at musically motivated intervals rather than relying on mechanical signal changes.

The Unmasking-Musicality Trade-off: While the DSP Baseline achieved the highest theoretical masking release (4.27), it did so at the complete expense of stability and aesthetics. Our model strikes an optimal balance, providing effective masking release (3.67) while preserving the spatial comfort expected in professional productions.

6. CONCLUSION AND FUTURE WORK

This paper introduced SEND, a dual-stream framework for MSED in immersive mixing. Evaluated on 1,000 professional immersive audio projects, SEND significantly outperforms baselines (Frame-F1: 0.7675, AUC-ROC: 0.8777), with ablation studies confirming the critical roles of its ST-CTGI and ST-TCN units for cross-track interaction and rhythmic synchronization.

Crucially, SEND captures macroscopic musical intentionality, achieving a SBA of 0.67 to seamlessly synchronize spatial transitions with overarching narratives. Subjective tests further

validated that SEND performs comparably to human experts, optimally balancing masking release, spatial stability, and musical appropriateness as a creative arrangement tool. Future work will extend this framework to synthesize continuous 3D trajectories within predicted windows, and optimize for low-latency, real-time spatialization in DAW plugins, gaming, and immersive soundscapes.

7. REFERENCES

[1] D. Turner and D. T. Murphy, "A deep learning approach to the prediction of time-frequency spatial parameters for use in stereo upmixing," in *Proceedings of the 2024 27th International Conference on Digital Audio Effects (DAFx)*, 2024, pp. 428–435.

[2] Z. Liang, R. Wang, X. Ye, and Q. Kong, "Immersiveflow: Stereo-to-7.1. 4 spatial audio generation with flow matching," *arXiv preprint arXiv:2601.12950*, 2026.

[3] J. D. Reiss, "An intelligent systems approach to mixing multitrack audio," in *Mixing music*. Routledge, 2016, pp. 246–264.

[4] P. D. Pestana and J. D. Reiss, "A cross-adaptive dynamic spectral panning technique," in *Proceedings of the 2014 17th International Conference on Digital Audio Effects (DAFx)*, 2014.

[5] S. Hafezi and J. D. Reiss, "Autonomous multitrack equalization based on masking reduction," *Journal of the Audio Engineering Society*, vol. 63, no. 5, pp. 312–323, 2015.

[6] E. P. Gonzalez and J. D. Reiss, "Automatic mixing: live downmixing stereo panner," in *Proceedings of the 2007 7th International Conference on Digital Audio Effects (DAFx)*, 2007, pp. 63–68.

[7] K. Shimada, Y. Koyama, S. Takahashi, N. Takahashi, E. Tsunoo, and Y. Mitsufuji, "Multi-accdoa: Localizing and detecting overlapping sounds from the same class with auxiliary duplicating permutation invariant training," in *Proceedings of the 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (IEEE ICASSP)*, 2022, pp. 316–320.

[8] P. Morgado, N. Vasconcelos, T. Langlois, and O. Wang, "Self-supervised generation of spatial audio for 360 video," *Advances in neural information processing systems*, vol. 31, 2018.

[9] A. Richard, D. Markovic, I. D. Gebru, S. Krenn, G. A. Butler, F. Torre, and Y. Sheikh, "Neural synthesis of binaural speech from mono audio," in *Proceedings of the 2021 International Conference on Learning Representations (ICLR)*, 2021.

[10] C. J. Steinmetz, J. Pons, S. Pascual, and J. Serrà, "Automatic multitrack mixing with a differentiable mixing console of neural audio effects," in *Proceedings of the 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (IEEE ICASSP)*, 2021, pp. 71–75.

[11] C. J. Steinmetz, N. J. Bryan, and J. D. Reiss, "Style transfer of audio effects with differentiable signal processing," *arXiv preprint arXiv:2207.08759*, 2022.

[12] J. Tseng, R. Castellon, and K. Liu, "Edge: Editable dance generation from music," in *Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 448–458.

[13] Z. Dong, W. Hao, J.-C. Wang, P. Zhang, and P. Polak, "Musedance: A diffusion-based music-driven image animation system," in *Proceedings of the 2026 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2026, pp. 3813–3824.

[14] R. M. Patil and C. Patil, "Unveiling the state-of-the-art: A comprehensive survey on voice activity detection techniques," in *Proceedings of the 2024 Asia Pacific Conference on Innovation in Technology (APCIT)*, 2024, pp. 1–5.

[15] F. Jamshidi, G. Pike, A. Das, and R. Chapman, "Machine learning techniques in automatic music transcription: A systematic survey," *arXiv preprint arXiv:2406.15249*, 2024.

[16] W.-T. Lu, J.-C. Wang, M. Won, K. Choi, and X. Song, "Spectnt: A time-frequency transformer for music audio," *arXiv preprint arXiv:2110.09127*, 2021.

[17] Y.-N. Hung, J.-C. Wang, X. Song, W.-T. Lu, and M. Won, "Modeling beats and downbeats with a time-frequency transformer," in *Proceedings of the 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (IEEE ICASSP)*, 2022, pp. 401–405.

[18] K. Toyama, T. Akama, Y. Ikemiya, Y. Takida, W.-H. Liao, and Y. Mitsufuji, "Automatic piano transcription with hierarchical frequency-time transformer," *arXiv preprint arXiv:2307.04305*, 2023.

[19] Y. Wang, Y. Liu, J. Liu, K. Niu, and Z. He, "Cagcrn: Real-time speech enhancement with a lightweight model for joint acoustic echo cancellation and noise suppression," in *Proceedings of the 2025 Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2025, pp. 768–772.

[20] Y. N. Dauphin, A. Fan, M. Auli, and D. Grangier, "Language modeling with gated convolutional networks," in *Proceedings of the 2017 International Conference on Machine Learning (ICML)*, 2017, pp. 933–941.

[21] Y. Luo and N. Mesgarani, "Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 27, no. 8, pp. 1256–1266, 2019.

[22] T. Khandelwal and R. K. Das, "Dynamic thresholding on fixmatch with weak and strong data augmentations for sound event detection," in *Proceedings of the 2022 13th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, 2022, pp. 428–432.

[23] J. Ebbers, R. Haeb-Umbach, and R. Serizel, "Threshold independent evaluation of sound event detection scores," in *Proceedings of the 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (IEEE ICASSP)*, 2022, pp. 1021–1025.

[24] S. Adavanne, P. Pertilä, and T. Virtanen, "Sound event detection using spatial features and convolutional recurrent neural network," in *Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (IEEE ICASSP)*, 2017, pp. 771–775.

[25] G. Parascandolo, H. Huttunen, and T. Virtanen, "Recurrent neural networks for polyphonic sound event detection in real life recordings," in *Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (IEEE ICASSP)*, 2016, pp. 6440–6444.