LANGUAGE DIRECTED MONO-TO-STEREO MUSIC UPMIXING

Anonymous authors
Paper under double-blind review

ABSTRACT

The same mono song can become many valid stereo mixes. A producer may keep the kick centered, spread the claps across the speakers, or widen a vocal only in the chorus; none of these choices is determined by the mono recording alone. We introduce *Width as a Command*, a language-directed system that makes such choices explicit. It turns a user's request into an editable plan for the width of individual musical elements, grounds that plan in the recording, and generates a stereo mix while preserving the original mono content. Our key representation describes width through rapid left–right motion, independent of an element's overall position. A generative model then realizes a plausible motion for each requested width without requiring the user to specify an exact stereo trajectory. Experiments on public music examine both automatic upmixing and authored width control; ablations indicate that left–right motion provides a more useful width description than side-channel energy. Making stereo imaging describable and editable offers an interface for music production and a path toward spatially annotated data for future music generation.

1 INTRODUCTION

Consider the same chorus in two different mixes. In one, the kick and bass hold a narrow center while guitars open around the vocal; in the other, the vocal feels intimate and the clap briefly stretches from one ear to the other. Both mixes can preserve the same notes, rhythm, timbre, and loudness. Both can also collapse to nearly the same mono signal. Yet they tell different spatial stories. A mix engineer would not ask which story is the uniquely correct reconstruction. They would ask which one serves the song.

This observation changes the nature of mono-to-stereo music upmixing. A downmix is a many-to-one projection: after the left and right channels are summed, the arrangement of musical elements across the stereo field is no longer identifiable. A vocal may remain compact or bloom in the chorus; a clap may be centered, doubled, or spread; a lead instrument may occupy either side. Pointwise reconstruction objectives assign these valid production choices to the same mono observation and encourage their conditional mean. Left–right fluctuations then cancel, the side channel shrinks, and the resulting image returns toward the center. We call this failure *stereo mean reversion*.

We formulate mono-to-stereo music upmixing as language-directed stereo-image authorship. The user describes an intended image—for example, "keep the kick tight, make the vocal naturally wide, and let the bright claps open the chorus"—and the system realizes that image while preserving the musical content of the mono input. Language is valuable here because stereo decisions are relational, semantic, and often local: a producer thinks about the vocal, the low end, or a transient in the chorus, not about an undifferentiated pair of output waveforms.

Our system turns this request into an explicit stereo script through a *chain-of-thought*. Inspired by the spatial attribute induction used to construct BOTH EARS WIDE OPEN (Sun et al., 2025), the chain identifies musical elements, resolves relative descriptions, and assigns width over time and frequency. The resulting element-aware width map is inspectable and editable: it states which parts of the song should be tight, natural, or very wide before any stereo audio is synthesized. This intermediate representation gives free-form language a precise acoustic destination and also supports direct human editing.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

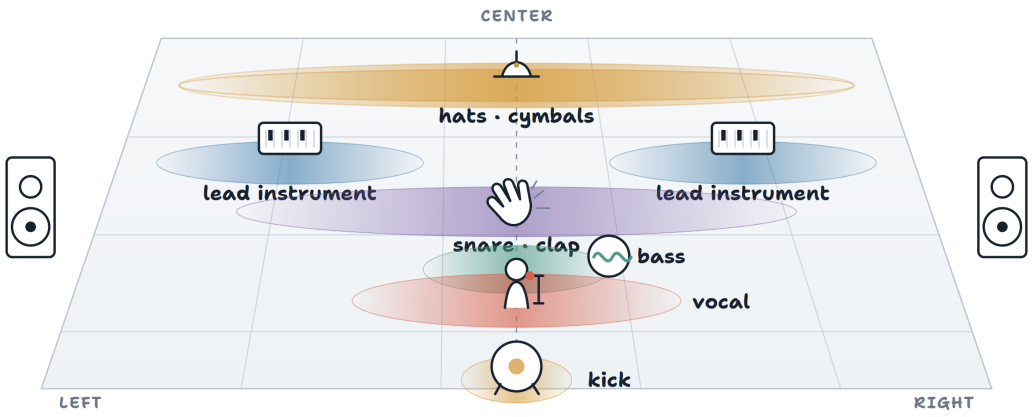

Figure 1: A stereo mix as a spatial composition. The horizontal coordinate denotes left–right position; the vertical coordinate schematizes perceived front–back depth; and each ellipse depicts an element's lateral spread. Level and monaural reverberation cues contribute to perceived depth (Kolarik et al., 2011; Prud'homme & Lavandier, 2020); the figure depicts it as context for element-wise stereo placement and width.

The acoustic control is built on parametric stereo (Purnhagen, 2004; Breebaart et al., 2005; Serrà et al., 2023). Inter-channel intensity difference (IID) and inter-channel coherence (IC) describe a compact, time-varying stereo image that can be rendered together with the original mono signal. We characterize width through fast, locally centered IID variation, which we call *width swing*. Centering removes a slow pan bias while retaining the alternating left–right structure that makes an element perceptually broad. The width map records the magnitude of this motion without prescribing its particular sign sequence, matching the creative ambiguity of the task.

A conditional backbone predicts a first parametric-stereo estimate from the mono signal and width map. A diffusion model supplies the structured IID motion that deterministic prediction tends to average away. A fixed decoder then renders the parameters into two channels with exact access to the mono content, and panning is composed as a separate post-model control. The same renderer supports authored commands and automatic operation through conditional dropout during training.

The same control surface may enrich future music generation. Music captions already describe instrumentation, style, and mood (Agostinelli et al., 2023; Manco et al., 2023; Doh et al., 2023); stereo-image decisions can be specified alongside these attributes. Rendering different scripts for the same mono song also creates aligned audio–caption–control examples for stereo-aware generation and captioning.

Our contributions are:

- We formulate language-directed mono-to-stereo music upmixing as stereo-image authorship and identify stereo mean reversion as a central failure mode of one-to-many stereo prediction.
- We introduce an element-aware width representation based on locally centered IID swing, connecting natural-language production intent to time–frequency stereo control.
- We develop a chain-of-thought system that compiles user language and musical activity into a stereo script, together with a mono-preserving parametric-stereo backbone and conditional diffusion renderer.
- We establish the width script as a controllable production interface and a source of aligned supervision for stereo-aware music generation and captioning.

2 Related Work

Stereo music upmixing. Early pseudo-stereo systems create inter-channel differences from a mono signal through phase, delay, or complementary filtering (Schroeder, 1958; Bauer, 1963; Orban, 1970). Later signal-processing designs develop decorrelation and downmix-compatible conversion (Faller, 2005; Fink et al., 2015; Uhle & Gampp, 2016). Source-aware upmixing uses inferred musical components or note events for placement (Lagrange et al., 2007; Fitzgerald, 2011; Delgado Castro & Szymanski, 2019). Learned systems range from early monaural-to-stereo networks (Chun et al., 2015) to latent spatial-style transfer (Yang et al., 2022). MusicHiFi predicts a side channel with an adversarial waveform generator while passing the mid channel through exactly (Zhu et al., 2024); DiffStereo applies a frequency-domain diffusion transformer to mono-conditioned stereo generation (Zhang et al., 2025).

Parametric stereo represents a stereo signal with a mono downmix and compact time–frequency spatial parameters; binaural cue coding and MPEG-4 parametric stereo established related cue-based representations (Baumgarte & Faller, 2003; Purnhagen, 2004; Breebaart et al., 2005). Serrà et al. (2023) formulate music upmixing as prediction of IID and IC, study regression, nearest-neighbor, autoregressive, and masked-token models, and connect the symmetric distribution of IID to generative modeling. Their two-parameter formulation, non-uniform IID/IC quantization, and fixed decoder provide the foundation for our renderer.

End-to-end music generation. Text-conditioned generation has developed through token-based music models (Agostinelli et al., 2023; Copet et al., 2023; Ziv et al., 2024) and diffusion or flow models (Evans et al., 2024a; Prajwal et al., 2024). Stereo-capable examples include stereophonic language modeling and open audio diffusion (Li et al., 2024; Evans et al., 2024b). Commercial systems such as Suno v6 and open systems such as ACE-Step 1.5 and Qwen-Music generate music from language, lyrics, and musical attributes (Gong et al., 2026; Xu et al., 2026). The latter two natively render 48-kHz stereo waveforms, making stereo a standard output capability of modern music generation. Their documented control vocabularies emphasize genre, mood, instrumentation, lyrics, structure, voice, and reference audio. The internal stereo image—where each element sits and how its width evolves over time—remains implicit in the rendered waveform and documented control vocabularies.

Language-conditioned binaural and spatial audio. Binaural audio and broader spatial-audio formats represent an auditory scene around a listener through cues such as azimuth, elevation, distance, motion, room acoustics, head-related filtering, or a multichannel sound field. Stereo music imaging denotes the artistic width and pan relationships within a conventional two-channel mix; these relationships need not describe a physically realizable three-dimensional scene. Visually guided stereo generation has been studied jointly with source separation (Zhou et al., 2020); binaural diffusion and neural Ambisonic upmixing model other spatial formats (Leng et al., 2022; Zang et al., 2024). Generative systems also model first-order Ambisonics and immersive soundscapes (Kushwaha et al., 2024; Heydari et al., 2024). Spatially aware language–audio representations such as ELSA connect acoustic semantics with direction descriptions (Devnani et al., 2024). Both Ears Wide Open introduces the BEWO-1M corpus and SpatialSonic for controllable spatial audio generation; its chain-of-thought pipeline induces scene attributes, source locations, and motion descriptions before simulation (Sun et al., 2025). Its attribute induction informs our language director. Text-guided spatialization has also been studied with a monaural reference: AudioSpa associates named sound events with requested directions (Feng et al., 2025), and the SpatialTAS framework learns absolute and relative binaural locations from textual prompts (Pan et al., 2025).

Semantic control for music mixing. Automatic music mixing systems predict production parameters or render processed audio from multitrack recordings (Steinmetz et al., 2021; Vanka et al., 2024). Spatial control has been studied through learned mixing styles (Yang et al., 2022) and explicit source placement. SPHERE post-trains an audio-language model to assign each available mono stem a musical role, gain, azimuth, and elevation, using a suite of centering, spread, and level-balance rewards (Guo et al., 2026). This line of work demonstrates that audio-language models can express production decisions as structured parameters. Music source separation provides an estimated element decomposition from a single mixture (Défossez et al., 2019; Luo & Yu, 2022; Rouard et al., 2023).

3 System Overview

Given a mono mix and a user request, our language director compiles element-wise width and pan rules. Mono-derived source activity grounds these rules in time and frequency, and an event gate forms the six-band width map. The backbone predicts IID and IC; diffusion samples the centered IID motion that realizes width. We then compose source-owned pan and render stereo with the input mono mid preserved. Figure 2 shows the complete path from language to audio.

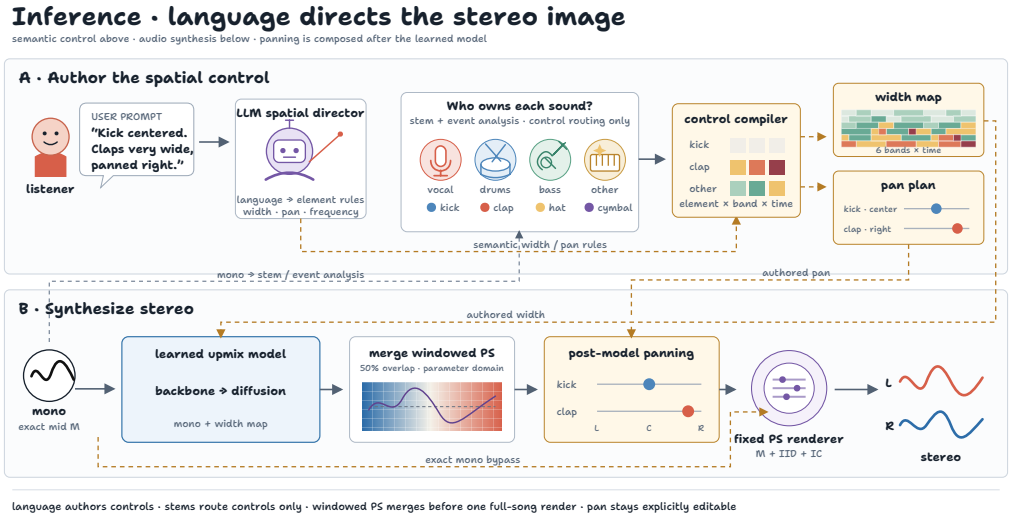

Figure 2: Inference pipeline. Language specifies element-wise width, source analysis routes the command, and a mono-derived event gate retains locally supported width before the learned renderer synthesizes centered width. For long-form inference, overlapping predictions are merged in IID/IC space before optional source-owned panning and a single full-song parametric-stereo rendering pass.

4 Width Representation

Parametric-stereo coordinates. We represent the stereo image with inter-channel intensity difference (IID) and inter-channel coherence (IC), following the compact parametric-stereo interface of Serrà et al. (2023), rooted in parametric stereo coding (Purnhagen, 2004; Breebaart et al., 2005). For the complex STFT coefficients $L_{f,t}$ and $R_{f,t}$,

$$G_{f,t} = 10 \log_{10} \frac{|L_{f,t}|^2 + \epsilon}{|R_{f,t}|^2 + \epsilon}. \tag{1}$$

Thus each time–frequency atom contains one signed IID value: positive and negative values denote opposite lateral biases. One atom holds one sign; width emerges from positive–negative IID motion across a local neighbourhood. IC describes complementary channel similarity. We retain the compact 31-level IID and eight-level IC interface of Serrà et al. (2023), measured at 172.27 frames/s so rapid motion remains visible. The temporal resolution study and its sampling analysis are in Appendix A.1.

Removing local position. Raw IID confounds width with panning: a coherent source held hard left has large $|G|$ but need not be wide. We form an energy-weighted six-band IID $I_b(t)$, quantize it on the same IID grid, and subtract its piecewise-constant local bias over non-overlapping P-frame blocks:

$$P_{b,k} = \frac{1}{P} \sum_{t \in k} q(I_b(t)), \qquad G^{c}_{f,t} = G_{f,t} - P_{b(f),k(t)}. \tag{2}$$

Centering preserves the IID differences within each band–block while moving their local origin toward zero (Figure 3). We use $P = 13$ (75.5 ms) throughout, selected on the MTG-Jamendo

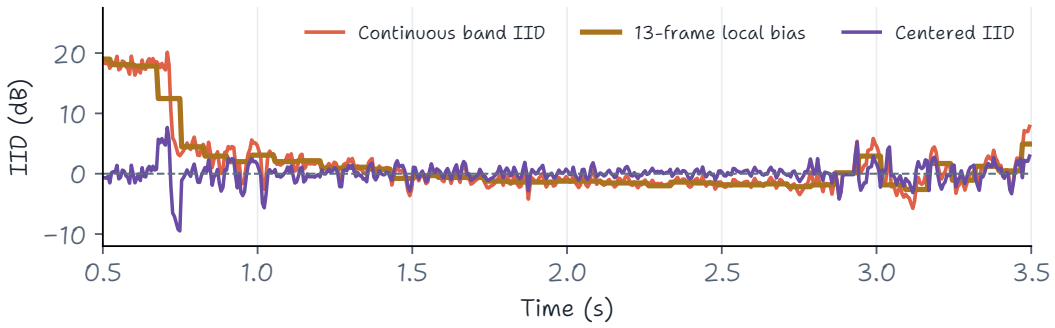

Figure 3: Local IID centering in a real stereo excerpt (0.8–2.5 kHz). A 13-frame bias carries the sustained lateral offset; the residual retains signed left–right motion around the local origin.

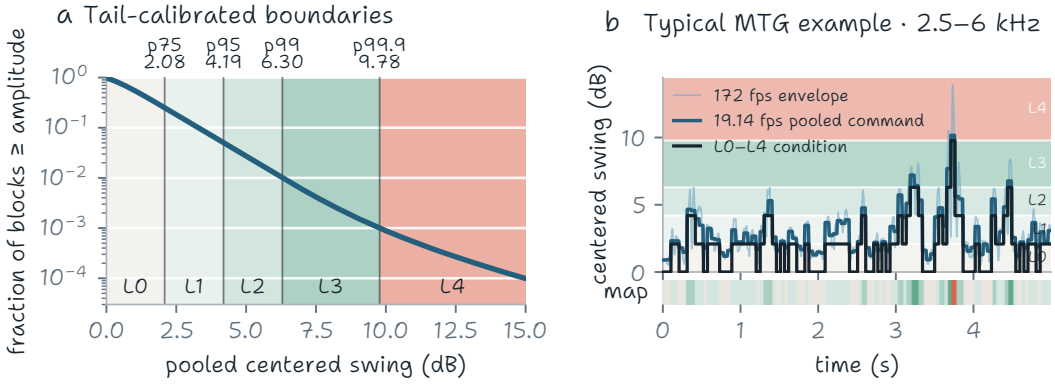

Figure 4: Full-corpus calibration of the five-level width vocabulary. **A:** the centered-swing survival curve and four percentile boundaries partition the 19.14-fps command map. **B:** an unedited five-second MTG validation excerpt whose 2.5–6-kHz band traverses every level. The envelope is measured at 172 fps, pooled to 19.14 fps, and quantized to the L0–L4 condition in the lower strip.

training corpus to suppress slow positional bias while retaining rapid IID motion. The selection curve, frequency response, and bootstrap analysis are in Appendix A.1.

From centered motion to a width command. We pool G^{c} into six signed band trajectories using mono-mid power, remove the residual below 5 Hz, and measure local RMS before pooling to a 19.14-frames/s command grid. This produces a slowly varying envelope of rapid signed motion. We quantize it to L0–L4 with values $\{0, 2.08, 4.19, 6.30, 9.78\}$ dB. Tail-weighted calibration allocates more levels to rare high-width events (Figure 4); the corpus statistics and threshold protocol are in Appendix A.2. The command specifies swing magnitude. Diffusion supplies a signed trajectory consistent with that magnitude. Language typically uses three ordinal levels and expands to four or five for finer stated relationships.

The native rate captures signed motion; the 13-frame window establishes its local origin; the 19.14-frames/s grid carries its editable magnitude. Appendix A.3 derives this measurement order.

Composing pan after width. The model generates G^{c} without a panning condition. After generation, an authored, source-owned offset is added in the same logarithmic coordinate,

$$G^{\mathrm{final}}_{f,t} = G^{\mathrm{c}}_{f,t} + P^{\mathrm{author}}_{f,t}, \tag{3}$$

as illustrated in Figure 2. The user can therefore place an element laterally while retaining its authored width.

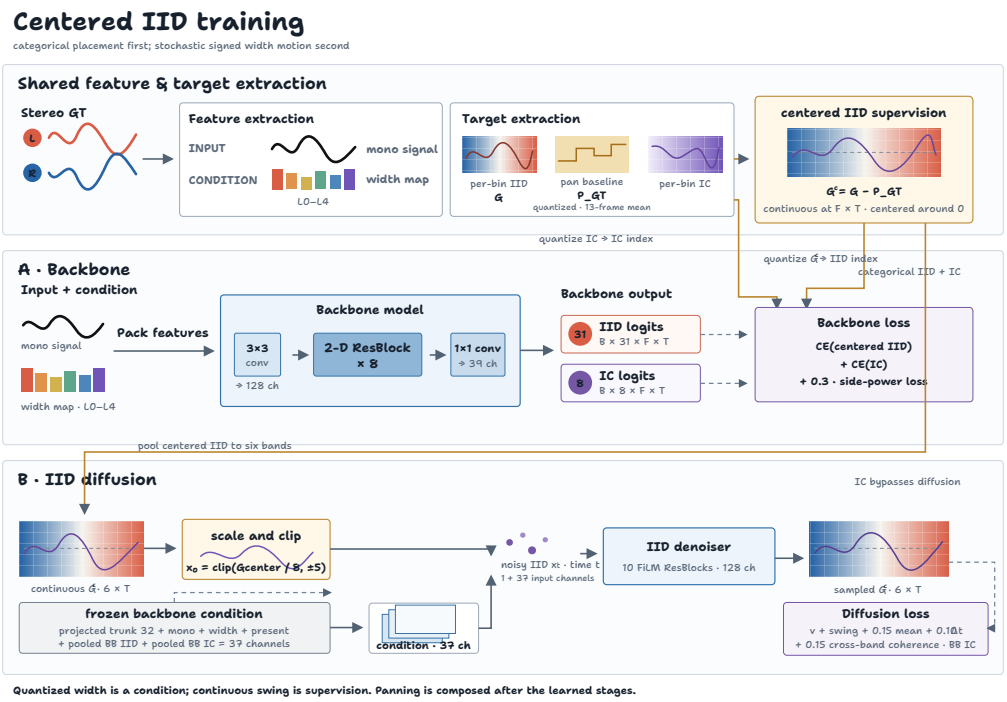

Figure 5: Two-stage training. The backbone predicts categorical per-bin centered IID and IC from mono content and a five-level width map. After the backbone is frozen, diffusion samples a continuous six-band centered-IID trajectory. The quantized map is an inference-time condition; continuous swing and stereo GT are used only as training supervision.

5 Method

We now specify the two learned stages in Figure 5. The backbone predicts categorical IID and IC at native time–frequency resolution. Diffusion samples a continuous centered-IID trajectory: different signed motions can realize the same width command, while their categorical expectation can cancel left and right excursions.

5.1 Training Data and Shared Extraction

We train on a 13,141-track stereo subset of MTG-Jamendo (Bogdanov et al., 2019), retaining tracks whose per-track license metadata exclude non-commercial and no-derivatives restrictions. The collection contains 919 hours of public music. A song-disjoint 12,878/263-track split yields 655,409 training and 12,678 validation windows; we use every contiguous five-second window at 44.1 kHz.

For every stereo window we compute a 1024-point STFT with hop 256. The exact mid $M = (L + R)/2$ supplies mono content. Stereo GT supplies per-bin IID and IC, and Equation 2 supplies the centered target G^{c}. We use six bands with boundaries at 120, 300, 800, 2500, and 6000 Hz. The width condition is obtained from G^{c} by the high-rate measurement and five-level quantization defined in Section 4.

5.2 Categorical Placement Backbone

The backbone receives mono log magnitude, the six-band L0–L4 width map, and a width-presence flag. A lightweight 2-D residual network predicts 31-way IID and eight-way IC distributions per time–frequency bin. Exact layers and tensor shapes appear in Appendix A.4.

Let y_G and y_C denote the quantized centered-IID and IC targets. Expected IID and IC under the predicted categorical distributions are passed through the fixed parametric-stereo equations to obtain

predicted centered-side power. The backbone objective is

$$\mathcal{L}_{\mathrm{B}} = \mathrm{CE}(\hat{y}_G, y_G) + \mathrm{CE}(\hat{y}_C, y_C) + 0.3\,\mathcal{L}_{\mathrm{side}}, \tag{4}$$

where $\mathcal{L}_{\mathrm{side}}$ matches six-band log side power computed from predicted and GT stereo parameters through the fixed renderer.

5.3 Centered-IID Trajectory Diffusion

After backbone training, its parameters are frozen. The diffusion target is the mono-power-weighted six-band pooling of continuous G^{c} at 172.27 frames/s. A diffusion denoiser (Ho et al., 2020) with a cosine noise schedule (Nichol & Dhariwal, 2021) and v-prediction (Salimans & Ho, 2022) conditions on mono energy, the width map, and frozen backbone features and expectations. Architecture, scaling, and schedules are in Appendix A.4.

The diffusion objective is

$$\mathcal{L}_{\mathrm{D}} = \mathcal{L}_v + \mathcal{L}_{\mathrm{swing}} + 0.15\mathcal{L}_{\mathrm{mean}} + 0.10\mathcal{L}_{\Delta t} + 0.15\mathcal{L}_{\mathrm{coh}}. \tag{5}$$

$\mathcal{L}_v$ is mono-energy-weighted velocity MSE. $\mathcal{L}_{\mathrm{swing}}$ renders the estimated IID with backbone-predicted IC, remeasures the continuous swing descriptor, and applies a wide-region-weighted smooth-L_1 loss against GT swing. $\mathcal{L}_{\mathrm{mean}}$ penalizes residual nine-frame local bias; $\mathcal{L}_{\Delta t}$ matches temporal IID differences; and $\mathcal{L}_{\mathrm{coh}}$ matches the 6×6 normalized cross-band correlation matrix. IC is learned by the backbone and passed to the renderer alongside diffusion's IID sample.

5.4 Conditional Dropout, Rendering, and Long-Form Inference

During both stages, we drop the width command for 50% of examples, following conditional/unconditional training (Ho & Salimans, 2022). Its values and presence flag share the dropout mask, enabling commanded and automatic inference in one model.

At inference, a 24-step deterministic sampler (Song et al., 2021) generates six-band centered IID; the frozen backbone supplies per-bin IC. The source-owned pan map is added to the generated IID as described by Equation 3. A fixed parametric-stereo renderer then broadcasts the IID trajectory to its frequency bands and constructs left and right channels as $L = M + S$ and $R = M - S$, preserving the supplied mono mid algebraically. This ordering makes width stochastic but keeps the authored lateral position directly editable.

Long recordings are analyzed in approximately 9.1-second windows with 50% overlap. Their predictions are merged in IID and IC parameter space before optional pan and a single full-song render against the input mono mid. Appendix A.5 specifies the merge.

6 Experiments

6.1 Evaluation Protocol

We evaluate on 100 songs from two public corpora. The first half comprises all 50 MUSDB18-HQ test songs (Rafii et al., 2019). The second half is selected from the official MTG-Jamendo split-0 test partition (Bogdanov et al., 2019) after excluding every track in our 13,141-song audio-model corpus. We use 50 unique-artist selections under CC BY-NC or BY-NC-SA licenses. From every song in both corpora we extract three fixed 9.1-second excerpts at 15%, 45%, and 75% of its available start range. MTG selection is a seeded genre round-robin and does not rank recording quality, stereo width, or model performance. The resulting 300 equal-length excerpts contain 45.5 minutes of audio. We pool excerpts directly; because every song contributes exactly three, this is also exactly song-balanced.

We group metrics by the failure they expose. Common-bandwidth, gain-aligned mid and side log-spectral distances (Gray & Markel, 1976) measure content preservation. Global side-to-mid level error, centered-swing error and correlation, IC error, and channel-correlation error measure complementary stereo-image properties. Fréchet-PS compares corpus distributions of six-band IID and IC. We also report *Side-FAD*: Fréchet Audio Distance (Kilgour et al., 2019) computed only on $S = (L - R)/2$, using all VGGish embeddings (Hershey et al., 2017) from the 45.5-minute evaluation set. Side-FAD measures distributional differences in the side channel.

Table 1: Main evaluation on 100 songs (300 excerpts). Lower is better except Swing r. Mid/side LSD use the shared 0–12-kHz band; width error is in dB. **Bold** is best within each panel, not across panels. † denotes a GT-derived coarse descriptor supplied at inference; all unmarked methods receive mono only. The baseline panel tests automatic upmixing and the oracle-command ceiling; the ablation panel isolates model choices.

Method	Mid LSD ↓	Side LSD ↓	Side FAD ↓	Width err. ↓	Swing MAE ↓	Swing r ↑	IC MAE ↓	ChCorr err. ↓	Fréchet PS ↓
Baseline comparison									
Mono duplicate	0.1474	58.683	7.48	91.566	1.155	.319	.400	.277	64.92
Schroeder	**0.00006683**	12.587	**.56**	7.986	.821	.449	.437	.371	31.13
PS-MTM	0.0003538	14.683	1.96	5.119	.660	.317	.322	.151	10.21
DiffStereo	11.157	15.505	4.85	5.189	.733	.332	.329	.164	15.46
Ours (auto)	0.0003742	12.203	1.56	7.364	.810	.405	.258	.162	**7.82**
Ours (oracle width)†	0.0003744	**10.219**	1.47	**4.588**	**.616**	**.679**	**.219**	**.121**	8.11
Core model ablation									
Backbone only†	0.0003747	**10.031**	2.61	5.111	**.473**	.657	.214	.127	13.23
Energy descriptor†	0.0003770	12.403	1.57	7.447	.627	.497	**.205**	.136	**8.00**
Energy (auto)	**0.0003736**	18.571	1.84	16.336	.836	.361	.301	.203	10.27
Ours (oracle width)†	0.0003744	10.219	**1.47**	**4.588**	.616	**.679**	.219	**.121**	8.11

6.2 Baseline Comparison

Mono duplication copies the input to both channels; Schroeder (Schroeder, 1958) is a signal-processing-only pseudo-stereo upmix that creates left–right differences through all-pass decorrelation. PS-MTM (Serrà et al., 2023) predicts quantized parametric-stereo cues, whereas DiffStereo (Zhang et al., 2025) generates stereo with a diffusion model. All four receive only mono audio. With its width command dropped, our automatic model is competitive with PS-MTM: it obtains lower Fréchet-PS (7.82 vs. 10.21) and Side-FAD (1.56 vs. 1.96), while PS-MTM has lower global width error (5.12 vs. 7.36 dB). DiffStereo's mid log-spectral distance (11.16) and Side-FAD (4.85) show larger content and side-distribution differences. Schroeder has the lowest Side-FAD (0.56) but much higher Fréchet-PS (31.13), illustrating the limits of a single distributional metric. With a coarse, GT-derived six-band width command at 19 frames/s, our conditioned model improves side log-spectral distance (10.22 vs. 14.68 dB), swing correlation (0.68 vs. 0.32), and IC error (0.22 vs. 0.32) relative to PS-MTM. This oracle-command setting measures how faithfully the model realizes width control; the command is not available to the mono-only baselines. The editable, band- and time-resolved width interface is the central capability.

6.3 Core Ablations

Backbone-only renders the six-band pooled expectation of backbone IID using the same oracle width command, predicted IC, and PS renderer as the full model. It has lower paired swing MAE (0.47 vs. 0.62) and slightly lower side-LSD (10.03 vs. 10.22). Across the 300 excerpts, its mean realized swing is 0.93 dB, compared with 1.16 dB in GT and 1.56 dB after diffusion: sampling adds motion but overshoots the mean swing amplitude. It nevertheless reduces global width error (5.11 to 4.59 dB), Side-FAD (2.61 to 1.47), and Fréchet-PS (13.23 to 8.11). Both variants share the backbone's predicted IC before rendering.

With oracle descriptors, centered swing yields lower width error than centered side-to-mid energy (4.59 vs. 7.45 dB) and higher swing correlation (0.68 vs. 0.50). The difference is larger with both commands dropped: swing versus energy gives 7.36 vs. 16.34 dB width error, 12.20 vs. 18.57 side-LSD, and 7.82 vs. 10.27 Fréchet-PS. Swing directly describes coherent signed IID motion, while side-to-mid energy also responds to pan and decorrelated content.

7 LANGUAGE DIRECTOR

The language director turns an open-ended request into an inspectable spatial script. Language specifies each element's behavior; mono-derived activity determines when and where that behavior enters the model's width map.

7.1 A TWO-STEP SPATIAL SCRIPT

The first step forms a complete *element plan* for vocal, bass, other accompaniment, kick, snare/clap, hi-hat, and cymbals. Each receives an L0–L4 base width and optional pan rule; unmentioned elements use natural-width defaults. Relative requests such as "pads wider than vocal" are resolved jointly on the ordinal scale.

The second step extracts *exceptions*: element-specific band/time rules and mix-wide modifiers. An element keeps one width unless language separates, for example, vocal air from vocal body. Kick and bass profiles keep their low-frequency body centered while assigning requested width to attack or harmonics. Absolute phrases set or cap levels; comparative phrases change level indices. The result is a machine-checkable spatial program.

7.2 MONO-DERIVED ELEMENT OWNERSHIP

From mono, a frozen separator estimates vocal, drums, bass, and other masks (Défossez et al., 2019; Rouard et al., 2023); a second frozen separator divides drums into kick, snare/clap, hi-hat, and cymbal masks (Mezza et al., 2024). Drum sub-masks are nested inside drums, so a snare command reaches only time–frequency energy jointly owned by drums and snare/clap. The *other* mask covers remaining instruments, pads, and ambience.

Masks are smoothed and pooled to six bands at 19.14 frames/s. Let $A_e(b,t)$ be element e's normalized soft ownership. The grid is fixed, while each element's support follows the song: its rule acts where that element is active.

7.3 COMPILING THE WIDTH MAP

Let $k_e(b,t) \in \{0,\dots,4\}$ be the element plan after its specific band/time exceptions, and let $v(k)$ be the calibrated L0–L4 command value. The element-owned map is

$$W^{\mathrm{elem}}_{b,t} = Q_5\left(\sum_e A_e(b,t)v(k_e(b,t))\right), \tag{6}$$

where Q_5 returns the nearest valid command value. Soft ownership smooths overlapping sources; absent elements have no effect.

Mix-wide absolute directives set or cap selected bands and times; relative directives shift level indices within L0–L4. Explicit element–band–time rules take precedence over conflicting global rules. The routed plan is quantized to five levels. A mono-derived event gate retains its locally salient width commands, forming a sparse $6 \times T_{19}$ map. Pan instructions become source-owned IID offsets after model sampling (Equation 3). Appendix B gives the event projection, full language prompt, JSON contract, deterministic compiler, and case-study protocol.

8 CONCLUSION

Mono-to-stereo music upmixing is a one-to-many production task: a mono song does not determine where its elements should sit or how widely they should spread. Our language-directed script makes these choices editable, and our mono-preserving model realizes them through locally centered IID motion with independent pan. The experiments examine automatic and commanded upmixing and show the value of motion-based width control.

The script can also describe existing stereo images, support stereo-image captioning, and generate aligned audio–caption–control examples for future music generators. Better element routing and stereo rendering are natural next steps. We hope this work makes stereo imaging a more explicit subject of music-generation research.

AI USE STATEMENT

Generative AI tools were used to assist with language polishing and editorial revision of the manuscript. In particular, they were used to improve clarity and concision, reorganize existing material, and help condense the presentation to meet the nine-page limit. All scientific ideas, methods, experiments, analyses, and conclusions were developed by the authors. The authors reviewed and verified all AI-assisted revisions and take full responsibility for the final manuscript.

ETHICS STATEMENT

We use two publicly released music datasets for non-commercial academic research: MTG-Jamendo for training and evaluation, and the MUSDB18-HQ official test set for evaluation (Bogdanov et al., 2019; Rafii et al., 2019). MTG-Jamendo recordings carry track-specific Creative Commons licenses, and MUSDB18-HQ contains recordings with differing rights and academic-use access conditions. The supplementary material provides dataset identifiers and evaluation timings, but no source recordings or audio excerpts. Stereo upmixing modifies an existing recording; use or distribution of a processed song remains subject to the rights attached to that recording.

REPRODUCIBILITY STATEMENT

The model and control representation are specified in Sections 4–5, and training settings and evaluation selection are documented in Appendix C. The supplementary ZIP includes the acoustic model source and selected inference weights, the renderer, the language director protocol and compiler, the frozen 100-song/300-excerpt manifest, and the metric implementation. The referenced music datasets and external language and separation models are obtained under their own distribution terms. The supplement records the exact inference command and the scope of the released training materials.

REFERENCES

Andrea Agostinelli et al. MusicLM: Generating music from text. *arXiv preprint arXiv:2301.11325*, 2023.

Benjamin B. Bauer. Some techniques toward better stereophonic perspective. *IEEE Transactions on Audio*, 11:88–92, 1963.

Frank Baumgarte and Christof Faller. Binaural cue coding—part I: Psychoacoustic fundamentals and design principles. *IEEE Transactions on Speech and Audio Processing*, 11(6):509–519, 2003. doi: 10.1109/TSA.2003.818109.

Dmitry Bogdanov, Minz Won, Philip Tovstogan, Alastair Porter, and Xavier Serra. The MTG-Jamendo dataset for automatic music tagging. In *Machine Learning for Music Discovery Workshop, International Conference on Machine Learning*, Long Beach, CA, USA, 2019. URL `http://hdl.handle.net/10230/42015`.

Jeroen Breebaart, Steven van de Par, Armin Kohlrausch, and Erik Schuijers. Parametric coding of stereo audio. *EURASIP Journal on Applied Signal Processing*, 2005(9):1305–1322, 2005. doi: 10.1155/ASP.2005.1305.

C. J. Chun, S. H. Jeong, S. Y. Park, and H. K. Kim. Extension of monaural to stereophonic sound based on deep neural networks. In *Audio Engineering Society Convention*, 2015.

Jade Copet, Felix Kreuk, Itai Gat, Tal Remez, David Kant, Gabriel Synnaeve, Yossi Adi, and Alexandre Défossez. Simple and controllable music generation. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Alexandre Défossez, Nicolas Usunier, Léon Bottou, and Francis Bach. Music source separation in the waveform domain. *arXiv preprint arXiv:1911.13254*, 2019.

Alejandro Delgado Castro and John Szymanski. Semi-automatic mono-to-stereo upmixing via separation of note events. In *Audio Engineering Society Conference on Immersive and Interactive Audio*, 2019.

Bhavika Devnani, Skyler Seto, Zakaria Aldeneh, Alessandro Toso, Elena Menyaylenko, Barry-John Theobald, Jonathan Sheaffer, and Miguel Sarabia. Learning spatially-aware language and audio embeddings. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

SeungHeon Doh, Keunwoo Choi, Jongpil Lee, and Juhan Nam. LP-MusicCaps: LLM-based pseudo music captioning. *arXiv preprint arXiv:2307.16372*, 2023.

Zach Evans, Cj Carr, Josiah Taylor, Scott H. Hawley, and Jordi Pons. Fast timing-conditioned latent audio diffusion. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 12652–12665, 2024a.

Zach Evans, Julian D. Parker, C. J. Carr, Zack Zukowski, Josiah Taylor, and Jordi Pons. Stable audio open. *arXiv preprint arXiv:2407.14358*, 2024b.

Christof Faller. Pseudostereophony revisited. In *Audio Engineering Society Convention*, 2005.

Linfeng Feng, Lei Zhao, Boyu Zhu, Xiao-Lei Zhang, and Xuelong Li. AudioSpa: Spatializing sound events with text. *arXiv preprint arXiv:2502.11219*, 2025.

Marco Fink, Sebastian Kraft, and Udo Zölzer. Downmix-compatible conversion from mono to stereo in time- and frequency-domain. In *International Conference on Digital Audio Effects*, 2015.

Derry Fitzgerald. Upmixing from mono: A source separation approach. In *International Conference on Digital Signal Processing*, 2011.

Junmin Gong, Yulin Song, Wenxiao Zhao, Sen Wang, Shengyuan Xu, Jing Guo, and Xuerui Yang. ACE-Step 1.5: Pushing the boundaries of open-source music generation. *arXiv preprint arXiv:2602.00744*, 2026.

Augustine H. Gray and John D. Markel. Distance measures for speech processing. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 24(5):380–391, 1976. doi: 10.1109/TASSP.1976.1162849.

Zixun Guo, Calvin Murdock, Sanjeel Parekh, W. Owen Brimijoin, Simon Dixon, Joshua Reiss, and Ishwarya Ananthabhotla. SPHERE: Automatic music upmixing via audio language model post-training with spatial heuristic rewards. *arXiv preprint arXiv:2608.30559*, 2026.

Shawn Hershey et al. CNN architectures for large-scale audio classification. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2017.

Mojtaba Heydari, Mehrez Souden, Bruno Conejo, and Joshua Atkins. ImmerseDiffusion: A generative spatial audio latent diffusion model. *arXiv preprint arXiv:2410.14945*, 2024.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Kevin Kilgour, Mauricio Zuluaga, Dominik Roblek, and Matthew Sharifi. Fréchet audio distance: A reference-free metric for evaluating music enhancement algorithms. In *Proceedings of Interspeech*, pp. 2350–2354, 2019. doi: 10.21437/Interspeech.2019-2219.

Andrew J. Kolarik, Silvia Cirstea, and Shahina Pardhan. Perceiving auditory distance using level and direct-to-reverberant ratio cues. *The Journal of the Acoustical Society of America*, 130:2545, 2011. doi: 10.1121/1.3655171.

Saksham Singh Kushwaha, Jianbo Ma, Mark R. P. Thomas, Yapeng Tian, and Avery Bruni. Diff-SAGe: End-to-end spatial audio generation using diffusion models. *arXiv preprint arXiv:2410.11299*, 2024.

Mathieu Lagrange, Luis Gustavo Martins, and George Tzanetakis. Semi-automatic mono to stereo up-mixing using sound source formation. In *122nd Audio Engineering Society Convention*, 2007.

Yichong Leng, Zehua Chen, Junliang Guo, Haohe Liu, Jiawei Chen, Xu Tan, Danilo P. Mandic, Lei He, Xiangyang Li, Tao Qin, Sheng Zhao, and Tie-Yan Liu. BinauralGrad: A two-stage conditional diffusion probabilistic model for binaural audio synthesis. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Xingda Li, Fan Zhuo, Dan Luo, Jun Chen, Shiyin Kang, Zhiyong Wu, Tao Jiang, Yang Li, Han Fang, and Yahui Zhou. Generating stereophonic music with single-stage language models. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1471–1475, 2024. doi: 10.1109/ICASSP48485.2024.10446643.

Yi Luo and Jianwei Yu. Music source separation with band-split RNN. *arXiv preprint arXiv:2209.15174*, 2022.

Ilaria Manco, Benno Weck, SeungHeon Doh, Minz Won, Yixiao Zhang, Dmitry Bogdanov, Yusong Wu, Ke Chen, Philip Tovstogan, Emmanouil Benetos, Elio Quinton, György Fazekas, and Juhan Nam. The song describer dataset: A corpus of audio captions for music-and-language evaluation. *arXiv preprint arXiv:2311.10057*, 2023.

Alessandro Ilic Mezza, Riccardo Giampiccolo, Alberto Bernardini, and Augusto Sarti. Toward deep drum source separation. *Pattern Recognition Letters*, 183:86–91, 2024. doi: 10.1016/j.patrec.2024.04.026.

Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8162–8171, 2021.

Robert Orban. A rational technique for synthesizing pseudo-stereo from monophonic sources. *Journal of the Audio Engineering Society*, 18(2):157–164, 1970.

Tianrui Pan, Jie Liu, Zewen Huang, Jie Tang, and Gangshan Wu. In-the-wild audio spatialization with flexible text-guided localization. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, pp. 1989–2001, 2025. doi: 10.18653/v1/2025.acl-long.98.

K. R. Prajwal, Bowen Shi, Matthew Le, Apoorv Vyas, Andros Tjandra, Mahi Luthra, Baishan Guo, Huiyu Wang, Triantafyllos Afouras, David Kant, and Wei-Ning Hsu. MusicFlow: Cascaded flow matching for text guided music generation. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 41052–41063, 2024.

Luna Prud'homme and Mathieu Lavandier. Do we need two ears to perceive the distance of a virtual frontal sound source? *The Journal of the Acoustical Society of America*, 148(3):1614–1623, 2020. doi: 10.1121/10.0001954.

Heiko Purnhagen. Low complexity parametric stereo coding in MPEG-4. In *International Conference on Digital Audio Effects*, pp. 163–168, 2004.

Zafar Rafii, Antoine Liutkus, Fabian-Robert Stöter, Stylianos Ioannis Mimilakis, and Rachel Bittner. MUSDB18-HQ: An uncompressed version of MUSDB18, 2019. URL `https://doi.org/10.5281/zenodo.3338373`.

Simon Rouard, Francisco Massa, and Alexandre Défossez. Hybrid transformers for music source separation. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2023. doi: 10.1109/ICASSP49357.2023.10097087.

Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. In *International Conference on Learning Representations*, 2022.

Manfred R. Schroeder. An artificial stereophonic effect obtained from a single audio signal. *Journal of the Audio Engineering Society*, 6(2):74–79, 1958.

Joan Serrà, Davide Scaini, Santiago Pascual, Daniel Arteaga, Jordi Pons, Jeroen Breebaart, and Giulio Cengarle. Mono-to-stereo through parametric stereo generation. In *Proceedings of the 24th International Society for Music Information Retrieval Conference*, pp. 304–311, 2023.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021.

Christian J. Steinmetz, Jordi Pons, Santiago Pascual, and Joan Serrà. Automatic multitrack mixing with a differentiable mixing console of neural audio effects. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2021. URL https://arxiv.org/abs/2010.10291.

Peiwen Sun, Sitong Cheng, Xiangtai Li, Zhen Ye, Huadai Liu, Honggang Zhang, Wei Xue, and Yike Guo. Both ears wide open: Towards language-driven spatial audio generation. In *International Conference on Learning Representations*, 2025.

Christian Uhle and Philipp Gampp. Mono-to-stereo upmixing. In *Audio Engineering Society Convention*, 2016.

Soumya Sai Vanka, Christian J. Steinmetz, Jean-Baptiste Rolland, Joshua D. Reiss, and George Fazekas. Diff-MST: Differentiable mixing style transfer. *arXiv preprint arXiv:2407.08889*, 2024.

Jin Xu, Kangdi Wang, Ruibin Yuan, Shun Lei, Xiong Wang, Xize Cheng, Xueyao Zhang, Yang Zhang, Yiheng Chen, Yongqi Wang, Yue Wang, Zhifang Guo, Zihan Liu, Zijian Lin, Dake Guo, Hangrui Hu, Lei Xie, Linhan Ma, Wei Xue, Wenxiang Guo, Xinfa Zhu, Xipin Wei, Yangze Li, Yuanjun Lv, Yuxuan Wang, Yunfei Chu, and Zhiyong Wu. Qwen-Music technical report. *arXiv preprint arXiv:2607.11699*, 2026.

Haici Yang, Sanna Wager, Spencer Russell, Mike Luo, Minje Kim, and Wontak Kim. Upmixing via style transfer: A variational autoencoder for disentangling spatial images and musical content. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022.

Yongyi Zang, Yifan Wang, and Minglun Lee. Ambisonizer: Neural upmixing as spherical harmonics generation. *arXiv preprint arXiv:2405.13428*, 2024.

Suqi Zhang, Zheqi Dai, Yongyi Zang, Yin Cao, and Qiuqiang Kong. DiffStereo: End-to-end mono-to-stereo audio generation with diffusion transformer. In *Proceedings of Interspeech*, pp. 3150–3154, 2025. doi: 10.21437/Interspeech.2025-1850.

Hang Zhou, Xudong Xu, Dahua Lin, Xiaogang Wang, and Ziwei Liu. Sep-Stereo: Visually guided stereophonic audio generation by associating source separation. In *European Conference on Computer Vision*, 2020.

Ge Zhu, Juan-Pablo Caceres, Zhiyao Duan, and Nicholas J. Bryan. MusicHiFi: Fast high-fidelity stereo vocoding. *IEEE Signal Processing Letters*, 2024. doi: 10.1109/LSP.2024.3432393.

Alon Ziv, Itai Gat, Gael Le Lan, Tal Remez, Felix Kreuk, Jade Copet, Alexandre Défossez, Gabriel Synnaeve, and Yossi Adi. Masked audio generation using a single non-autoregressive transformer. In *International Conference on Learning Representations*, 2024.

A Representation and Model Details

A.1 Temporal resolution and centering-window selection

The 1024-point STFT with hop 256 measures IID at 172.27 frames/s. Figure 6 analyzes the same real excerpt at progressively coarser 75%-overlap STFT resolutions. The associated corpus hop study retains approximately 100%, 31%, 9%, and 1% of native 5–85-Hz swing power at hops 256, 512, 1024, and 2048. Lower frame rate and its longer analysis window jointly suppress rapid signed excursions.

We select the centering block length from a song-balanced analysis of all 13,141 MTG-Jamendo tracks. For frame rate f_s, removing a P-sample block mean from random-phase sinusoidal IID modulation of frequency ν retains the expected energy fraction

$$R_P(\nu) = 1 - \left[\frac{\sin(\pi\nu P/f_s)}{P\sin(\pi\nu/f_s)}\right]^2. \tag{7}$$

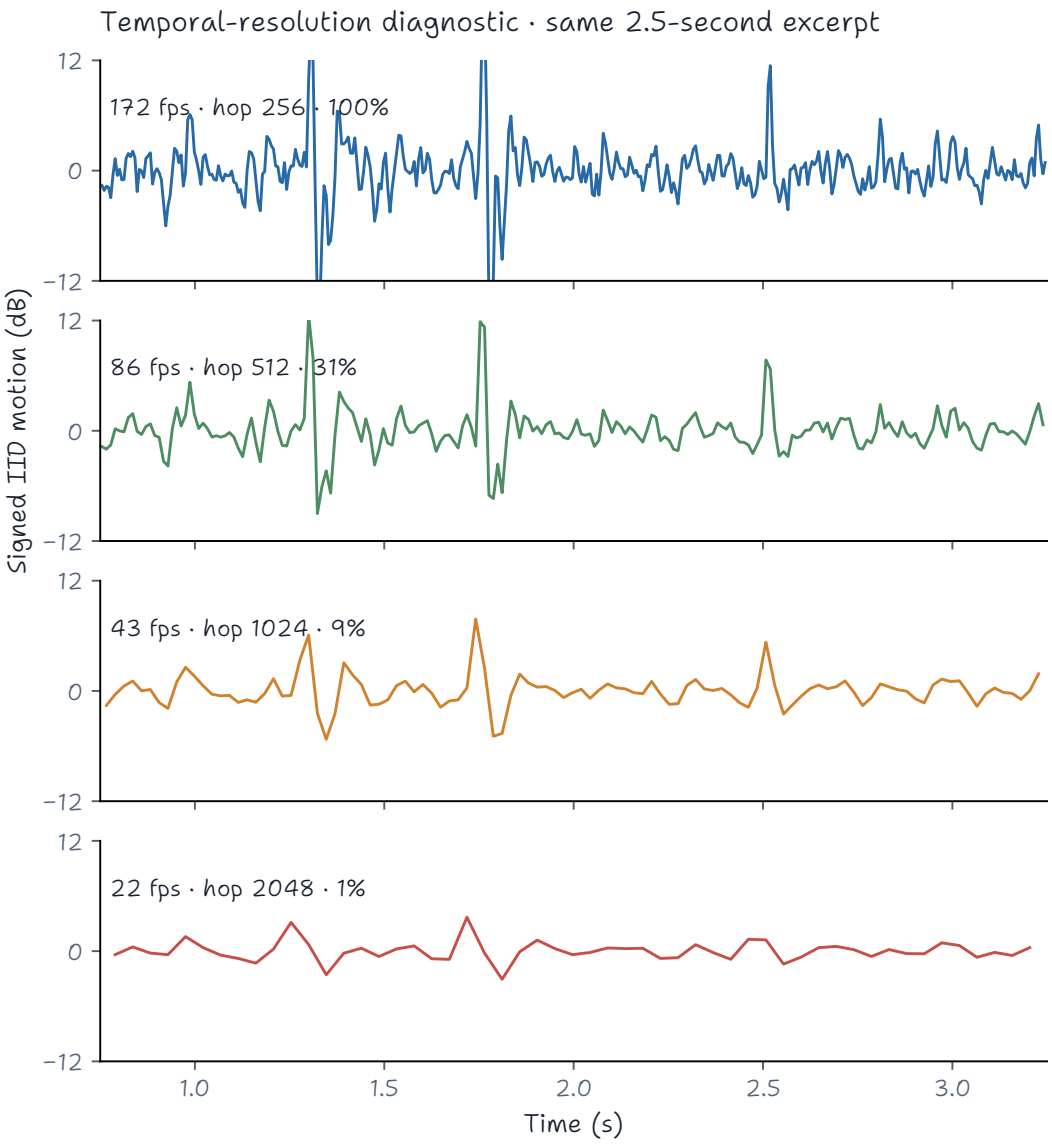

Figure 6: The same 2.5-second IID interval at four STFT resolutions. Percentages give recovered 5–85-Hz swing power relative to hop 256 in the associated aggregate hop study.

We define 0.5–5 Hz as the slow-bias range and 5–85 Hz as the width-motion range for this selection, then choose the P closest to zero retention in the former and full retention in the latter. Figure 7 shows the two curves. At $P = 13$ (75.46 ms), mean slow-bias removal is 89.31% and mean motion retention is 90.54%. All 2,000 song-level bootstrap resamples select the same value under these ranges.

Subtracting $P_{b,k}$ preserves every pairwise IID difference within the band–block. It can also be expressed as opposite channel gains, $L_c = 10^{-P/40}L$ and $R_c = 10^{P/40}R$, whose IID is $G - P$. The operation removes a mixture-level local bias; source-specific pan is authored separately after generation.

A.2 Full-corpus width-command calibration

The continuous centered IID is pooled into six signed trajectories with mono power, high-passed above 5 Hz, and measured by nine-frame local RMS before pooling to 19.14 frames/s. Mid energy

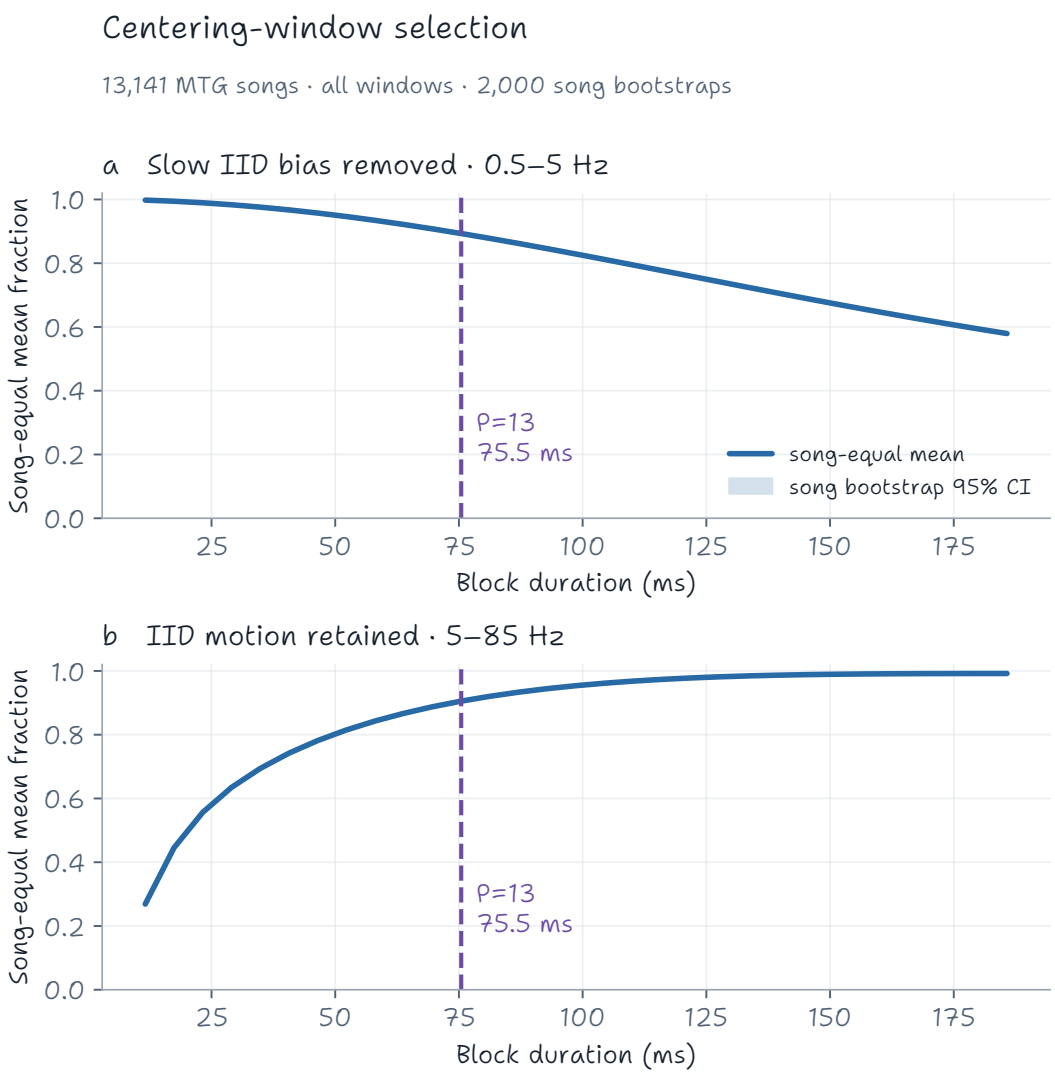

Figure 7: Full-corpus centering-window selection. Each song aggregates all of its windows before contributing equally to the mean and 95% song-level bootstrap interval. The vertical line marks $P = 13$.

supplies a relative audibility gate; IID itself is invariant to global gain. From the 655,409 training windows and 377.5 million six-band command cells, the 75th, 95th, 99th, and 99.9th percentiles give the nonzero boundaries 2.08, 4.19, 6.30, and 9.78 dB. Together with explicit L0, the five levels occupy approximately 75%, 20%, 4%, 0.9%, and 0.1% of these cells. The tail-weighted codebook preserves distinct control levels for uncommon high-width events.

A.3 Swing measurement before command-rate pooling

The 19.14-frames/s command in Figure 4 is produced by nonlinear measurement of native-rate IID. The operations are ordered as

$$G^{c}_{172} \xrightarrow{>5\,\mathrm{Hz}} H_{172} \xrightarrow{\text{local RMS}} A_{172} \xrightarrow{\text{nine-frame pool}} W_{19}, \tag{8}$$

where G^{c} is signed centered IID, A is its non-negative swing amplitude, and W is the editable command.

To see this directly, let D_P denote the mean within a block of P samples. A blockwise idealization of the local swing estimator is

$$\mathcal{S}_P(G) = \sqrt{D_P(G^2) - D_P(G)^2} = \sqrt{\mathrm{Var}_P(G)}. \tag{9}$$

The coarse sequence $D_P(G)$ retains only the first moment; the second moment $D_P(G^2)$ has already been discarded. Consequently there is no function of $D_P(G)$ alone that recovers $\mathcal{S}_P(G)$. As a constructive counterexample, both $G_k = 0$ and $G_k = A\cos(2\pi k/P + \phi)$ have zero block mean over one complete cycle, yet their swing amplitudes are respectively 0 and $A/\sqrt{2}$. Squaring (or local RMS) must therefore precede pooling: $D_P(G^2) \neq D_P(G)^2$ in general.

Sampling provides a second, independent reason. The native rate is $f_s = 172.27$ fps, whereas a nine-frame command grid is $f_c = f_s/9 = 19.14$ fps with a Nyquist frequency of only 9.57 Hz. It cannot represent most of the declared 5–85-Hz motion band. Moreover, the nine-sample box average has magnitude response

$$|H_9(\nu)| = \left| \frac{\sin(9\pi\nu/f_s)}{9\sin(\pi\nu/f_s)} \right|, \tag{10}$$

which is exactly zero at $\nu = f_s/9 = 19.14$ Hz: a one-cycle-per-command-block left–right swing is annihilated even though its width is large. The representative example in Figure 6 shows the same loss under direct low-rate analysis. Thus 172 fps is the *identification rate*; 19.14 fps is only the rate of the slowly varying amplitude command produced after sign and phase have been removed.

A.4 Acoustic-model architecture and conditioning

The backbone input is $X_{\mathrm{B}} \in \mathbb{R}^{3\times F\times T}$: normalized mono log magnitude, the six-band L0–L4 width map expanded to frequency bins, and the binary width-presence flag. A 3×3 convolution projects to 128 channels. Eight 2-D residual blocks use temporal and frequency dilations $[1, 2, 4, 8, 16, 8, 4, 2]$; a 1×1 head produces 39 logits at each bin and frame, partitioned into 31 IID classes and eight IC classes. The IID grid spans -50 to 50 dB with finer spacing near zero; the IC grid has eight values from 1 to -0.91902.

The frozen backbone supplies a 32-channel projected trunk, pooled expected IID and IC, and its width-mask state to diffusion. The remaining condition channels are six-band mono log energy and the quantized width map, giving 37 condition channels. The noisy one-channel IID state makes 38 channels on the $6 \times T$ grid. A 3×3 input convolution, ten 128-channel dilated 2-D residual blocks with timestep FiLM, and a 1×1 single-channel head implement cosine v-prediction. The continuous IID target is scaled by 8 dB and clipped to $[-5, 5]$.

Both stages jointly drop width values and their presence flag in 50% of training examples. In diffusion the same mask is applied to the explicit width channel and the width input of the frozen backbone, preserving the automatic pathway through both condition routes.

A.5 Long-form parameter merging

Long recordings use approximately 9.1-second windows with 50% overlap, aligned to the 256-sample STFT grid. Let $w_j(t)$ be complementary raised-cosine overlap weights, and G_j, C_j the IID and IC predicted by window j. We merge

$$\bar{G}(t) = \frac{\sum_j w_j(t) G_j(t)}{\sum_j w_j(t)}, \qquad \bar{C}(t) = \tanh\left(\frac{\sum_j w_j(t)\,\mathrm{atanh}(\tilde{C}_j(t))}{\sum_j w_j(t)} \right), \tag{11}$$

where $\tilde{C}_j = \mathrm{clip}(C_j, -1+\epsilon, 1-\epsilon)$. IID is interpolated in decibels and IC in Fisher-z space. Authored pan is composed after merging, followed by one rendering pass with the complete input mono track as mid.

A.6 Six-band parametric-stereo corpus reference

Figure 8 shows the marginal distribution of quantized IID and IC, while Figure 9 shows their six bandwise joint distributions, in the full 13,141-song MTG-Jamendo acoustic-model collection. We

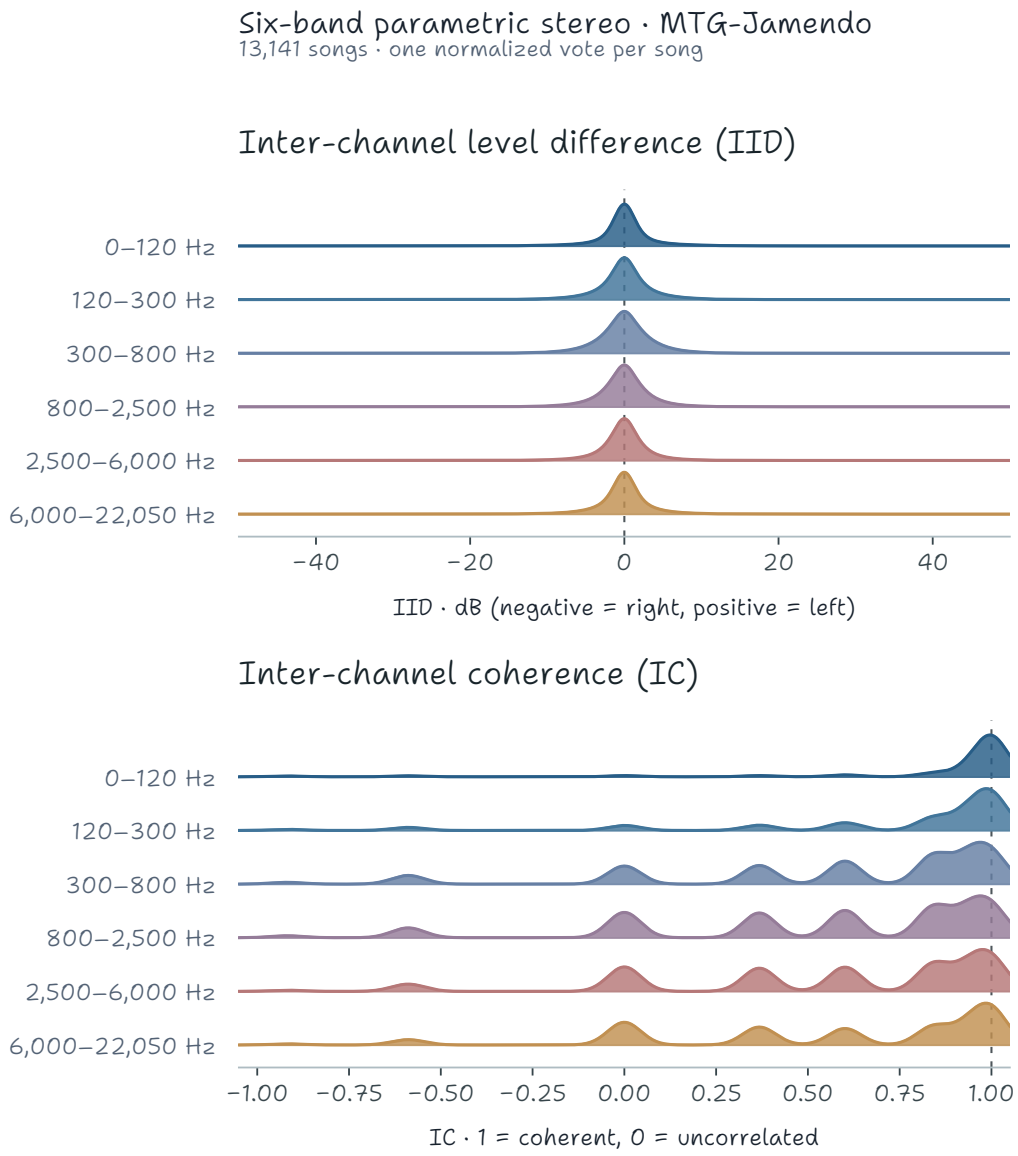

Figure 8: Song-balanced six-band IID and IC distributions across the complete MTG-Jamendo acoustic-model collection. Each ridge has independent peak scaling for legibility; shape and spread, not ridge height, are comparable across bands. The plotted smoothing does not affect histogram counts or discrete-code entropy.

analyze all 668,087 five-second windows with the same 44.1-kHz, 1024-point STFT, 256-sample hop, six frequency bands, and 31/8-level codebooks used by the model. Each song first contributes a channel-energy-weighted histogram for each band; histograms are normalized within song and band before the songs are averaged. This makes the reference song-balanced. The ridgelines are smoothed and rescaled only for display. This distribution describes the population of stereo cues available for learning; it is not the calibration used to choose the five width-command levels.

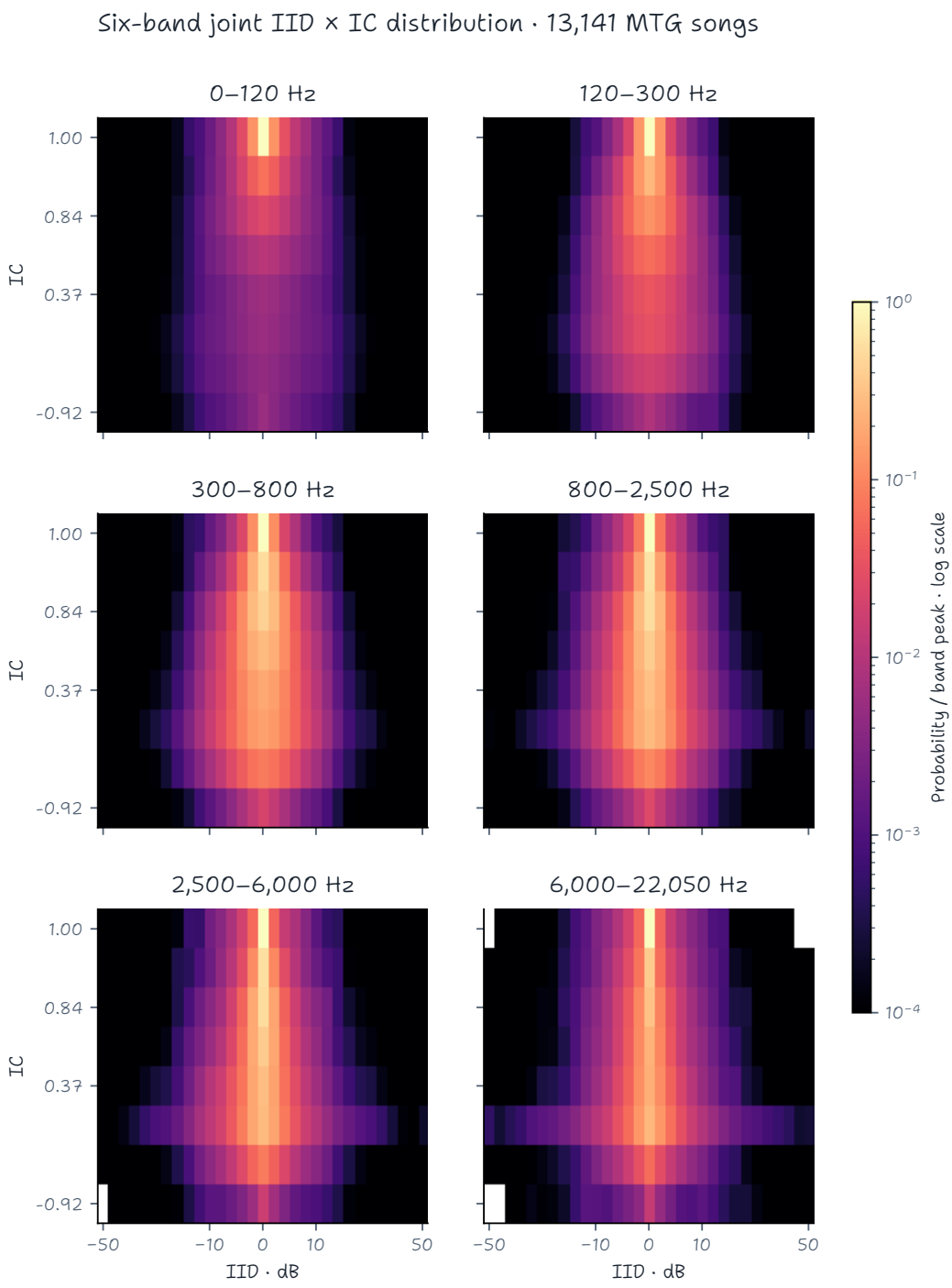

Figure 9: Joint IID–IC distributions for all six frequency bands. Each panel has an independent peak normalization and the same logarithmic relative-probability color scale. The unsmoothed cells retain the dependence between left/right level difference and channel coherence.

B LANGUAGE DIRECTOR PROTOCOL AND CASE STUDIES

B.1 REPRODUCIBLE LLM INSTANTIATION

We instantiate the director with Qwen3.5-27B using the released instruction chat template and a four-bit weight-only checkpoint. Its input contains the user's request, the controlled vocabulary below, and optional externally supplied section timestamps. Decoding is deterministic (greedy, temperature zero) with a fixed model revision and a maximum of 2,048 new tokens. The output is one JSON object conforming to the director schema.

The chain of thought is serialized as an explicit, task-specific intermediate program with two visible parts: a complete element plan and a sparse list of exceptions. The semantic decision is inspectable, and the acoustic compiler is fully deterministic.

The system instruction is:

> You are a stereo-width director. Convert the user request into the supplied JSON schema. Emit exactly one row for each of `vocal`, `bass`, `other`, `kick`, `snare_clap`, `hihat`, and `cymbals`. Width is ordinal: L0 is mono/tight, L1 narrow, L2 medium, L3 wide, and L4 very wide. First assign every element a base width and an independent pan instruction. Then add only explicitly requested frequency or time exceptions and mix-wide modifiers. Resolve comparative statements jointly while preserving their ordering. A width adjective for kick applies to its attack above 800 Hz; its body below 300 Hz remains L0. A width adjective for bass applies to harmonics above 300 Hz; its fundamental below 120 Hz remains L0 and 120–300 Hz is capped at L1. General widening leaves the 0–120-Hz band unchanged. Explicit requests for wide sub-bass or kick fundamental are recorded in `unresolved` for confirmation. Do not invent instruments, event times, section boundaries, or frequency qualifications. A named section may be used only when its timestamps are supplied; otherwise place it in `unresolved`. Return JSON only.

The director emits enumerated JSON fields that are validated before audio generation. The shortened schema below shows the machine contract; the released artifact includes the full JSON Schema.

```
{
  "schema_version": "1.0",
  "element_plan": [
    {"element": "vocal", "base_width": "L2",
     "pan": {"mode": "center"},
     "overrides": [
       {"bands": [5], "time": {"kind": "all"},
        "operation": "set", "width": "L3"}
     ]}
  ],
  "relations": [
    {"lhs": "snare_clap", "relation": "wider_than",
     "rhs": "vocal", "margin_levels": 1}
  ],
  "global_modifiers": [
    {"bands": [0], "time": {"kind": "all"},
     "operation": "cap_max", "width": "L0"}
  ],
  "unresolved": []
}
```

Band indices refer to the six fixed analysis bands defined in Section 4. A time scope is either `all`, an interval in seconds, a source event class, or a named section with externally provided start and end times. Pan is one of `unchanged`, `center`, `left`, or `right`, with an optional bounded strength. Width and pan are separate fields: a wide centered source and a narrow lateral source are both legal.

B.2 VALIDATION, REPAIR, AND DETERMINISTIC COMPILATION

The validator requires exactly one row for each of the seven controlled elements, legal enum values, and in-range band and time indices. It rejects unknown sources and unsupported production controls. A malformed response is returned to Qwen once with the machine-readable validation errors and an instruction to repair those errors. A second invalid response is counted as a director failure. Valid but acoustically absent elements are retained in the program and later become no-ops through ownership masking.

Compilation then proceeds without an LLM. Relative constraints are applied iteratively on the integer lattice $\{0, \dots, 4\}$. An unmentioned element takes a frozen default. Its base level is expanded through an element-specific six-band profile and across the command frames. The kick profile fixes bands below 300 Hz to L0 and caps 300–800 Hz at L1; the bass profile fixes 0–120 Hz to L0 and caps 120–300 Hz at L1. Element-specific band/time exceptions overwrite this profile. Given the mono-derived soft ownership $A_e(b,t)$, the compiler forms Equation 6, applies global rules, clips indices to L0–L4, and forms the initial quantized map. Precedence is

$$\text{explicit element}\times\text{band}\times\text{time} \;>\; \text{element exception} \;>\; \text{global modifier} \;>\; \text{base/default}. \tag{12}$$

An absolute rule uses `set`, `cap_min`, or `cap_max`; a comparative rule adds an integer number of levels. For example, "make the highs one step wider" applies $+1$ after element routing, whereas "everything below 120 Hz must be mono" caps the corresponding lowest band at L0. The ownership masks are smoothed before pooling to produce continuous spectral and temporal boundaries. Pan is compiled to a separate source-owned IID edit and is composed after model sampling.

Concretely, for a requested pan strength $a_e \in [0,1]$, let $s_e = +1$ for left, -1 for right, and 0 for center/unchanged. With $\widetilde{A}_e(b,t)$ denoting mono-derived ownership at the native 172-fps time grid, the implemented offset is

$$P^{\text{author}}_{b,t} = G_5\left(\sum_e \widetilde{A}_e(b,t)\, 9\,\text{dB}\; s_e a_e\right), \tag{13}$$

where G_5 smooths over five native frames. The resulting six-band IID offset is broadcast to STFT bins before the fixed renderer. Center and unchanged both contribute zero; they do not cancel a generated IID deviation.

B.3 SPARSE EVENT PROJECTION

The frozen language compiler ends with an inference-only event projection. The element-routed map first uses a conservative L1 bed for L3/L4 requests, with source-salient accents promoted to L2/L3; lower requests retain their element level. The unconditioned model's own width map is used as a floor only where the element plan asks to widen. Explicit mix-wide rules are then applied, with element-specific exceptions retaining priority. This produces a candidate map $W^{\text{bed}}_{b,t}$. No GT width or stereo channel is used in this step.

Let $P_{b,t}$ be mono power pooled to six bands and held over each nine-frame command block ($\simeq 19.14$ frames/s). Within each clip and band, blocks below -20 dB relative to that band's peak are inactive. For active blocks, write $z_{b,t} = \log(1 + P_{b,t})$ and define the positive local novelty score

$$s_{b,t} = G_{0.7}\Big([z_{b,t} - G_3(z_b)_t]_+ + 0.45\,[z_{b,t} - z_{b,t-1}]_+\Big), \tag{14}$$

where G_σ denotes one-dimensional Gaussian smoothing in command-frame units. For each band, we retain the highest-scoring positive-novelty blocks, up to $\lceil 0.25 N^{\text{active}}_b \rceil$, and set every other width command to L0:

$$W^{\text{final}}_{b,t} = \begin{cases} W^{\text{bed}}_{b,t}, & t \in \text{TopNovelty}_{0.25}(b), \\ \text{L0}, & \text{otherwise.} \end{cases} \tag{15}$$

The mask is expanded from the command grid to the model's native time grid. This is a sparse realization of the semantic width plan: a request for wide accompaniment does not prescribe sustained

width through all accompaniment frames. It is the exact inference rule in the released language-director script; the acoustic checkpoint is unchanged.

On the frozen 100-song, 300-excerpt GT audit, 77.9% of command cells are L0, yet 74.3% of cells above the same -20-dB activity threshold are also L0. Thus low energy alone does not explain the natural map's sparsity. Across 1,800 six-band GT trajectories, only two have a contiguous $\geq$L2 interval of at least one second.

B.4 WORKED SEMANTIC CASE

Table 2 records the visible intermediate state for the request: "Keep kick and bass mono; give the vocal natural space; make every snare and clap the widest element; keep cymbals narrow, but let synths and ambience above 3 kHz bloom wide." The intermediate program uses element-level parameters. Song-specific ownership projects each row into a time–frequency map.

Table 2: Element-plan state for a language-director case. "High L3" is a sparse exception; blank pan fields are unchanged. Ownership projection turns this semantic program into the $6 \times T_{19}$ acoustic map.

Element	Base width	Explicit exception	Pan
vocal	L2	—	center
bass	L0	—	center
other	L2	high bands L3	unchanged
kick	L0	—	center
snare/clap	L4	—	unchanged
hi-hat	L1	—	unchanged
cymbals	L1	—	unchanged

Our case suite tests four distinct language operations: (i) absolute element widths, (ii) relative rankings involving three or more elements, (iii) element-specific spectral or temporal exceptions, and (iv) global rules that partly conflict with an explicit element exception. For every case we retain the user request, raw Qwen response, validation/repair trace, normalized JSON, element table, six-band ownership maps, compiled width map, and rendered audio. We report JSON-valid rate, constraint-satisfaction rate, and deterministic repeatability separately from audio quality.

B.5 GT-DESCRIPTION ROUND TRIP

We additionally evaluate an oracle-assisted round trip that measures the language bottleneck. From a reference stereo clip we compute the same centered, five-level GT width map used by the conditioned audio model. For each element, mono-derived ownership and mono energy aggregate this map into level histograms over band and time. A weighted median supplies the candidate base level; a band or contiguous time interval is proposed as an exception only when it differs from that base by at least one level and has sufficient active support. Qwen receives these structured summaries and writes a natural spatial-mixing request. A fresh director call then parses that request through the standard JSON protocol, and the deterministic compiler produces a new width map.

The round trip is measured before audio generation by ownership- and mono-energy-weighted level accuracy, level MAE, and rank correlation between the original and reconstructed maps. We then render the reconstructed map with the frozen audio model and report the same stereo metrics as the main evaluation. The row is labeled *GT-described language round trip* and measures oracle-assisted representational loss. Its detailed results and case studies appear in the appendix; a main-table version is placed in a separate oracle-application block.

The full evaluation freezes prompts, model revision, decoding parameters, schema, defaults, mask estimator, thresholds, and sampling seeds. Raw and normalized Qwen outputs are both retained. This protocol distinguishes three possible failures: semantic parsing (language to JSON), grounding (JSON plus ownership to width map), and stereo synthesis (width map plus mono to audio).

C Training and Evaluation Details

The acoustic training set contains 13,141 MTG-Jamendo songs from the declared training and validation license selection. Stereo audio was decoded at 44.1 kHz and cut into 668,087 complete five-second windows. Both stages use a 1,024-point STFT, 256-sample hop, six bands split at 120, 300, 800, 2,500, and 6,000 Hz, and a 13-frame local IID-centering window. The five width levels use boundaries 2.08375, 4.19125, 6.29625, and 9.77625 dB.

The selected backbone was trained for 8,380 steps with a per-device batch size of 12 on three GPUs; the diffusion stage ran for 3,000 steps with a per-device batch size of 36. Both use width-condition dropout probability 0.5, learning rate 5×10^{-5}, weight decay 0.01, and mixed bfloat16 precision. The backbone combines IID and IC cross-entropy with a centered-side term weighted 0.3. Diffusion combines velocity prediction with rendered-width, local-mean, derivative, and cross-band-coherence terms weighted 1.0, 0.15, 0.10, and 0.15, respectively. Both selected stages were warm-started from preceding project checkpoints. The released weights contain the final diffusion model and its frozen backbone; precursor checkpoints and optimizer histories are outside the supplement.

The frozen evaluation uses all 50 MUSDB18-HQ official-test songs and 50 MTG-Jamendo official split-0 test songs, with no MTG track-ID overlap with acoustic training. Three 9.1-second excerpts per song begin at 15%, 45%, and 75% of the valid start-time range. The supplementary manifest records each source ID and exact start time. Inference uses 24 sampling steps and seed 20260925; stereo outputs are 32-bit float WAVs. The supplementary metric script computes the paired spectral and parametric-stereo measures reported in Table 1. Side-FAD additionally uses an external VGGish feature extractor. Benchmark audio is not redistributed.