

PHYSICSSLIDERS: LEARNING PHYSICAL CONTROLS FOR AUDIO FOLEY GENERATION

Peiwei Chang¹, Yongqiang Chen¹, Yuelan Cheng¹, Liting Gao¹, Yishun Lu²,
Li Zhang³, Mark D. Plumbley⁴, Philip J.B. Jackson¹, Wenwu Wang¹

¹Centre for Vision, Speech and Signal Processing, University of Surrey, UK

²Department of Engineering Science, University of Oxford, UK

³Department of Computer Science, Royal Holloway, University of London, UK

⁴Department of Informatics, King’s College London, UK

ABSTRACT

Physical properties provide intuitive controls for sounds made by interacting objects. The challenge is to make a generative model respond predictably to these controls, both individually and in combination. We present PhysicsSliders (PXS), trained on simulated contact sounds with 15 material, geometry and contact parameters. Two separately trained variants accept direct parameter inputs or numerical captions. PXS controls hardness, size, speed and damping, supports room descriptions and produces spectrally distinct outputs at fixed targets. When controlling damping and size together, specifying thickness improves frequency accuracy and joint target delivery. Measurements from real recordings define effective-damping targets, which simulator inversion maps to generation inputs without real-audio fine-tuning. Across 18 impact classes from five real-audio datasets, PXS-Physical follows these targets with mean grouped Spearman correlation $\rho = 0.976$ and class-median error 0.094 dex (a factor of 1.24).

Index Terms— controllable audio generation, Foley synthesis, physical modelling, flow matching, physical control

1. INTRODUCTION

Contact sounds convey what objects are made of, how they move and how they interact. Recreating these sounds for media is part of Foley production. Controllable audio generation must go beyond event labels to respond predictably to changes in objects and interactions. For games, virtual environments and animation, this could adapt sounds to changing scenes without selecting a recording for every event. Editors could explore alternatives for the same physical setting or change room acoustics. Material, geometry and excitation provide natural controls because they shape attack, resonances and decay.

Existing audio-generation methods offer complementary forms of control. Text prompts specify sound events and scenes [1, 2]. Temporal-event conditioning controls event timing [3]. Acoustic-feature conditioning follows loudness, brightness and pitch trajectories [4], while latent-space guidance targets intensity, pitch and beats [5]. Continuous semantic controls further support attribute editing and content preservation [6]. These approaches give users expressive control over the acoustic result, but an edit to an object’s size or contact hardness must be expressed through those acoustic or semantic coordinates. In contrast, physics-aware methods connect sound to objects and interactions through modal priors [7], geometry and strike position [8], or physical-state descriptions [9] and

estimated mass and velocity [10]. Simulation-trained explosion synthesis also demonstrates that neural models can inherit controls from a physically inspired synthesiser [11]. Single-factor video-to-audio benchmarks further assess physical responses under controlled interventions [12]. These advances establish physical conditioning and directional responses. Target-driven audio Foley generation additionally requires meeting specified acoustic values, including two at once. This calls for evaluating absolute target accuracy alongside response direction, and identifying which object properties must be supplied for joint control.

We introduce PhysicsSliders (PXS) with direct-parameter (PXS-Physical) and numerical-caption (PXS-Text) interfaces. Both share a flow-matching architecture [13] but train separately on simulation. Our contributions are: **(1) Measurement-based generation:** we map effective-damping targets from five real-audio datasets to generation inputs through simulator inversion, and demonstrate target delivery without real-audio fine-tuning. **(2) Joint target control:** we evaluate simultaneous damping and frequency targets, then isolate the role of thickness information by masking it in the same generator. This distinguishes directional response from joint target accuracy. **(3) Physical and linguistic interfaces:** we learn hardness, size, speed and damping responses that transfer to held-out materials, combine parameter inputs with room descriptions, and produce spectrally distinct outputs at fixed targets.

2. METHOD

PXS-Physical accepts parameters with or without text; PXS-Text accepts text alone (Fig. 1).

Physical supervision. Online modal simulation generates 2-second, 44.1-kHz plate and bar sounds from Hertzian impacts and random surface excitation (45 % hit, 15 % bounce, 25 % scrape, 15 % roll). The 15-dimensional parameter vector θ describes material and damping (wave speed c_L , density, source loss η , external damping multiplier κ), geometry (size L , thickness ratio h/L), excitation (effective contact modulus E , mass, radius, speed v , position), and motion (restitution, roughness, normal force, rolling radius). For fixed shape, mode k has frequency $f_k \propto c_L(h/L)/L$ and amplitude-decay rate $d_k = \pi f_k \eta \kappa + 1.5 \text{ s}^{-1}$. E controls hardness; κ adds damping such as hand muting. Room acoustics, microphone response, noise and RMS gain are randomised. Room 60-dB decay time T_{60} is log-uniform over 0.08–0.8 s. PXS-Physical captions describe the same simulated room as dry below 0.12 s or reverberant above 0.45 s, with no room phrase between these values.

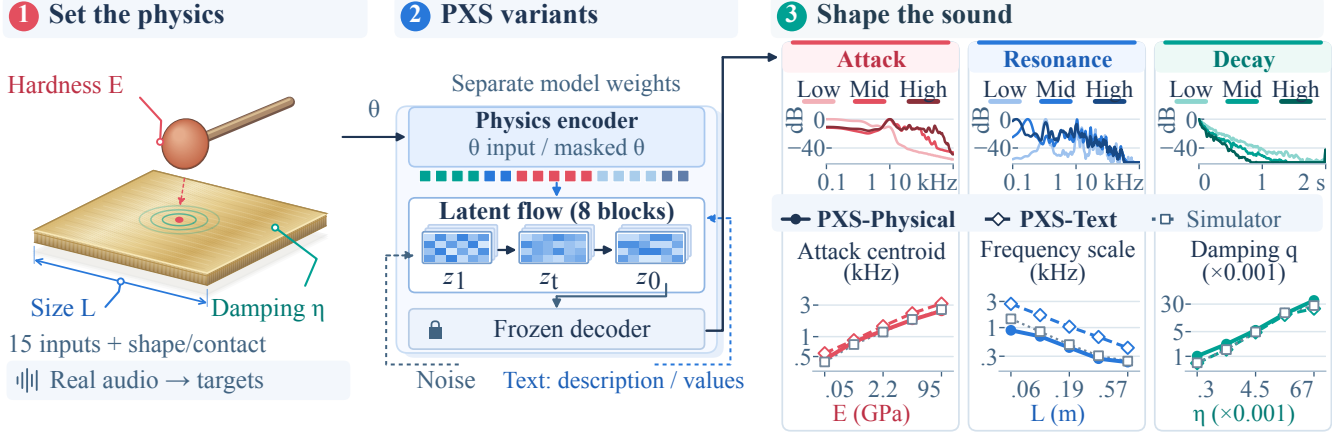


Fig. 1. PXS variants train separately with a shared architecture and frozen decoder. Paired labels show PXS-Physical first. Objects and latents are schematic. Upper-right brass impacts vary one input at fixed sampling noise. Spectra (attack 30 ms, resonance 500 ms) and RMS envelopes are peak-normalised. Low/mid/high use $E = \{0.05, 2.24, 100\}$ GPa, $L = \{0.126, 0.200, 0.316\}$ m and $\eta = \{1.71, 4.30, 10.8\} \times 10^{-3}$. Lower curves show five-level impact sweeps with paired noise. PXS curves geometrically average valid-reading medians (32 samples) across three held-out materials and three runs. Simulator curves average over materials. q is measured effective damping.

Plate aspect ratio is uniform on $[1.15, 1.6]$ but is not an input. Size, thickness ratio and hardness are log-uniform. Loss sampling mixes material-specific draws (65%) with full-range coverage (35%), $\eta \in [2.82 \times 10^{-4}, 7.10 \times 10^{-2}]$. Speed is log-uniform over 0.2–5 m/s for impacts and 0.03–1 m/s for sliding; $\kappa = 1$ in 25% of draws, otherwise log-uniform on $[1, 16]$. Training uses steel, glass, ceramic, stone, wood and plastic, $L = 0.06$ –0.6 m and $h/L = 0.005$ –0.10. Held-out aluminium, brass and concrete have wave speeds 5100, 3430 and 3540 m/s. Aluminium overlaps training support; brass and concrete lie between training bands.

Conditioned audio generation. An eight-layer, 54.8-million-parameter 1-D flow transformer is trained from scratch using Stable Audio Open (SAO)’s frozen pretrained Ooblock variational autoencoder (VAE) [1]. Parameter tokens use $a_i = (T_i(\theta_i) - \mu_i)/\sigma_i$, standardised by frozen simulation means μ_i and standard deviations σ_i . T_i is the natural log (identity for position, restitution and roughness exponent). Features $\gamma(a_i)$ concatenate a_i with sine and cosine terms at πf , $f \in \{0.25, 0.5, 1\}$. Binary mask m_i selects a relevant supplied value or learned null token n_i . With identity embedding p_i and shared multilayer perceptron V , the token and input at latent position τ are

$$e_i = m_i \{V(\gamma(a_i)) + p_i\} + (1 - m_i)n_i, \quad (1)$$

$$H_\tau^{(0)} = W_z z_{t,\tau} + p_\tau + W_r [\bar{e}; m \odot a; m].$$

Shape and contact type add tokens e_{16}, e_{17} . A learned map P pools all 17: $\bar{e} = P(\frac{1}{17} \sum_{i=1}^{17} e_i)$. Here $t \in [0, 1]$ is flow time, W_z, W_r are affine maps, p_τ embeds position, $[\cdot]$ concatenates and $\odot$ multiplies elementwise. Physics tokens and frozen FLAN-T5-large [14] text embeddings enter cross-attention; pooled physics and time embeddings modulate blocks. Flow matching (FM) learns conditional velocity v_ϕ from audio latent z_0 to Gaussian noise (ϕ : weights, C : conditions):

$$z_t = (1 - t)z_0 + t\epsilon, \quad \epsilon \sim \mathcal{N}(0, I), \quad (2)$$

$$\mathcal{L}_{\text{FM}} = \mathbb{E} \|v_\phi(z_t, t, C) - (\epsilon - z_0)\|_2^2.$$

Each variant uses three independent AdamW runs (50,000 updates, batch 48, learning rate 10^{-4}), evaluated with final exponential mov-

ing average weights. During training, PXS-Physical drops text with probability 15%, physics with 14.5% (both: 6%), and individual coordinates with 10%. Physics dropout includes shape and contact type. PXS-Text always masks physics and drops text with 15% probability. Sampling uses 50 Euler steps from Gaussian noise with classifier-free guidance (CFG) [15] of 3, then the frozen decoder.

From measurements to requests. Users set θ or request effective damping $q = \hat{\eta} = h_c(x)$ from recording x , with readout h_c for excitation c . For impacts, Hilbert envelopes track resonance amplitudes; sustained excitation masks free decay, so unbiased autocorrelation envelopes measure decay with lag. Peak detection uses 8192-sample Hann spectra, onset-aligned for impacts or Welch-averaged with 50% overlap for sustained sounds. We retain up to six impact or eight sustained peaks within 45 or 20 dB of the spectral maximum, respectively, excluding peaks below 150 Hz. Rectangular Fourier masks isolate each resonance. Impact half-widths $b_0 = 0.015 f_k + 4$ Hz expand to $\max(b_0, 5d_k^{(1)}/\pi)$ Hz using initial decay-rate estimate $d_k^{(1)}$. Invalid refits retain the first estimate. Fits start 10 ms after the envelope peak, requiring 20-dB headroom initially and 15 dB for refits above the final 20% envelope median. They stop at a 30-dB fall, 6 dB above this floor or 1.5 s. Sustained bands have half-width $\max(0.05 f_k, 25)$ Hz. After discarding the first 10% of audio, autocorrelation fits start at $\max(2/f_k, 0.002)$ s and stop below 0.15 normalised amplitude or after one quarter of the remaining duration. Valid fits last at least 10 ms with $d_k > 0$. Their median $d_k/(\pi f_k)$ defines q , missing if no peak is valid.

Excitation, rooms, modal frequencies and κ affect q , so it need not equal source loss η . For object/contact template $\mathcal{T}$, simulator S maps between them using 24 log-spaced training-range losses $\mathcal{E}$ and three excitation/recording realisations ν_b per loss:

$$\eta^*(q; \mathcal{T}) = \arg \min_{\eta \in \mathcal{E}} \left| \log_{10} \frac{\text{med}_{b=1}^3 h_c(S(\eta; \mathcal{T}, \nu_b))}{q} \right|. \quad (3)$$

Here med takes the median of valid readings. Template loss η^* specifies input θ_i^* for target i , with draws y_{id} indexed by d . PXS-Text receives η^* in captions. Templates use 0.30-m plates ($h/L = 0.008$, $\kappa = 1$), class-assigned materials (wood if unspecified), and hits for

Table 1. Effective-damping control. (a) Tracking (n : classes; error: dex). (b) Impact requests with two or more of four outputs on target (%).

(a) Physical and numeric-text control							(b) Multiple on-target outputs ($n = 18$)		
Generator	Impulsive ($n = 18$)			Sustained ($n = 12$)			Generator	$1.5\times$	$2\times$
	$\rho \uparrow$	$g \rightarrow 1$	Error $\downarrow$	$\rho \uparrow$	$g \rightarrow 1$	Error $\downarrow$			
PXS-Physical	.976	.77	.094	.829	.19	.293	PXS-Physical	83	96
PXS-Text	.954	.78	.058	.844	.27	.360	PXS-Text	85	96
							Simulator (fixed aspect)	90	100
							Simulator (varied aspect)	92	100

Table 2. Encoding ablation: model/simulator slope ratios (two runs; ideal 1). Smooth: Eq. 1; High-freq.: $f = 2^j$, $j = 0, \dots, 7$, no scalar. Direct: value/mask path. Bold: PXS-Physical.

Token features	Direct	Hardness	Size	Damping
Smooth	Yes	0.95	0.76	0.97
High-freq.	Yes	1.11	0.65	0.89
Smooth	No	0.94	0.65	0.97
High-freq.	No	0.20	0.19	0.25

Table 3. Two-second zero-shot grouped rank. Missing: incomplete class responses (scored zero). p_H : one-sided class-paired Wilcoxon signed-rank tests (PXS greater), Holm-corrected across eight comparisons.

Generator	Impulsive ($n = 18$)			Sustained ($n = 12$)		
	$\rho \uparrow$	Missing	p_H	$\rho \uparrow$	Missing	p_H
PXS-Physical	.976	0	–	.829	0	–
SAO numeric	.022	3 <	.001	-.108	2	.001
SAO ordinal	.278	5	.002	.033	3	.010
TangoFlux numeric	-.028	17 <	.001	.250	1	.026
TangoFlux ordinal	.100	16 <	.001	.233	1	.026

impulsive targets or scrapes otherwise.

3. EXPERIMENTS

We assess real-audio target accuracy (Table 1) and joint control (Fig. 2). Evaluation includes 10 ESC-50 [16], 4 FSD50K [17], 6 AudioSet [18], 5 NSynth [19] and 5 RealImpact [20] classes.

Targets and grouping. We retain 18 impulsive and 12 sustained classes after excluding two with insufficient range and two AudioSet/FSD50K duplicates. After trimming endpoints separated by over 2δ while retaining at least eight recordings, each class supplies eight evenly indexed targets from sorted valid $\log_{10} q$. Ordering and gain require at least three groups, each spanning at most δ from its first log-value. Dex denotes base-10 log units: 0.10 dex is a factor of 1.26. Calibrated material- and excitation-specific thresholds span 0.065–0.491 dex and set ranking resolution, not target tolerance. Impact thresholds average within-strike microphone-pair log-difference 95th percentiles from RealImpact. Sustained thresholds use Arni recordings [21], taking the root-sum-of-squares of percentiles within and across configurations. We evaluate all 240 requests, including those outside simulator support.

Evaluation metrics. Each target’s median $\tilde{q}_i$ requires at least three of four valid outputs. Gain g is the slope of log group-mean predictions against log group-mean targets, fitted with an intercept

(ideal 1). Error is $|\log_{10}(\tilde{q}_i/q_i)|$, and Spearman ρ measures group ordering. Flat or incomplete class responses score zero correlation and gain. We average median target errors and gains over three runs, then report class medians and mean rank. Selection and qualification use no model outputs. At tolerance $\lambda \in \{1.5, 2\}$, requests succeed if at least two of four outputs have $h_c(y_{id}) \in [q_i/\lambda, \lambda q_i]$. Invalid readings fail. We average success over requests and training runs or simulator noise banks, then take class medians.

Measurement validation. Of 36 dry impulses with known decay, 32 yield valid fits and 29 are within $1.5\times$. An alternative readout is $\ln(1000)/(\pi f_{\text{scale}} T_{60})$, using Schroeder decay time T_{60} [22] and frequency scale f_{scale} , the geometric mean of up to three lowest decay-validated resonances. Both readouts are valid for all 144 PXS-Physical impact requests (three runs) and independent simulator templates. Ungrouped correlation and $1.5\times$ request-median agreement are 0.941 and 78.9 % for h_c , versus 0.986 and 40.7 % for the alternative. Ordering is stable, but absolute agreement depends on the decay definition.

3.1. Target delivery and physical responses

Both learned interfaces track damping targets from real impacts, with PXS-Physical gain 0.77 (95 % class-bootstrap confidence interval [0.65, 0.86]). Classes have 3–7 groups (median 4). Without grouping, correlations remain 0.920 for PXS-Physical and 0.911 for PXS-Text. By contrast, sustained responses cover less of the target range. In a paired scrape sweep, PXS-Physical’s effective-damping floor is 12.7 times the simulator value. Impacts show no such floor.

Sampling at fixed targets. For 144 impact requests (Table 1b), we vary only latent noise, fixing parameters and masking text for PXS-Physical, or fixing the caption for PXS-Text. Simulator references share parameters and recording seeds, with aspect fixed at 1.32 or uniform on [1.15, 1.6].

Onset-aligned, unit-RMS audio is compared by $D(x, y) = \sqrt{\text{mean}_{\mathcal{M}}[(A_x - A_y)^2]}$, where $A_x = 10 \log_{10}(P_x + 10^{-8})$ uses mel power P_x from 64 Slaney bands (20 Hz–16 kHz), 2048-sample Hann windows and 512-sample hops at 44.1 kHz. Before selecting on-target pairs, shared mask $\mathcal{M}$ retains bins within 40 dB of the 12-output mean-power map’s peak (four per arm). Pair distances are averaged within each request and run. PXS-Physical and both simulator arms yield $1.5\times$ -accurate pairs in 323 of 432 request–run evaluations. Class-median distances are 10.27 dB for PXS-Physical, 7.15 dB for fixed-aspect simulation and 15.03 dB for varied-aspect simulation.

Physical responses. We vary E, L, η, v, κ individually over five levels on hits and bounces across the three held-out materials (32 paired-noise outputs per setting and run). Log-spaced speed and κ span 0.21–4.75 m/s and 1.05–15.2, respectively. Model-to-simulator log–log slope ratios measure directional responses (ideal 1). Simulator qualification requires at least two curves with ≥ 3 valid levels, mean direction-corrected Kendall $\tau \geq 0.5$, mean log–log fit

Table 4. Metrics: VGGish Fréchet audio distance (FAD) [23]; Kullback–Leibler (KL) divergence of pretrained audio neural networks (PANNs) [24] softmax outputs (reference to generated, nats); contrastive language–audio pretraining (CLAP) [25] audio–text cosine for material-and-strike descriptions. References are 16 independent renders per θ^* or real target recordings. SAO captions specify material, plate and strike.

Generator	FAD ↓		CLAP ↑		KL _{r g} ↓
	Sim. ref.	Real ref.	Semantic	Sim. ref.	
Simulator	0.13	4.23	0.486	0.070	
Simulator + VAE	1.13	3.62	0.456	0.436	
PXS-Physical	4.05	3.43	0.433	0.709	
PXS-Text	2.67	2.82	0.437	0.925	
SAO q	4.16	2.37	0.404	2.896	
SAO η^*	4.10	2.40	0.415	2.865	

$R^2 \geq 0.8$ and absolute mean slope ≥ 0.02 . Hardness and speed use attack centroid (first 30 ms); size uses f_{scale} , κ uses T_{60} and η uses h_c . PXS-Physical’s speed and κ ratios span 0.72–1.41 and 0.89–1.04 (five and four of six material–contact combinations qualify, respectively). With matched training and inputs, either smooth encoding or the direct path preserves hardness and damping responses; removing both weakens them (Table 2).

3.2. Conditioning interfaces

Trained text. PXS-Text learns from captions such as “a glass plate of size 0.30 meters struck at 1.00 meters per second with a striker of hardness 10.0 gigapascal, material loss factor 0.0006”.

Zero-shot generators. External numeric prompts use “{event}, material loss factor {q}”, inserting raw q without inversion and preserving distinct rounded target values. Ordinal prompts map ordered target groups to evenly spaced, rounded indices among “ringing and resonant”, “lightly damped”, “damped”, “heavily damped” and “completely muted”. SAO 1.0 uses 100 steps, CFG 7 and negative prompt “low quality”. TangoFlux uses 50 steps and CFG 4.5. Both request 2 s with four shared seeds. We retain the first 2 s in mono, matching corpus bandwidth and –30 dB RMS. PXS-Physical leads eight failure-aware ordering comparisons under this protocol (Table 3).

We test timing sensitivity with predefined 2-second energy windows from 10-second outputs. With matched plate captions, SAO completes all 18 impact classes (correlations –0.085 for q and –0.158 for η^*). TangoFlux completes 14/18 numeric and 10/18 ordinal classes, with complete-class mean correlations –0.534 and 0.535.

SonicGauss [8] receives geometry and contact coordinates, while PXS-Physical uses reference damping and proxy plates. We compare four energy-selected 2-s outputs at each of two validation positions on four ObjectFolder-Real objects [26] (one PXS run). Errors average within positions, then objects, using 30 valid SonicGauss and 32 PXS outputs: 0.440 versus 0.171 dex for damping and 0.366 versus 0.520 dex for frequency.

Audio distribution and semantic alignment. Table 4 compares 576 outputs per model at matched duration, RMS and bandwidth, using three-run PXS averages and fixed-aspect simulation. With four-output medians, VAE reconstruction raises simulator class-median error to real targets from 0.032 to 0.047 dex but lowers real-reference FAD.

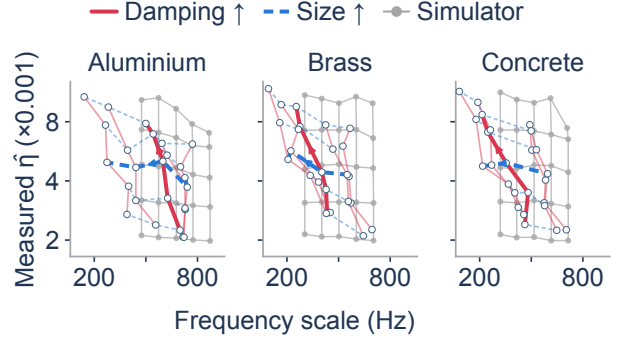


Fig. 2. Damping–size grid (log axes). Open points geometrically average PXS-Physical four-output medians over three runs. Grey points are eight-output simulator medians. Arrows increase damping (red) or size (blue). Bold lines fix the other control at its midpoint; faint lines show full grids.

3.3. Joint control

A 5×5 grid (Fig. 2) spans η $1.7\text{--}10.8 \times 10^{-3}$ and L 0.13–0.32 m on the three held-out materials. Before generation, all materials pass simulator checks: ≥ 6 valid readings from eight outputs per cell, strictly monotone axes spanning at least half the requested log range, and ≥ 0.05 dex separation in $(\log_{10} q, \log_{10} f_{\text{scale}})$. Training-format captions specify material, L, v, E, η . Physical inputs add material constants, $h/L = 0.008$ and contact geometry. Both follow requested directions (slope ratios 0.82–0.98).

Joint success requires at least two of four outputs within $1.5 \times$ of both eight-output simulator medians. PXS-Physical leads PXS-Text in every run: 68.0 % versus 15.6 % over 225 setting–run evaluations. Frequency error is lower (0.126 versus 0.372 dex), while damping errors are similar (0.055 versus 0.053 dex). Cross-talk slopes (dex/dex) for PXS-Physical and PXS-Text are –0.26 and –0.33 for frequency versus loss, and +0.34 and +0.26 for damping versus size (simulator: zero and +0.08).

Masking only h/L at fixed weights and noise lowers success in every run (27.6 % overall), despite a size response of 0.86. Errors are 0.245 dex for frequency and 0.058 dex for damping. With 32-output simulator medians, supplying thickness raises $1.5 \times$ joint success from 51.1 % to 80.0 % and from 57.8 % to 95.6 % at $h/L = 0.016$ and 0.032 (corner and centre settings), with gains in all three runs.

Room text and size. PXS-Physical combines five sizes with dry, reverberant or no room text for brass and concrete ($\eta = 0.04$, 360 shared-noise impacts). Reverberant versus dry text raises onset-aligned tail-to-attack energy (80–600 versus 0–80 ms) in all 30 comparisons, with context medians of 5.75–10.10 dB. All 18 size slopes remain negative, with frequency-scale shifts of +0.097–0.240 dex.

4. CONCLUSION

Simulation-trained PXS provides direct-parameter and numerical-caption interfaces that deliver effective-damping targets from real impacts without real-audio fine-tuning. Joint tests distinguish directional response from meeting two acoustic targets: specifying thickness improves frequency accuracy and joint delivery. Parameter inputs combine with room descriptions, while repeated sampling produces spectrally distinct sounds at fixed targets. Future work targets sustained range compression, cross-talk and perceptual evaluation.

5. ACKNOWLEDGEMENTS

This work was supported by the UKRI AI Centre for Doctoral Training in AI for Digital Media Inclusion [grant number EP/Y030915/1] and by the Engineering and Physical Sciences Research Council (EPSRC) [grant number EP/Y028805/1]. For the purpose of open access, the authors have applied a Creative Commons Attribution (CC BY) licence to any Author Accepted Manuscript version arising from this submission.

6. REFERENCES

- [1] Zach Evans, Julian D. Parker, CJ Carr, Zack Zukowski, Josiah Taylor, and Jordi Pons, “Stable Audio Open,” *arXiv preprint arXiv:2407.14358*, 2024.
- [2] Chia-Yu Hung, Navonil Majumder, Zhifeng Kong, Ambuj Mehrish, Amir Zadeh, Chuan Li, Rafael Valle, Bryan Catanzaro, and Soujanya Poria, “TangoFlux: Super fast and faithful text to audio generation with flow matching and CLAP-ranked preference optimization,” in *Proc. ICLR*, 2026.
- [3] Yoonjin Chung, Junwon Lee, and Juhan Nam, “T-FOLEY: A controllable waveform-domain diffusion model for temporal-event-guided foley sound synthesis,” in *Proc. IEEE ICASSP*, 2024.
- [4] Hugo Flores García, Oriol Nieto, Justin Salamon, Bryan Pardo, and Prem Seetharaman, “Sketch2Sound: Controllable audio generation via time-varying signals and sonic imitations,” in *Proc. IEEE ICASSP*, 2025.
- [5] Zachary Novack, Zack Zukowski, CJ Carr, Julian Parker, Zach Evans, Josiah Taylor, Taylor Berg-Kirkpatrick, Julian McAuley, and Jordi Pons, “Low-resource guidance for controllable latent audio diffusion,” in *Proc. IEEE ICASSP*, 2026.
- [6] Rotem Ezra, Hedi Zisling, Nimrod Berman, Ilan Naiman, Alexey Gorkor, Liran Nochumsohn, Eliya Nachmani, and Omri Azencot, “FreeSliders: Training-free, modality-agnostic concept sliders for fine-grained diffusion control in images, audio, and video,” *arXiv preprint arXiv:2511.00103*, 2025.
- [7] Kun Su, Kaizhi Qian, Eli Shlizerman, Antonio Torralba, and Chuang Gan, “Physics-driven diffusion models for impact sound synthesis from videos,” in *Proc. IEEE/CVF CVPR*, 2023, pp. 9749–9759.
- [8] Chunshi Wang, Hongxing Li, and Yawei Luo, “SonicGauss: Position-aware physical sound synthesis for 3D gaussian representations,” in *Proc. ACM MM*, 2025, pp. 10886–10895.
- [9] Chen Yang, Ganye Wen, Bin Huang, Jiayi Lyu, Zehai Niu, Linlin Shen, and Jinbao Wang, “FillGauss: Fine-grained filling-aware impact sound generation for 3D gaussian splatting,” *arXiv preprint arXiv:2607.17773*, 2026.
- [10] Hyun-Bin Oh, Yuhta Takida, Toshimitsu Uesaka, Tae-Hyun Oh, and Yuki Mitsufuji, “PAVAS: Physics-aware video-to-audio synthesis,” in *Proc. IEEE/CVF CVPR*, 2026, pp. 14481–14491.
- [11] Yisu Zong and Joshua D. Reiss, “Learning control of neural sound effects synthesis from physically inspired models,” in *Proc. IEEE ICASSP*, 2025.
- [12] Tingle Li, Siddharth Gururani, Kevin J. Shih, Gantavya Bhatt, Sang gil Lee, Zhifeng Kong, Arushi Goel, Gopala Anumanchipalli, and Ming-Yu Liu, “Benchmarking single-factor physical video-to-audio generation,” in *Proc. IEEE/CVF CVPR*, 2026, pp. 1939–1949.
- [13] Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le, “Flow matching for generative modeling,” in *Proc. ICLR*, 2023.
- [14] Hyung Won Chung, Le Hou, Shayne Longpre, et al., “Scaling instruction-finetuned language models,” *J. Mach. Learn. Res.*, vol. 25, no. 70, pp. 1–53, 2024.
- [15] Jonathan Ho and Tim Salimans, “Classifier-free diffusion guidance,” *arXiv preprint arXiv:2207.12598*, 2022.
- [16] Karol J. Piczak, “ESC: Dataset for environmental sound classification,” in *Proc. ACM Multimedia*, 2015, pp. 1015–1018.
- [17] Eduardo Fonseca, Xavier Favory, Jordi Pons, Frederic Font, and Xavier Serra, “FSD50K: An open dataset of human-labeled sound events,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 30, pp. 829–852, 2022.
- [18] Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter, “Audio Set: An ontology and human-labeled dataset for audio events,” in *Proc. IEEE ICASSP*, 2017, pp. 776–780.
- [19] Jesse Engel, Cinjon Resnick, Adam Roberts, Sander Dieleman, Mohammad Norouzi, Douglas Eck, and Karen Simonyan, “Neural audio synthesis of musical notes with WaveNet autoencoders,” in *Proc. ICML*, 2017, vol. 70 of *Proceedings of Machine Learning Research*, pp. 1068–1077.
- [20] Samuel Clarke, Ruohan Gao, Mason Wang, Mark Rau, Julia Xu, Jui-Hsien Wang, Doug L. James, and Jiajun Wu, “ReallImpact: A dataset of impact sound fields for real objects,” in *Proc. IEEE/CVF CVPR*, 2023, pp. 1516–1525.
- [21] Karolina Prawda, Sebastian J. Schlecht, and Vesa Välimäki, “Calibrating the Sabine and Eyring formulas,” *J. Acoust. Soc. Am.*, vol. 152, no. 2, pp. 1158–1169, 2022.
- [22] M. R. Schroeder, “New method of measuring reverberation time,” *J. Acoust. Soc. Am.*, vol. 37, no. 3, pp. 409–412, 1965.
- [23] Kevin Kilgour, Mauricio Zuluaga, Dominik Roblek, and Matthew Sharifi, “Fréchet Audio Distance: A reference-free metric for evaluating music enhancement algorithms,” in *Proc. Interspeech*, 2019, pp. 2350–2354.
- [24] Qiuqiang Kong, Yin Cao, Turab Iqbal, Yuxuan Wang, Wenwu Wang, and Mark D. Plumbley, “PANNS: Large-scale pre-trained audio neural networks for audio pattern recognition,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 2880–2894, 2020.
- [25] Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov, “Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation,” in *Proc. IEEE ICASSP*, 2023.
- [26] Ruohan Gao, Yiming Dou, Hao Li, Tanmay Agarwal, Jeanette Bohg, Yunzhu Li, Li Fei-Fei, and Jiajun Wu, “The ObjectFolder benchmark: Multisensory learning with neural and real objects,” in *Proc. IEEE/CVF CVPR*, 2023, pp. 17276–17286.