TRANSFER ASYMMETRY IN VEHICLE PASS-BY AUDIO SYNTHESIS: A BIDIRECTIONAL EVALUATION OF PASSBYDIT AND DOPPLERSIM

Anonymous authors
Paper under double-blind review

ABSTRACT

Synthetic audio offers a potential remedy for scarce labeled vehicle pass-bys, but acoustic realism alone does not establish its usefulness as training data. We study this distinction through transfer asymmetry in vehicle speed estimation. We introduce PassByDiT, a physics-conditioned diffusion transformer, and DopplerSim, a source-spectrum recovery and rendering pipeline, and compare both with Pyroadacoustics under matched pass-by conditions. PassByDiT achieves the closest agreement with real recordings across most spectral and temporal measures, while the physics simulators align energy peaks more precisely with the requested closest-point-of-approach timing. A real-trained speed estimator achieves similar mean absolute errors on real and PassByDiT audio (7.05 and 7.80 km/h, respectively), indicating that PassByDiT preserves motion cues usable by a model trained on real recordings. PassByDiT also provides the strongest synthetic-to-real speed-estimation performance among the evaluated synthesis methods. Bidirectional evaluation nevertheless reveals transfer asymmetry: agreement observed when testing a real-trained estimator on synthetic audio does not imply equivalent training utility in the reverse direction. Together, these findings establish complementary strengths of learned and physics-based synthesis and distinguish motion control, acoustic agreement, and training utility as separate dimensions of evaluation.

1 INTRODUCTION

The sound of a passing vehicle evolves with its motion, carrying information about speed, distance, and trajectory. As the vehicle approaches and recedes, radial velocity produces a changing Doppler shift, while distance and source behavior shape the received level and spectrum. These cues allow roadside microphones to support vehicle detection, classification, localization, and speed estimation without relying on visual observations (Szwoch & Kotus, 2021; Abeßer et al., 2021; Rocha et al., 2021; Djukanović et al., 2021). Learning to infer motion from those cues, however, requires diverse recordings paired with reliable speed and closest-point-of-approach (CPA) measurements.

Such data is difficult to collect. Motion annotations require reference sensors synchronized with the audio, and roadside acquisition introduces equipment, access, safety, and logistical constraints. Coverage must also span vehicles, speeds, roads, and recording geometries if a model is to generalize beyond a narrow setup. Publicly available corpora with reliable motion annotations therefore remain small; VS13, for example, contains just 400 recordings from 13 vehicles (Djukanović et al., 2022). The central problem is not whether sound contains motion information, but whether enough labeled real audio can be obtained to learn that relationship reliably.

Synthetic pass-bys offer a way to expand this training data. Existing approaches follow two broad routes: physical renderers apply motion and propagation to a source signal, while learned generators model acoustic structure from recordings (Pieren et al., 2015; Kong et al., 2021; Liu et al., 2023). Each route solves only part of the problem. Pyroadacoustics provides explicit trajectory control but depends on a supplied waveform; changing the trajectory does not create the engine and tire behavior associated with a different operating speed, and using a recorded pass-by retains motion and

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.

environmental effects already present in that recording (Damiano & van Waterschoot, 2022). General audio generators learn richer sound structure, but their standard conditioning does not directly specify numerical speed and CPA timing or establish that these quantities remain recoverable from the output (Kong et al., 2021; Liu et al., 2023; Zhang et al., 2024). Existing synthesis can therefore produce controlled or plausible audio without yet establishing that the result is suitable supervision for real-world motion inference.

Our premise is that synthetic pass-bys reduce dependence on scarce real training data only when they preserve the relationship between sound and speed needed on real recordings. We pursue two complementary solutions. DopplerSim uses the original recording geometry to recover an intrinsic source spectrum, synthesizes multiple emitter signals, and propagates them under a target trajectory. PassByDiT learns recorded acoustic structure with a diffusion transformer conditioned on vehicle metadata and frame-wise motion signals describing CPA timing, Doppler factors, and amplitude envelopes. DopplerSim separates source recovery from motion rendering, while PassByDiT generates from numerical motion conditions without requiring a source recording at inference.

We assess whether the generated pass-bys follow the requested motion timing, resemble real recordings, and provide useful training data for speed estimation. We compare DopplerSim, PassByDiT, and Pyroadacoustics on matched VS13 conditions, training and testing speed estimators across all real and synthetic domains. Following prior work on evaluating generated data (Esteban et al., 2017; Shmelkov et al., 2018), we examine transfer in both directions: from real training data to synthetic test audio, and from synthetic training data to real test audio. The methods show complementary strengths. DopplerSim and Pyroadacoustics place energy peaks closer to the requested CPA, while PassByDiT achieves the closest agreement with real recordings on most acoustic measures. A real-trained estimator also performs similarly on PassByDiT and real audio. The reverse direction is more challenging: estimators trained on synthetic audio perform well within their respective domains but lose accuracy when tested on real recordings. We describe this dependence on transfer direction as *transfer asymmetry*.

Our contributions are:

- Two complementary approaches to motion-conditioned vehicle pass-by synthesis: DopplerSim, which recovers a source spectrum and renders multiple moving emitters, and PassByDiT, which generates audio through physics-conditioned diffusion.
- A matched comparison with real VS13 recordings and Pyroadacoustics that evaluates motion timing, acoustic agreement, and synthetic-to-real transfer for speed estimation, revealing an asymmetry between transfer directions.

The central finding is that motion control and acoustic agreement alone do not establish the value of synthetic training data. That value must also be assessed through the performance of models trained on generated audio and tested on real pass-bys.

2 Related work

Source representations for controllable pass-by synthesis. Rendering a vehicle under a new trajectory requires a source representation that can be separated from the motion in the original recording. Existing vehicle auralization methods approach this through spectral source models or separate rolling and propulsion components (Pieren et al., 2015; Damiano et al., 2024). Such representations allow source construction and propagation to be controlled separately. Pyroadacoustics provides a flexible propagation framework for a supplied waveform (Damiano & van Waterschoot, 2022), leaving the choice of excitation to the user. When that excitation is a recorded roadside pass-by, it already contains Doppler variation, distance-dependent attenuation, and environmental effects. Applying target motion therefore requires attention to those existing effects; this is a source-preparation issue rather than a limitation of propagation alone. DopplerSim addresses this setting by approximately compensating for the original motion, estimating a source spectrum, and resynthesizing emitter signals before rendering the target trajectory. Its contribution lies in this source-recovery and rendering pipeline.

Learning acoustic structure under numerical motion control. An explicit renderer provides a direct relationship between trajectory and received sound, but reproducing the acoustic detail of

recorded vehicles also depends on the source model. Learned audio generation offers a complementary approach. DiffWave and AudioLDM demonstrate how diffusion can capture acoustic structure from recordings (Kong et al., 2021; Liu et al., 2023), while FoleyCrafter illustrates the value of controlling event timing alongside sound content (Zhang et al., 2024). For labeled vehicle pass-bys, however, the required controls are numerical speed and CPA geometry, which must shape the sound throughout the event. PassByDiT draws on diffusion transformers and FiLM conditioning (Peebles & Xie, 2023; Perez et al., 2018) to introduce motion-derived temporal signals alongside vehicle metadata. This design allows the model to learn recorded acoustic detail while receiving an explicit description of the requested motion. Whether it uses that description effectively remains an empirical question, since conditioning does not enforce physical consistency.

From generated audio to useful supervision. The motivation for these synthesis approaches is the limited availability of recordings with motion labels. Acoustic speed-estimation research and VS13 establish a setting in which roadside audio is paired with reference speeds (Djukanović et al., 2021; 2022). Generating more labeled examples could support this task, but acoustic agreement and training usefulness answer different questions. Fréchet audio distance compares audio distributions (Kilgour et al., 2019); it does not directly test whether a model trained on generated examples predicts speed accurately from real recordings. Bidirectional task evaluation provides that distinction by testing both real-trained models on synthetic data and synthetic-trained models on real data (Esteban et al., 2017; Shmelkov et al., 2018). Evidence of synthetic-to-real domain gaps in sound-event detection further motivates evaluating training utility directly (Zheng et al., 2021). We adopt this established evaluation principle to examine whether the motion and acoustic properties reproduced by each generator provide supervision that transfers to real pass-bys.

Positioning of this work. Our work connects source preparation, motion-conditioned generation, and downstream training utility within a matched vehicle-audio study. DopplerSim and PassByDiT offer explicit and learned approaches to constructing pass-bys under requested motion conditions, with Pyroadacoustics serving as a waveform-driven propagation comparison. The evaluation then tests whether their strengths in timing and acoustic agreement are accompanied by effective speed-estimation training. This connection is the focus of the study: determining which properties of generated pass-bys remain useful when those recordings become the training data.

3 VEHICLE PASS-BY SYNTHESIS

Given vehicle information, speed s, and closest-point-of-approach time t_{CPA}, we generate a pass-by under a specified geometry. CPA denotes geometric closest approach and need not coincide with the recorded energy peak. DopplerSim estimates a source spectrum from a reserved recording and renders new emitter signals under the target trajectory. PassByDiT generates a spectrogram conditioned on vehicle metadata and frame-wise motion signals. Source selection and training partitions are specified in Section 4.

3.1 DOPPLERSIM: SOURCE-SPECTRUM RECOVERY AND RENDERING

DopplerSim uses the original recording geometry to approximately compensate for Doppler shift and distance attenuation before synthesizing the target pass-by. Its inputs are a source recording, the original and target motion geometries, receiver position, vehicle length, and emitter count. Figure 1 summarizes the procedure.

Trajectory and propagation. We represent the vehicle by N emitters distributed uniformly along its length. For emitter offset o_k, speed $v = s/3.6$, lateral distance h, and vehicle-centre CPA time t_{CPA}, position and range at emission time u are

$$x_k(u) = v(u - t_{\mathrm{CPA}}) + o_k, \qquad R_k(u) = \sqrt{x_k(u)^2 + h^2}. \tag{1}$$

Sound emitted at time u_k arrives at observer time t:

$$t = u_k + \frac{R_k(u_k)}{c}, \qquad c = 343\,\mathrm{m/s}. \tag{2}$$

Solving this relation at each observer sample gives the retarded time $u_k(t)$. Its variation with t produces the Doppler shift.

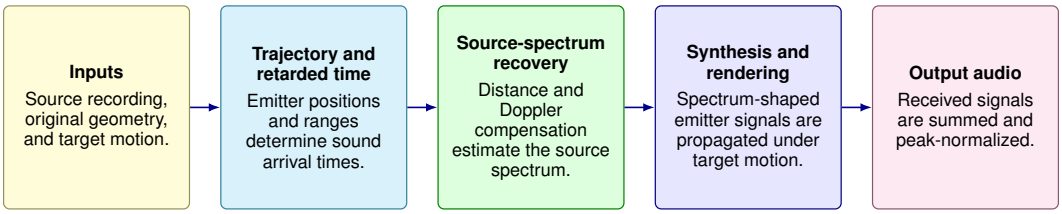

Figure 1: DopplerSim estimates a source spectrum using the original geometry, then synthesizes and propagates emitter signals under the target geometry.

Source-spectrum recovery. The recording is transformed into a time–frequency representation Y. Using the original geometry, we compensate for pressure spreading and remap observed frequencies:

$$A_{k,i} = R_k(u_k(t_i))\,|Y(f_{\text{obs}}, t_i)|, \qquad f_{\text{src}} = \frac{f_{\text{obs}}}{D_k(t_i)}, \tag{3}$$

where $D_k(t_i)$ is the emitter's Doppler factor. The corrected power spectra are interpolated onto a common frequency grid and averaged over valid frames and emitters to estimate the source PSD $\widehat{P}(f)$. This is a time-averaged spectral estimate, not a separation of individual sources; it discards the recording's phase and temporal texture.

Target rendering. The estimated PSD shapes an independent noise signal s_k for each emitter. Recomputing retarded time and range under the target geometry gives

$$y(t) = \frac{1}{\sqrt{N}} \sum_{k=1}^{N} \frac{s_k(u_k(t))}{R_k(u_k(t))}. \tag{4}$$

Retarded-time sampling produces the target Doppler shift, and $1/R_k$ applies distance attenuation. The factor $1/\sqrt{N}$ scales the independent source contributions. Individual source normalization and final peak normalization mean that the output preserves relative envelope shape rather than calibrated sound-pressure level.

The benchmark uses 100 emitters and a nominal target distance of 0.5 m. Section 4 describes the evaluated files and exclusions; Appendix B provides implementation and provenance details.

3.2 PassByDiT: Physics-conditioned diffusion

PassByDiT conditions a diffusion transformer on vehicle metadata and temporal motion signals. Three effective source positions, centre, front, and rear, describe motion across the vehicle without resolving individual physical radiators. Figure 2 shows the architecture.

Motion representation. For offsets $o \in \{0, +\ell_f, -\ell_r\}$, speed $v = s/3.6$, frame time τ, and lateral distance d, we compute

$$x_o(\tau) = v(\tau - t_{\text{CPA}}) + o, \qquad r_o(\tau) = \sqrt{d^2 + x_o(\tau)^2},$$
$$D_o(\tau) = \frac{1}{1 + (v/c)x_o(\tau)/r_o(\tau)}, \qquad A_o(\tau) = \frac{1}{r_o(\tau)}. \tag{5}$$

The relative CPA time is $\tau_o = \tau - t_{\text{CPA}} + o/v$. For the positive speeds evaluated here, the front source reaches CPA before the centre, followed by the rear source.

These quantities form eight conditioning channels:

$$P = [\tau_c, \tau_f, \tau_r, A_c, A_{\text{env}}, D_c, D_f - D_r, A_r - A_f]^\top, \quad A_{\text{env}} = 0.4A_c + 0.3A_f + 0.3A_r. \tag{6}$$

The subscripts c, f, r identify the source positions. The channels encode event progression, amplitude variation, centre-source Doppler shift, and differences across the vehicle. They are evaluated on the frame-time grid and guide generation without enforcing exact propagation.

Each channel is standardized using training-split statistics. A ten-dimensional metadata vector contains scaled speed and CPA time, engine and transmission codes, standardized power and production year, and a four-way body-class indicator.

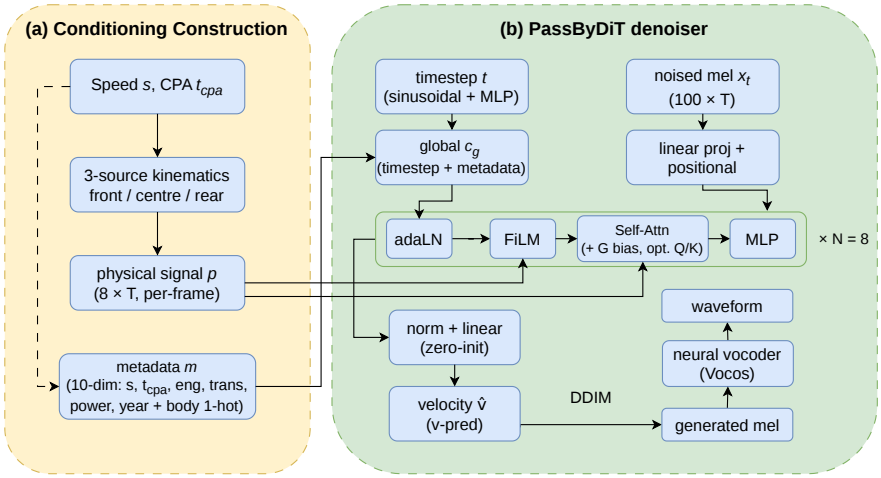

Figure 2: PassByDiT generates a log-mel spectrogram from vehicle metadata and frame-wise motion signals. Vocos decodes the spectrogram into a waveform.

Conditioning architecture. Each log-mel frame is projected to a 384-dimensional token and processed by eight transformer blocks with twelve attention heads. Global metadata modifies the diffusion-timestep embedding and controls adaptive layer normalization. A convolutional physics encoder supplies frame-wise FiLM scales and shifts, while a learned physics affinity modifies attention:

$$\ell_{ij}^{(h)} = \frac{q_i^{(h)\top} k_j^{(h)}}{\sqrt{d_h}} + G_{ij} + g_h \, \widehat{\phi}_i^{(h)\top} \widehat{\phi}_j^{(h)}. \tag{7}$$

Here, i, j index frames, h indexes heads, and d_h is the head dimension. The evaluated configuration, Model M, uses metadata, FiLM, and learned affinity, with the optional analytic bias disabled ($G_{ij} = 0$).

The physics vectors $\widehat{\phi}$ have unit norm, so their inner product lies in $[-1, 1]$. The gate g_h is initialized to zero and learned without constraints; the complete affinity term is therefore not bounded by this normalization.

Training and inference. Given a clean spectrogram x_0, noise ϵ, and cumulative noise schedule $\bar{\alpha}_t$, the noised input and diffusion-velocity target are

$$x_t = \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon, \qquad v_t = \sqrt{\bar{\alpha}_t}\epsilon - \sqrt{1 - \bar{\alpha}_t} x_0. \tag{8}$$

Here, t indexes diffusion steps, and v_t is distinct from vehicle velocity. Training minimizes mean squared error between the predicted and target v_t (Salimans & Ho, 2022), using a cosine schedule with 1,000 steps (Nichol & Dhariwal, 2021).

At inference, the model generates a spectrogram through 50 deterministic DDIM steps (Song et al., 2021). Vocos decodes the 100-bin log-mel output into a waveform at 24 kHz (Siuzdak, 2024). Appendix A gives the training settings, and Appendix L reports the earlier conditioning ablations.

4 Evaluation Protocol

4.1 Evaluation Questions

EQ1. Does training on synthetic audio instead of real audio reduce speed-estimation performance on real recordings? We answer this by comparing the real-test errors of synthetic-trained and real-trained estimators.

EQ2. How does an estimator trained only on synthetic audio perform on an unseen vehicle relative to matched real training and a constant predictor?

EQ3. How do simulated pass-bys compare with real ones across spectrum, envelope, timing, and loudness? Significant per-metric gaps identify where a simulator's approximation breaks down.

4.2 System implementation

To test these questions we need a common estimator and a shared data pipeline; we describe the estimator here and the data in the next subsection. We train an SE-ResNet speed estimator with channel recalibration per domain (Hu et al., 2018), using AdamW, batch size 16, up to 150 epochs, patience 30, and learning rate 5×10^{-4} with cosine decay, normalized only on each estimator's own training partition. PassByDiT is fit on 24 kHz recordings using Vocos features, with output converted to 16 kHz for evaluation; the estimator receives 128-bin log-mel arrays (FFT 2,048, hop 512). Remaining settings, including seeds, are in Appendix A.

4.3 Data

This estimator and simulator pipeline draws on two datasets. VS13 contains 400 vehicle pass-by recordings from 13 vehicle models. The benchmark uses 378 after excluding nine recordings outside the 9.98–10.10 s duration window and reserving one recording per vehicle at 60–63 km/h as simulator source material. These 13 recordings are excluded from all fitting, validation, and test targets because they retain their own envelope, Doppler shift, and environmental cues. DopplerSim estimates an intrinsic spectrum from each source and resynthesizes 100 emitters under the target geometry, while Pyroadacoustics propagates the reserved waveform directly (Damiano & van Waterschoot, 2022).

Each pairing fixes vehicle, speed, CPA time, duration, and sample rate; the 0.5 m lateral offset and CPA timestamp are nominal rather than independently measured. All files are standardized to mono, 16 kHz, and 10 s. We use five vehicle-stratified folds with 76/76/76/75/75 test conditions, 61 validation conditions, and 241/241/241/242/242 training conditions. For every train–test pairing, a fresh estimator is fit on that fold's training partition, with synthetic examples produced only from the corresponding fold-specific generator. All 13 vehicles recur across folds, so this measures unseen conditions from known vehicles rather than unseen-vehicle generalization; the leave-one-vehicle-out experiment below addresses that case directly.

4.4 Baselines and comparisons

Against this data, we define the comparisons each question requires. The main comparison is a frozen 16-pairing transfer matrix across real, PassByDiT, DopplerSim, and Pyroadacoustics, with 378 pooled predictions per cell and identical folds, architecture, and per-cell normalization. For unseen-vehicle evaluation, we compare against a constant baseline that predicts the median training-fold speed.

4.5 Metrics and evaluation protocol

Alongside these speed-estimation comparisons, EQ3 calls for acoustic metrics rather than estimator error. Eleven such metrics compare generated and matched real recordings: normalized PSD, one-third-octave error, MR-STFT distance, envelope correlation, Wasserstein distance, duration, pre/post-CPA balance, peak timing relative to both the real peak and CPA, centroid, loudness, and CLAP distance. Definitions are in Appendix E.

Acoustic metric differences are averaged within vehicle and tested using a two-sided exact sign-flip test over all $2^{13} = 8{,}192$ sign assignments, with Holm correction across 33 comparisons (Holm, 1979). Transfer-matrix confidence intervals use 1,000 paired group-bootstrap draws over condition groups and reflect condition-sampling uncertainty under the fitted models. The unseen-vehicle experiment uses an exact sign test over 13 paired PassByDiT-versus-real differences.

4.6 Experiments

The transfer matrix answers EQ1 by comparing each training source's in-domain error with its error on real recordings. The 13-fold leave-one-vehicle-out experiment answers EQ2 by withholding

one vehicle from both PassByDiT and estimator training, then comparing synthetic training with matched real training and the constant predictor. The acoustic metrics answer EQ3 by identifying which properties of generated pass-bys match real recordings and where systematic gaps remain.

Appendix L reports a separate PassByDiT conditioning ablation under an earlier split and frontend.

5 Results and Discussion

5.1 Results for Each Evaluation Question

EQ1: Can synthetic audio replace real training audio? The transfer matrix answers this question negatively (Table 1). Every estimator trained on a synthetic domain performs substantially worse on real recordings than the real-trained estimator, despite achieving low error within its own domain. Among the synthetic training sources, PassByDiT transfers most effectively, cutting real-test error by roughly 43% compared with Pyroadacoustics and 59% compared with DopplerSim. The reverse direction is less demanding for PassByDiT: features learned from real recordings remain useful on its generated audio. This directional difference is the central transfer asymmetry.

Table 1: Five-fold transfer MAE (km/h), 378 held-out predictions per cell. Rows are estimator training domains, columns are test domains.

Train	Real	PassByDiT	DopplerSim	Pyroad
Real	7.047	7.803	24.408	39.259
PassByDiT	14.511	2.538	38.304	76.457
DopplerSim	35.734	34.544	2.034	39.103
Pyroad	25.593	25.110	28.872	4.574

EQ2: Does synthetic training generalize to an unseen vehicle? The leave-one-vehicle-out point estimates place PassByDiT between matched real training and the constant predictor (Table 2). Its apparent improvement over the constant is descriptive because no paired significance test for that comparison is reported. Relative to real training, the paired interval is wide and the sign test detects no reliable difference across vehicles. The experiment therefore shows that some speed information transfers to unseen vehicles, but does not establish parity with real training.

Table 2: Leave-one-vehicle-out summary. MAE and differences are in km/h; averaging is over 13 held-out vehicles.

Quantity	Value
Real-trained macro MAE	10.91
PassByDiT-trained macro MAE	15.23
Training-median macro MAE	17.97
PassByDiT minus real	4.31
Vehicle-bootstrap 95% interval	$[-3.85, 16.16]$
Exact sign-test p	1.0

EQ3: How closely do generated pass-bys match real recordings? The answer depends on which property is measured. DopplerSim and Pyroadacoustics place their energy peaks closest to the requested CPA, showing precise timing control, but their events are much narrower than recorded pass-bys. PassByDiT is less precise at the requested timestamp but is closer to real audio on most spectral and temporal measures (Table 3). The corrected paired tests support this advantage on nine of eleven metrics against each simulator; full effects and adjusted probabilities are reported in Appendix J. No method therefore maximizes both control fidelity and recorded-event agreement.

5.2 Component Contributions

The earlier conditioning ablation isolates acoustic contributions under a separate split and frontend (Appendix L). FiLM provides the clearest improvement in CPA-region reconstruction and temporal

Table 3: Selected acoustic comparisons with matched real recordings. Lower is better except envelope correlation.

Metric	PassByDiT	DopplerSim	Pyroad
Peak-to-CPA error (s)	0.2190	0.0090	0.0087
PSD-shape MAE (dB)	3.0247	5.4090	4.5774
Envelope correlation	0.8521	0.5754	0.5701
Energy-duration error (s)	1.2444	4.2282	4.2356
MR-STFT distance	1.5206	3.3272	4.7275

alignment, while learned affinity adds only a small further change. Because that experiment is not compatible with the current transfer matrix, it supports conclusions about acoustic reconstruction rather than replacement performance.

The frontend-reconstruction control asks whether the mel/Vocos pipeline alone explains the transfer gap. It produces a much smaller penalty than full PassByDiT training, and matching the reconstruction pipeline does not close the gap (Appendix G). Frontend shift is therefore a contributor, but not a sufficient explanation. Neither control isolates the remaining cause of the synthetic-to-real error.

5.3 Additional Findings

A second completed generator and estimator fit preserves both the direction and approximate size of the PassByDiT replacement gap, supporting repeatability across the completed runs without characterizing broader retraining variance (Appendix G). Acoustic similarity does not determine replacement performance: frozen embeddings place DopplerSim closer to real audio than Pyroadacoustics, whereas the transfer matrix reverses their ordering on real-test error. PassByDiT ranks best under both views, but the disagreement between the simulators shows that distributional proximity alone cannot rank training utility. The PassByDiT gap also varies more clearly by vehicle than by speed band: it is non-monotonic across speed, while a small number of vehicles account for much larger excess errors. Lower correlation with true speed and a positive prediction bias show that this pattern is not merely a constant offset; Appendix I reports the complete strata.

5.4 Qualitative Analysis and Failure Cases

The simulator outputs concentrate energy into short, high-crest-factor events rather than reproducing the gradual shoulders of recorded pass-bys (Appendix Table 5). Increasing DopplerSim's lateral distance broadens the event and improves acoustic agreement, but does not recover the real envelope shape (Appendix C). Adding a stationary ambient floor fails: it reduces envelope concentration while worsening spectral and temporal agreement (Appendix D). These cases suggest that geometry can partially correct event width, whereas a flat noise floor cannot replace missing time-varying source structure. They diagnose concrete export behavior but do not by themselves establish the cause of the transfer gap.

5.5 Overall Interpretation

The results support the paper's premise that synthetic data must preserve the sound–speed relationship needed on real recordings, not merely produce controllable or acoustically similar outputs. PassByDiT comes closest to real audio and transfers best among the synthetic domains, yet it still does not match real-data training. The physics simulators sharpen the distinction: precise CPA control and low in-domain error coexist with poor replacement performance.

The demonstrated conclusion is the asymmetric and vehicle-structured transfer gap. The proposed mechanisms are less certain. Frontend shift explains only part of the gap, and the DopplerSim diagnostics implicate missing temporal structure, but neither analysis establishes a complete causal account. The comparison also concerns specific source-construction pipelines rather than physics-based synthesis in general.

5.6 LIMITATIONS

The main folds contain only 13 recurring vehicles and therefore measure held-out conditions rather than broad unseen-vehicle or unseen-site generalization. The leave-one-vehicle-out result is too uncertain to confirm the pattern independently. Acoustic tests are also based on 13 vehicle-level means, while transfer intervals condition on the fitted models and observed vehicles. The benchmark uses one generated realization per condition, does not test sampling diversity, and does not establish suitability for enforcement or safety-critical use. Finally, the diagnostic interventions motivate explanations for the transfer gap but do not validate them causally.

6 CONCLUSION

Scarce motion-labeled recordings motivate synthetic vehicle audio as a training resource for acoustic speed estimation. We introduced DopplerSim and PassByDiT as complementary approaches to motion-conditioned synthesis, together with a matched evaluation of motion control, acoustic agreement, and training utility across 378 conditions from 13 vehicles.

PassByDiT achieves stronger agreement with real recordings on nine of eleven acoustic metrics against each tested simulator and provides the best synthetic-to-real transfer. Training on PassByDiT reduces real-test MAE by approximately 43% relative to Pyroadacoustics and 59% relative to DopplerSim. The physics simulators provide more precise energy-peak alignment with the requested CPA, highlighting complementary strengths in explicit rendering and learned synthesis.

These results establish PassByDiT as a promising synthetic training source within the evaluated setting, while identifying the remaining challenge. Real-trained estimators transfer well to PassByDiT audio, but synthetic training does not yet match the accuracy achieved with real training data. This asymmetry shows why acoustic agreement and motion control must be evaluated alongside downstream transfer. Our comparison makes both progress toward useful synthetic supervision and the remaining replacement gap measurable.

Future work should investigate richer temporal source models, transfer-aware generation, and synthetic augmentation of limited real data, with validation across more vehicles, recording sites, and generated realizations. The central lesson is that useful synthetic pass-bys must preserve not only the acoustics of vehicle motion, but also the sound–speed relationship that models need to generalize to real recordings.

AI USE STATEMENT

Generative AI tools were used to assist with revision and editing of the manuscript. The authors reviewed the resulting text and remain responsible for the technical claims, experimental results, analyses, and final submitted content.

ETHICS STATEMENT

Synthetic roadside audio can support research without requiring every scene to be recorded again. It can also reproduce dataset biases or create false confidence in a speed estimator. This study does not establish accuracy for enforcement or safety-critical deployment. Real roadside recordings may include bystanders or other identifying sounds, so data use and release require attention to recording consent, privacy, and source-audio licenses. No new human listening study is reported here.

REPRODUCIBILITY STATEMENT

Section 3 and Appendices A and B specify PassByDiT Model M, the Vocos frontend, evaluation seeds, the Pyroadacoustics configuration, and DopplerSim. The current transfer export fixes the folds, manifests, estimator hashes, and 6,048 predictions for 16 train–test cells. Transfer metrics were recomputed directly from these prediction tables. The DopplerSim package contains all 400 exported WAV files, including the 378 benchmark targets, together with target metadata and source and output hashes. Appendix M states the remaining regeneration boundary.

REFERENCES

Jakob Abeßer, Saichand Gourishetti, András Kátai, Tobias Clauß, Prachi Sharma, and Judith Liebetrau. IDMT-traffic: An open benchmark dataset for acoustic traffic monitoring research. In *2021 29th European Signal Processing Conference (EUSIPCO)*, pp. 551–555, 2021. doi: 10.23919/EUSIPCO54536.2021.9616080. URL https://doi.org/10.23919/EUSIPCO54536.2021.9616080.

Mikołaj Bińkowski, Danica J. Sutherland, Michael Arbel, and Arthur Gretton. Demystifying MMD GANs. In *International Conference on Learning Representations*, 2018. URL https://arxiv.org/abs/1801.01401.

Stefano Damiano and Toon van Waterschoot. Pyroadacoustics: A road acoustics simulator based on variable length delay lines. In *Proceedings of the 25th International Conference on Digital Audio Effects*, pp. 216–223, 2022. URL https://www.dafx.de/paper-archive/2022/papers/DAFx20in22_paper_9.pdf.

Stefano Damiano, Luca Bondi, Andre Guntoro, and Toon van Waterschoot. A framework for the acoustic simulation of passing vehicles using variable length delay lines. *EURASIP Journal on Audio, Speech, and Music Processing*, 2024(1):49, 2024. doi: 10.1186/s13636-024-00372-4. URL https://doi.org/10.1186/s13636-024-00372-4.

Slobodan Djukanović, Jiří Matas, and Tuomas Virtanen. Acoustic vehicle speed estimation from single sensor measurements. *IEEE Sensors Journal*, 21(20):23317–23324, 2021. doi: 10.1109/JSEN.2021.3110009. URL https://doi.org/10.1109/JSEN.2021.3110009.

Slobodan Djukanović, Nikola Bulatović, and Ivana Čavor. A dataset for audio-video based vehicle speed estimation. *arXiv preprint arXiv:2212.01651*, 2022. URL https://arxiv.org/abs/2212.01651.

Cristóbal Esteban, Stephanie L. Hyland, and Gunnar Rätsch. Real-valued (medical) time series generation with recurrent conditional GANs. *arXiv preprint arXiv:1706.02633*, 2017. URL https://arxiv.org/abs/1706.02633.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13:723–773, 2012. URL https://www.jmlr.org/papers/v13/gretton12a.html.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979. URL https://www.jstor.org/stable/4615733.

Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7132–7141, 2018. URL https://openaccess.thecvf.com/content_cvpr_2018/html/Hu_Squeeze-and-Excitation_Networks_CVPR_2018_paper.html.

Kevin Kilgour, Mauricio Zuluaga, Dominik Roblek, and Matthew Sharifi. Fréchet audio distance: A reference-free metric for evaluating music enhancement algorithms. In *Interspeech*, pp. 2350–2354, 2019. doi: 10.21437/Interspeech.2019-2219. URL https://www.isca-archive.org/interspeech_2019/kilgour19_interspeech.html.

Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. DiffWave: A versatile diffusion model for audio synthesis. In *International Conference on Learning Representations*, 2021. URL https://arxiv.org/abs/2009.09761.

Haohe Liu, Zehua Chen, Yi Yuan, Xinhao Mei, Xubo Liu, Danilo Mandic, Wenwu Wang, and Mark D. Plumbley. AudioLDM: Text-to-audio generation with latent diffusion models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 21450–21474, 2023. URL https://proceedings.mlr.press/v202/liu23f.html.

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. URL https://proceedings.mlr.press/v139/nichol21a.html.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4195–4205, 2023. URL https://openaccess.thecvf.com/content/ICCV2023/html/Peebles_Scalable_Diffusion_Models_with_Transformers_ICCV_2023_paper.html.

Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville. FiLM: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018. URL https://ojs.aaai.org/index.php/AAAI/article/view/11671.

Reto Pieren, Thomas Bütler, and Kurt Heutschi. Auralisation of accelerating passenger cars. In *Proceedings of EuroNoise*, 2015. URL https://www.conforg.fr/euronoise2015/proceedings/data/articles/000015.pdf.

Gabriela Dantas Rocha, Julio Cesar B. Torres, Mariane Rembold Petraglia, and Michael Vorländer. Direction of arrival estimation of partial sound sources of vehicles with a two-microphone array. *Acta Acustica*, 5:18, 2021. doi: 10.1051/aacus/2021011. URL https://doi.org/10.1051/aacus/2021011.

Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. In *International Conference on Learning Representations*, 2022. URL https://arxiv.org/abs/2202.00512.

Konstantin Shmelkov, Cordelia Schmid, and Karteek Alahari. How good is my GAN? In *Proceedings of the European Conference on Computer Vision*, pp. 213–229, 2018. doi: 10.1007/978-3-030-01216-8_14.

Hubert Siuzdak. Vocos: Closing the gap between time-domain and fourier-based neural vocoders for high-quality audio synthesis. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/6db0903efdfe9b1bbafb015c10990b78-Abstract-Conference.html.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021. URL https://arxiv.org/abs/2010.02502.

Grzegorz Szwoch and Józef Kotus. Acoustic detector of road vehicles based on sound intensity. *Sensors*, 21(23):7781, 2021. doi: 10.3390/s21237781. URL https://doi.org/10.3390/s21237781.

Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Marianna Nezhurina, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. *arXiv preprint arXiv:2211.06687*, 2022. URL https://arxiv.org/abs/2211.06687.

Yiming Zhang, Yicheng Gu, Yanhong Zeng, Zhening Xing, Yuancheng Wang, Zhizheng Wu, and Kai Chen. FoleyCrafter: Bring silent videos to life with lifelike and synchronized sounds. *arXiv preprint arXiv:2407.01494*, 2024. URL https://arxiv.org/abs/2407.01494.

Xu Zheng, Yan Song, Li-Rong Dai, Ian McLoughlin, and Lin Liu. An effective mutual mean teaching based domain adaptation method for sound event detection. In *Interspeech*, pp. 556–560, 2021. doi: 10.21437/Interspeech.2021-281. URL https://www.isca-archive.org/interspeech_2021/zheng21_interspeech.html.

A BENCHMARK IMPLEMENTATION AND SPLIT CONSTRUCTION

Appendices A, B, and E–K describe the five-fold benchmark. Appendices C and D report separate DopplerSim diagnostics and are not pooled with that benchmark. Appendix L reports the earlier three-seed acoustic factorial under its original split and frontend.

Eligibility and reserved sources. The duration gate accepts recordings between 9.98 and 10.10 s. It removes nine of 400 VS13 clips: Peugeot307 at 30, 33, 35, 50, 56, 82, 85, and 101 km/h, and RenaultScenic at 57 km/h. One eligible recording per vehicle is reserved as source material for DopplerSim and Pyroadacoustics. The reserved files, including those used in the lateral-offset probe in Appendix C, are `CitroenC4Picasso_63`, `KiaSportage_61`, `MercedesGLA_61`, `OpelInsignia_61`, and the 60 km/h clips of Mazda3, MercedesAMG550, NissanQashqai, Peugeot208, Peugeot3008, Peugeot307, RenaultCaptur, RenaultScenic, and VWPassat. Removing these 13 sources leaves 378 conditions. Each source is an existing pass-by, with changes due to its original motion and recording geometry.

Grouping and folds. The code joins records with the same vehicle and integer speed, identical waveform hashes, or shared pass identifiers. Outer test folds contain 76, 76, 76, 75, and 75 conditions. Each has a separate 61-condition validation set, leaving fit sets of 241, 241, 241, 242, and 242. Assignment uses seed 1234 and inner seeds $1234 + 100 + k$. Constructed groups are disjoint across partitions. No session mapping is available. Vehicles occur in multiple partitions.

Generator fitting. A fresh Model M is fitted in each fold, enabling metadata, FiLM, and learned physics affinity, with analytic attention bias disabled. Audio is 24 kHz, with 100 mel bins, FFT size 1,024, hop 256, and 938 frames. Mel features use the Vocos feature extractor. Mel and physics normalization, and metadata power and year statistics, use only the fit partition. Training uses 200 epochs, batch size 2, AdamW with learning rate 2×10^{-4} and weight decay 10^{-4}, gradient clipping at 1, and EMA decay 0.9995. The cosine diffusion schedule has 1,000 steps. Generation uses 50 DDIM steps and the EMA checkpoint selected by validation loss. The primary run uses base training seed 1234, offset by fold; the completed robustness run replaces this base with 2234 for generator and estimator fitting. In every run, generation resets to seed 1234 per condition, holding the initial stochastic realization fixed across conditions. Envelope-loss weight is zero. The present protocol therefore evaluates one realization per condition; independent and multiple realizations are reserved for a diversity study.

Each fold's generator synthesizes the fit, validation, and test conditions needed by the speed estimator. Its real test targets occur in neither generator fitting nor validation. Synthetic fit conditions are generated by a model that has seen their real counterparts. This differs from the held-out synthetic test conditions.

Speed estimation. Audio is monaural, 16 kHz, and 10 s. Endpoint cropping and padding do not align CPA. Waveforms are peak-normalized before prediction. The estimator input is $128 \times 313 \times 1$ log-mel features, using FFT size 2,048, hop 512, and power converted to dB relative to its maximum. Scalar normalization is fitted on the fit partition. The Keras SE-ResNet uses the PyTorch backend, batch size 16, at most 150 epochs, patience 30, AdamW with initial learning rate 5×10^{-4} and weight decay 10^{-4}, and cosine scheduling. No waveform augmentation is used. Four training domains and five folds give 20 estimators, reused across the four test domains.

Pyroadacoustics. The reference configuration feeds the corresponding reserved real VS13 pass-by waveform to a straight omnidirectional source. The target path uses a 0.5 m lateral offset; source and receiver heights are 1.2 m, temperature is 20°C, and humidity is 50%. Air absorption is enabled, ground reflection is disabled, and interpolation uses sinc. The buffer can extend for propagation. Pyroadacoustics produces a synthetic received waveform by propagating the recorded excitation through this moving-source configuration. Acoustic scoring uses stored file levels; peak normalization is applied later for speed estimation.

Spectrogram displays. Paper comparison plots use the final 16 kHz WAVs, a periodic Hann window of 1,024 samples, hop 256, and $20 \log_{10} |X|$ with magnitude floor 10^{-8}. They show 0–8 kHz

on a shared -100 to $+10$ dB scale without waveform normalization, boundary extension, or STFT padding. The notebook's interactive preview can instead use automatically scaled PSD plots; those are not the common-scale comparison or the estimator input. A higher native sample rate does not change the bandwidth of the final evaluated WAV. For CLAP, this common waveform is upsampled to 48 kHz, without recovering content removed above 8 kHz.

B DopplerSim implementation details

The interactive implementation accepts the original geometry $(v_1, h_1, t_{\mathrm{CPA},1})$, vehicle length and emitter count, and target geometry $(v_2, h_2, t_{\mathrm{CPA},2}, T_{\mathrm{out}})$. Speeds are converted to SI units internally. Uploaded audio retains its native sampling rate during source inversion; output is mono 16-bit PCM at 44.1 kHz. The straight renderer uses fixed $c = 343\,\mathrm{m/s}$, exact per-sample retarded-time roots, linear fractional-time interpolation, and $1/R$ pressure spreading. It does not add ground reflection, atmospheric absorption, reverberation, contour broadening, or a softened-distance gain.

Straight-path geometry. The implementation solves Equation 2 as a quadratic at every observer sample. Squaring the travel-time relation can introduce extraneous roots, so a candidate is retained only when $u_k < t$, propagation distance is positive, and geometric range agrees with $c(t - u_k)$ within relative tolerance 10^{-4}. Samples without a valid root remain silent. Radial velocity and the retarded-time Jacobian are

$$v_{r,k}(t) = v\frac{x_k(u_k(t))}{R_k(u_k(t))}, \qquad D_k(t) = \frac{du_k}{dt} = \frac{c}{c + v_{r,k}(t)}. \tag{9}$$

Approach gives $v_r < 0$ and $D > 1$, while departure gives $v_r > 0$ and $D < 1$. A single emitter's geometric CPA is heard approximately h/c after its emission-time CPA.

Source inversion. The uploaded recording is converted to mono, peak-normalized, and transformed with a centered Hann STFT using a 4,096-point FFT and 512-sample hop. The squared, frequency-remapped magnitudes from Equation 3 are interpolated onto the common frequency grid and averaged over valid frames and emitters. Because each emitter is inferred from the same observed mixture, this is a spectral estimate rather than source demixing: it retains average timbre while discarding original phase and time-varying source structure. The observed PSD is retained separately for diagnostics.

Source synthesis and output scaling. For each emitter, independent white noise n_k is shaped by the shared intrinsic PSD:

$$s_k(u) = \frac{\mathcal{F}^{-1}\left[\mathcal{F}\{n_k\}\sqrt{\widehat{P}_{\mathrm{int}}}\right](u)}{\|\cdot\|_\infty + \epsilon}, \tag{10}$$

where $\widehat{P}_{\mathrm{int}}$ is interpolated in $\log(f+1)$ and log-power coordinates. Linear interpolation reads s_k at fractional retarded times. The $1/\sqrt{N}$ sum in Equation 4 follows the standard convention for combining independent-source signals rather than a calibrated preservation of power; after this summation the waveform is scaled to peak magnitude 0.95, so absolute power is not tracked through the pipeline. Unfixed emitter noise draws make a fresh render stochastic. Each reported benchmark file is one frozen draw, identified below.

The source PSD depends only on the recording and original geometry. Batch generation caches it by source path, source speed, original distance and CPA time, vehicle length, and emitter count, then reuses it across target geometries. Each render also saves the observed and inverted PSDs, source STFT, retarded time, range, radial velocity, and Doppler factor. The Phase 1 export pairs the generated STFT with the instantaneous centreline state

$$\mathbf{s}(t) = [x(t), v_x(t), y(t), v_y(t)] = [v_2(t - t_{\mathrm{CPA},2}), v_2, h_2, 0]. \tag{11}$$

This label clock is intentionally distinct from acoustic retarded time: geometric CPA is labelled at $t_{\mathrm{CPA},2}$, while its sound arrives later by approximately h_2/c.

The current code also supports free 2D and 3D paths. Consecutive path points define a polyline, cumulative arc length sets constant-speed timing, and emitter offsets follow the local tangent. Retarded

time is recovered by inverting the sampled arrival-time field $u + R(u)/c$. These capabilities are not part of the straight-path benchmark. The benchmark uses the straight-renderer export identified by recipe `supplied_get_doppler_audio_array_straight_defaults_seed1234_v1` and revision `4e5cdbd`; each benchmark WAV is fixed by its SHA-256 hash.

C DopplerSim lateral-offset check

The main benchmark uses a shared nominal offset $h = 0.5\,\text{m}$. Figures 3–4 test how larger roadside offsets change DopplerSim acoustics. Each vehicle is inverted once from the reserved file listed in Appendix A; that fixed inversion is then re-propagated at $h \in \{0.5, 1, 2, 3\}\,\text{m}$.

Figure 3 shows one matched condition (KiaSportage, 72 km/h). Increasing h spreads energy in time, but the synthetic event remains a narrower, more symmetric burst than the real recording. Figure 4 makes the same point on peak-normalized envelopes: the real shoulders stay higher for several seconds, while $h = 0.5\,\text{m}$ decays fastest and $h = 3\,\text{m}$ is closer without matching the real width. Threshold duration at $-14\,\text{dB}$ relative to the peak therefore rises with h, yet remains well below the real $-14\,\text{dB}$ duration of $5.43\,\text{s}$; a $-40\,\text{dB}$ gate spans essentially the full $10\,\text{s}$ clip for both real and synthetic audio and is not a useful duration contrast here.

Across the 13 vehicles, mean paired MR-STFT distance falls from 4.991 at $h = 0.5\,\text{m}$ to 3.796 at $h = 3\,\text{m}$, with twelve vehicles improving. Larger offsets change event width and spectral agreement on this 13-condition probe. The real envelope remains broader, and the main benchmark keeps the nominal $0.5\,\text{m}$ offset.

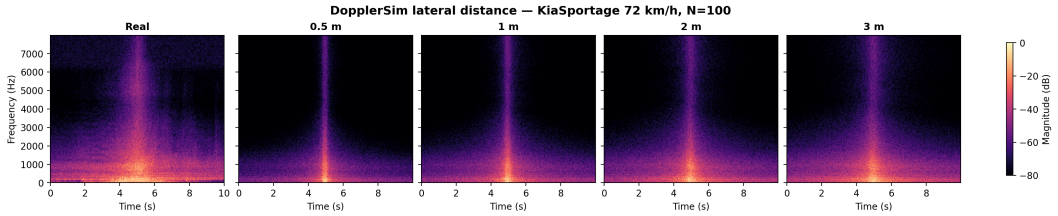

Figure 3: Matched spectrograms for KiaSportage at 72 km/h. The real clip has a broader rise and fall than DopplerSim at $h \in \{0.5, 1, 2, 3\}\,\text{m}$. This is an illustration of one condition, not an aggregate over the 378-clip benchmark.

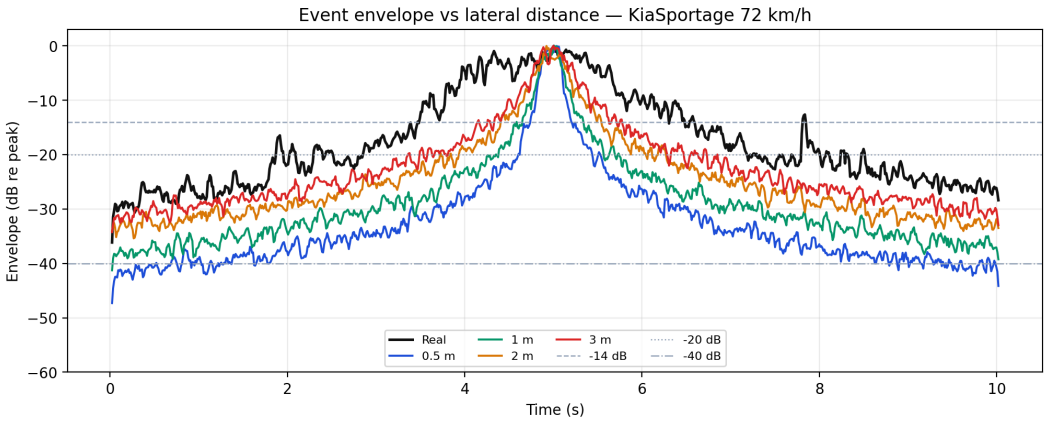

Figure 4: Peak-normalized RMS envelopes for the same KiaSportage 72 km/h condition. Increasing h raises the synthetic shoulders, but they remain below the real envelope away from the peak.

D DopplerSim ambient-floor intervention

To test whether the concentrated DopplerSim envelope can be broadened directly, we compare the baseline renderer with an intervention that adds an ambient floor at -39.3 dBFS. Source inversion, reserved recordings, trajectories, and all other renderer settings are held fixed. Table 4 reports means over 13 matched vehicle conditions.

The intervention reduces envelope coefficient of variation from 2.4179 to 1.8062, making energy less concentrated around the peak. Acoustic agreement does not improve: mean MR-STFT distance rises from 5.1192 to 17.9250, and mean envelope correlation falls from 0.6959 to 0.6797. The added energy is a flat tail. Recorded pass-bys have time-varying shoulders. This intervention does not identify the cause of the transfer gap.

Table 4: DopplerSim ambient-floor intervention, averaged over 13 matched vehicle conditions. Lower is better except envelope correlation.

Condition	MR-STFT	Envelope corr.	Envelope CV
Baseline	5.1192	0.6959	2.4179
Ambient floor	17.9250	0.6797	1.8062

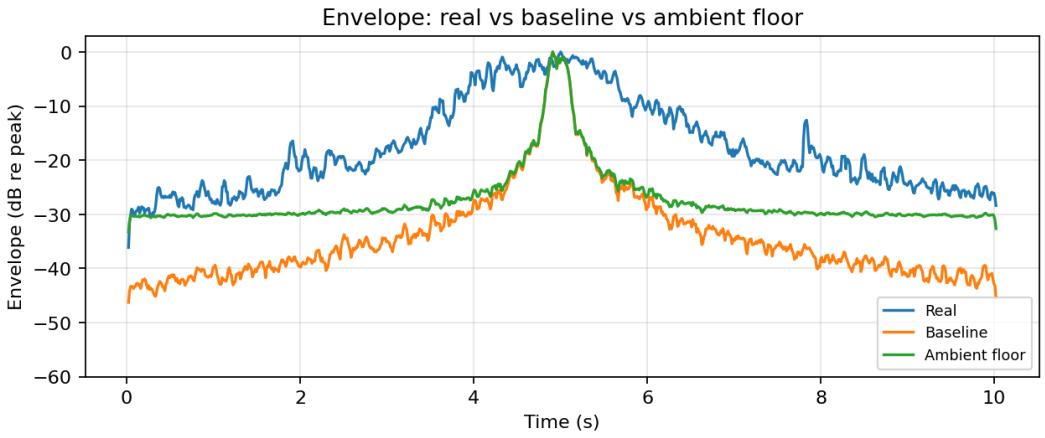

Figure 5: Representative peak-normalized envelopes for real audio, baseline DopplerSim, and the -39.3 dBFS ambient-floor intervention. The intervention raises the low-energy tails and lowers envelope variability, but it produces an approximately flat floor rather than the broader time-varying shoulders of the real recording.

E Acoustic measures and statistical analysis

Pairs share vehicle and speed labels and use the same CPA annotation as the target timestamp; physical geometry and realized acoustic CPA are not independently matched. Acoustic scoring uses stored levels, rather than peak-normalized estimator inputs.

Spectrum. Welch PSD estimation uses 8,192 samples with overlap 4,096. PSD mass is normalized within 50–7,500 Hz before mean absolute dB difference. One-third-octave centres are $1000\,2^{k/3}$ Hz, $k = -10, \ldots, 8$. Integration includes fractional FFT-bin overlap with band edges. Band totals are normalized to unit mass before mean absolute dB difference.

Envelope and timing. RMS uses 50 ms frames every 10 ms, smoothed over ten frames. Wasserstein distance treats time points as locations and RMS values as weights. Correlation uses full unshifted envelopes. Local peaks are found within 1.5 s of annotated CPA. Real-reference error

compares synthetic and real peak times; annotation error compares the synthetic peak with CPA. Energy duration spans 5% to 95% of cumulative squared RMS, and its error is the absolute difference from the real duration.

Energy and local spectrum. Pre/post error is the absolute discrepancy in the dB ratio of squared waveform energy before and after CPA. MR-STFT averages FFT sizes 256, 512, 1,024, and 2,048 with hops of one quarter of each size. Each resolution adds spectral convergence, $\|S - \widehat{S}\|_F / \|S\|_F$, to mean absolute log-magnitude error with floor 10^{-7}. The centroid uses power-weighted STFT spectra with FFT size 1,024 and hop 256, interpolated to RMS times. Centroid RMSE uses the real RMS mask above 30 dB below its peak. Integrated loudness uses pyloudnorm; signed and absolute LUFS differences are exported.

Tests and intervals. Paired clip differences are averaged within each vehicle. The exact two-sided sign-flip test evaluates all $2^{13} = 8192$ assignments over these 13 vehicle-level means, assuming independent vehicle-level differences with sign symmetry under the null. By construction, a positive advantage favors the first method. Holm correction covers all 33 comparisons jointly; 24 pass the corrected 0.05 threshold. Effects weight vehicles equally, whereas absolute means weight clips equally.

Transfer intervals use 1,000 bootstrap draws of condition groups within the fixed vehicle mixture, conditional on fitted models, sources, and observed vehicles; this resamples condition groups rather than vehicles, and is a different procedure from the vehicle-level sign-flip tests above. They exclude variation from retraining or new vehicles. Pooled R^2 is calculated over all held-out predictions, not averaged across folds.

E.1 How to read the metrics

Two sounds can have similar overall spectra yet differ in when their energy appears. Conversely, two envelopes can have the same shape even when one recording is much darker or quieter. The metric families deliberately separate these cases.

Frequency content and detail. Normalized PSD asks whether energy is distributed across frequencies in the same way; its normalization removes overall level. One-third-octave bands provide a coarser view that is less tied to individual FFT bins. MR-STFT checks both frequency and time at several resolutions, exposing local differences that an average spectrum can hide. Centroid reduces the spectrum to a brightness trajectory, so agreement there alone can conceal missing harmonics or texture.

Event shape and control. Envelope correlation compares the rise and fall but is insensitive to a uniform amplitude scale. Wasserstein distance says how much the temporal energy distribution must move to agree. Duration checks event spread, and pre/post balance checks its asymmetry. Peak error against real audio asks whether synthesis matches the observed event. Peak error against the CPA annotation instead asks whether synthesis follows the control timestamp. A narrow artificial spike can score well on the latter while poorly matching real audio.

Level, distributions, and task relevance. Absolute LUFS difference measures loudness mismatch; signed difference reveals whether the method is quieter or louder. Embedding distances summarize broader differences between collections, at the cost of less direct acoustic interpretation and small-sample sensitivity. Speed-transfer error has a different purpose: it tests whether cues useful to a real-trained estimator survive synthesis. None of these measures establishes listening quality on its own. Agreement across them is more informative than selecting a single favorable score.

Separate diagnosis. The separate diagnostic uses 50 ms RMS frames every 10 ms. Its event duration spans the first to last frames at or above 20% of the peak, including one hop, and can bridge low-energy gaps. Envelope variability is standard deviation divided by mean. These definitions differ from energy duration above.

Table 5: Median descriptors for the four evaluated domains. Threshold duration uses 20% of the peak envelope; RMS level and crest factor describe stored waveforms.

Descriptor	Real	PassByDiT	DopplerSim	Pyroad
Threshold duration (s)	5.125	4.495	0.300	0.290
RMS level (dB)	-20.29	-21.51	-29.10	-28.65
Crest factor (dB)	19.04	20.61	28.63	29.31

F Full cross-domain speed results

Each cell contains 378 predictions, giving 6,048 predictions across the 16 train–test directions. Negative R^2 means that squared error exceeds that of the evaluated targets' constant mean. It does not imply that the audio contains no speed information, since the low within-domain errors show that speed remains learnable within each synthetic domain.

Table 6: MAE for all transfer directions in the current frozen bundle (km/h).

Train	Test	MAE
Real	Real	7.047
Real	PassByDiT	7.803
Real	DopplerSim	24.408
Real	Pyroad	39.259
PassByDiT	Real	14.511
PassByDiT	PassByDiT	2.538
PassByDiT	DopplerSim	38.304
PassByDiT	Pyroad	76.457
DopplerSim	Real	35.734
DopplerSim	PassByDiT	34.544
DopplerSim	DopplerSim	2.034
DopplerSim	Pyroad	39.103
Pyroad	Real	25.593
Pyroad	PassByDiT	25.110
Pyroad	DopplerSim	28.872
Pyroad	Pyroad	4.574

G Component and robustness diagnostics

The reconstruction control uses separately trained estimators and is not an additive decomposition of the PassByDiT penalty. Its results show that the common mel/Vocos frontend contributes to domain shift but does not account for most of the replacement gap.

Table 7: Frontend-reconstruction control (MAE in km/h).

Train	Test	MAE
Real	Reconstructed real	11.64
Reconstructed real	Real	8.05
Reconstructed real	Reconstructed real	5.74
PassByDiT	Reconstructed real	25.49

The primary and second completed seeded fits use the same folds, data, architecture, and optimization settings. Their similar replacement penalties show that the direction and approximate size of the PassByDiT gap recur in these two runs. Two seeds do not characterize broader retraining variance.

Table 8: Central transfer cells across two completed seeded runs (MAE in km/h). Penalty is PassByDiT→Real minus Real→Real.

Seed	R→R	R→P	P→P	P→R	Penalty
1234	7.05	7.80	2.54	14.51	7.46
2234	5.90	6.47	2.94	13.13	7.23

H PassByDiT leave-one-vehicle-out diagnostic

The leave-one-vehicle-out result is macro-averaged over the same 13 held-out vehicles. Real-to-real, PassByDiT-to-real, and training-median MAE are 10.91, 15.23, and 17.97 km/h. The paired PassByDiT-minus-real difference is 4.31 km/h with vehicle-bootstrap 95% interval $[-3.85, 16.16]$ km/h. PassByDiT is worse than real training for 7 of 13 vehicles, with a two-sided exact sign-test value of $p = 1.0$. The interval includes zero, so this diagnostic does not establish a reliable vehicle-level difference.

Per-vehicle MAE summaries and paired errors are included in the released result tables. The primary comparison includes all 13 vehicles without post-hoc exclusions.

I Speed and vehicle strata

These tables compare the two estimators on the same real test recordings. Excess error is PassByDiT-trained absolute error minus real-trained absolute error. These are descriptive strata of the same experiment.

Table 9: Reverse-transfer MAE by speed band (km/h).

Speed band	n	Real-trained	PassByDiT-trained	Excess
30–49	96	5.04	12.13	7.09
50–69	97	6.57	14.02	7.45
70–89	110	6.81	16.38	9.56
90–105	75	10.58	15.46	4.88

Table 10: Reverse-transfer MAE by vehicle (km/h). These are exploratory strata, not independent significance tests.

Vehicle	n	Real-trained	PassByDiT-trained	Excess
CitroenC4Picasso	22	6.37	20.25	13.88
KiaSportage	32	5.82	10.59	4.78
Mazda3	31	8.06	13.95	5.89
MercedesAMG550	29	7.37	15.90	8.54
MercedesGLA	33	5.28	11.02	5.74
NissanQashqai	28	5.81	10.11	4.30
OpelInsignia	26	5.19	10.85	5.66
Peugeot208	28	7.48	9.21	1.73
Peugeot3008	30	6.70	18.12	11.42
Peugeot307	20	9.42	9.07	-0.34
RenaultCaptur	32	6.61	9.74	3.12
RenaultScenic	33	10.69	31.16	20.46
VWPassat	34	7.03	16.35	9.32

J FULL ACOUSTIC RESULTS

Absolute means weight clips equally. Test effects weight vehicles equally. Positive advantages favor the first named method. All exact and corrected probabilities are retained, including comparisons above the corrected significance threshold.

Table 11: Absolute acoustic scores, averaged over conditions. These clip-weighted means differ from the vehicle-weighted advantages in the exact tests.

Metric	PassByDiT	DopplerSim	Pyroad
PSD-shape MAE (dB)	3.0247	5.4090	4.5774
One-third-octave MAE (dB)	1.6504	2.8151	2.7989
Envelope Wasserstein (s)	0.3669	0.8472	1.2642
Real-reference peak error (s)	0.1186	0.2301	0.2299
Peak-to-annotation error (s)	0.2190	0.0090	0.0087
Envelope correlation	0.8521	0.5754	0.5701
Energy-duration error (s)	1.2444	4.2282	4.2356
Pre/post energy error (dB)	1.5315	3.2450	3.0272
MR-STFT distance	1.5206	3.3272	4.7275
Centroid RMSE (Hz)	173.6817	168.0123	207.3537
Absolute LUFS difference	1.1838	4.6005	4.3586

Table 12: PassByDiT versus DopplerSim. Positive advantage favors the first method. All exact tests use 13 vehicle means and all 8,192 sign assignments.

Metric	Advantage	Exact p	Holm p
PSD-shape MAE (dB)	2.365889	0.000244	0.008057
One-third-octave MAE (dB)	1.158790	0.000244	0.008057
Envelope Wasserstein (s)	0.472588	0.000244	0.008057
Real-reference peak error (s)	0.110743	0.000488	0.008057
Peak-to-annotation error (s)	-0.209295	0.000244	0.008057
Envelope correlation	0.278003	0.000244	0.008057
Energy-duration error (s)	2.954264	0.000244	0.008057
Pre/post energy error (dB)	1.695484	0.000488	0.008057
MR-STFT distance	1.808997	0.000244	0.008057
Centroid RMSE (Hz)	-5.922640	0.705322	1.000000
Absolute LUFS difference	3.467545	0.000244	0.008057

Table 13: PassByDiT versus Pyroad. Positive advantage favors the first method. All exact tests use 13 vehicle means and all 8,192 sign assignments.

Metric	Advantage	Exact p	Holm p
PSD-shape MAE (dB)	1.547720	0.000244	0.008057
One-third-octave MAE (dB)	1.147671	0.000488	0.008057
Envelope Wasserstein (s)	0.880469	0.000244	0.008057
Real-reference peak error (s)	0.110511	0.000488	0.008057
Peak-to-annotation error (s)	-0.209761	0.000244	0.008057
Envelope correlation	0.282134	0.000244	0.008057
Energy-duration error (s)	2.947873	0.000244	0.008057
Pre/post energy error (dB)	1.495438	0.001465	0.016113
MR-STFT distance	3.207583	0.000244	0.008057
Centroid RMSE (Hz)	33.591447	0.040039	0.360352
Absolute LUFS difference	3.344699	0.000244	0.008057

Table 14: Pyroad versus DopplerSim. Positive advantage favors the first method. All exact tests use 13 vehicle means and all 8,192 sign assignments.

Metric	Advantage	Exact p	Holm p
PSD-shape MAE (dB)	0.818169	0.000488	0.008057
One-third-octave MAE (dB)	0.011119	0.966064	1.000000
Envelope Wasserstein (s)	-0.407881	0.000244	0.008057
Real-reference peak error (s)	0.000233	0.671875	1.000000
Peak-to-annotation error (s)	0.000466	0.523438	1.000000
Envelope correlation	-0.004131	0.480713	1.000000
Energy-duration error (s)	0.006391	0.971924	1.000000
Pre/post energy error (dB)	0.200046	0.177979	1.000000
MR-STFT distance	-1.398586	0.000244	0.008057
Centroid RMSE (Hz)	-39.514087	0.001465	0.016113
Absolute LUFS difference	0.122846	0.912354	1.000000

K Frozen audio-embedding comparisons

CLAP learns audio representations through contrastive audio-language pretraining (Wu et al., 2022). The run uses `laion/clap-htsat-unfused`, 48 kHz resampling, and 512-dimensional embeddings without PCA.

Fréchet audio distance compares embedding means and covariances (Kilgour et al., 2019). Here the encoder is CLAP rather than the original VGGish encoder. The implementation evaluates the covariance cross term through singular values of the product of centred data matrices. This is an exact low-rank calculation for the empirical inputs, not a correction for finite-sample bias.

Maximum mean discrepancy compares kernel means (Gretton et al., 2012). RBF bandwidth is 1. The polynomial score follows the KID construction (Bińkowski et al., 2018), with kernel $(x^\top y/512 + 1)^3$. The code excludes within-set diagonals and the matched-pair cross diagonal, comparing independent condition draws in the cross term. Unbiased estimates can be negative and should not be clipped. Intervals use 200 conditional bootstrap draws.

On FAD-CLAP, RBF MMD, and polynomial KID, the order from smaller to larger distance is PassByDiT, DopplerSim, then Pyroad. That order matches most spectral and temporal scores. It does not hold for centroid RMSE or peak-to-annotation error, and it does not establish perceptual equivalence. With 378 examples and 512 dimensions, covariances are rank-deficient. Smaller vehicle and speed strata are exploratory. Intervals do not correct finite-sample bias or include retraining variation.

Table 15: Frozen CLAP distribution scores and conditional 95% bootstrap intervals. Each method uses 378 paired conditions; the embedding dimension is 512.

Metric	Method	Value	95% CI
FAD-CLAP	PassByDiT	0.255390	[0.249678, 0.264345]
RBF MMD^2	PassByDiT	0.180789	[0.176093, 0.185727]
Polynomial KID	PassByDiT	0.001221	[0.001189, 0.001256]
FAD-CLAP	Pyroad	0.391747	[0.384475, 0.403695]
RBF MMD^2	Pyroad	0.272014	[0.264017, 0.278996]
Polynomial KID	Pyroad	0.001929	[0.001870, 0.001982]
FAD-CLAP	DopplerSim	0.361394	[0.354274, 0.371878]
RBF MMD^2	DopplerSim	0.245907	[0.239658, 0.251780]
Polynomial KID	DopplerSim	0.001703	[0.001656, 0.001745]

L Earlier conditioning ablations

This complementary experiment used 400 VS13 recordings, a fixed vehicle-stratified split with 58 test clips, and seeds 0, 1, and 42. Its frontend and split differ from the current five-fold benchmark,

so we use it to compare the acoustic behavior of the B/F/M conditioning sequence rather than as a pooled transfer result.

Adding FiLM to metadata conditioning reduces CPA-region error from 0.5848 to 0.4797 and raises best-shift correlation from 0.7953 to 0.9137. The full Model M remains close to F while attaining the lowest mean log-mel and MR-STFT errors of the three. It is not best on multi-scale Fréchet distance, where B is lower, or on cosine similarity, where F is higher. Because the split and frontend differ from the primary benchmark, these results characterize acoustic reconstruction only and are not combined with the current transfer matrix.

Table 16: Earlier spectral measures for B/F/M. Mean ± standard deviation over three seeds.

Model	MS-FD ↓	log-mel L_1 ↓	cosine ↑
B	3.572 ±0.427	0.5584 ±0.0043	0.9678 ±0.0009
F	4.319 ±0.548	0.5122 ±0.0004	**0.9691** ±0.0015
M	4.291 ±0.654	**0.5117** ±0.0013	0.9689 ±0.0007

Table 17: Earlier reconstruction and temporal measures for B/F/M. Mean ± standard deviation over three seeds.

Model	MR-STFT ↓	CPA L_1 ↓	best-shift ↑
B	0.7889 ±0.0031	0.5848 ±0.0066	0.7953 ±0.003
F	0.7593 ±0.0008	**0.4797** ±0.0009	0.9137 ±0.002
M	**0.7585** ±0.0005	0.4800 ±0.0019	0.9145 ±0.001

M PROVENANCE AND EVALUATION SCOPE

The release records the split and training seeds, five-fold manifests, source exclusions, model hashes, and 6,048 out-of-fold predictions for the primary 16-cell transfer matrix. Each cell contains the same 378 condition identifiers and fold assignments. The saved folds have disjoint constructed groups, and all transfer targets exclude the 13 simulator source recordings.

The DopplerSim release contains 400 WAV files and per-file sidecars. All 378 benchmark targets match their recorded SHA-256 hashes, and each sidecar identifies the source hash, target speed and CPA, 100-emitter configuration, nominal 0.5 m distance, recipe, and code revision. The 13 reserved real pass-bys serve as source material for both physics simulators: DopplerSim estimates and resynthesizes their spectra, whereas Pyroadacoustics propagates the recorded waveforms directly.

The main transfer matrix uses Model M, one fitted generator per fold, one generated realization per condition, and the source-access protocol in Section 4. The geometry and ambient-floor studies are separate paired acoustic diagnostics and are not pooled with the transfer matrix.