Real-time Prediction of Urban Sound Propagation with Conditioned Normalizing Flows

Achim Eckerle
Stralsund University
achim.eckerle@hochschule-stralsund.de

Martin Spitznagel
IMLA, Offenburg University
martin.spitznagel@hs-offenburg.de

Janis Keuper
IMLA, Offenburg University
keuper@imla.ai

Abstract

Accurate and fast urban noise prediction is pivotal for public health and for regulatory workflows in cities, where the Environmental Noise Directive mandates regular strategic noise maps and action plans, often needed in permission workflows, right-of-way allocation, and construction scheduling. Physics-based solvers are too slow for such time-critical, iterative "what-if" studies. We evaluate conditional Normalizing Flows (Full-Glow) for generating for generating standards-compliant urban sound-pressure maps from 2D urban layouts in real time ($\approx$ 102 ms per 256$\times$256 map on a single RTX 4090), enabling interactive exploration directly on commodity hardware. On datasets covering Baseline, Diffraction, and Reflection regimes, our model accelerates map generation by >2000$\times$ over a reference solver while improving NLoS accuracy by up to 24% versus prior deep models; in Baseline NLoS we reach 0.65 dB MAE with high structural fidelity. The model reproduces diffraction and interference patterns and supports instant recomputation under source or geometry changes, making it a practical engine for urban planning, compliance mapping, and operations (e.g., temporary road closures, night-work variance assessments).

1 Introduction

Urban noise is both a public-health and regulatory concern: WHO guidelines link chronic exposure to sleep disturbance and cardiovascular risks, and the EU Environmental Noise Directive mandates recurrent city-scale noise maps and action plans.[1, 2] Consequently, urban planners need physically reliable predictions at interactive latencies for permitting and operations. From barrier design to regulating construction projects such as urban drilling, decisions hinge on accurate models of sound propagation in complex cityscapes [3, 4]. Although physics-based solvers (ray tracing, FEM) provide high fidelity, their computational cost renders them impractical for the iterative, large-scale "what-if" analyses required in modern city planning [5].

This computational bottleneck has spurred interest in AI-driven alternatives. Deep learning models, particularly from the U-Net [6] or GAN [7] families, can generate sound maps orders of magnitude faster. However, they often trade physical consistency for speed, struggling to accurately model complex wave phenomena like multi-path reflections and diffraction, which are ubiquitous in dense urban canyons [8].

We propose leveraging conditional Normalizing Flows (NFs), a class of generative models known for their mathematical rigor and stable training [9, 10]. Their unique invertible architecture allows

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: UrbanAI.

for exact likelihood computation, making them highly suitable for modeling complex physical distributions [11]. Specifically, we adopt the Full-Glow architecture [12] to perform an image-to-image transformation from 2D building layouts to sound pressure maps. Our contributions are: (1) a successful application of a fully conditional NF to model distinct urban acoustic phenomena; (2) a quantitative demonstration that our approach significantly outperforms previous deep learning methods in physical accuracy, especially in occluded urban spaces; and (3) validation that NFs can accelerate these simulations by a factor of over 2000 while maintaining high physical fidelity.

2 Related Work

Physics-Based Urban Acoustics. The gold standard for sound simulation remains physics-based solvers. Geometric acoustics methods like ray-tracing are effective for high-frequency sounds, modeling reflections and shadowing [3]. For greater precision, wave-based approaches like the Finite-Element-Method (FEM) solve the underlying wave equations but with a severe computational overhead [5]. Open-source frameworks like NoiseModelling, which implements the CNOSSOS-EU standard, serve as a valuable reference for physically grounded simulations but are too slow for large-scale generative tasks [13].

Deep Learning for Physics Simulation. AI models have emerged as powerful accelerators. U-Net architectures [6] are a common baseline for image-to-image tasks but can produce blurry or physically inconsistent results. Generative Adversarial Networks (GANs), such as pix2pix [7], can generate sharp, realistic outputs but often suffer from training instability and mode collapse [14]. Denoising Diffusion Models (DDPMs) produce high-quality samples but their iterative inference process is computationally intensive, limiting their utility in time-sensitive applications [15].In the urban-acoustics setting, PhysicsGen[8] provides benchmarked deep baselines on the same dataset, which we use for comparison alongside the public benchmark results[16].

Normalizing Flows. NFs provide a compelling alternative by modeling probability densities explicitly through a series of invertible transformations [9]. This allows for stable maximum-likelihood training and exact inference. The Glow model [17] introduced key architectural innovations like invertible 1x1 convolutions, making NFs practical for high-resolution images. The Full-Glow model [12] advances this by conditioning every transformation layer on an input, making it exceptionally well-suited for image-to-image tasks where strong structural guidance is needed. This deep conditioning is what we leverage to enforce physical constraints in the generation of urban sound maps.

3 Method

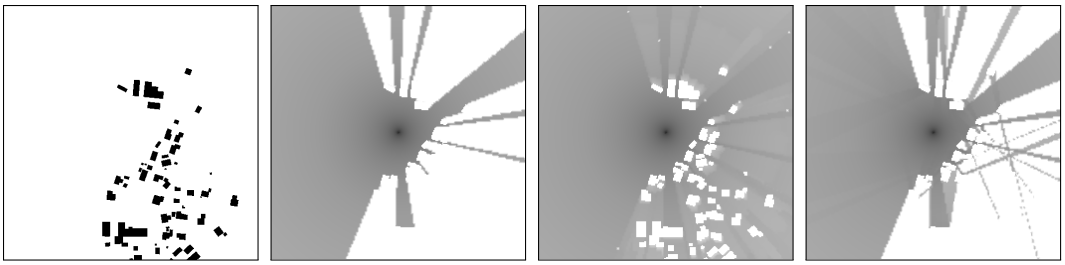
Data. We use the *Urban Sound Data* benchmark[16, 18] as our data source, comprising **25,000** paired samples of OSM-based building masks (inputs) and simulated sound-pressure maps (targets) at 256×256 resolution. Simulations follow CNOSSOS-compliant settings via NoiseModelling and are provided in three variants: Baseline, Diffraction (edge diffraction at building corners), and Reflection (up to multiple orders), with predefined train/validation/test splits [13, 3]. Intensities are normalized to $[0, 1]$; conditioning variables (when present) are min–max scaled. Where indicated, we compare against PhysicsGen[8].

Fully conditioned Glow. The architecture extends Glow [10] with conditioning injected into *all* invertible steps (ActNorm, invertible 1×1-convolution, affine coupling), following the Full-Glow design principle [12]. Each flow block receives features from the source pathway (building layout) through a lightweight conditioning network. LU-parameterization keeps $\log|\det(\cdot)|$ tractable in 1×1-convolutions. Coupling transforms partition channels $(\mathbf{x}_1, \mathbf{x}_2)$ and predict scale $\mathbf{s}$ and translation $\mathbf{t}$ from $\mathbf{x}_1$ plus conditioning $\mathbf{c}$:

$$\mathbf{y}_1 = \mathbf{x}_1, \quad \mathbf{y}_2 = \mathbf{s}(\mathbf{x}_1, \mathbf{c}) \odot \mathbf{x}_2 + \mathbf{t}(\mathbf{x}_1, \mathbf{c}) \tag{1}$$

Likelihood objective. Flows maximize exact data likelihood using the change-of-variables formula

$$\log p(\mathbf{x}) = \log p(\mathbf{z}) + \sum_{k=1}^{K} \log\Big|\det\big(\frac{\partial h_k}{\partial h_{k-1}}\big)\Big|, \tag{2}$$



(a) OSM Layout (b) Baseline Sim. (c) Diffraction Sim. (d) Reflection Sim.

Figure 1: Example data pair: (a) Input urban layout from OSM, and corresponding ground truth simulations for (b) Baseline, (c) Diffraction, and (d) Reflection scenarios.

with standard normal base density $p(\mathbf{z}) = \mathcal{N}(0, I)$ [19, 20]. For the conditional mapping $p(\mathbf{x}_{\text{out}} \mid \mathbf{x}_{\text{in}})$, the target flow is conditioned on the source representation (buildings).

Training setup. Training was conducted for 1.2M iterations (60 epochs) on 19,908 training samples. Images are processed in a 4-scale multi-scale flow with $[8, 8, 8, 8]$ steps per scale. Batch size is 1 due to memory constraints (≈14GB VRAM per sample). Adam optimizer (β_1=0.9, β_2=0.999) with initial learning rate 10^{-4}, followed by linear decay to 5×10^{-6} from iteration 1M to 1.2M. All experiments performed on NVIDIA RTX 4090 (24GB VRAM), with total training time of ≈108 hours per model. Gradient checkpointing reduces peak memory usage by 30-40%.

Metrics. Mean Absolute Error (MAE) and weighted Mean Absolute Percentage Error (wMAPE) are computed separately for Line-of-Sight (LoS) and Non-Line-of-Sight (NLoS) regions, determined via ray-tracing from the central sound source. NLoS regions represent acoustically shadowed areas where direct sound paths are blocked by buildings. wMAPE is computed as $\sum_i |y_i - \hat{y}_i| / \sum_i |y_i|$ with 30 dB threshold to avoid division by near-zero values.

4 Results

Table 1: MAE, wMAPE, and runtime across tasks and conditions (smaller is better). Other results are reported from the public benchmark [16].

Model	**Metric**	**Baseline**		**Reflection**		**Diffraction**	
		LoS	NLoS	LoS	NLoS	LoS	NLoS
Sim.	MAE	0.00	0.00	0.00	0.00	0.00	0.00
	wMAPE	0.00	0.00	0.00	0.00	0.00	0.00
	Runtime (ms)		204 700		251 000		206 000
UNet	MAE	2.29	1.73	2.29	5.72	0.94	3.27
	wMAPE	12.91	37.57	12.75	80.46	4.22	22.36
	Runtime (ms)		0.14		0.138		0.14
Pix2Pix	MAE	**1.73**	1.19	2.14	4.79	0.91	3.36
	wMAPE	9.36	6.75	11.30	30.67	3.51	18.06
	Runtime (ms)		0.14		0.138		0.14
DDPM	MAE	2.42	3.26	2.74	7.93	1.59	3.27
	wMAPE	15.57	51.08	17.85	80.38	8.25	20.30
	Runtime (ms)		3986.35		3986.35		3986.35
Full Glow (ours)	MAE	1.84	**0.65**	**2.06**	**3.64**	**0.79**	**2.63**
	wMAPE	**8.83**	**4.52**	**8.98**	**22.69**	**2.43**	**11.12**
	Runtime (ms)		101.70		102.30		107.62

Best values per column are highlighted in light green. All metrics are averaged over 1,245 test samples for each scenario.

Quantitative Analysis. Table 1 shows a quantitative comparison against prior deep learning models from [8] on the same dataset splits. Our model sets a new benchmark in almost all scenarios. For the **Baseline** condition, it achieves an NLoS-MAE of only **0.65 dB**, a 45% improvement over the best competing model (Pix2Pix). This indicates an exceptional ability to model basic acoustic shadowing. In the more complex **Diffraction** scenario, our model again leads with an NLoS-MAE of **2.63 dB**. Most notably, in the challenging **Reflection** scenario, where multi-path interference is key, our model achieves an NLoS-MAE of **3.64 dB**, a 24% improvement over Pix2Pix. These strong NLoS results confirm the model's superior ability to capture complex wave phenomena in occluded urban spaces.

Qualitative and Structural Analysis. As shown in Figure 2, the sound maps generated by Full-Glow are visually almost indistinguishable from the ground truth simulations. The model correctly reproduces the sharp acoustic shadows in the Baseline case, the characteristic soft-edged fans of diffraction, and the complex interference patterns in the Reflection scenario. The absolute error maps confirm that errors are small and localized, avoiding the systematic blurring seen in other models. The high structural similarity is further confirmed by SSIM scores, with mean values of **0.92** for Baseline, **0.96** for Diffraction, and **0.85** for Reflection, indicating excellent preservation of the sound field's spatial structure.

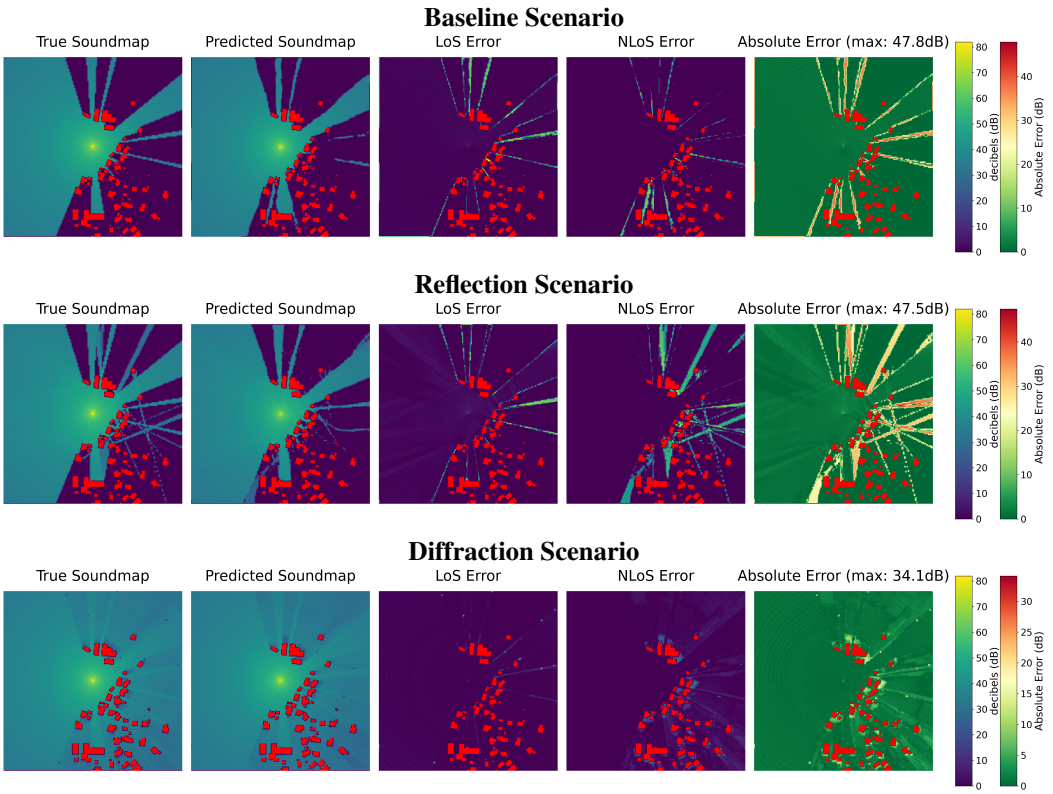


Figure 2: Visual comparison of our model's predictions (center column) against the ground truth (left column) for the Baseline, Reflection, and Diffraction scenarios. The absolute error maps (right column) confirm high physical fidelity across all cases.

Statistical Significance. Figure 3 extends our quantitative analysis by providing 95% confidence intervals across all test samples. The narrow confidence intervals around our Full-Glow model's performance demonstrate that the improvements reported in Table 1 are statistically robust and not driven by outliers. Notably, the confidence intervals for NLoS errors do not overlap between our method and competing approaches in any scenario, confirming statistical significance. The consistently larger confidence intervals in the Reflection scenario across all models reflect the inherent stochasticity of multi-path interference patterns, yet our approach maintains the tightest bounds even in this challenging regime.

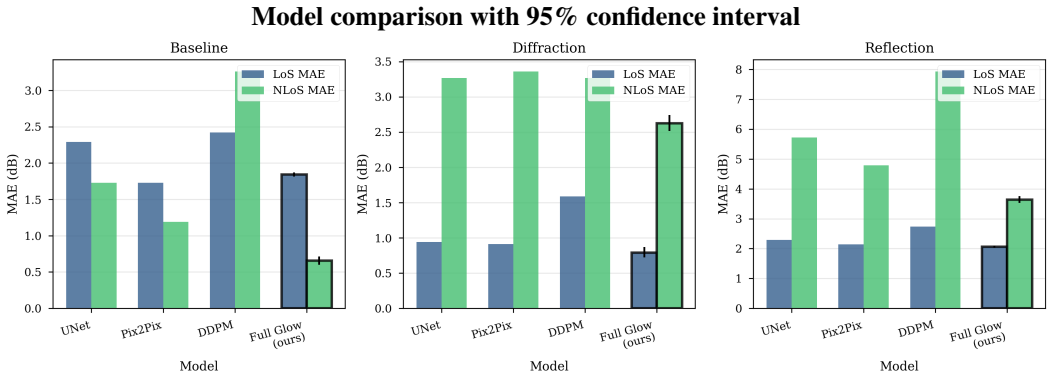


Figure 3: Model comparison with 95% confidence intervals across all three acoustic scenarios, computed over 1,245 test samples per scenario. The non-overlapping confidence intervals confirm the statistical significance of our Full-Glow model's performance gains, particularly in acoustically shadowed (NLoS) regions.

5 Conclusion

This paper reports a fully conditioned normalizing-flow approach for urban sound propagation. On three scenarios (Baseline, Diffraction, Reflection), the method achieves accurate LoS/NLoS metrics and while providing large inference-time speedups over classical simulation. They offer a compelling balance of generative speed, model stability, and physical accuracy, making them a highly promising tool for practical applications in urban planning, noise assessment, and beyond. Limitations include sensitivity to multi-effect complexity (reflections remain hardest) and high training memory.

Funding Acknowledgement

The authors acknowledge the financial support by the German Federal Ministry of Education and Research (BMBF) in the program "Forschung an Fachhochschulen in Kooperation mit Unternehmen (FH-Kooperativ)" within the joint project "KI-Bohrer" under grant 13FH525KX1 `https://www.ki-bohrer.de/`.

References

[1] World Health Organization. Regional Office for Europe. *Environmental Noise Guidelines for the European Region*. WHO Regional Office for Europe, Copenhagen, 2018. ISBN 9789289053563. URL https://www.who.int/europe/publications/i/item/9789289053563.

[2] European Parliament and Council of the European Union. Directive 2002/49/ec of 25 june 2002 relating to the assessment and management of environmental noise. Official Journal of the European Communities, L 189, 18 July 2002, pp. 12–26, 2002. URL https://eur-lex.europa.eu/eli/dir/2002/49/oj.

[3] Erik M. Salomons. *Computational Atmospheric Acoustics*. Springer, 2001.

[4] *ISO 9613-2: Acoustics — Attenuation of sound during propagation outdoors — Part 2: General method of calculation*. International Organization for Standardization, 1996.

[5] Lawrence E. Kinsler, Austin R. Frey, Alan B. Coppens, and James V. Sanders. *Fundamentals of Acoustics*. Wiley, 4 edition, 2000.

[6] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015.

[7] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, 2017.

[8] Martin Spitznagel, Jan Vaillant, and Janis Keuper. Physicsgen: Can generative models learn from images to predict complex physical relations? In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pages 11125–11134, June 2025.

[9] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real nvp. In *ICLR*, 2017.

[10] Durk Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1×1 convolutions. In *NeurIPS*, 2018.

[11] Lynton Ardizzone et al. Analyzing inverse problems with invertible neural networks. *ICLR*, 2019.

[12] Mehdi Sorkhei et al. Full-glow: Fully conditional glow for conditional image synthesis. In *Proceedings of a CV/ML venue*, 2021. Preprint/Workshop version.

[13] NoiseModelling Developers. Noisemodelling v4.x: Cnossos-compliant environmental noise simulation. https://noise-planet.org/noisemodelling.html, 2025.

[14] Antonia Creswell, Tom White, Vincent Dumoulin, Kai Arulkumaran, Biswa Sengupta, and Anil A. Bharath. Generative adversarial networks: An overview. *IEEE Signal Processing Magazine*, 35(1):53–65, 2018.

[15] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020.

[16] Urban Sound Data Project. Urban sound data benchmark, 2025. URL https://www.urban-sound-data.org/. Accessed: 2025-08.

[17] Durk P. Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1×1 convolutions. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

[18] Martin Spitznagel and Janis Keuper. Urban sound propagation: a benchmark for 1-step generative modeling of complex physical systems, 2024.

[19] Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *ICML*, 2015.

[20] George Papamakarios, Eric Nalisnick, Danilo Rezende, Shakir Mohamed, and Balaji Lakshminarayanan. Normalizing flows for probabilistic modeling and inference. *JMLR*, 22(57):1–64, 2021.