forum acusticum 2026
8-12 September · Graz, Austria

URBAN SOUNDSCAPE GENERATION FROM VIDEO: DOMAIN-SPECIFIC FINE-TUNING OF A MULTIMODAL AI MODEL

Yuqi Liang[1], Francesco Aletta[1], Andrew Mitchell[2], Jian Kang[1]

[1] *UCL Institute for Environmental Design and Engineering, The Bartlett, University College London, Central House, 14 Upper Woburn Place, London WC1H 0NN, UK*
[2] *Connected Places Catapult, London*

This paper is a revised version of the authors' submitted manuscript that was peer-reviewed by at least two anonymous reviewers.

Abstract

Soundscape research has received growing attention in recent years, particularly in relation to machine learning-based prediction and multimodal applications. Recent multimodal audio generation models have shown promising performance on general audiovisual content, but their applicability to urban soundscape generation has received limited attention. This paper presents a pilot study on adapting the pre-trained MMAudio model for video-conditioned urban soundscape generation. Urban scene recordings were segmented into 8 s clips using overlapping temporal sampling with a stride of 4 s, resulting in a dataset of 6,946 clips. A lightweight adaptation pipeline was developed on top of the original MMAudio framework, including custom clip loading, streaming TensorDict memmap feature extraction, and full fine-tuning in 16 kHz mode on a single NVIDIA L4 GPU. The study establishes a practical baseline workflow for urban soundscape generation under limited computational resources. Preliminary results suggest that domain-specific fine-tuning is feasible and can move generated outputs closer to the target urban soundscape domain, while also highlighting limitations related to data scale, simple textual conditioning, feature completeness, and the need for fuller evaluation. The work provides an initial step toward video-conditioned urban soundscape generation.

Keywords: *Urban soundscape, Video-conditioned audio generation, Fine-tuning, MMAudio*

Corresponding author: *yuqi.liang.17@ucl.ac.uk*

1. Introduction

Soundscape research explores how people perceive and experience the acoustic environment[1], and the field has a wide range of applications, such as urban planning and environmental psychology. Urban soundscapes are shaped by the interplay of acoustic events, visual context, environmental background, and human perception. Integrating multimodal data (particularly audio and visual data) has significantly advanced soundscape research. When auditory and visual stimuli are combined, attention paid to the visual cues can change the subjective perception of sound and vice versa[2]. Studies demonstrate a strong correlation between visual factors and soundscape perception[3], highlighting the importance of visual information. Modelling these relationships remains a longstanding challenge in soundscape research, particularly because it requires integrating multiple types of data, including visual, acoustic, and perceptual information.

Early work in audio generation focused primarily on waveform synthesis or spectrogram-based generation from audio-only inputs[4], [5]. More recent approaches extend this paradigm to multimodal settings, such as text-to-audio or video-to-audio generation, where visual or semantic cues provide contextual constraints on the generated sound. These models aim not only to produce acoustically plausible signals, but also to ensure semantic consistency and temporal synchronization between modalities. MMAudio[6] is a recent multimodal video-to-audio generation framework that addresses several limitations of earlier approaches. Rather than relying solely on relatively small-scale video–audio datasets, MMAudio adopts a joint training strategy that leverages large-scale text–audio data alongside video–audio pairs. This design enables the model to learn semantically meaningful audio representations while maintaining alignment

with visual input. In addition, MMAudio incorporates a conditional synchronization mechanism that aligns visual features with audio latent representations at the frame level, improving temporal coherence between generated audio and video content. The model is trained using a flow-matching[7] objective and achieves competitive performance in terms of audio quality, semantic alignment and audio–visual synchrony, while remaining computationally efficient. However, these results have mainly been demonstrated on general audiovisual data, and it remains unclear how well such a model adapts to the urban soundscape, where scenes often contain ambient background sound and multiple simultaneous sources.

To investigate this issue, we fine-tune MMAudio on a custom urban dataset using a resource-constrained workflow based on pre-extracted memmap features. The dataset was derived from the International Soundscape Database (ISD)[8], a large-scale resource compiled from soundscape assessment campaigns conducted across Europe and China. ISD follows the standardized SSID Protocol[9], which integrates in situ soundscape experience questionnaires, ambisonics recordings, binaural recordings, monaural recordings, video captures, acoustic data, and other environmental information and in accordance with the ISO 12913[10] standard for soundscape data collection. The contributions of this paper are threefold. First, it presents a practical pipeline for adapting MMAudio to an urban soundscape dataset. Second, it establishes a feature extraction and fine-tuning workflow that can be implemented on limited computational resources. Third, it provides an initial comparison between ground-truth audio and the outputs generated by the pretrained and fine-tuned models, together with a discussion of the main limitations and future directions. The overall aim is not to claim a final solution, but to establish an initial baseline for video-conditioned urban soundscape generation.

2. Method

2.1. Dataset preparation

A custom urban video dataset derived from ISD[8] was prepared for this study. The source material consisted of approximately 30 long-form urban soundscape recordings, each with a duration of around 14–20 min. These recordings include typical public urban scenes such as streets, markets, pedestrian areas, and other shared public spaces. To construct model-ready samples, the videos were segmented into 8 s clips using a sliding-window approach with a stride of 4 s for data augmentation.

The primary experiment reported in this paper uses 6,946 overlapping clips. The data were divided into training, validation, and test subsets using TSV files that define clip identifiers and labels. The final split used in the main experiment contains 5,556 training clips, 695 validation clips, and 695 test clips. During the current stage of the study, a single generic text label, “urban soundscape”, was used as the textual conditioning signal.

Temporal sampling density is an important design choice in this dataset construction process. The use of a stride shorter than the clip duration produces overlapping neighboring clips from the same source recording, thereby increasing the number of available training samples. In practice, this acts as a denser temporal sampling strategy and may be regarded as a simple form of data augmentation, although adjacent samples remain highly correlated [11]. For this reason, overlap must be handled carefully when defining training and evaluation splits. In the present pilot study, stride is treated primarily as a dataset construction variable rather than as a central experimental contribution.

2.2. Feature extraction

Following the MMAudio training workflow, audiovisual features were extracted in advance and stored as TensorDict memmap tensors for efficient fine-tuning. The extraction pipeline used the official MMAudio framework together with lightweight custom modifications for folder-based clip loading and dataset-specific handling. The extracted representations include audio latent means and standard deviations from the MMAudio VAE[12], CLIP-based visual features[13], Synchformer-based synchronization features[14], and CLIP text embeddings. Of the 6,946 allocated clips, 6,860 were successfully processed. Feature completeness varied across modalities, and 5,210 clips contained valid features for all modalities and were therefore used for training.

To avoid excessive RAM usage during large-scale extraction, a streaming memmap strategy was adopted: each batch was encoded, written directly to disk, and then discarded before the next batch was processed. This approach enabled stable extraction without requiring the full dataset to be kept in memory. From a practical perspective, this step was essential for making the workflow feasible in a small-scale experimental setting.

2.3. Model adaptation and fine-tuning

A full fine-tuning strategy was adopted using the pretrained mmaudio_small_16k. The task is formulated as conditional audio generation from video input, with urban soundscape content as the target domain. The core MMAudio architecture was left unchanged. Instead, the adaptation focused on the surrounding training pipeline, including data loading, sequence configuration, extracted feature compatibility, and sample generation robustness.

Several practical modifications were necessary to stabilize the workflow. These included aligning sync_seq_len with the extracted synchronization features, supporting both caption and label columns in TSV files, ensuring consistency between training and evaluation configurations, and safely skipping non-critical failures during video joining and automatic metric evaluation. These changes do not modify the

generative backbone itself, but they make the official framework usable for a urban soundscape dataset under constrained experimental conditions.

3. Experimental setup

Fine-tuning was conducted in 16 kHz mode using the pretrained mmaudio_small_16k checkpoint. Training relied on pre-extracted memmap features and the main experiments were run in Google Colab using a single NVIDIA L4 GPU. The batch size was set to 2, which means that one epoch corresponds to approximately 2778 iterations given 5556 training clips.
Because the current extraction output does not provide a fully valid feature set for every allocated sample and modality, the present study should be understood as a pilot implementation focused on practical feasibility rather than exhaustive optimization. The present submission therefore emphasizes the training workflow, loss behavior, and a lightweight comparison between ground-truth, pretrained, and fine-tuned outputs. Fuller evaluation, including broader evaluation metrics and perceptual listening assessment, is reserved for future work.

4. Results

This section presents the main observations from the pilot study, focusing on the training process and qualitative comparison between ground-truth, pretrained, and fine-tuned outputs. Rather than full evaluation, the aim here is to provide initial evidence on whether the adapted MMAudio pipeline can be used for urban soundscape generation from video under limited computational resources.

4.1. Training process

The training process provides initial evidence that domain adaptation is feasible. Figure 1 shows the training and validation loss during fine-tuning and is intended to illustrate overall optimization behavior rather than to claim definitive convergence. In the present setting, training loss is logged more frequently than validation loss, while the validation curve serves as the main indicator of generalization behavior.

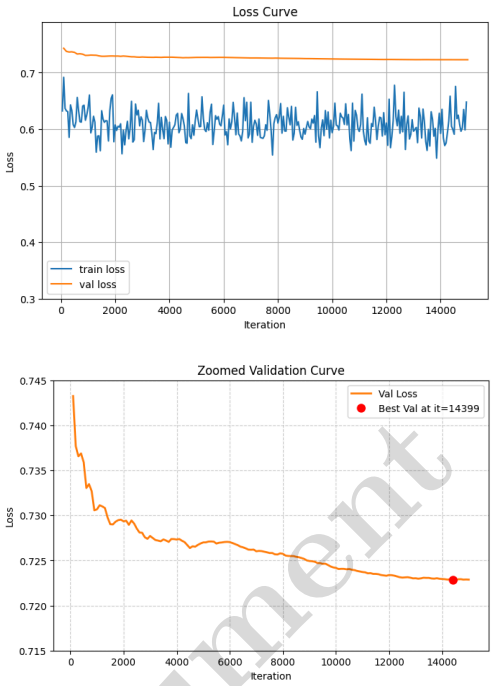

Figure 1: Training and validation loss curves for MMAudio fine-tuning on the urban soundscape dataset. (a) Full loss curves showing both training and validation loss over 15k iterations. (b) Zoomed view of the validation loss, with the best checkpoint (iteration 14,399) marked in red.

The loss curves suggest that the model adapts to the urban soundscape training distribution, while the validation curve provides a more informative signal of generalization beyond the seen clips. The model converges rapidly within the first 2,000 iterations, with validation loss stabilizing around 0.719 after approximately 10,000 iterations. Because the present study uses a modest dataset and a small batch size, some fluctuation in validation loss is expected. Therefore, the most meaningful checkpoint is not necessarily the final one, but rather the one with the lowest validation loss or the best overall balance between quantitative metrics and qualitative output quality.

4.2. Initial comparison of generated audio

An initial comparison is also presented between the ground-truth audio and the outputs generated by the pretrained and fine-tuned MMAudio models. This is shown using spectrogram-based visual comparison in Figure 2. The purpose of this figure is not to replace formal evaluation, but to provide an initial indication of whether fine-tuning moves the generated outputs closer to the acoustic characteristics of the target urban soundscape.

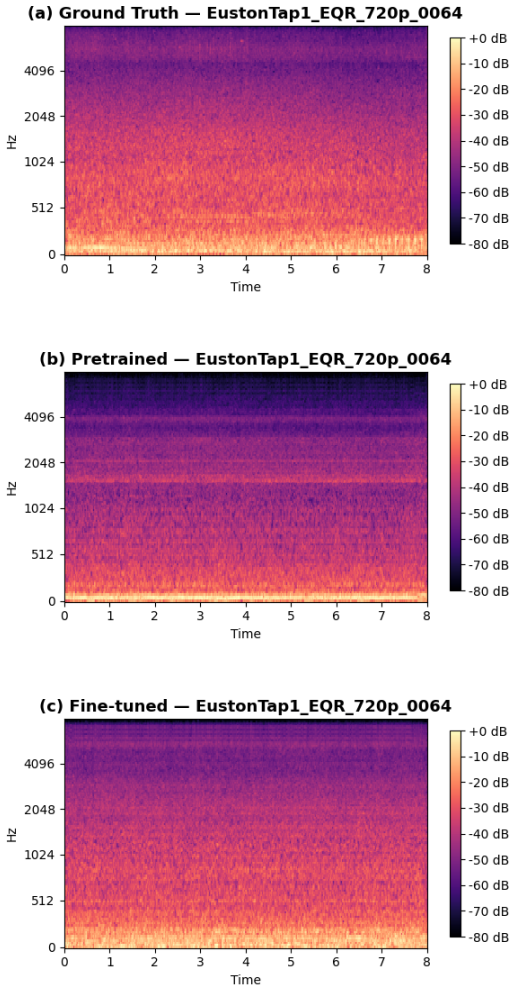

Figure 2: Spectrogram comparison of the ground-truth audio and the audio generated by the pretrained and fine-tuned MMAudio models for one representative clip from the test set.

Table 1: Mean spectrogram-distance values over 30 randomly selected test clips. Lower values indicate closer agreement with the ground-truth spectrogram.

Metric	Pretrained	Fine-tuned	Δ
L1	11.610	11.066	−0.544
MSE	217.378	203.379	−13.999
LSD	13.319	12.539	−0.780

Figure 2 presents a preliminary spectrogram comparison between the ground-truth audio, the audio generated by the pretrained MMAudio model, and the audio generated by the fine-tuned model for a representative test clip. All three signals show a low-frequency-dominant structure, which is typical of urban ambient soundscapes. Compared with the pretrained output, the fine-tuned result exhibits a smoother broadband spectral distribution and fewer pronounced narrow-band horizontal patterns in the mid- and high-frequency regions, making it visually closer to the ground truth.

To complement this visual comparison, a lightweight spectrogram-distance analysis was conducted on 30 samples randomly selected from the test set. For each sample, distances between the generated output and the ground-truth spectrogram were computed using L1 distance, mean squared error (MSE), and log-spectral distance (LSD), where lower values indicate closer agreement with the target spectrum. Table 1 reports the mean values of these distances over the 30 sampled test clips. Across all three measures, the fine-tuned model achieved smaller average distances than the pretrained baseline. Although clear differences from the target signal remain, these results provide initial quantitative evidence that fine-tuning moves the generated audio closer to the target urban soundscape domain.

5. Discussion

The present findings should be interpreted cautiously. First, the dataset remains relatively small compared with the scale of the original pretraining regime, making overfitting a realistic concern during longer fine-tuning runs. Second, the text conditioning used here is deliberately simple, relying on a single generic caption. This reduces semantic diversity and limits controllability. Third, the current evaluation remains limited. In an urban soundscape context, conventional automatic audio metrics do not necessarily capture scene-level plausibility [15], environmental realism, or perceptual suitability, so lightweight comparisons are useful but not sufficient on their own.

A further issue concerns the role of datasets in soundscape AI. As highlighted in our recent literature review, existing soundscape datasets often differ widely in recording methods, scale, perceptual framework, and accessibility[16]. The present pilot study reflects this broader problem: beyond model design itself, the practical usability of a dataset for multimodal generation depends strongly on how consistently its audio, visual, and metadata components can be reorganized into a trainable format. From this perspective, database scale, structure, and comparability remain central challenges for future soundscape AI research.

Within this scope, the current results nevertheless suggest that domain-specific fine-tuning is meaningful. The loss behavior indicates that the pretrained model can adapt to the urban soundscape training distribution, while the spectrogram comparison suggests that the fine-tuned outputs become closer to the ground-truth than the untuned pretrained baseline. For a pilot study of this scope, this kind of initial comparison is valuable because it addresses a practical question: whether fine-tuning produces outputs that are at least more consistent with the target data distribution.

At the same time, the current study should not be taken as a definitive evaluation of urban soundscape generation quality. The spectrogram comparison is informative as a first qualitative analysis, but it cannot fully capture perceptual realism, audiovisual

consistency, or scene-level appropriateness. A more complete assessment would ideally combine broader evaluation metrics with listening-based perceptual evaluation.

Overall, the present results suggest that pretrained multimodal audio generation models are promising for soundscape-related applications, but further work is needed in three main directions: larger and more diverse datasets, richer caption conditioning strategies, and evaluation protocols that are better aligned with environmental soundscape perception.

6. Conclusion

This paper presented a pilot study on adapting the pretrained MMAudio model to urban soundscape generation from video. Using a custom urban dataset derived from the International Soundscape Database (ISD) and a lightweight fine-tuning workflow, the study established a practical baseline for domain-specific multimodal audio generation under limited computational resources. The main contribution is therefore methodological and practical: it shows that the official MMAudio framework can be adapted to urban soundscape data through a practical and workable pipeline.

The preliminary findings suggest that domain-specific fine-tuning can move generated audio closer to the target urban soundscape domain, but they also underline the current limitations of small-scale experiments, simple textual conditioning and limited evaluation. The work should therefore be understood as an initial baseline rather than a final generative solution. Future work will extend the present study through fuller objective evaluation, richer caption conditioning, and larger datasets.

References

[1] B. Schulte-Fortkamp, A. Fiebig, J. A. Sisneros, A. N. Popper, and R. R. Fay, *Soundscapes: Humans and Their Acoustic Environment*, vol. 76. Cham, Switzerland: Springer International Publishing, 2023.

[2] M. Southworth, "The Sonic Environment of Cities," *Environ. Behav.*, vol. 1, no. 1, p. 49, 1969, doi: 10.1177/001391656900100104.

[3] X. Ren and J. Kang, "Effects of the visual landscape factors of an ecological waterscape on acoustic comfort," *Applied Acoustics*, vol. 96, pp. 171–179, 2015, doi: 10.1016/j.apacoust.2015.03.007.

[4] A. van den Oord and others, "WaveNet: A Generative Model for Raw Audio," *arXiv preprint arXiv:1609.03499*, 2016, doi: 10.48550/arXiv.1609.03499.

[5] S. Mehri and others, "SampleRNN: An Unconditional End-to-End Neural Audio Generation Model," *arXiv preprint arXiv:1612.07837*, 2017, doi: 10.48550/arXiv.1612.07837.

[6] H. K. Cheng, M. Ishii, A. Hayakawa, T. Shibuya, A. Schwing, and Y. Mitsufuji, "Taming Multimodal Joint Training for High-Quality Video-to-Audio Synthesis," *arXiv preprint arXiv:2412.15322*, Dec. 2024, doi: 10.48550/arXiv.2412.15322.

[7] Y. Lipman, R. T. Q. Chen, H. Ben-Hamu, M. Nickel, and M. Le, "Flow Matching for Generative Modeling," *arXiv preprint arXiv:2210.02747*, 2022.

[8] A. Mitchell and others, "The International Soundscape Database: An integrated multimedia database of urban soundscape surveys – questionnaires with acoustical and contextual information," *Zenodo*, Nov. 2021, doi: 10.5281/zenodo.5914762.

[9] A. Mitchell *et al.*, "The Soundscape Indices (SSID) Protocol: A Method for Urban Soundscape Surveys—Questionnaires with Acoustical and Contextual Information," *Applied Sciences*, vol. 10, no. 7, p. 2397, 2020, doi: 10.3390/app10072397.

[10] International Organization for Standardization, "ISO/TS 12913-2:2018 Acoustics—Soundscape: Part 2: Data collection and reporting." Geneva, Switzerland, 2018.

[11] A. Guzhov, F. Raue, J. Hees, and A. Dengel, "ESResNet: Environmental Sound Classification Based on Visual Domain Models," *arXiv preprint arXiv:2004.07301*, 2020.

[12] D. P. Kingma and M. Welling, "Auto-Encoding Variational Bayes," in *International Conference on Learning Representations (ICLR)*, 2014.

[13] A. Radford *et al.*, "Learning Transferable Visual Models From Natural Language Supervision," in *Proceedings of the 38th International Conference on Machine Learning (ICML)*, 2021, pp. 8748–8763.

[14] V. Iashin, W. Xie, E. Rahtu, and A. Zisserman, "Synchformer: Efficient Synchronization from Sparse Cues," *arXiv preprint arXiv:2401.16423*, 2024.

[15] J. Y. Jeon and H. I. Jo, "Effects of audio-visual interactions on soundscape and landscape perception and their influence on satisfaction with the urban environment," *Building and Environment*, vol. 169, p. 106544, 2020, doi: 10.1016/j.buildenv.2019.106544.

[16] Y. Liang, A. Mitchell, J. Kang, and F. Aletta, "A Review of Soundscape Datasets: Challenges and Prospects for Multimodal Research," *IEEE Transactions on Affective Computing*, 2026, doi: 10.1109/TAFFC.2026.3659084.