Supplementary Material

1. Overview

The supplementary material includes the components:
- Additional Visualizations
- Dataset Details
- Dataset Analysis
- Visual Explanation of Geo-Contextual Metrics Suite
- User Study Details
- Additional Experiments

2. Additional Visualizations

We provide additional landscape images in Figure 1, generated by SounDiT and covering diverse visual scene attributes on both SonicUrban and SoundingSVI.

3. Dataset Details

The construction of our proposed datasets is motivated by the limitations of existing audio-visual datasets, such as VGGSound [6], IntoTheWild [12], and Landscape [13]. These datasets primarily focus on object sounds, human voices, natural phenomena, and have limited types of scenes, as illustrated in Figure 2. Therefore, these datasets are unable to comprehensively capture geo-contextual sounding and visual environments, such as parks vs. beaches, and urban vs. rural regions. As a result, audio2image models trained on these datasets often produce unrealistic representations of the real-world environmental settings, such as those artistic-style images, as illustrated in Figure 3. To support the GeoS2L problem, we propose two large-scale, geo-contextual multimodal datasets, **SoundingSVI** and **SonicUrban**, each consisting of paired soundscape–landscape data, which enable multimodal generative modeling within real-world geographic settings. In this section, we introduce the dataset construction pipelines and data processing methods presented in Figure 4, and conduct a comprehensive dataset analysis to illustrate the diversity of these two datasets.

3.1. SoundingSVI Construction Details

Data preparation. Geotagged soundscapes and Street View Images (SVI) were collected to construct soundscape-landscape pairs, as illustrated in Figure 4. Geotagged soundscape recordings were retrieved from the Aporee platform [16]. For each soundscape recording, SVIs from different directions and timestamps were downloaded via Google based on longitude and latitude. It should be noted that each soundscape might be associated with multiple street view imagery candidates. In total, 51,059 raw soundscapes representing 3,470 hours of audio and 553,522 SVIs were downloaded.

Data format. The soundscape recordings were sampled at 44.1kHz. SVIs were retrieved at a resolution of 400×400 pixels.

Data cleaning. After retrieving raw soundscape recordings, they were further segmented into 10-second audio clips. As some clips are dominated by human voices, which may not reflect real-world environmental soundscapes, a human voice detection method was employed to filter clips comprised of over 40% human speech or singing [18]. After filtering, each audio clip was paired with multiple SVI candidates from the same location. A sound source localization model was then applied to each pair to identify the most relevant image across SVI candidates [17], ultimately resulting in 169,221 soundscape-landscape pairs. To provide additional scene prompts during training, a scene classification model was subsequently applied to each pair [7]. The predicted scene was then input into the SounDiT as an optional scene prompt.

3.2. SonicUrban Construction Details

Data preparation. Unedited YouTube videos were used to construct soundscape-landscape pairs for SonicUrban (Figure 4). Videos were retrieved by searching combinations of city names and keywords, such as "city walk no commentary; New York", "urban exploration no talking; London", and "urban exploration no voiceover; Paris". A word cloud image of all keywords is shown in Figure 5, highlighting search terms. To ensure these videos reflect real-world sounding and visual environments, we manually removed videos that included editing, subtitles, background music, speech-dominated audio, non-environmental sounds, static scenes, or fixed camera angles, resulting in a total of 1,167 street roaming videos. We then segmented each video into 10-second audio clips and evenly extracted 10 frames from each clip.

Data format. The audio clips were segmented at a sample rate of 44.1 kHz. Frames were extracted at a resolution of 854×480 pixels.

Data cleaning. Following a similar data cleaning process as SoundingSVI, a human voice detection model was first applied to filter speech-dominated audio clips [18] and a sound source localization model was used to select the most relevant frame for each remaining audio clip [17]. Finally, a scene classification model was employed to predict scene type for each pair to provide additional geographic semantics during training [7]. This process resulted in 236,674 soundscape-landscape pairs with corresponding scene prompts. The predicted scene was then input into SounDiT as an optional scene prompt.

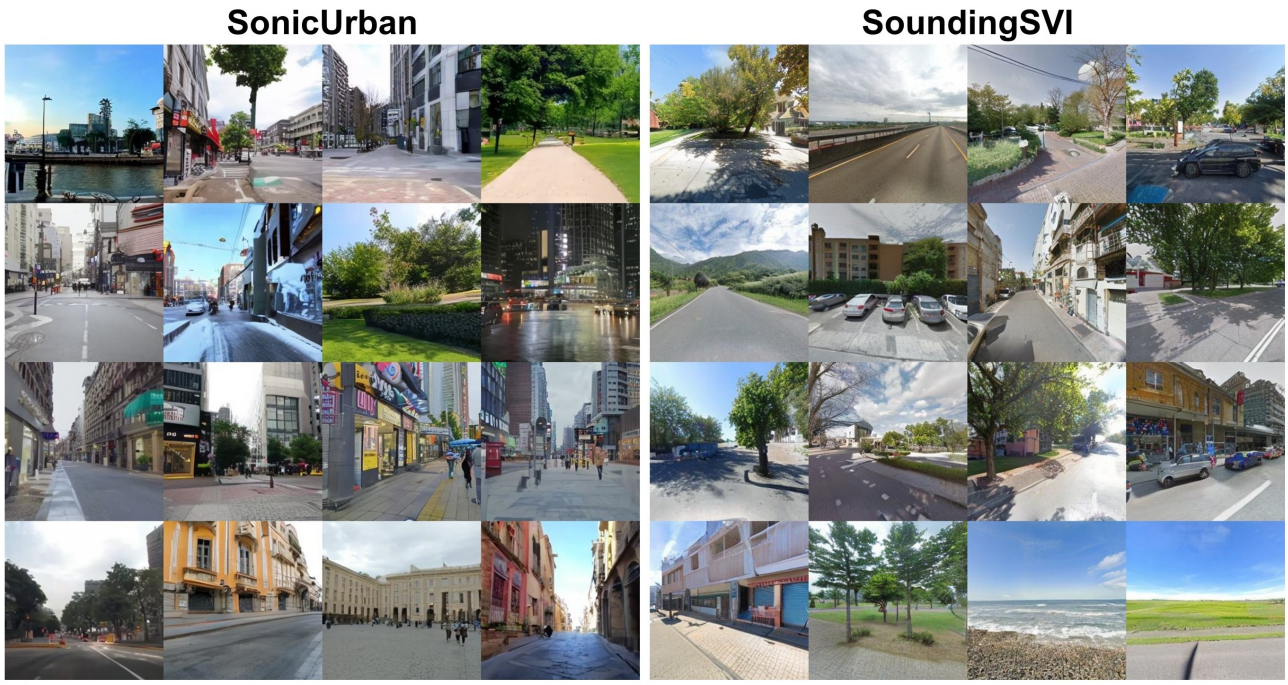

Figure 1. **Additional landscape images** generated by SounDiT across different visual scene attributes on both SonicUrban and SoundingSVI.

Figure 2. **Example samples from current audio–visual dataset.** Categories include object sounds (e.g., violin and crackling fire), human voices (e.g., people whispering and baby laughing), and weather sounds (e.g., rain and thunder). Limited attention has been paid to geographic contexts and environmental scenes.

Figure 3. **Unrealistic images with limited geographic contexts** generated using AudioToken trained on the VGGSound dataset.

3.3. Data Processing Methods

Human Voice Detection. While most clips capture meaningful environmental soundscapes, some audio clips are dominated by human speech or singing. To filter such clips, we employ a pre-trained Voice Activity Detection (VAD) model, Silero VAD [18], to estimate the duration of human voice within each clip. Silero VAD is trained on recordings spanning over 6,000 languages, enabling robust detection across diverse linguistic and acoustic environments. In our study, each 10-second audio clip is divided into 512 uniform segments. For each segment, the model detects a set of start–end time pairs corresponding to detected voice activities. We aggregate the time represented by these pairs across all segments to arrive at an overall voice activity rate and exclude clips exceeding 40% human voice. This filtering process ensures audio clips more accurately reflect real-world environmental soundscapes and better support applications in geography, urban planning, and environmental sciences.

Sound Source Localization. Each soundscape in Sound-

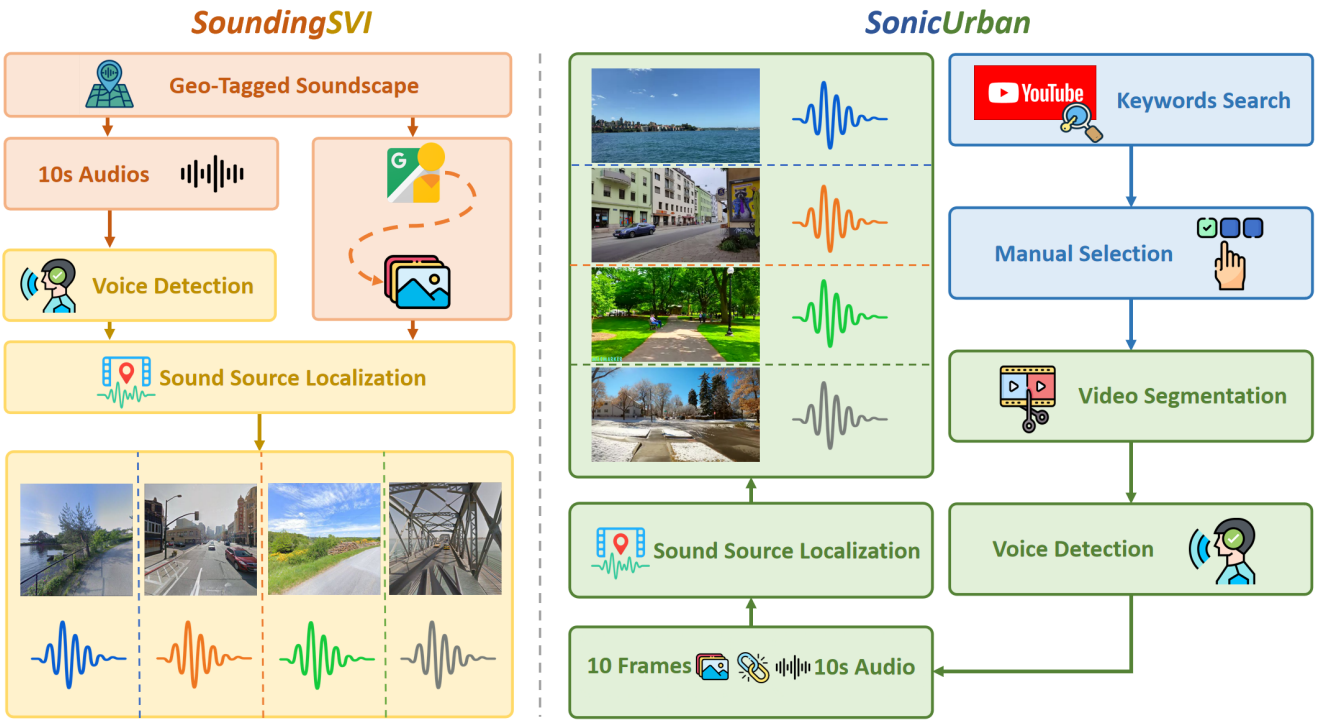

Figure 4. **Dataset construction pipelines for the SonicUrban and SoundingSVI datasets. Left (SoundingSVI):** Geotagged soundscape recordings were retrieved from the Aporee platform, and Street View Image (SVI) candidates were obtained from Google based on location. Raw recordings were segmented and filtered by voice detection to remove human-voice dominated clips. Sound source localization methods were then employed to pair each audio clip with its most relevant SVI. **Right (SonicUrban):** Unedited YouTube videos were collected via keyword search and manual selection. Each video was segmented into clips, and multiple frames were extracted per clip. Human detection and sound source localization methods were applied to construct soundscape-landscape pairs to ensure soundscapes reflect real-world sounding environments and maintain coherence.

Figure 5. **Word cloud of keywords for retrieving the YouTube videos.** To build SonicUrban, we paired city names with keywords, such as "city walk no commentary; New York", "urban exploration no talking; London", and "urban exploration no voiceover; Paris", to identify unedited YouTube videos that capture real-world geographic settings.

ingSVI and SonicUrban is associated with multiple candidate images. In SoundingSVI, nearby candidate images are selected based on the soundscape recording's coordinates; While in SonicUrban, each soundscape clip is paired with ten frames evenly extracted from the video. To se-

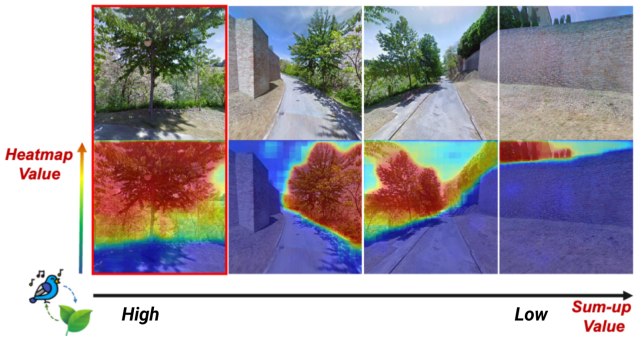

Figure 6. **Example of sound source localization.** The heatmap highlights image regions associated with the sound source, with color intensity from blue (lower association) to red (higher association) indicating the strength of this linkage. In this example, a bird chirping soundscape file is introduced, and its four nearby landscape image candidates are selected. As bird chirping is typically associated with trees, the image with the largest tree proportion, which also yields the highest heatmap sum-up score, is selected as the most relevant.

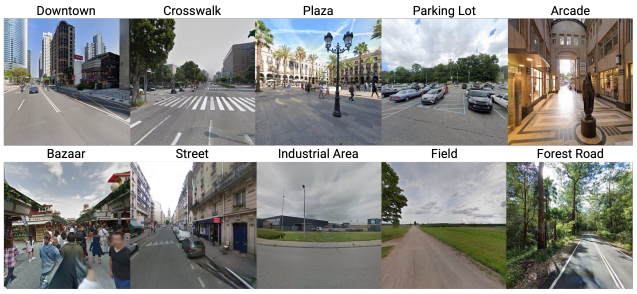

Figure 7. **Ten sample scene categories defined by Place365 dataset,** including downtown, crosswalk, plaza, parking lot, arcade, bazaar, street, industrial area, field, and forest road.

lect the most contextually relevant image from all candidate images, we employ a pre-trained audio-visual model, ACL-SSL [17]. We utilize this model to generate a sound localization heatmap for each image-audio pair, which highlights regions in the image closely related to the sound source. Then, we compute a similarity score by summing the heatmap values of each audio-image pair, and then select the candidate image with the highest score. By doing so, we select the candidate image that has the most relevant geographic context to the soundscape. Figure 6 illustrates this process with a soundscape dominated by bird chirping and four candidate images.

Scene Classification. In SounDiT, a scene prompt serves as an additional geographic scene semantic to enhance geo-contextual coherence of generated landscape images. To generate these prompts, we use a ResNet model [7], pre-trained on Place365 dataset [21], which enables the prediction of scenes from 365 categories, such as downtown, crosswalk, and plaza. Figure 7 provides several sample images across multiple scenes. For each soundscape-landscape pair, the landscape image is processed by this model, and the top-1 scene prediction is selected as the scene prompt. Moreover, we leverage the same method to perform the scene-level evaluation to compute the PSS scores. In SounDiT, scene prompts serve as additional geographic semantics to enhance the geo-contextual coherence of generated landscape images. To generate these prompts, we utilize an efficient open-source Vision-Language Model (VLM)—Qwen-2.5-VL [1]. Adopting the taxonomy of the Places365 dataset [21], we prompt Qwen-2.5-VL to predict scene categories for the landscape component of each soundscape-landscape pair. By enforcing strict output formatting and conducting iterative verification, we ensure a reliable scene prompt for each landscape image. Figure 7 displays sample images across multiple scene categories.

3.4. Benchmark Datasets

Although the current datasets for the GeoS2L task exhibit several limitations (as discussed in Sec. 3), we further evaluate our SounDiT models by constructing additional benchmarks from VGGSound, intothe wild and landscape.

Geo-Contextual VGGSound Subset. The original VGGSound dataset [6] contains a total of over 210,000 videos, consisting of over 500 hours of footage across 310 distinct classes such as people, animal, and music. Since most of these categories are not associated with geography and environmental sciences, we construct a geo-contextual benchmark for the GeoS2L task by curating a geo-contextual subset of the VGGSound. To accomplish this, we select 16 classes (e.g., footsteps on snow or wind rustling leaves) likely associated with geographic environments from the VGGSound dataset. In total, 6,713 soundscape-landscape image pairs are maintained.

Combined Landscape and IntoTheWild Dataset. The Landscape dataset [12] comprises a total of 1178 videos over 11 unique scenes. However, most videos are long, static recordings, limiting their scene diversity. In contrast, the IntoTheWild dataset [13] includes a total of 94 videos from three unique classes, forest, rain, and snow, captured from videos taken during hikes. Although smaller in total duration, IntoTheWild offers greater geographic diversity. Thus, we merge the two datasets to form a total of 1889 soundscape-landscape pairs, providing a more comprehensive benchmark that covers multiple geographic environmental settings.

Table 1. **Spatial coverage and scene diversity of our proposed datasets and benchmark datasets.**

Dataset	Cities	Countries	Scene Diversity
SoundingSVI	–	90	344
SonicUrban	131	67	363
VEGAS [22]	–	–	–
VGGSound [6]	–	–	15
IntoTheWild [13]	–	–	3
Landscape [12]	–	–	9
SoundingEarth [8]	–	90	–

4. Dataset Analysis

Compared to existing datasets, our proposed datasets demonstrate substantial advantages in both spatial coverage and scene diversity, as summarized in Table 1. In this section, we provide more details regarding these two aspects.

4.1. Spatial Coverage

Our proposed datasets exhibit broad spatial coverage. **SoundingSVI** includes soundscape recordings and SVIs from 90 countries across six continents, and **SonicUrban** covers YouTube city-walk videos from 131 cities across 67 countries. Both datasets significantly expand the spatial footprint compared to prior benchmark datasets, as pre-

sented in Figure 8. In SoundingSVI, major cities in the United States, Europe, Asia, and Australia are represented by over 3,000 soundscape-image pairs, including Los Angeles, Boston, London, Stockholm, Amsterdam, Tokyo, Seoul, Bangkok, and Sydney. Figure 9 shows that the SonicUrban datasets exhibit broad geographic coverage. The soundscape-landscape pairs in the SonicUrban dataset cover 18 cities in the United States, 14 cities in China, 6 cities in the United Kingdom, and cover 4 cities in Australia, Germany, and Italy.

4.2. Scene Diversity

In addition to broader spatial coverage, our proposed datasets offer significantly greater scene diversity compared to existing audio-visual benchmarks. While prior benchmarks provide a few categories of scenes with geographic contexts like forests, our datasets capture a diverse set of geographic contexts and environmental settings. Figure 10 presents street, residential neighborhood, promenade, and crosswalk scenes and comprises over 10,000 soundscape-landscape pairs in both datasets. In SoundingSVI, rural scenes, like forest roads and fields, also exceed 10,000 pairs, indicating strong coverage in both urban and rural settings. More samples from different scenes of both datasets are shown in Figure 11 and 12.

5. Visual Explanation of Geo-Contextual Suite

We provide a visual explanation of the proposed Geo-Contextual Suite, which evaluates the alignment between generated and ground-truth landscapes at three complementary levels. As illustrated in Figure 14, we first compare element-level structures to capture fine-grained spatial composition. We then assess scene-level consistency to ensure that the overall functional and visual character of a place is preserved. Finally, we incorporate human-perception–level attributes to reflect how people experience these environments in real-world.

5.1. Element-level Evaluation

$PSS_{element}$ is computed to evaluate the element-level similarity between generated images and their ground truths. Figure 14 illustrates the element-level evaluation process by comparing $PSS_{element}$ between generated landscape image and its ground truth. Pairs with high-similarity scores demonstrate strong alignment regarding their geographic elements and ratios, indicating the generated image effectively captures the key environmental elements.

5.2. Scene-level Evaluation

PSS_{scene} is designed to assess the consistency of the entire environmental scene between generated images and their ground truths. Figure 15 illustrates the scene-level evaluation process based on the overlaps between the predicted scene categories for both generated images and ground truths. For example, the top 3 predicted scene categories are listed, with shared scenes highlighted.

5.3. Human Perception-level Evaluation

$PSS_{perception}$ is proposed to measure whether the generated and ground truth images evoke similar human perceptions of place. For example, whether a specific neighborhood makes people feel safe or not. Six dimensions of human perceptions are assessed using a DenseNet121 model [10], pre-trained on the MIT Place Pulse dataset [15], including: "Safe", "Beautiful", "Depressing", "Lively", "Wealthy", and "Boring". Figure 16 illustrates human perception-level evaluation examples.

6. Implementation Details

Data Split. We select 10% of each proposed dataset as the test set. For SonicUrban, to ensure spatial diversity captured by YouTube videos across different cities, we sampled 10 % soundscape-landscape pairs from each city individually.

Training Details. All experiments were conducted using NVIDIA H100(98GB), A6000 (48 GB) and A100 (40 GB) GPUs. Our SounDiT model was trained for 400,000 gradient-update steps with a batch size of 128. To simplify hyperparameter tuning and ensure stable convergence, we used the AdamW optimizer with a fixed learning rate of 1×10^{-4}, a weight decay of 0.03, and $\epsilon = 1 \times 10^{-10}$. Additionally, we applied mixed-precision training (FP16) to improve computational efficiency, which supports to perform most operations in half-precision format and fall back to single-precision when needed.

Inference Details. During inference, all landscape images are generated at a resolution of 256×256 pixels. Classifier-free guidance with a scale of 4.0 is applied to both the soundscape and scene conditions to enhance geo-contextual coherence while preserving diversity. For the denoising process, we employ the SA-Solver sampling algorithm with a 30-step denoising schedule, which offers an effective trade-off between visual fidelity and generation speed.

7. User Study Details

While quantitative metrics such as FID and PSS are critical for evaluating model performance in generating landscape images, they may overlook how well generated landscape images maintain geographic and environmental contexts—an aspect crucial to geographic, ecological, and urban planning applications. In such scenarios, human evaluation remains important to assess whether the generated landscape images are both visually and geo-contextually coherent. To this end, we conducted a user study and recruited 17 volunteers to assess the model outputs, which could complement

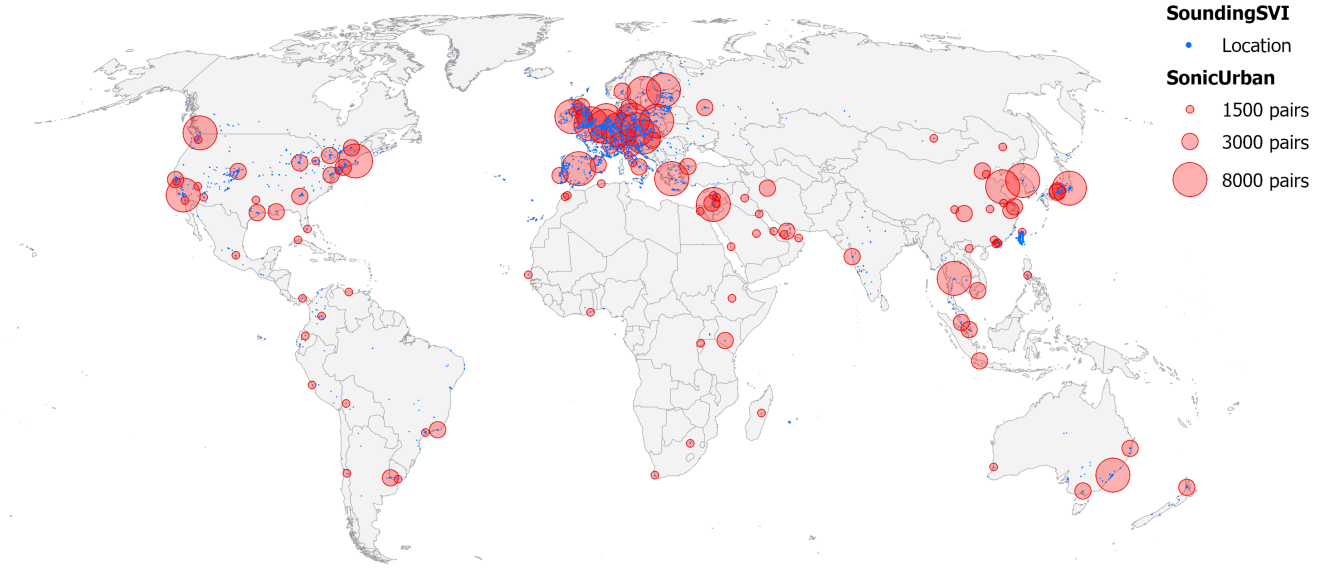

Figure 8. **Global distribution of soundscape-landscape pairs in SoundingSVI and SonicUrban datasets.** Blue points mark individual geotagged soundscape locations in SoundingSVI. Red circles scale with the number of paired soundscape–image clips per city in SonicUrban. Together, the two overlays illustrate the complementary geographic coverage and diversity of our constructed datasets.

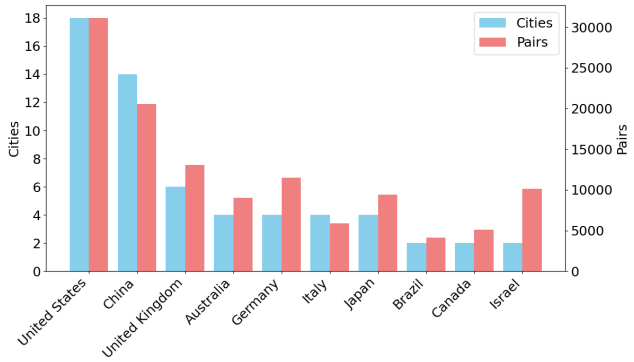

Figure 9. **Country-level distribution of the SonicUrban, demonstrating its widespread spatial coverage.** Bars show the number of cities sampled per country (left axis) and total soundscape-landscape pairs collected (right axis). Top ten countries are listed, with the United States leading at 18 cities (∼31K pairs), followed by China (14 cities, ∼22K pairs), the United Kingdom (6 cities, ∼8K pairs), Australia (6 cities, ∼10K pairs), and additional contributions from Germany, Italy, Japan, Brazil, Canada, and Israel.

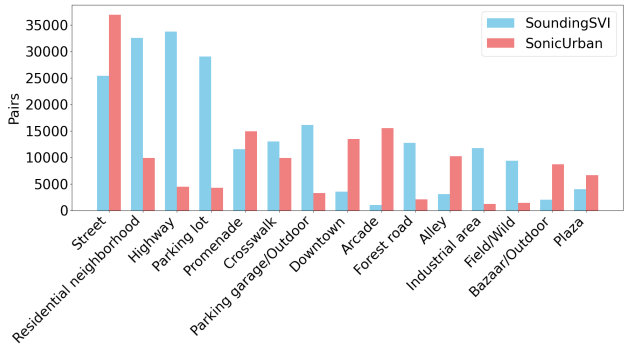

Figure 10. **Scene distribution of SoundingSVI and SonicUrban,** showing the number of soundscape-landscape pairs per scene category. In both datasets, street, residential neighborhood, promenade, and crosswalk account for a large portion of these datasets, indicating broad coverage of urban scenes. SoundingSVI includes a substantial number of rural scenes, like forest road and field.

our evaluation. Each participant is asked to complete the following two tasks.

7.1. Soundscape-to-Landscape Matching.

To assess whether the generated landscape images accurately reflect the geographic characteristics embedded in environmental soundscapes, participants were presented with 15 audio clips, each accompanied by four candidate images. They were asked to select the image that best contextually aligned with the soundscape. To evaluate SounDiT performance across real-world environmental settings, 15 audio clips featuring diverse sounding environments, such as bird chirping and bustling traffic, were randomly sampled from the validation set. For each clip, one candidate image was generated by SounDiT based on the audio, while the remaining three were generated from unrelated geographic

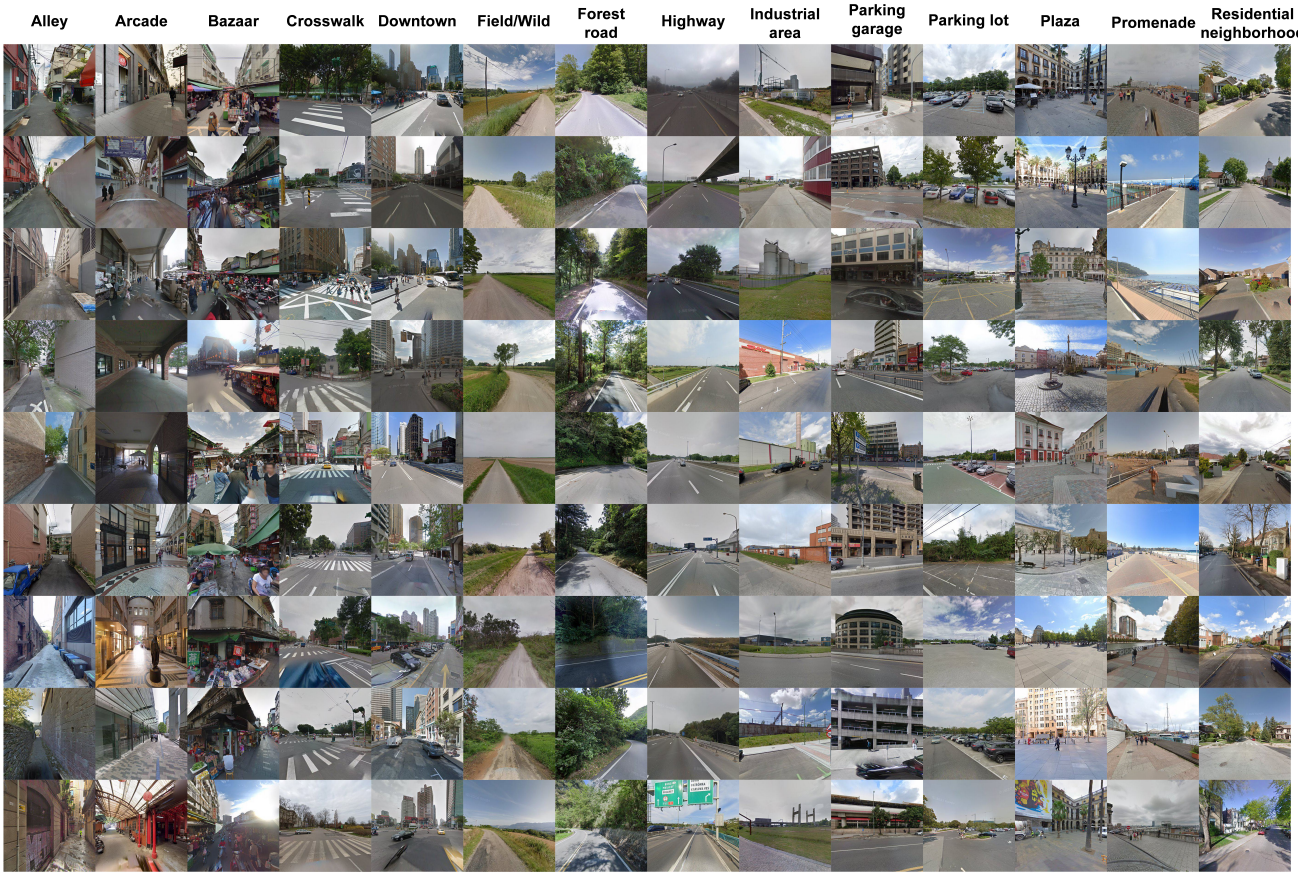

Figure 11. **Various scene samples from SoundingSVI including** alley, arcade, bazaar/outdoor, crosswalk, downtown, field/wild, forest road, highway, industrial area, parking garage/outdoor, parking lot, plaza, promenade, residential neighborhood, street.

contexts. As illustrated in Figure 17 Q1, the example includes an audio clip dominated by lake ripple, with candidate images representing field/wild, overpass, water, and residential area.

7.2. Landscape Matching.

This task evaluates the geographic coherence of the generated landscape images compared to real-world environmental settings. Participants were shown 15 pairs of ground truth landscape images, each accompanied by three candidate images, and asked to select the image that best matched the geographic context of ground truth images. The 15 pairs were randomly sampled from the validation set, representing diverse geographic contexts, such as urban and rural settings. For each pair, one candidate image was generated by SounDiT based on the soundscape associated with the ground truth, while the other two were generated from unrelated geographic contexts. As illustrated in Figure 17 Q2, the example includes two ground truth images of a street, with candidate images representing street (correct match), beach and highway.

The user study was conducted through a custom web interface which collected participant responses. Results show that the average matching accuracy across Task 1 and Task 2 were 90.26% (Standard Deviation (SD) = 0.085) and 82% (SD = 0.073), respectively. These results indicate a strong perceptual alignment between environmental soundscapes and their corresponding generated landscape images, highlighting the model's effectiveness in the integration of geographic knowledge to generate geographically contextual features in soundscapes in practice.

8. Additional Experiment

8.1. Benchmark Experiments

We evaluate SounDiT on the IntoTheWild&Landscape and Geo-Contextual VGGSound benchmark subsets. As shown in Figure 18, our model generates scene-consistent, high-quality visual results. The quantitative results in Table 2 further demonstrate that SounDiT achieves state-of-the-art performance on these benchmark datasets.

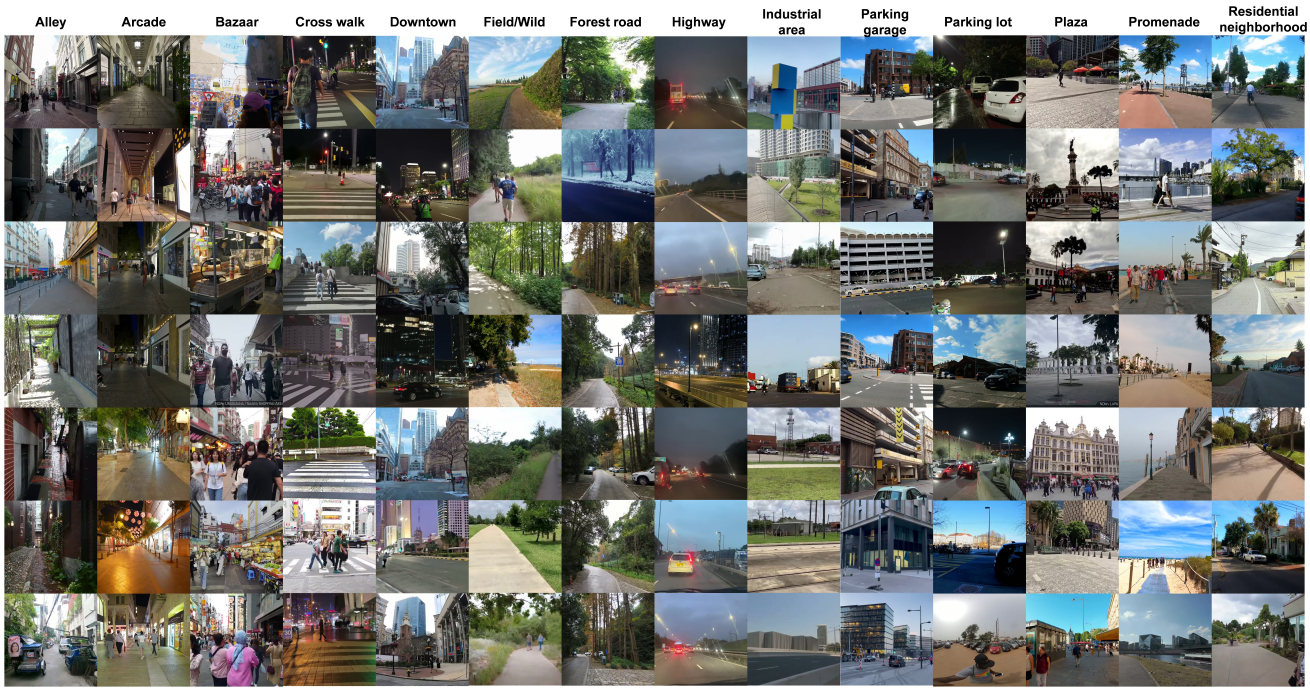

Figure 12. **Various scene samples from SonicUrban including** alley, arcade, bazaar/outdoor, crosswalk, downtown, field/wild, forest road, highway, industrial area, parking garage/outdoor, parking lot, plaza, promenade, residential neighborhood, street.

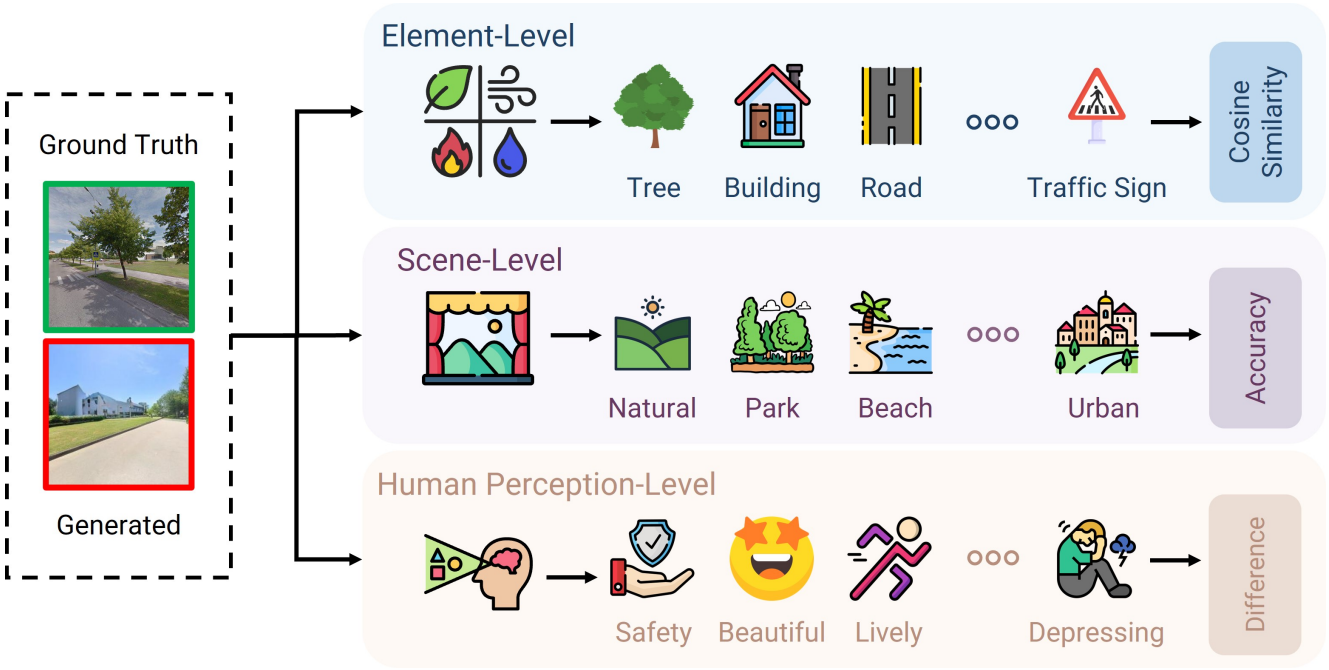

Figure 13. **The ground truth and output landscape images are assessed at the element-, scene-, and human-perception levels following geography and urban planning practices.**

8.2. Visual Ablation

Figure 19 presents visual ablations of the MoE module with different numbers of experts (E = 2, 4, 6, 8). Increasing the

Table 2. **Quantitative evaluation results across SounDiT and baseline models on two benchmark datasets.** † indicates pre-trained models. We compare our SounDiT model with baselines using general metrics (FID, AIS, IIS) and our PSS metrics.

		General Metrics			Place Similarity Score		
Dataset	Method	FID↓	AIS↑	IIS↑	Element↑	Scene↑	Perception↓
IntoTheWild & Landscape (1.8K)	CoDi†	121.218	0.701	0.706	0.447	0.698	0.817
	Sound2Scene	140.244	0.617	0.741	0.479	0.640	0.769
	AudioToken†	209.515	0.600	0.579	0.219	0.280	0.913
	AudioToken (SD1)	138.785	0.585	0.695	0.364	0.534	0.846
	AudioToken (SD2)	118.536	0.637	0.740	0.512	0.724	0.855
	GlueGen	195.625	0.584	0.645	0.362	0.582	0.815
	PixArt + MHCA	105.647	0.687	0.744	0.521	0.741	0.757
	SounDiT (Ours)	**97.471**	**0.718**	**0.764**	**0.618**	**0.803**	**0.684**
Geo-Contextual VGGSound Subset (6.7K)	CoDi†	129.185	**0.661**	**0.639**	0.317	0.382	0.828
	Sound2Scene	122.460	0.600	0.596	0.287	0.378	0.770
	AudioToken†	159.366	0.578	0.564	0.196	0.204	0.942
	AudioToken (SD1)	163.540	0.532	0.555	0.207	0.311	0.789
	AudioToken (SD2)	157.706	0.583	0.587	0.251	0.404	0.803
	GlueGen	156.777	0.536	0.547	0.233	0.249	0.839
	PixArt + MHCA	113.111	0.575	0.588	0.336	0.392	0.767
	SounDiT (Ours)	**90.831**	0.637	0.602	**0.394**	**0.411**	**0.718**

Figure 14. **Examples of element-level evaluation** $PSS_{element}$. The top-3 predicted elements and their ratios are presented for both ground truth and generated images.

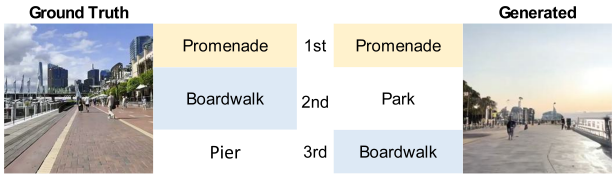

Figure 15. **Examples of scene-level evaluation** PSS_{scene}. The example includes the generated image and its ground truth, along with their top 3 predicted scene categories.

number of experts leads to more faithful and diverse generations, which is consistent with the quantitative improvements reported in our ablation study.

9. Broader Impacts

This work has significant implications across computer science, geography, urban planning, and environmental sciences. By linking the auditory and visual environments of places, our approach not only advances our knowledge of human-environment interactions, but also supports real-world practices. First, as two critical aspects of space, geographers and environmental scientists could better understand how different environmental elements interact and in-

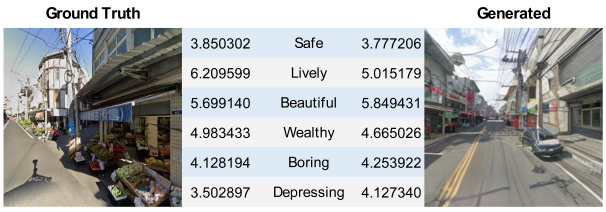

Figure 16. **Examples of perception-level evaluation** $PSS_{perception}$. For the example, perceptions of each ground truth image are listed, along with scores of its generated image.

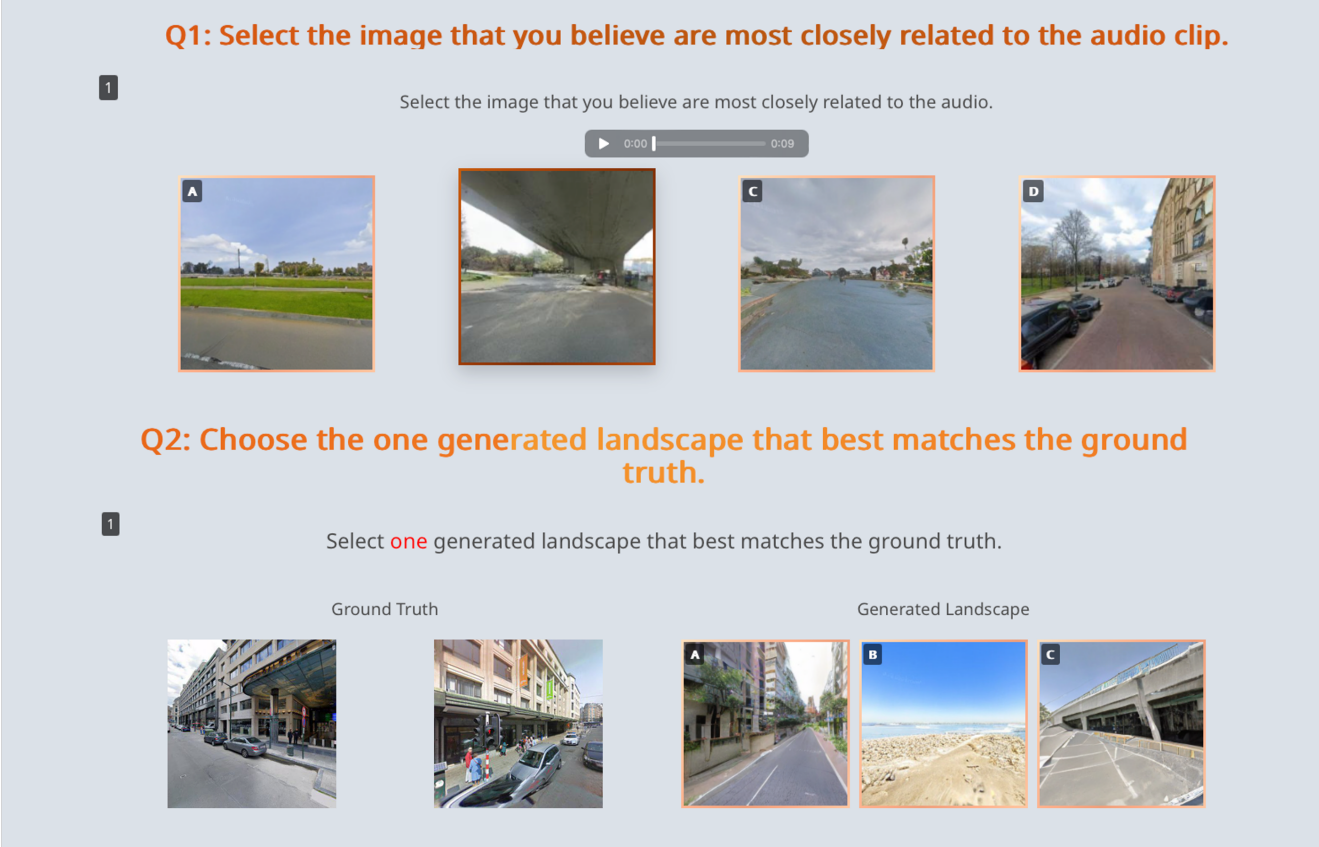

Figure 17. **User study webpage.** Examples of two tasks are presented. For task 1, participants are asked to select the image that best matches the soundscape. For task 2, participants are asked to select the image that best matches the ground truth landscape.

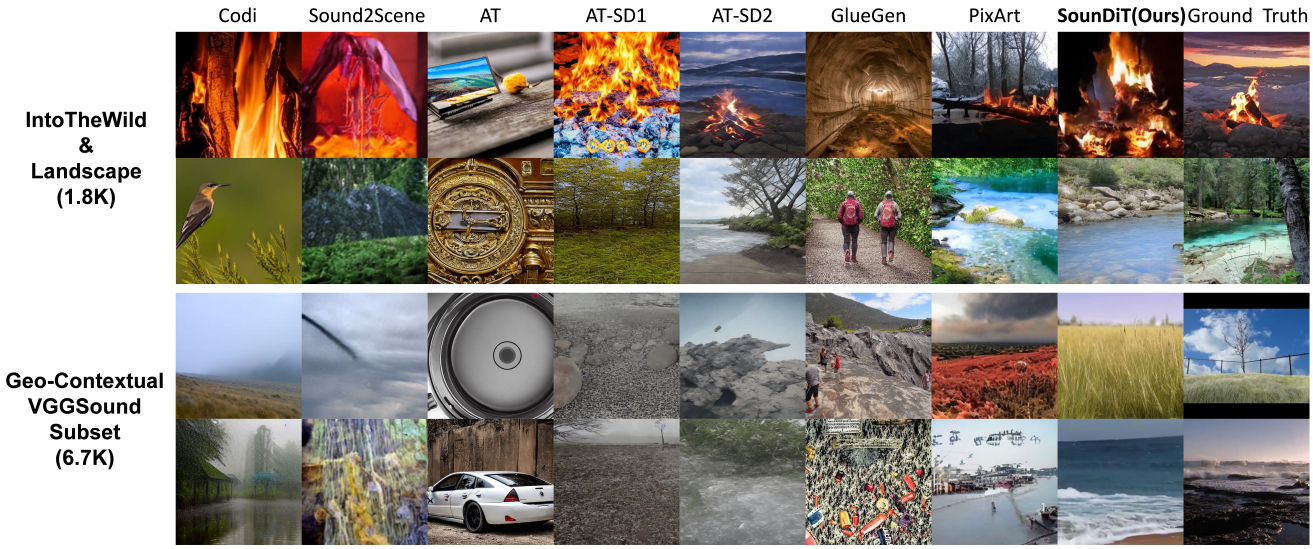

Figure 18. Results on the **benchmark datasets**: IntoTheWild&Landscape and the Geo-Contextual VGGSound subset.

fluence human behaviors and ecological processes [4, 14, 19, 20]. Second, it helps planners to design spaces that are both aesthetically pleasing and acoustically beneficial. By analyzing how sound propagates and influences areas, bet-

Figure 19. Visual comparison across different **expert configurations** in the MoE module, where E denotes the number of experts.

ter noise mitigation strategies might be developed to design better neighborhoods and benefit human well-being in urban planning practices. Third, incorporating geo-contextual sounds may improve user experience in navigation applications [3, 5, 11] , video games, and virtual reality settings [2, 9]. Specifically, for individuals with visual impairments, integrating acoustic signals provides essential navigational cues and a deeper understanding of their surroundings.

Beyond its technical contributions, we envision this work as a foundational step and open avenues for future technical advancements at the intersection of geography, environmental science, and generative AI. We advocate for interdisciplinary collaboration to better align models with real-world problems in the service of societal and ecological benefits. By formalizing the GeoS2L task, our study lays a foundation for future research that is both socially relevant and environmentally grounded, pointing toward new research frontiers at the intersection of generative AI and spatial data science.

References

[1] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 4

[2] Durand R Begault and Leonard J Trejo. 3-d sound for virtual reality and multimedia. Technical report, 2000. 11

[3] Stephen A Brewster. Using nonspeech sounds to provide navigation cues. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 5(3):224–259, 1998. 11

[4] Peter A Burrough, Rachael A McDonnell, and Christopher D Lloyd. *Principles of geographical information systems*. Oxford university press, 2015. 10

[5] Changan Chen, Unnat Jain, Carl Schissler, Sebastia Vicenc Amengual Gari, Ziad Al-Halah, Vamsi Krishna Ithapu, Philip Robinson, and Kristen Grauman. Soundspaces: Audio-visual navigation in 3d environments. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*, pages 17–36. Springer, 2020. 11

[6] Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. Vggsound: A large-scale audio-visual dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 721–725. IEEE, 2020. 1, 4

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1, 4

[8] Konrad Heidler, Lichao Mou, Di Hu, Pu Jin, Guangyao Li, Chuang Gan, Ji-Rong Wen, and Xiao Xiang Zhu. Self-supervised audiovisual representation learning for remote

sensing data. *International Journal of Applied Earth Observation and Geoinformation*, 116:103130, 2023. 4

[9] Joo Young Hong, Bhan Lam, Zhen-Ting Ong, Kenneth Ooi, Woon-Seng Gan, Jian Kang, Jing Feng, and Sze-Tiong Tan. Quality assessment of acoustic environment reproduction methods for cinematic virtual reality in soundscape applications. *Building and environment*, 149:1–14, 2019. 11

[10] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017. 5

[11] Rohan Kapoor, Subramanian Ramasamy, Alessandro Gardi, Ron Van Schyndel, and Roberto Sabatini. Acoustic sensors for air and surface navigation applications. *Sensors*, 18(2): 499, 2018. 11

[12] Seung Hyun Lee, Gyeongrok Oh, Wonmin Byeon, Chanyoung Kim, Won Jeong Ryoo, Sang Ho Yoon, Hyunjun Cho, Jihyun Bae, Jinkyu Kim, and Sangpil Kim. Sound-guided semantic video generation. In *European Conference on Computer Vision*, pages 34–50. Springer, 2022. 1, 4

[13] Tingle Li, Yichen Liu, Andrew Owens, and Hang Zhao. Learning visual styles from audio-visual associations. In *European Conference on Computer Vision*, pages 235–252. Springer, 2022. 1, 4

[14] Doreen Massey. *Space, Place, and Gender*. University of Minnesota Press, 1994. 10

[15] Nikhil Naik, Jade Philipoom, Ramesh Raskar, and César Hidalgo. Streetscore-predicting the perceived safety of one million streetscapes. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 779–785, 2014. 5

[16] Udo Noll. radio aporee. *Berlin: Udo Noll*, 2013. 1

[17] Sooyoung Park, Arda Senocak, and Joon Son Chung. Can clip help sound source localization? In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5711–5720, 2024. 1, 4

[18] Silero Team. Silero vad: pre-trained enterprise-grade voice activity detector (vad), number detector and language classifier, 2024. 1, 2

[19] Yi-Fu Tuan. *Space and place: The perspective of experience*. U of Minnesota Press, 1977. 10

[20] Monica G Turner, Robert H Gardner, Robert V O'neill, and Robert V O'Neill. *Landscape ecology in theory and practice*. Springer, 2001. 10

[21] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017. 4

[22] Yipin Zhou, Zhaowen Wang, Chen Fang, Trung Bui, and Tamara L Berg. Visual to sound: Generating natural sound for videos in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3550–3558, 2018. 4