SounDiT: Geo-Contextual Soundscape-to-Landscape Generation

Junbo Wang[1,2,†] Haofeng Tan[1,3,†] Bowen Liao[4] Albert Jiang[1] Teng Fei[5]
Qixing Huang[1] Bing Zhou[2] Zhengzhong Tu[6] Shan Ye[7,‡,] Yuhao Kang[1,‡,]
[1]The University of Texas at Austin [2]University of Tennessee, Knoxville
[3]University of South Carolina [4]Arizona State University [5]University of Canterbury
[6]Texas A&M University [7]University of Wisconsin-Madison
[†]Equal contributions. [‡]Corresponding Author: Yuhao Kang (yuhao.kang@austin.utexas.edu).

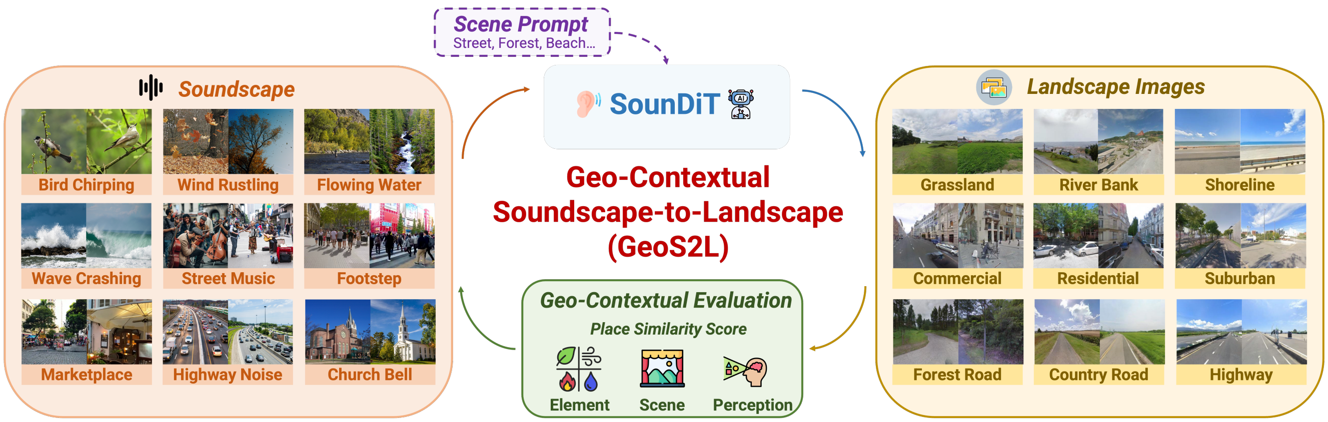

Figure 1. **Geo-contextual soundscape-to-landscape (GeoS2L) generation** aims to synthesize realistic landscape images from environmental soundscapes. We introduce large-scale geo-contextual datasets, the SounDiT model, and the Place Similarity Score evaluation framework to support this task.

Abstract

Recent audio-to-image models have shown impressive performance in generating images of specific objects conditioned on their corresponding sounds. However, these models fail to reconstruct real-world landscapes conditioned on environmental soundscapes. To address this gap, we present Geo-contextual Soundscape-to-Landscape (GeoS2L) generation, a novel and practically significant task that aims to synthesize geographically realistic landscape images from environmental soundscapes. To support this task, we construct two large-scale geo-contextual multi-modal datasets, SoundingSVI and SonicUrban, which pair diverse environmental soundscapes with real-world landscape images. We propose SounDiT, a diffusion transformer (DiT)-based model that incorporates environmental soundscapes and geo-contextual scene conditioning to synthesize geographically coherent landscape images. Furthermore, we propose the Place Similarity Score (PSS), a practically-informed geo-contextual evaluation framework to measure consistency between input soundscapes and generated landscape images. Extensive experiments demonstrate that SounDiT outperforms existing baselines in the GeoS2L, while the PSS effectively captures multi-level generation consistency across element, scene,and human perception. Project page: `https://gisense.github.io/SounDiT-Page/`

1. Introduction

"The earth has music for those who listen."

— William Shakespeare

Environmental soundscapes encode rich geographic information, often in greater detail than the specific object acting as the sound source [1, 4, 6, 67]. A burst of birdsong, for example, implies a leafy trail or an urban green space, while the sound of traffic may evoke a busy streetscape, beyond a single car. Recognizing this strong geo-contextual connection between auditory soundscape and visual landscape, researchers in geography [16, 43, 75], urban planning [17, 31, 39], and environmental psychology [28, 56], have long studied the impacts of soundscapes. For example, how analyzing specific sound types can inform and improve urban design[36], how to mitigate noise through environmental design [7, 8], or which soundscapes are associated with human restoration and well-being [76, 84]. Previous practices in these fields have often been limited to descriptive statistics of the soundscape environment (e.g., decibel levels, sound types) [15, 37], which are not always intuitive and fail to capture the co-located visual characteristics of the scene. The recent emergence of deep generative models for Audio-to-Image (A2I) synthesis [5, 63, 71, 80, 81], which translates auditory information directly into visual representations, offers a promising opportunity to model the geo-contextual soundscape-landscape relationships, unlocking significant practical applications.

However, we observe three noticeable gaps when applying prior A2I in geographic contexts and planning practices for associating soundscapes and landscapes. First, most A2I studies rely on generic audio-visual datasets such as object audio [10, 60, 87], human voice [52, 57], weathers [46], and limited scene types [45]. Consequently, these models link an audio signal to its source (e.g., generating an image of a bird for a bird call) rather than its environment, overlooking geo-contextual insights vital for real-world applications, such as parks vs. beaches and urban vs. rural settings [82]. For geographers and urban planners, the primary interest is not the object making the sound (that bird, that car), but the environmental setting that informs environmental design and noise mitigation strategies. Despite explorations of diffusion-based models in geographic audio-to-image settings [88, 89], these methods still take soundscapes as the sole input and output landscape images, without geographic context. Second, Diffusion Transformer (DiT) [62] techniques have recently demonstrated impressive results across various tasks such as image [11, 12, 48] and video generations [42, 51, 54, 77]. However, they remain unexplored in A2I generation and have yet to incorporate geographic knowledge to support applications in geography and urban planning. Furthermore, common evaluation metrics employed in prior studies, such as the Fréchet Inception Distance (FID) [33], may fall short in capturing geographic semantic alignment between the input soundscape and generated images. Ideally, the evaluation metrics should consider environmental characteristics and geographic contexts: the place setting depicted in the input soundscape and those represented in the generated images should be similar to ensure geographic contextual coherence and relevance.

Given these challenges, we introduce and formalize a new, practically-motivated task: Geo-Contextual Soundscape-to-Landscape (GeoS2L) generation (Figure 1), to synthesize realistic landscape images from environmental soundscapes, optionally conditioned on scene context. The GeoS2L problem extends traditional A2I synthesis by incorporating geo-contextual insights with significant practical implications. Here, *landscape* refers to geographical environments shaped by natural and built features [55]. A *soundscape* denotes the acoustic environment as perceived by humans at places, comprising sounds from both natural and anthropogenic sources (e.g., bird calls and vehicle noises) [69]. A *scene* indicates place types such as beach, forest, street, to guide image generation. Our contributions span three key areas: First, we introduce two large-scale, multi-modal geo-contextual datasets: *SoundingSVI* and *SonicUrban*, composed of over 169K and 237K soundscape-landscape image pairs. These datasets cover diverse geographic locations and scenes, supporting geo-contextual realistic synthesis. Second, we develop *SounDiT*, an innovative generative model that support GeoS2L problem. It introduces three novel, efficient components to integrate geo-contextual information: a MoE Soundscape Conditioning module, a Scene Low-Rank Content Mixer (SLRCM), and a Scene AdaLn (S-AdaLn). These modules effectively inject geographic contexts to environmental soundscapes, enabling the generation of geographically coherent landscape images. Third, we design a novel practically-motivated geo-contextual evaluation framework: *Place Similarity Scores (PSS)*. PSS includes three metrics: element-level, scene-level, and perception-level similarity to assess geographic contexts and environmental consistency of generated landscape images. Extensive experiments demonstrate that our model achieves state-of-the-art performance and delivers strong utility on downstream applications like geography, environment, and urban planning with significant practical impacts. Moreover, our datasets and evaluation metric provide a robust benchmark for the GeoS2L task.

2. Related Work

Audio-to-Image Synthesis aims to translate acoustic features into coherent and meaningful images and has gained increasing attention with the advancement of various multimodal alignment foundation models [71, 80] and generative techniques [25, 66]. Current A2I models are broadly divided into GAN-based (e.g., sound2scene [71]) and Diffusion-based (e.g., AudioToken [81], and GluenGen [63]). These models mainly rely on generic datasets which lack geo-contextual information, often limited to object audios [60], human voice [57], weather [46], and broad scene types [45]. As a result, they often produce stylized or unrealistic outputs and fail to capture the geographic coherence between soundscapes and landscapes.

Landscapes and Soundscapes represent two dimensions for characterizing places [20, 50]. Landscapes offer visual representations of the physical layout and built environments, and are commonly captured via street view imagery (SVI) and remote sensing (RS) images [45, 46]. In parallel, soundscapes refer to the acoustic environment perceived at specific places. Existing datasets, such as BirdCLEF [38], BirdSet [65], SoundingEarth [30]), AudioSet [21], Urban-

Sound8K [68], SONYC [9], EnigenScape [27], and EMO-Soundscape [18], have been utilized to support real-world practices including monitoring biodiversity [64], noise mitigation [35], and human perception analysis [19]. However, prior A2I studies have overlooked such geographic and environmental insights, limiting their real-world applications. To address this gap, our work integrates geo-contextual information for multimodal generative modeling by constructing datasets and evaluation frameworks.

3. Methodology

3.1. Preliminary

Diffusion Model. We adopt a latent-based diffusion model, consisting of a forward diffusion and backward denoising process in the latent space $\mathbf{z}$. The forward diffusion $\mathbf{q}$ starts from the initial latent $\mathbf{z}_0$, gradually adding Gaussian noise over timestep T based on a fixed variance α and defined as:

$$q(\mathbf{z}_t \mid \mathbf{z}_{t-1}) = \mathcal{N}\big(\mathbf{z}_t; \sqrt{\alpha_t}\,\mathbf{z}_{t-1}, (1-\alpha_t)\,\mathbf{I}\big). \quad (1)$$

The backward denoising process p_θ is trained to predict the less noisy latent:

$$p_\theta(\mathbf{z}_{t-1} \mid \mathbf{z}_t) = \mathcal{N}\big(\mathbf{z}_{t-1};\, \mu_\theta(\mathbf{z}_t, t, v),\, \Sigma_\theta(\mathbf{z}_t, t, v)\big). \quad (2)$$

where v is the conditional signal and μ_θ, Σ_θ are predicted through a denoising network $\epsilon_\theta(\mathbf{z}_t, t, v)$.

Classifier-Free Guidance. To control the influence of the conditioning signal v during backward denoising, we employ classifier-free guidance (CFG) to predict noise $\hat{\boldsymbol{\epsilon}}$ based on both conditional and unconditional signals with specific guidance scale s.

$$\hat{\boldsymbol{\epsilon}} = \boldsymbol{\epsilon}_\theta(\mathbf{z}_t, t, \varnothing) + s\big[\boldsymbol{\epsilon}_\theta(\mathbf{z}_t, t, v) - \boldsymbol{\epsilon}_\theta(\mathbf{z}_t, t, \varnothing)\big]. \quad (3)$$

3.2. GeoS2L Task Formulation

We formulate our task as *Geo-Contextual Soundscape-to-Landscape (GeoS2L)*, extending A2I synthesis with geographic contexts. GeoS2L aims to synthesize landscape images from environmental soundscapes guided by an optional scene prompt. We define a *landscape* l as the geographical–ecological environment shaped by natural and built features, a *soundscape* s as the human-perceived acoustic environment at a place, including natural and anthropogenic sources (e.g., bird calls and traffic noise). However, real-world soundscape recordings often lack sufficient information to fully capture the visual characteristics of the surrounding environment. For example, a bird call may occur in either rural or urban parks. To provide greater flexibility and user control, we incorporate an optional *scene prompt* c as the semantic geographic context (e.g., park, beach, street). This prompt serves as an additional geo-contextual conditioning signal to facilitate generative process toward a specific, context-aware landscape. Given a dataset $D = \{(s_i, c_i, l_i) \mid i = 1, \ldots, N\}$, our goal is to learn a generator $G:\ (s_i, c_i) \mapsto \hat{l}_i$ such that the synthesized image $\hat{l}_i$ is both visually realistic and geo-contextually consistent with the ground truth l_i, quantified by the relevance function $R(s, c, l)$. This can formulated as

$$\mathcal{L} \;=\; \mathbb{E}_{(s_i,c_i,l_i)\sim D}\Big[\,R(s_i, c_i, l_i) \;-\; R\big(s_i, c_i, \hat{l}_i\big)\Big]. \quad (4)$$

This enables the model to capture cross-modal correlations among acoustic, semantic, and visual cues, producing landscape images that reflect both the soundscape and the intended scene context.

3.3. SounDiT

To effectively and efficiently inject soundscape and scene context, each SounDiT block introduces (1) a Mixture-of-Experts (MoE) Soundscape Conditioning module, (2) a Scene Low-Rank Mixer (SLRCM) module, and a Scene AdaLN (S-AdaLN) module to improve scene consistency.

Multimodal Encoders. Following latent diffusion models (LDM) [66], landscape images are reconstructed within the latent space. Given a landscape image $l \in \mathbb{R}^{H\times W\times 3}$, a VAE encoder–decoder $(\mathcal{E}_L, \mathcal{D}_L)$ is employed to encode it into the latent embedding $\mathbf{e}_l = \mathcal{E}_L(l) \in \mathbb{R}^{h\times w\times c}$ and reconstruct the latent embedding back into a image $\hat{l} = \mathcal{D}_L(\mathbf{e}_l)$. To integrate acoustic features and geo-contextual semantics into denoising process, a multi-modal encoder is adopted to encode the soundscapes s and the scene prompt c into a shared latent space, resulting in a soundscape embedding $\mathbf{e}_s$ and a scene embedding $\mathbf{e}_c$.

SounDiT block. The SounDiT block operates in four stages. First, we apply multi-head self-attention layer conditioned by the timestep embedding e_{t} via AdaLN-Zero [26], preserving compatibility with pretrained DiT backbones. Second, we design a lightweight **Scene Low Rank Context Mixer (SLRCM)** module to inject scene context. SLRCM introduces a low-rank residual path inside each SounDiT block. Given an input token $x \in \mathbb{R}^{B\times N\times D}$ and scene embedding $\mathbf{e}_c \in \mathbb{R}^{B\times D_c}$, SLRCM modulates tokens through a rank-r linear operator parameterized by $\mathbf{e}_c$. We specify the low-rank content mixer $A(\mathbf{e}_c)$ that transforms token features along an r-dimensional path, together with a sample-wise strength $s(\mathbf{e}_c)$ that controls the overall contribution of this path:

$$A(\mathbf{e}_c) = W_q\,\mathrm{Diag}\big(\tanh(\phi(\mathbf{e}_c))\big)\,W_v, \quad (5)$$

where $W_q \in \mathbb{R}^{D\times r}$ and $W_v \in \mathbb{R}^{r\times D}$ are learned low-rank projections. The mapping $\phi: \mathbb{R}^{D_c} \to \mathbb{R}^r$ converts the scene embedding into an r-dimensional gating vector, and the diagonal operator applies this gate element-wise along the rank-r channel, yielding a compact, diagonally gated low-rank transform. In addition, we define a positive, sample-

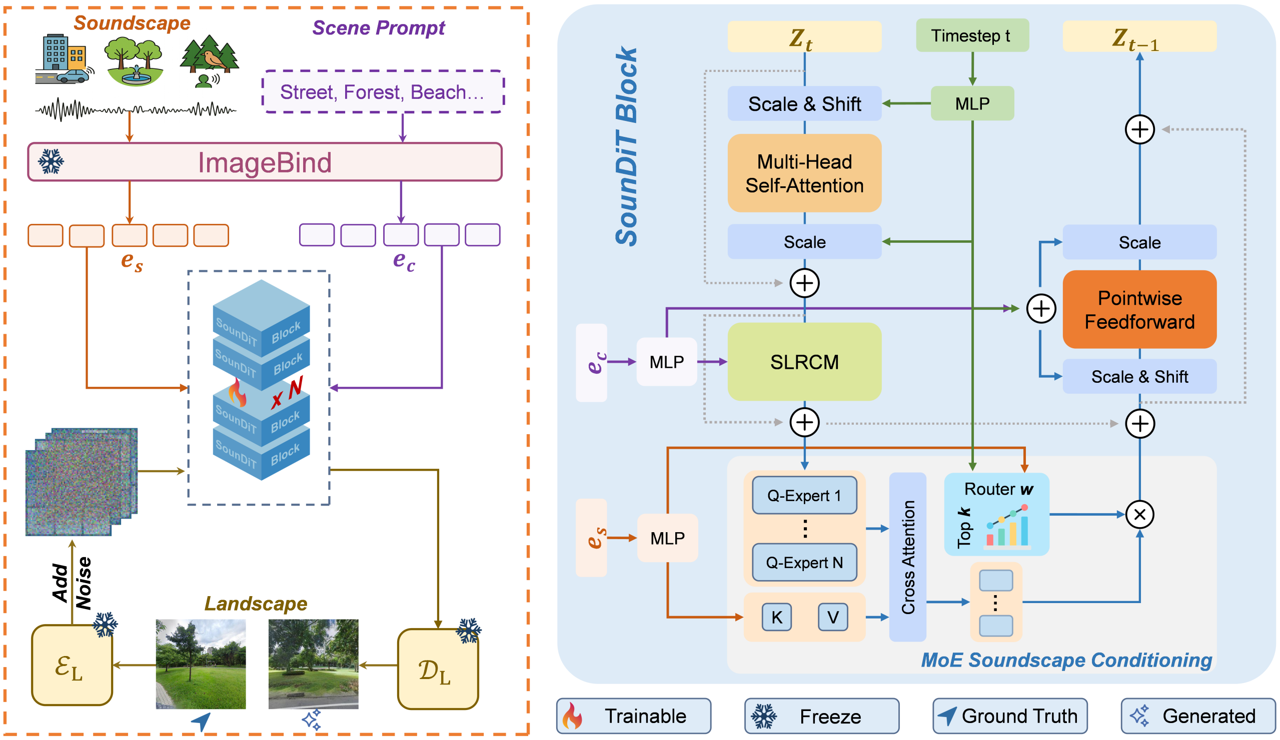

Figure 2. **The *SounDiT* framework.** *SounDiT* encodes soundscape and scene inputs into a shared space with pre-trained encoders, and performs denoising in latent space with DiT blocks equipped with AdaLN-Zero timestep conditioning. Scene information is injected through a Scene Low-Rank Mixer (SLRCM) and a lightweight Scene AdaLN (S-AdaLN) path to enhance geo-contextual consistency, while a Mixture-of-Experts (MoE) Soundscape conditioning module aligns soundscape features with visual landscape tokens.

wise strength to regulate the residual magnitude:

$$s(\mathbf{e}_c) = g(\mathbf{e}_c)\,\mu\,\tanh(\alpha), \tag{6}$$

where $g : \mathbb{R}^{D_c} \to \mathbb{R}_{>0}$ produces a per-sample scale (e.g., via softplus), μ is a global guidance scalar, and α is a learnable scalar bounded. Given these definitions, the token update becomes:

$$x' \;=\; x \;+\; s(\mathbf{e}_c)\,\mathrm{LN}(x)\,A(\mathbf{e}_c), \tag{7}$$

where $\mathrm{LN}(\cdot)$ denotes layer normalization. SLRCM thus provides a compact, diagonally gated low-rank modulation that preserves the pre-trained attention structure while injecting scene priors with low computational costs.

Third, to further integrate multi-level soundscape features, we introduce **MoE Soundscape Conditioning** module, composed of M experts that share the key K_s and value V_s while maintaining expert-specific, low-rank query Q. Given input token $x \in \mathbb{R}^{B\times N\times D}$,we compute $K_s = s\,W_K, V_s = s\,W_V$ once and for each expert $m \in \{1,\dots,M\}$ form a low-rank query Q_m

$$Q_m = f(x)\,W_{q\downarrow}^{(m)} W_{q\uparrow}^{(m)}, \tag{8}$$

where f is a per-token Layer Norm. Each expert then produces

$$Z_m \;=\; \mathrm{MHA}\big(Q_m,\, K_s,\, V_s\big)\, W_O, \tag{9}$$

where W_O is shared across experts, Z_m refers to soundscape-conditioned token features, and MHA is multi-head attention.

Routing weights are computed by a temperature-scaled dot product between the audio summary and learnable prototypes $\mathbf{P} \in \mathbb{R}^{d\times M}$ with time augmentation introduced through a projected timestep embedding:

$$w = \mathrm{Softmax}\Big(\tfrac{1}{\tau}\,\mathbf{e}_s^\top \big(\mathbf{P} + \mathbf{W}_t\, e_t\, \mathbf{1}^\top\big)\Big), \tag{10}$$

where e_t is the timestep embedding projected by $\mathbf{W}_t$. Finally, we aggregate the experts via a top-k soft mixture and apply a global audio gate to update tokens in each SounDiT block:

$$x' \;=\; x \;+\; \gamma \sum_{m\in\mathcal{K}} \mathrm{softmax}\big(w_m/\tau_m\big)_m\, Z_m, \tag{11}$$

where $\gamma = \tanh(\mathbf{e}_s)$ is a scalar gate, yielding a smooth, zero-centered bounded activation. By enabling expert specialization under a fixed K/V budget, the MoE provides

Table 1. **Comparison of audio-visual datasets.** † denotes partial/limited geo-context.

Dataset	Modalities	# Size	Geo-Contextual	Viewpoint
ACAV100M [44]	Video	100M	–	Ground
VEGAS [87]	Video	28K	–	Ground
VGGSound [10]	Video	200K	†	Ground
IntoTheWild [46]	Video	142	✓	Ground
Landscape [45]	Video	1747	✓	Ground
SoundingEarth [30]	Audio & Image	50K	✓	Overhead
SoundingSVI (Ours)	Audio & Image	169K	✓	Ground
SonicUrban (Ours)	Video	237K	✓	Ground

scalable conditioning that improves geo-contextual consistency across diverse soundscapes.

Finally, to enhance geo-contextual consistency, we introduce a practical-informed **S-AdaLN**, an AdaLN-Zero extension, which derives scale–shift parameters from the timestep embedding $\mathbf{e}_t$ and the scene embedding $\mathbf{e}_c$ via learnable bounded mixing. The modulated tokens are processed by a pointwise feed-forward network, followed by a gated residual add, yielding the block output for noise-residual prediction. By hierarchically fusing visual, scene, and soundscape cues, SounDiT generates geo-contextually coherent landscapes.

4. Datasets

Prior A2I studies have primarily relied on large-scale, general-domain audio-visual datasets, such as VEGAS [87], VGGSound [10], and ACAV100M [44]. These datasets primarily focus on sound-source localization and general audio-visual correspondence learning, lacking the specific geographical context required for soundscape-to-landscape generation. For instance, geographers may be more interested in generating a wetland implied by a frog call, rather than an image of the frog itself. A car horn might suggest a congested urban street, rather than a static vehicle. Some datasets, such as Landscape [45] and IntoTheWild [46], have included paired soundscape and image data, but offer limited spatial coverage and scene diversity. To address this gap, we introduce two new multimodal datasets: *SoundingSVI* and *SonicUrban*, designed to support GeoS2L generation with large-scale, geographically diverse, and context-rich data.

SoundingSVI. The SoundingSVI dataset was constructed from geotagged soundscape recordings from the Aporee platform [59]. For each raw soundscape recording, we retrieved nearby Google Street View Image (SVI) using its latitude and longitude to capture the surrounding visual landscapes. This follows a similar approach to the *Sounding Earth* dataset [30], which paired geotagged audio with remote sensing imagery. The raw recordings were then segmented into 10-second clips, with a human voice detection model [74] used to filter clips dominated by human voice. Notably, multiple candidate SVIs captured from different angles and timestamps might be retrieved for clips from the same recording. A sound source localization model [61] was further applied to match each clip with the most relevant landscape image. To ensure temporal consistency and account for environmental changes, we refined the dataset by discarding soundscape-SVI pairs with significant time gaps. Furthermore, to obtain annotated place semantics, a Vision-Language Model (VLM), Qwen2.5-VL-7b, [2] was used to annotate each pair with a scene context (e.g., street and residential neighborhood), which serve as the scene prompts for our model. In total, SoundingSVI dataset comprises 169,221 soundscape-landscape pairs in 90 countries.

SonicUrban. The SonicUrban dataset leverages unedited videos as a source of temporally and spatially consistent soundscape-landscape pairs. Following a prior practice [88], we searched and manually checked YouTube videos that contain real-world soundscapes and landscapes, using a combination of city names and keywords, like "City walk; New York". Each video was processed into 10-second audio clips, and 10 evenly distributed frames were extracted for each clip. The same human voice detection model [74] and sound source localization model [61] used in SoundingSVI were applied to filter human-voice dominated clips and identify the most representative frame for each audio clip, respectively. In total, the constructed SonicUrban consists of 236,674 soundscape-landscape pairs across 131 cities and 97 countries. The same VLM [2] was further employed to generate a scene prompt for each pair as a geographic scene condition.

5. Geo-Contextual Evaluation Suite

Prior A2I studies commonly rely on evaluation metrics such as Fréchet Inception Distance (FID) [33], Audio-Image Similarity (AIS) [33], and Image-Image Similarity (IIS) [5], to assess visual or audio fidelity. However, such metrics do not assess whether the generated image is coherent with the underlying geographic setting or the intended target scene. Since soundscapes and landscapes co-occur in the same space, they are expected to share similar environmental characteristics and place settings. Following this hypothesis, we introduce **Place Similarity Score (PSS)**, a practically informed geo-contextual evaluation that quantifies the underlying place setting reflected in images across the element-, scene-, and human-perception levels. These three levels of measurements of environmental characteristics have been widely employed in geographic and urban planning practices [41, 53].

Element-level PSS. This metric measures the agreement in geographic elements (e.g., trees [47], sky [24], water bodies [70], traffic signs [3], and buildings [13]) between the ground-truth and generated landscapes [40]. We use

DeepLabV3 [14] pre-trained on ADE20K [86] to segment $K = 150$ predefined elements. For each image i, let $\mathbf{e}_i \in \mathbb{R}^K$ and $\hat{\mathbf{e}}_i \in \mathbb{R}^K$ be the normalized element-ratio vectors derived from the ground-truth image l_i and the generated image $\hat{l}_i$, respectively. Formally, The element-level score is calculated as:

$$\mathrm{PSS}_{\text{elem}} = \frac{1}{n} \sum_{i=1}^{n} \frac{\mathbf{e}_i^\top \hat{\mathbf{e}}_i}{\|\mathbf{e}_i\|_2 \, \|\hat{\mathbf{e}}_i\|_2}, \tag{12}$$

where $\mathbf{e}_i = (e_{i1}, \ldots, e_{iK})^\top$ and $\hat{\mathbf{e}}_i = (\hat{e}_{i1}, \ldots, \hat{e}_{iK})^\top$; n is the number of evaluation images.

Scene-level PSS. This metric is designed to evaluate the consistency of the entire environmental scene (e.g., forest, beach, or residential area) between the generated landscape images and their corresponding ground truth images [40]. To infer scene categories, we employ a ResNet50 model [29], pre-trained on the Places365 dataset [85], to identify 365 predefined scene categories. For each soundscape s, we predict the top-k scene labels and evaluate whether the generated image $\hat{l}_i$ aligns with the same scene category as its ground truth image l_i.

$$\mathrm{PPS}_{\text{Scene}} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\left\{ P_i^{\text{k}} \cap T_i^{\text{k}} \neq \varnothing \right\}, \tag{13}$$

where $k = 1$ or 5. n is the total number of images, P_i^{k} denotes the top-k predicted category set for the i-th generated image, T_i^{k} denotes the top-k ground truth category set for the i-th real image.

Human Perception-level PSS. This metric is designed to evaluate whether the generated and real landscape images evoke similar human subjective perceptions and feelings, such as whether a place makes people feel "safe" or "lively" [83]. We utilize a DenseNet121 model [34], pre-trained on the MIT Place Pulse dataset [58], to measure six dimensions of human perceptions of environment, including safe, beautiful, depressing, lively, wealthy, and boring. For each image i, we compute six-dimensional place-perception scores for the ground-truth image l_i and the generated image $\hat{l}_i$. The perception-level PSS is defined as:

$$\mathrm{PSS}_{\text{perc}} = \frac{1}{n} \sum_{i=1}^{n} \left\| \mathbf{R}(l_i) - \mathbf{R}(\hat{l}_i) \right\|_1, \tag{14}$$

where $\mathbf{R}(\cdot) \in \mathbb{R}^6$ denotes the six-dimensional perception score vector, and n is the number of test images.

By integrating metrics across element-, scene-, and human perception levels, we provide a holistic geo-contextual evaluation framework. Informed by geographic domain knowledge and urban planning practices, we extend beyond measuring visual quality to assessing whether generated landscape images are contextually aligned with the environmental characteristics in input soundscapes.

6. Experiment

6.1. Implementation Details

Model Architecture and Training. We adopt the pre-trained Variational Autoencoder (VAE) modules from Latent Stable Diffusion [66] as our landscape encoder–decoder pair, which were originally trained on the COCO dataset [49]. For soundscape and scene prompt conditioning, we employ the ImageBind-Huge [23], pre-trained on two million AudioSet clips [22] covering a wide range of environmental, chosen for its accuracy–efficiency balance among several encoders [78]. Our SounDiT is trained using a learning rate of 1×10^{-4}. Experiments are conducted on NVIDIA H100, A100 and A6000 GPUs. During training, we set the soundscape guidance scalar to $\mu = 1.0$. For the scene condition, we employ a learnable scaling parameter α initialized to 0 to stabilize early training by starting from an identity mapping. At inference stage, we apply classifier-free guidance with a scale of 4.0 for both the soundscape and the scene prompt.

Baseline Models. We compare against state of the art audio to image models: CoDi [73], Sound2Scene [71], AudioToken [81], GlueGen [63] and a PixArt [11] variant augmented with multihead cross attention for the soundscape condition (PixArt+MHCA). For CoDi, we utilized its official pre-trained model. For AudioToken, we employed the pre-trained model without additional training, and trained it on our datasets using Stable Diffusion 1 and 2 as its backbone, respectively. The other baselines are training from scratch based on their released code and training settings.

Benchmark Evaluation Metrics. To evaluate the model performance, we adopted FID [33], AIS [33], and IIS [5], effective in previous audio-to-image generation studies [5, 72]. FID compares Gaussian distributions of features extracted from real and generated images, using a pre-trained Inception model. AIS measures audio-image semantic similarity using Wav2CLIP [79]. IIS evaluates image-image semantic similarity using CLIP-based embedding [32].

6.2. Main Results

Figure 3 presents landscape images generated by our proposed SounDiT and baseline models across SoundingSVI and SonicUrban. The results generated by SounDiT show greater geo-contextual similarity to ground truth images, highlighting potential applications of SounDiT in geography and urban planning. To comprehensively evaluate model performance, we adopt general metrics (FID, AIS, IIS) in addition to our proposed geo-contextual PSS framework. As reported in Table 2, SounDiT consistently outperforms baselines on both SoundingSVI and SonicUrban datasets, with notable FID improvements (34→16 and 41→11), and achieves superior or competitive results across all metrics. SounDiT consistently ranks highest under PSS,

Table 2. **Quantitative evaluation results across SounDiT and baseline models on two proposed datasets and 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 (PSS)**		
Dataset	Method	**FID↓**	**AIS↑**	**IIS↑**	**Element↑**	**Scene↑**	**Perception↓**
SoundingSVI (169K)	CoDi† [73]	87.027	**0.641**	0.612	0.400	0.302	0.800
	Sound2Scene [71]	78.137	0.586	0.623	0.467	0.374	0.735
	AudioToken† [81]	142.441	0.600	0.519	0.248	0.184	0.884
	AudioToken (SD1) [81]	105.472	0.509	0.572	0.527	0.553	0.797
	AudioToken (SD2) [81]	149.473	0.504	0.559	0.249	0.311	0.845
	GlueGen [63]	74.250	0.500	0.566	0.400	0.461	0.776
	PixArt + MHCA [11]	34.108	0.518	0.578	0.474	0.390	0.743
	SounDiT (Ours)	**16.839**	0.538	**0.753**	**0.572**	**0.753**	**0.729**
SonicUrban (237K)	CoDi† [73]	83.904	0.503	0.497	0.273	0.310	0.856
	Sound2Scene [71]	49.681	0.503	0.498	0.332	0.324	0.829
	AudioToken† [81]	127.337	0.497	0.501	0.203	0.199	0.976
	AudioToken (SD1) [81]	95.210	0.505	0.503	0.419	0.484	0.979
	AudioToken (SD2) [81]	165.014	0.502	0.504	0.298	0.455	0.980
	GlueGen [63]	64.434	0.498	0.499	0.405	0.481	0.805
	PixArt + MHCA [11]	41.456	0.517	0.592	0.411	0.396	0.796
	SounDiT (Ours)	**11.553**	**0.520**	**0.706**	**0.520**	**0.739**	**0.759**

indicating its ability to ensure geographic coherence in generated images. These results validate the effectiveness of PSS in measuring geo-contextual coherence.

Practical Applications. To enhance real-world utility and user control, SounDiT supports scene-conditioned generation. Given a fixed soundscape, modifying the scene prompt generates geo-contextual, semantically coherent images that preserve both auditory information and scene context (Fig. 4). This supports practical applications such as soundscape-guided urban design, informing evidence-based design strategies to promote public health, safety, and environmental comfort.

User Study. In addition to these quantitative metrics, we further conducted a user study to assess whether the generated landscape images align with human perceptions of geographic contexts. Specifically, 17 participants were asked to complete two matching tasks: (1) selecting the generated image that best aligned with the given soundscape; and (2) identifying which generated landscape image most closely matched a ground-truth landscape. The average matching accuracy is 86.13%, indicating a strong perceptual alignment between soundscapes and their corresponding generated images. This demonstrates the effectiveness of SounDiT for GeoS2L.

Table 3. **Component ablations on SoundingSVI (169K)** with two experts in the MoE Soundscape Conditioning module. We assess the individual contributions of *Scene Low-Rank Content Mixer (SLRCM)* and the *S-AdaLN*.

Variant	FID↓	AIS↑	IIS↑	PSS_{Scene} ↑
Full Model	**19.195**	**0.538**	**0.750**	**0.734**
w/o SLRCM + S-AdaLN	25.375	0.511	0.539	0.428
w/o SLRCM	20.335	0.534	0.728	0.704
w/o S-AdaLN	23.435	0.529	0.629	0.572

Table 4. **Expert scalability of MoE cross-attention on SoundingSVI (169K).** We vary the number of audio experts $M \in \{2, 4, 6, 8\}$ while keeping other settings fixed.

Experts	2	4	6	8
FID↓	19.195	18.304	17.278	**16.839**
PSS_{Scene} ↑	0.734	0.741	0.742	**0.753**

6.3. Ablations

To validate our architectural design and quantify the impact of our proposed modules of SounDiT, we conduct two types of ablation experiments: component ablations and component scalability ablations. First, to assess the contribution of scene prompt and two scene conditioning modules, we perform experiments by first removing both SLRCM and S-AdaLN, and then removing each module individually. As shown in Table 3, removing either module results in a noticeable drop in performance. These results indicate that injecting scene condition both before and after the MoE Soundscape Conditioning module improves geo-contextual

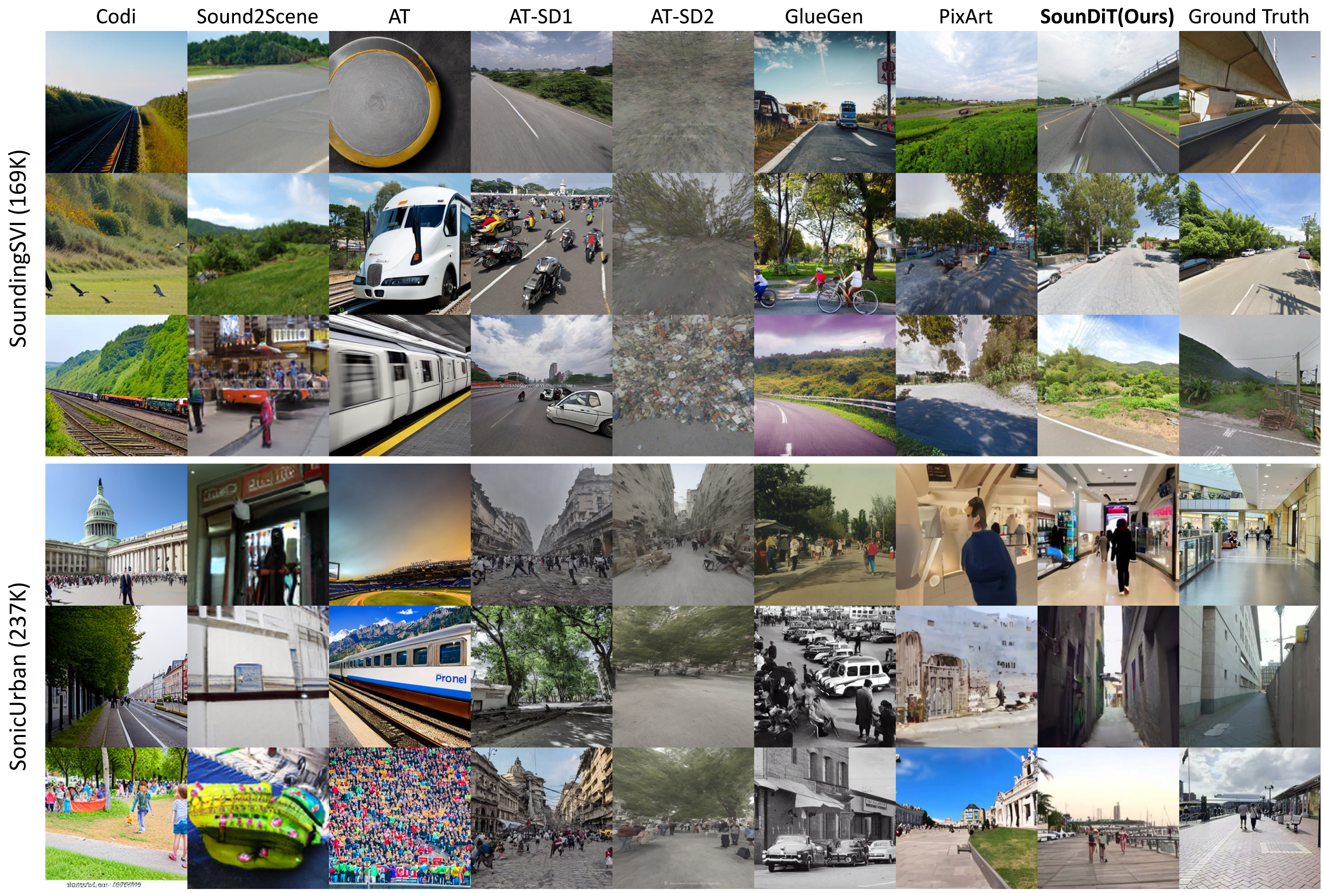

Figure 3. **Visual comparison of landscape images** generated by baseline models (CoDi, Sound2Scene, AudioToken(AT), GlueGen, PixArt) and our proposed SounDiT across SoundingSVI and SonicUrban. Ground truth landscape images are provided in the last column. Our proposed SounDiT significant improvements over baseline models in generating geographical context.

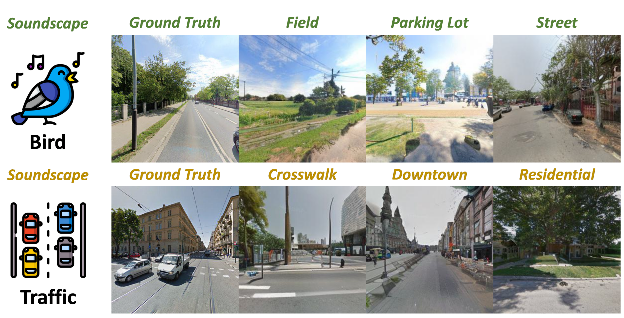

Figure 4. **Scene-conditioned generation.** With the same soundscape input, SounDiT generates visually distinct yet acoustically consistent images.

alignment and overall quality. Second, we test the scalability of MoE Soundscape Conditioning by varying the number of experts within this module (Table 4). Results show that the model with more experts achieves better performance in FID and scene consistency.

7. Conclusion

In this work, we present Geo-Contextual Soundscape-to-Landscape (GeoS2L), a practically-informed novel generation problem that synthesizes realistic landscape images from environmental soundscapes. To support GeoS2L, we construct SoundingSVI and SonicUrban, two large-scale, multi-modal, geo-contextual datasets, capturing diverse sounding and visual environments. We propose SounDiT, a scalable diffusion-based framework that captures both sounding environment and scene context by MoE Soundscape Conditioning and Scene Low-Rank Content Mixer. We further design a novel geo-contextual evaluation framework, the Place Similarity Score (PSS), which assesses geographic and environmental consistency. Extensive experiments demonstrate that SounDiT consistently outperforms existing baselines and can generate realistic landscape images. The proposed geo-contextual datasets, model, and evaluation framework effectively integrate geographic knowledge into generative modeling, establishing a solid benchmark for the GeoS2L task.

Acknowledgements

We acknowledge ACCESS, supported by the U.S. National Science Foundation, for allocation SOC250002. We also acknowledge resources from the NAIRR Pilot (NAIRR250075). We acknowledge the Texas Advanced Computing Center (TACC) at The University of Texas at Austin for providing HPC resources.

References

[1] Alaa Algargoosh, Babak Soleimani, Sile O'Modhrain, and Mojtaba Navvab. The impact of the acoustic environment on human emotion and experience: A case study of worship spaces. *Building Acoustics*, 29(1):85–106, 2022. 1

[2] Shuai Bai. Qwen2. 5-vl technical report. Technical report, Alibaba Damo Academy, 2025. 5

[3] Vahid Balali, Armin Ashouri Rad, and Mani Golparvar-Fard. Detection, classification, and mapping of us traffic signs using google street view images for roadway inventory management. *Visualization in Engineering*, 3:1–18, 2015. 5

[4] Arnold Berleant. *Living in the landscape: Toward an aesthetics of environment*. University press of Kansas, 1997. 1

[5] Burak Can Biner, Farrin Marouf Sofian, Umur Berkay Karakaş, Duygu Ceylan, Erkut Erdem, and Aykut Erdem. Sonicdiffusion: Audio-driven image generation and editing with pretrained diffusion models. *arXiv preprint arXiv:2405.00878*, 2024. 2, 5, 6

[6] Barry Blesser and Linda-Ruth Salter. Spaces speak, are you listening. *Experiencing aural architecture*, 232, 2007. 1

[7] AL Brown and Andreas Muhar. An approach to the acoustic design of outdoor space. *Journal of Environmental planning and Management*, 47(6):827–842, 2004. 2

[8] Alan Lex Brown. Soundscape planning as a complement to environmental noise management. In *InterNoise14: Improving the World through Noise Control*. Australian Acoustical Society, 2014. 2

[9] Mark Cartwright, Ana Elisa Mendez Mendez, Jason Cramer, Vincent Lostanlen, Graham Dove, Ho-Hsiang Wu, Justin Salamon, Oded Nov, and Juan Bello. SONYC urban sound tagging (SONYC-UST): A multilabel dataset from an urban acoustic sensor network. In *Proceedings of the Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE)*, pages 35–39, 2019. 3

[10] 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. 2, 5

[11] Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, et al. Pixart-$alpha$: Fast training of diffusion transformer for photorealistic text-to-image synthesis. *arXiv preprint arXiv:2310.00426*, 2023. 2, 6, 7

[12] Junsong Chen, Chongjian Ge, Enze Xie, Yue Wu, Lewei Yao, Xiaozhe Ren, Zhongdao Wang, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-σ: Weak-to-strong training of diffusion transformer for 4k text-to-image generation. In *European Conference on Computer Vision*, pages 74–91. Springer, 2024. 2

[13] Long Chen, Yi Lu, Yu Ye, Yang Xiao, and Linchuan Yang. Examining the association between the built environment and pedestrian volume using street view images. *Cities*, 127: 103734, 2022. 5

[14] Liang-Chieh Chen. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017. 6

[15] Xiaochao Chen and Jian Kang. Natural sounds can encourage social interactions in urban parks. *Landscape and Urban Planning*, 239:104870, 2023. 2

[16] Denis Cosgrove. *Geography and vision*. IB Tauris, 2012. 1

[17] Bert De Coensel, Annelies Bockstael, Luc Dekoninck, Dick Botteldooren, Brigitte Schulte-Fortkamp, Jian Kang, and Mats E Nilsson. The soundscape approach for early stage urban planning: a case study. In *INTER-NOISE and NOISE-CON Congress and Conference Proceedings*, pages 3294–3303. Institute of Noise Control Engineering, 2010. 1

[18] Jianyu Fan, Miles Thorogood, and Philippe Pasquier. Emo-soundscapes: A dataset for soundscape emotion recognition. In *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 196–201. ISSN: 2156-8111. 3

[19] Xingyue Fang, Tian Gao, Marcus Hedblom, Naisheng Xu, Yi Xiang, Mengyao Hu, Yuxuan Chen, and Ling Qiu. Soundscape perceptions and preferences for different groups of users in urban recreational forest parks. *Forests*, 12(4):468, 2021. 3

[20] Susan Fuller, Anne C. Axel, David Tucker, and Stuart H. Gage. Connecting soundscape to landscape: Which acoustic index best describes landscape configuration? 58:207–215. 2

[21] Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 776–780. ISSN: 2379-190X. 2

[22] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 776–780. IEEE, 2017. 6

[23] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15180–15190, 2023. 6

[24] Fang-Ying Gong, Zhao-Cheng Zeng, Fan Zhang, Xiaojiang Li, Edward Ng, and Leslie K Norford. Mapping sky, tree, and building view factors of street canyons in a high-density urban environment. *Building and Environment*, 134:155–167, 2018. 5

[25] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014. 2

[26] Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Accurate, large minibatch sgd: Training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*, 2017. 3

[27] Marc Ciufo Green and Damian Murphy. EigenScape: A database of spatial acoustic scene recordings. 7(11):1204. Number: 11 Publisher: Multidisciplinary Digital Publishing Institute. 3

[28] Catherine Guastavino. Categorization of environmental sounds. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, 61(1): 54, 2007. 1

[29] 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. 6

[30] 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. 2, 5

[31] Walihabib Hemmat, Atiq Mohammad Hesam, and Hamza Atifnigar. Exploring noise pollution, causes, effects, and mitigation strategies: a review paper. *European Journal of Theoretical and Applied Sciences*, 1(5):995–1005, 2023. 1

[32] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*, 2021. 6

[33] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 2, 5, 6

[34] 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. 6

[35] Jing Huang, Teng Fei, Yuhao Kang, Jun Li, Ziyu Liu, and Guofeng Wu. Estimating urban noise along road network from street view imagery. *International Journal of Geographical Information Science*, 38(1):128–155, 2024. 3

[36] Jin Yong Jeon, Pyoung Jik Lee, Jin You, and Jian Kang. Acoustical characteristics of water sounds for soundscape enhancement in urban open spaces. *The Journal of the Acoustical Society of America*, 131(3):2101–2109, 2012. 1

[37] Hyun In Jo and Jin Yong Jeon. Urban soundscape categorization based on individual recognition, perception, and assessment of sound environments. *Landscape and urban planning*, 216:104241, 2021. 2

[38] Stefan Kahl, Fabian-Robert Stöter, Hervé Goëau, Hervé Glotin, Bob Planqué, Willem-Pier Vellinga, and Alexis Joly. Overview of birdclef 2019: Large-scale bird recognition in soundscapes. In *Working Notes of CLEF 2019 - Conference and Labs of the Evaluation Forum, Lugano, Switzerland, September 9-12, 2019*. CEUR-WS.org, 2019. 2

[39] Jian Kang. Soundscape in city and built environment: current developments and design potentials. *City and Built Environment*, 1(1):1, 2023. 1

[40] Yuhao Kang, Fan Zhang, Song Gao, Hui Lin, and Yu Liu. A review of urban physical environment sensing using street view imagery in public health studies. *Annals of GIS*, 26(3): 261–275, 2020. 5, 6

[41] Rachel Kaplan and Stephen Kaplan. *The experience of nature: A psychological perspective*. Cambridge university press, 1989. 5

[42] Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024. 2

[43] John B Krygier. Sound and geographic visualization. In *Modern cartography series*, pages 149–166. Elsevier, 1994. 1

[44] Sangho Lee, Jiwan Chung, Youngjae Yu, Gunhee Kim, Thomas Breuel, Gal Chechik, and Yale Song. Acav100m: Automatic curation of large-scale datasets for audio-visual video representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10274–10284, 2021. 5

[45] 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. 2, 5

[46] 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. 2, 5

[47] Xiaojiang Li, Chuanrong Zhang, Weidong Li, Robert Ricard, Qingyan Meng, and Weixing Zhang. Assessing street-level urban greenery using google street view and a modified green view index. *Urban Forestry & Urban Greening*, 14(3):675–685, 2015. 5

[48] Zhimin Li, Jianwei Zhang, Qin Lin, Jiangfeng Xiong, Yanxin Long, Xinchi Deng, Yingfang Zhang, Xingchao Liu, Minbin Huang, Zedong Xiao, et al. Hunyuan-dit: A powerful multi-resolution diffusion transformer with fine-grained chinese understanding. *arXiv preprint arXiv:2405.08748*, 2024. 2

[49] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13*, pages 740–755. Springer, 2014. 6

[50] Jiang Liu, Jian Kang, Tao Luo, and Holger Behm. Landscape effects on soundscape experience in city parks. 454-455:474–481. 2

[51] Yixin Liu, Kai Zhang, Yuan Li, Zhiling Yan, Chujie Gao, Ruoxi Chen, Zhengqing Yuan, Yue Huang, Hanchi Sun, Jianfeng Gao, et al. Sora: A review on background, technology, limitations, and opportunities of large vision models. *arXiv preprint arXiv:2402.17177*, 2024. 2

[52] Steven R Livingstone and Frank A Russo. The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english. *PloS one*, 13(5): e0196391, 2018. 2

[53] Kevin Lynch. The image of the environment. *The image of the city*, 11:1–13, 1960. 5

[54] Xin Ma, Yaohui Wang, Gengyun Jia, Xinyuan Chen, Ziwei Liu, Yuan-Fang Li, Cunjian Chen, and Yu Qiao. Latte: Latent diffusion transformer for video generation. *arXiv preprint arXiv:2401.03048*, 2024. 2

[55] Jeff Malpas. *The Place of Landscape: Concepts, Contexts, Studies*. The MIT Press, 2011. 2

[56] Albert Mehrabian and James A Russell. *An approach to environmental psychology.* the MIT Press, 1974. 1

[57] Arsha Nagrani, Joon Son Chung, and Andrew Zisserman. Voxceleb: a large-scale speaker identification dataset. *arXiv preprint arXiv:1706.08612*, 2017. 2

[58] 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. 6

[59] Udo Noll. radio aporee. *Berlin: Udo Noll*, 2013. 5

[60] Andrew Owens, Phillip Isola, Josh McDermott, Antonio Torralba, Edward H Adelson, and William T Freeman. Visually indicated sounds. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2405–2413, 2016. 2

[61] 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. 5

[62] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205, 2023. 2

[63] Can Qin, Ning Yu, Chen Xing, Shu Zhang, Zeyuan Chen, Stefano Ermon, Yun Fu, Caiming Xiong, and Ran Xu. Gluegen: Plug and play multi-modal encoders for x-to-image generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23085–23096, 2023. 2, 6, 7

[64] Jeppe H Rasmussen, Dan Stowell, and Elodie F Briefer. Sound evidence for biodiversity monitoring. *Science*, 385 (6705):138–140, 2024. 3

[65] Lukas Rauch, Raphael Schwinger, Moritz Wirth, René Heinrich, Denis Huseljic, Marek Herde, Jonas Lange, Stefan Kahl, Bernhard Sick, Sven Tomforde, and Christoph Scholz. BirdSet: A large-scale dataset for audio classification in avian bioacoustics. page arXiv:2403.10380. 2

[66] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 2, 3, 6

[67] Arana Ruiz et al. Urban soundscapes: A guide to listening for landscape architecture and urban design. 2024. 1

[68] Justin Salamon, Christopher Jacoby, and Juan Pablo Bello. A dataset and taxonomy for urban sound research. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 1041–1044. Association for Computing Machinery. 3

[69] R. Murray Schafer. *The Tuning of the World: Toward a Theory of Soundscape Design*. University of Pennsylvania Press, Philadelphia, 1977. 2

[70] MK Steele and JB Heffernan. Morphological characteristics of urban water bodies: mechanisms of change and implications for ecosystem function. *Ecological Applications*, 24 (5):1070–1084, 2014. 5

[71] Kim Sung-Bin, Arda Senocak, Hyunwoo Ha, Andrew Owens, and Tae-Hyun Oh. Sound to visual scene generation by audio-to-visual latent alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6430–6440, 2023. 2, 6, 7

[72] Kim Sung-Bin, Kim Jun-Seong, Junseok Ko, Yewon Kim, and Tae-Hyun Oh. Soundbrush: Sound as a brush for visual scene editing. *arXiv preprint arXiv:2501.00645*, 2024. 6

[73] Zineng Tang, Ziyi Yang, Chenguang Zhu, Michael Zeng, and Mohit Bansal. Any-to-any generation via composable diffusion. *Advances in Neural Information Processing Systems*, 36, 2024. 6, 7

[74] Silero Team. Silero vad: pre-trained enterprise-grade voice activity detector (vad), number detector and language classifier, 2024. 5

[75] Yi-fu Tuan. Images and mental maps. *Annals of the Association of American geographers*, 65(2):205–212, 1975. 1

[76] Konrad Uebel, Melissa Marselle, Angela J Dean, Jonathan R Rhodes, and Aletta Bonn. Urban green space soundscapes and their perceived restorativeness. *People and Nature*, 3(3): 756–769, 2021. 2

[77] Kai Wang, Shijian Deng, Jing Shi, Dimitrios Hatzinakos, and Yapeng Tian. Av-dit: Efficient audio-visual diffusion transformer for joint audio and video generation. *arXiv preprint arXiv:2406.07686*, 2024. 2

[78] Zehan Wang, Ziang Zhang, Hang Zhang, Luping Liu, Rongjie Huang, Xize Cheng, Hengshuang Zhao, and Zhou Zhao. Omnibind: Large-scale omni multimodal representation via binding spaces. *arXiv preprint arXiv:2407.11895*, 2024. 6

[79] Ho-Hsiang Wu, Prem Seetharaman, Kundan Kumar, and Juan Pablo Bello. Wav2clip: Learning robust audio representations from clip. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022. 6

[80] Shiqi Yang, Zhi Zhong, Mengjie Zhao, Shusuke Takahashi, Masato Ishii, Takashi Shibuya, and Yuki Mitsufuji. Visual echoes: A simple unified transformer for audio-visual generation. *arXiv preprint arXiv:2405.14598*, 2024. 2

[81] Guy Yariv, Itai Gat, Lior Wolf, Yossi Adi, and Idan Schwartz. Audiotoken: Adaptation of text-conditioned diffusion models for audio-to-image generation. *arXiv preprint arXiv:2305.13050*, 2023. 2, 6, 7

[82] Jing Yuan, Yu Zheng, and Xing Xie. Discovering regions of different functions in a city using human mobility and pois. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 186–194, 2012. 2

[83] Fan Zhang, Bolei Zhou, Liu Liu, Yu Liu, Helene H Fung, Hui Lin, and Carlo Ratti. Measuring human perceptions of a large-scale urban region using machine learning. *Landscape and Urban Planning*, 180:148–160, 2018. 6

[84] Yuan Zhang, Jian Kang, and Joe Kang. Effects of soundscape on the environmental restoration in urban natural environments. *Noise and Health*, 19(87):65–72, 2017. 2

[85] 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. 6

[86] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ade20k dataset. *International Journal of Computer Vision*, 127(3):302–321, 2019. 6

[87] 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. 2, 5

[88] Yonggai Zhuang, Yuhao Kang, Teng Fei, Meng Bian, and Yunyan Du. From hearing to seeing: Linking auditory and visual place perceptions with soundscape-to-image generative artificial intelligence. *Computers, Environment and Urban Systems*, 110:102122, 2024. 2, 5

[89] Yonggai Zhuang, Yanni Gui, and Teng Fei. Soundscape-to-panorama: spatialize auditory perception by linking acoustic environment to panorama. *International Journal of Digital Earth*, 18(1):2545584, 2025. 2