Sat2Sound: A Unified Framework for Zero-Shot Soundscape Mapping

Subash Khanal[1], Srikumar Sastry[1], Aayush Dhakal[1], Adeel Ahmad[1,2], Abby Stylianou[3], Nathan Jacobs[1]

[1]Washington University in St. Louis [2]Taylor Geospatial [3]Saint Louis University

CVPR JUNE 3-7, 2026 DENVER COLORADO

Overview

- Soundscape mapping predicts environmental sounds anywhere on Earth, enabling biodiversity monitoring, immersive media, and geospatial understanding.
- Existing methods rely on paired satellite imagery and geotagged audio but fail to capture semantic diversity and are limited to coarse image-level alignment.
- We use a VLM to generate semantically rich acoustic descriptions from satellite imagery and learn a codebook-based multimodal representation across four modalities — scaling soundscape maps from a single image to continental scale.

Original Audio Caption: "birds are chirping"
Synthetic Caption: "From the location captured in the aerial view image, we can expect to hear sounds of birds chirping, leaves rustling, and the gentle flow of water from the pond."

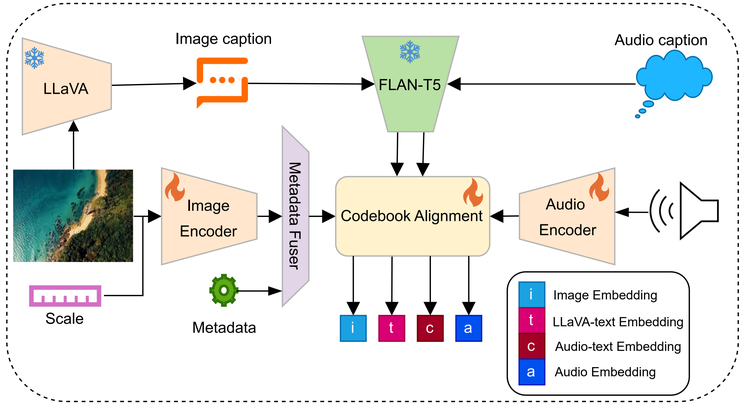
Method

$$r_m^i = \max_j \left(p_j^i \cdot C_m\right) \qquad w_m^i = \frac{\exp(r_m^i)}{\sum_{n=1}^{M} \exp(r_n^i)}$$

$$f^y = \sum_{m=1}^{M} w_m^y \cdot C_m, \quad y \in \{i, a, c, t\}$$

$$\mathcal{L}_{u,v}^{\dagger} = \mathcal{L}_{u,v} + \alpha \cdot \mathcal{L}_{u,v}^{\text{pseudo}} \qquad \mathcal{L}_{\text{tri}} = (\mathcal{L}_{i,a}^{\dagger} + \mathcal{L}_{i,c}^{\dagger} + \mathcal{L}_{a,c}^{\dagger})/3$$

$$\mathcal{L}_{total} = \mathcal{L}_{tri} + \mathcal{L}_{i,a+c}^{\dagger} + \mathcal{L}_{i,t}^{\dagger}$$



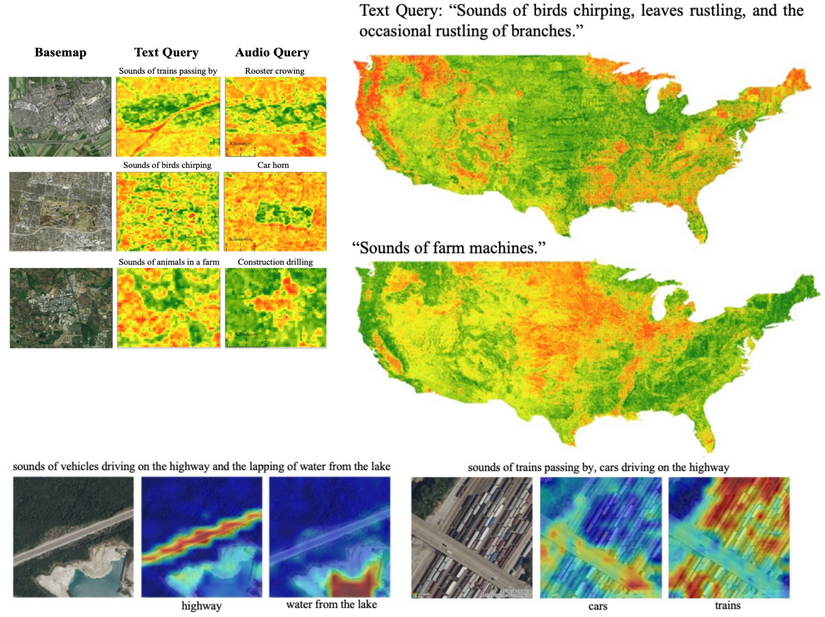
Results

Cross-Modal Retrieval

Dataset	Method	Metadata	Image-to-Audio R@10%	Image-to-Audio MR	Audio-to-Image R@10%	Audio-to-Image MR
GeoSound-Bing	GeoCLAP [17]	✗	0.399	1500	0.403	1464
	PSM [18]	✗	0.423	1401	0.428	1344
	Ours	✗	**0.534**	**872**	**0.535**	**850**
	PSM [18]	✓	0.828	261	0.829	248
	Ours	✓	**0.871**	**168**	**0.875**	**164**
GeoSound-Sentinel	GeoCLAP [17]	✗	0.459	1179	0.465	1141
	PSM [18]	✗	0.474	1101	0.485	1061
	Ours	✗	**0.549**	**802**	**0.556**	**778**
	PSM [18]	✓	0.802	294	0.804	283
	Ours	✓	**0.868**	**191**	**0.872**	**183**
SoundingEarth	GeoCLAP [17]	✗	0.454	667	0.449	694
	PSM [18]	✗	0.514	547	0.518	543
	Ours	✗	**0.570**	**438**	**0.562**	**463**
	PSM [18]	✓	0.563	454	0.569	447
	Ours	✓	**0.626**	**358**	**0.621**	**372**

Location-to-soundscape synthesis

Approach	Score	#Params	TFLOPS
Generative	3.77±0.51	7.57B	49.03
Retrieval	3.52±0.48	130M	0.14



Conclusion

- We present a unified multimodal framework that aligns audio, audio captions, satellite images, and VLM-generated image captions via contrastive learning over a shared codebook of soundscape concepts.
- Sat2Sound achieves state-of-the-art cross-modal retrieval on GeoSound and SoundingEarth, with interpretable local alignment between image patches and soundscape concepts.
- We introduce location-based soundscape synthesis via an efficient retrieval-based approach (Sat2Sound → Text2Audio), achieving audio quality comparable to cascaded generative method at a fraction of the compute.