# Sat2Sound: A Unified Framework for Zero-Shot Soundscape Mapping

Supplementary Material

## 8. Experimental Details

**Datasets:** We experiment with two datasets: *GeoSound* and *SoundingEarth*. *GeoSound* contains $294019/5000/9931$ train/validation/test samples and uses both $0.6m$ GSD (Ground Sample Distance) *Bing* image tiles ($1500 \times 1500$) and $10m$ GSD *Sentinel-2* image tiles ($1280 \times 1280$). *SoundingEarth* with $0.2m$ GSD *Google Earth* satellite image tiles of size ($1024 \times 1024$) contains $41469/3242/5801$ train/validation/test samples.

**Input Processing:** We process our three input modalities: audio, text, and image as follows:

*Audio:*We convert all input audio to mono, randomly sample a 10-second segment, and resample it to 32,000 Hz. The audio then undergoes STFT (window size 1024, hop length 320), followed by conversion to a 64-band Mel spectrogram (50–14,000 Hz), yielding a tensor of shape $1001 \times 64$, where $N^a = 1001$ denotes the number of temporal frames and $F = 64$ denotes the number of Mel frequency bins.

*Text:* Both audio captions and image captions are tokenized using the `google/flan-t5-large` tokenizer with `model_max_length` of 512.

*Image:* For the GeoSound dataset, we center-crop satellite images using a scale factor ($s$) from $\{1, 3, 5\}$, multiplied by the source-specific tile sizes (256 px for Sentinel-2 and 300 px for Bing). During training, to learn a unified multi-scale embedding space, we uniformly sample $s$ from $\{1, 3, 5\}$. For the SoundingEarth dataset, we apply a single-scale center crop of 256 px (i.e., scale $= 1$). In both cases, the cropped images are resized to $224 \times 224$ pixels and augmented with color jitter and normalization during training. The image encoder patchifies the image with a $16 \times 16$ patch size, producing 196 tokens ($N^{i,s}$).

**Metadata:** Following PSM [18], Sat2Sound is also trained with metadata (geolocation, month, hour, audio source, and audio caption source) in addition to satellite imagery and associated audio and text. For the GeoSound dataset used in our work, geotagged audio was collected from four sources: Freesound, Aporee, iNaturalist, and Flickr. Each metadata component is embedded into a 1024-dimensional vector and fused using Sat2Sound's transformer-based metadata fusion module. To prevent overfitting, we apply a dropout rate of 0.5, independently dropping each metadata component during training.

**Audio Captions:** The audio caption can either come from the user-uploaded textual description or be generated using recent SOTA audio-to-text generation models such as Pengi [10] or Qwen-Audio [6], with the caption selection based on the caption's CLAP score [41] with the ground-truth audio. For the *GeoSound* dataset, this resulted in 58.7% of audio captions from Pengi, 23.8% from Qwen-Audio, and 17.5% from human-annotated text.

**Image Captions:** For the cropped satellite images at each scale, we generate detailed soundscape captions using LLaVA [23], a powerful open-source Vision Language Model that has been proven effective in captioning satellite images. Specifically, we query `llava-hf/llava-1.5-7b-hf` on `HuggingFace` using the following prompt: *"What types of sounds can we expect to hear from the location captured by this aerial view image? Describe in up to two sentences."*

**Encoders:** Following [18], we fine-tune the pre-trained checkpoint for the SATMAE-Base [9] to encode satellite imagery while updating its positional embeddings with scale-aware GSDPE [29] to encode the scale of the satellite image. For audio, we fine-tune the pre-trained audio encoder of MGA-CLAP [21], which generates frame-level audio embeddings. The textual modality is processed using a frozen FLAN-T5 [30] model, which extracts token embeddings from texts for each sample.

**Hyper-parameters:** We set the embedding dimension of Sat2Sound ($d$) to 1024 and the number of concepts in the codebook ($M$) to 16000. We train our model using the `AdamW` optimizer with cosine-annealing with warm restarts as the learning rate scheduler with the following parameters: learning rate of 5e-5, weight decay of 0.2, and betas of (0.9, 0.98). We set the pseudo-positives contribution to loss ($\alpha$) in Equation 7 to 0.1. We train Sat2Sound for 20 epochs with train-batch size ($B$) of 128. For evaluation, we select the checkpoint that achieves the best I2A-R@10% performance on our validation set.

**Compute Infrastructure:** All experiments were conducted on an `NVIDIA H100 80GB` GPU, using 16 workers to enable faster data loading. We employed full-precision training throughout.

**Human Study:** In this study, 16 participants were shown a Bing satellite image at scale 1 for 20 locations on Earth. These 20 locations were selected by clustering SatCLIP's [**?** ] geolocation embeddings of all the samples in our gallery, with the centroid of each cluster serving as a test location. Each satellite image was paired with two 10-second synthetic audios generated using TangoFlux[15] with the inference parameters: `steps = 50` and `guidance = 4.5`. One audio was generated using the top-1 retrieved image caption by Sat2Sound, and the second using the directly generated LLaVA caption passed to TangoFlux.

## 9. Ablation Studies

### 9.1. Loss Ablation

We conduct an ablation study on different components of the loss to assess their impact on the overall training objective (Equation 9). We observe that the addition of the composite audio-based loss ($\mathcal{L}^{\dagger}_{i,a+c}$) slightly improves the performance of the standard audio-image cross-modal retrieval as observed in Table 4 and noticeably improves for composed audio-image cross modal retrieval as observed in Table 5. Furthermore, the inclusion of an additional image-text loss ($\mathcal{L}^{\dagger}_{i,t}$) does not degrade performance and provides the benefit of accurate retrieval of text as reflected in Table 2, Figure 5, and our demo available in our code repository[2].

### 9.2. Metadata Ablation

This experiment is designed to quantify the impact of different metadata components on the cross-modal retrieval performance of Sat2Sound. To this end, we evaluate our best-performing Sat2Sound models—trained with either Bing or Sentinel imagery—under varying combinations of metadata availability at inference time. As observed in Table 6, the most contributing metadata is *audio source*. This is consistent with results in prior work, PSM [18].

In addition to the independent metadata ablation presented in Table 6, we conduct a more detailed analysis of *composite* metadata combinations, assuming the availability of the *audio source* metadata as a base component (Table 7). This experiment evaluates the incremental contribution of time, geolocation, and month metadata when added cumulatively. The results show a clear pattern: model performance improves consistently as more metadata components are incorporated. Notably, for Bing imagery, the I2A-R@10% increases from 0.787 (only *audio source*) to 0.866 when time (hour and month) and geolocation (latitude, longitude) are added—demonstrating the benefit of simple, readily available metadata. In contrast, the performance gain from incorporating the *audio caption source* is modest, suggesting that any audio captioning model or user-written text can be used, as long as a reasonable audio caption is provided to query our model.

### 9.3. Codebook Size Ablation

We conduct an ablation of the codebook size of our framework. As seen in Tables 8 and 9, performance remains fairly consistent across different codebook sizes. We speculate that the sparsification operation [24] of the attention weights (Equation 4) encourages our framework to select only the relevant concepts, making the framework independent of the codebook size.

Our choice of codebook-based learning is motivated by the intuition that a fixed set of soundscape concepts can be shared across modalities. This approach offers a more interpretable way to align local features between the satellite image and corresponding soundscape elements. As shown in Figure 4, this local alignment serves as a valuable byproduct from an interpretability perspective, further discussed in Section 12.

[2] https://github.com/mvrl/sat2sound

## 10. Simpler Baselines

In this section, we compare the performance of Sat2Sound with existing off-the-shelf multimodal embedding spaces. As shown in the image-text cross-modal retrieval results (Table 10), existing pre-trained image-text models underperform compared to Sat2Sound. We attribute this to the mismatch between the soundscape descriptions generated by LLaVA from satellite images and the textual data these models were originally trained on. In contrast, Sat2Sound is explicitly trained on these captions, giving it a clear advantage and resulting in significantly better performance than the compared vision-language baselines. A similar trend is observed in Table 11 for image-audio cross-modal retrieval. These findings underscore the limitations of existing state-of-the-art multimodal embedding spaces for soundscape mapping, highlighting the need for a specialized framework like Sat2Sound tailored to this task.

## 11. Multi-scale Cross-Modal Retrieval

Sat2Sound is trained on multi-scale satellite imagery for the GeoSound dataset. The results presented in the main paper are for satellite imagery at scale 1. In this section, we present results for two additional scales: 3 and 5, using both Sentinel and Bing imagery from the GeoSound dataset. Additionally, for both datasets (GeoSound and SoundingEarth), we provide results for composed retrieval settings where the audio caption embedding is added either only to the audio query embedding (indicated as *audio* in the tables) or to both the audio query and image query embeddings (indicated as *query* in the tables), as done in PSM [18]. As observed in Tables 12, 13, and 14, Sat2Sound outperforms existing baselines by a noticeable margin in almost all of the settings.

## 12. Analyzing codebook concepts

As illustrated in Figure 4, the codebook learned by Sat2Sound can be used to generate fine-grained soundscape maps for regions covered by a single satellite image. In this section, we qualitatively explore what the codebook has learned. Specifically, for our gallery of image captions, we first obtain the corresponding codebook attention weights (Equation 4) and group together samples that share a similar set of highly activated codebook concepts. For a subset of these groups, we randomly sample examples to examine the behavior and semantic meaning captured by different

Table 4. Ablation of different loss components. Evaluated for cross-modal image-to-audio retrieval on GeoSound with Bing imagery at scale 1.

| trimodal | L(a+c) | L(i,t) | I2A-R@10% | I2A-MdR | A2I-R@10% | A2I-MdR |
|---|---|---|---|---|---|---|
| ✓ | | | 0.866 | 192 | 0.871 | 179 |
| ✓ | | ✓ | 0.852 | 206 | 0.852 | 203 |
| ✓ | ✓ | | 0.873 | 182 | 0.876 | 169 |
| ✓ | ✓ | ✓ | 0.871 | 168 | 0.875 | 164 |

Table 5. Ablation of different loss components. Evaluated for cross-modal composite audio-to-image retrieval on GeoSound with Bing imagery at scale 1.

| trimodal | L(a+c) | L(i,t) | I2A-R@10% | I2A-MdR | A2I-R@10% | A2I-MdR |
|---|---|---|---|---|---|---|
| ✓ | | | 0.935 | 88 | 0.949 | 75 |
| ✓ | | ✓ | 0.922 | 105 | 0.937 | 94 |
| ✓ | ✓ | | 0.960 | 71 | 0.962 | 62 |
| ✓ | ✓ | ✓ | 0.955 | 70 | 0.958 | 64 |

sets of codebook concepts. Representative samples are visualized in Figure 6.

## 13. Linear Probing Experiments

Sat2Sound learns a multimodal embedding space between audio and satellite imagery. We evaluate these embeddings on two downstream tasks: audio classification and satellite image classification, using linear probing on the audio and image embeddings, respectively.

For audio classification, we compare the Sat2Sound audio encoder against four strong baselines across three bird sound classification benchmarks: BirdCLEF-2022, BirdCLEF-2023, and BirdCLEF-2024. The baselines include CLAP [40], MGA-CLAP [21], ImageBind [12], and TaxaBind [32]. CLAP and MGA-CLAP represent state-of-the-art audio–text models, while ImageBind and TaxaBind align multiple modalities within a shared embedding space. As shown in Table 15, Sat2Sound achieves the best performance on two of the three benchmarks and the second-best on the remaining one. We attribute this to Sat2Sound being trained on the GeoSound dataset, where a significant portion of the samples originate from iNaturalist, providing diverse coverage of bird sounds. For satellite image classification, we evaluate Sat2Sound image embeddings on EuroSAT, UC-Merced, and RESISC-45, comparing them with SatMAE [9] and SatMAE++ [26]. Sat2Sound performs comparably but does not surpass either baseline, both of which use ViT-Large backbones compared to Sat2Sound's ViT-Base encoder. Nonetheless, Sat2Sound offers broader multimodal capabilities beyond pure image classification.

Table 6. Metadata ablation to evaluate Sat2Sound models trained on GeoSound dataset with satellite imagery at scale 1.

| Imagery | latlong | month | time | a-source | c-source | I2A-R10% | I2A-MR | A2I-R10% | A2I-MR |
|---|---|---|---|---|---|---|---|---|---|
| Sentinel | ✓ | | | | | 0.603 | 613 | 0.627 | 566 |
| Sentinel | | ✓ | | | | 0.585 | 682 | 0.604 | 613 |
| Sentinel | | | ✓ | | | 0.641 | 537 | 0.669 | 477 |
| Sentinel | | | | ✓ | | **0.805** | **318** | **0.808** | **306** |
| Sentinel | | | | | ✓ | 0.562 | 763 | 0.590 | 672 |
| Bing | ✓ | | | | | 0.627 | 547 | 0.644 | 499 |
| Bing | | ✓ | | | | 0.577 | 697 | 0.594 | 638 |
| Bing | | | ✓ | | | 0.629 | 564 | 0.651 | 521 |
| Bing | | | | ✓ | | **0.787** | **325** | **0.793** | **320** |
| Bing | | | | | ✓ | 0.551 | 785 | 0.566 | 742 |

Table 7. Composite metadata ablation to evaluate Sat2Sound models trained on GeoSound dataset with satellite imagery at scale 1.

| Imagery | a-source | time | latlong | month | c-source | I2A-R10% | I2A-MR | A2I-R10% | A2I-MR |
|---|---|---|---|---|---|---|---|---|---|
| Sentinel | ✓ | | | | | 0.805 | 318 | 0.808 | 306 |
| Sentinel | ✓ | ✓ | | | | 0.819 | 295 | 0.820 | 278 |
| Sentinel | ✓ | ✓ | ✓ | | | 0.850 | 239 | 0.851 | 227 |
| Sentinel | ✓ | ✓ | ✓ | ✓ | | 0.862 | 200 | 0.865 | 192 |
| Sentinel | ✓ | ✓ | ✓ | ✓ | ✓ | **0.868** | **191** | **0.872** | **183** |
| Bing | ✓ | | | | | 0.787 | 325 | 0.793 | 320 |
| Bing | ✓ | ✓ | | | | 0.807 | 289 | 0.811 | 283 |
| Bing | ✓ | ✓ | ✓ | | | 0.854 | 209 | 0.857 | 202 |
| Bing | ✓ | ✓ | ✓ | ✓ | | 0.866 | 175 | 0.871 | 169 |
| Bing | ✓ | ✓ | ✓ | ✓ | ✓ | **0.871** | **168** | **0.875** | **164** |

Table 8. Codebook ablation for Image-Text retrieval on GeoSound dataset with Bing imagery (scale=1) and corresponding image captions.

| Codebook Size | I2T-R10% | I2T-MR | T2I-R10% | T2I-MR |
|---|---|---|---|---|
| 4000 | 0.905 | 167 | 0.915 | 145 |
| 8000 | 0.899 | 163 | 0.917 | 146 |
| 16000 | 0.908 | 160 | 0.914 | 136 |
| 32000 | 0.902 | 165 | 0.914 | 148 |

Table 9. Codebook ablation for Image-Audio retrieval on GeoSound dataset with Bing imagery (scale=1) and corresponding audio.

| Codebook Size | I2A-R10% | I2A-MR | A2I-R10% | A2I-MR |
|---|---|---|---|---|
| 4000 | 0.868 | 167 | 0.870 | 161 |
| 8000 | 0.875 | 171 | 0.876 | 163 |
| 16000 | 0.871 | 168 | 0.875 | 164 |
| 32000 | 0.874 | 164 | 0.879 | 160 |

Table 10. Image-Text retrieval comparison with additional baselines. Results on GeoSound with Bing Imagery (scale=1).

| Model | I2T-R10% | I2T-MR | T2I-R10% | T2I-MR |
|---|---|---|---|---|
| CLIP [28] | 0.528 | 1420 | 0.461 | 1999 |
| SigLIP [47] | 0.340 | 3307 | 0.368 | 2652 |
| SigLIP2 [36] | 0.449 | 2641 | 0.397 | 2400 |
| Ours(w/o meta) | 0.881 | 183 | 0.900 | 166 |
| Ours(w meta) | **0.908** | **160** | **0.914** | **136** |

Table 11. Image-Audio retrieval comparison with additional baselines. Results on GeoSound with Sentinel Imagery (scale=1).

| Model | I2A-R10% | I2A-MR | A2I-R10% | A2I-MR |
|---|---|---|---|---|
| ImageBind [12] | 0.214 | 3675 | 0.231 | 3541 |
| TaxaBind [32] | 0.235 | 3448 | 0.250 | 3400 |
| Ours(w/o meta) | 0.549 | 802 | 0.556 | 778 |
| Ours(w meta) | **0.868** | **191** | **0.872** | **183** |

Table 12. Image-Audio retrieval results for SoundingEarth with different composed audio-image settings.

| Method | Composed | I2A-R10% | I2A-MR | A2I-R10% | A2I-MR |
|---|---|---|---|---|---|
| *Without Metadata* | | | | | |
| GeoCLAP | query | 0.523 | 533 | 0.470 | 641 |
| PSM | query | 0.687 | 234 | 0.560 | 451 |
| Ours | query | **0.847** | **94** | **0.564** | **448** |
| GeoCLAP | audio | 0.478 | 624 | 0.470 | 641 |
| PSM | audio | 0.558 | 462 | 0.560 | 451 |
| Ours | audio | **0.567** | **443** | **0.564** | **448** |
| *With Metadata* | | | | | |
| PSM | query | 0.690 | 264 | 0.608 | 371 |
| Ours | query | **0.855** | **91** | **0.862** | **129** |
| PSM | audio | 0.606 | 380 | 0.608 | 371 |
| Ours | audio | **0.855** | **127** | **0.862** | **129** |

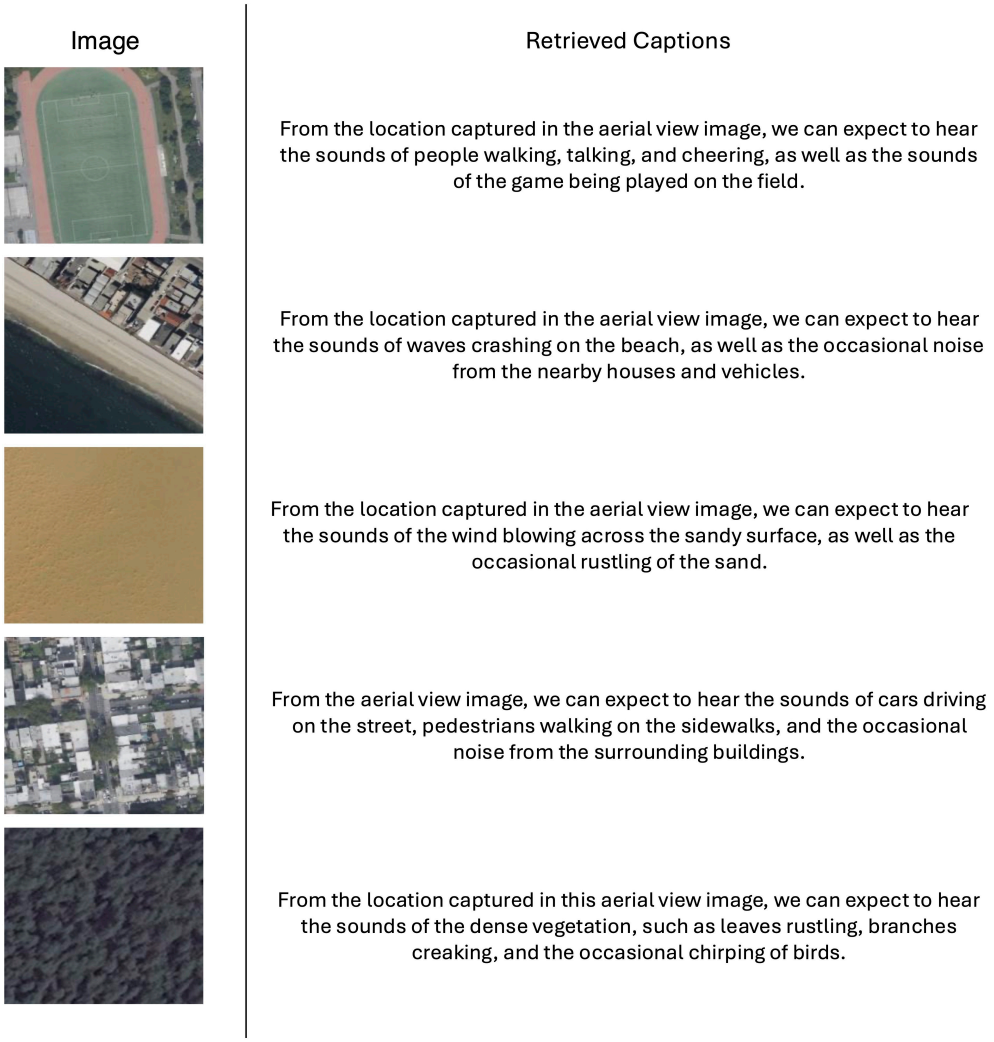


Figure 5. Examples of Top-1 retrieved LLaVA captions for a Bing image by Sat2Sound from our gallery, which is the *test*-set of the GeoSound dataset.

Table 13. Image-Audio retrieval results for GeoSound with Bing imagery at different scales.

| Scale | Method | Composed | I2A-R10% | I2A-MR | A2I-R10% | A2I-MR |
|---|---|---|---|---|---|---|
| *Without Metadata* | | | | | | |
| 1 | GeoCLAP | query | 0.577 | 712 | 0.468 | 1141 |
| | PSM | query | 0.754 | 204 | 0.510 | 952 |
| | Ours | query | **0.903** | **82** | **0.540** | **836** |
| | GeoCLAP | audio | 0.464 | 1159 | 0.468 | 1141 |
| | PSM | audio | 0.503 | 980 | 0.510 | 952 |
| | Ours | audio | **0.535** | **864** | **0.540** | **836** |
| 3 | GeoCLAP | none | 0.408 | 1441 | 0.420 | 1389 |
| | PSM | none | 0.440 | 1302 | 0.443 | 1266 |
| | Ours | none | **0.560** | **777** | **0.561** | **779** |
| | GeoCLAP | query | 0.577 | 707 | 0.483 | 1056 |
| | PSM | query | 0.753 | 207 | 0.529 | 880 |
| | Ours | query | **0.908** | **79** | **0.567** | **737** |
| | GeoCLAP | audio | 0.477 | 1092 | 0.483 | 1056 |
| | PSM | audio | 0.523 | 891 | 0.529 | 880 |
| | Ours | audio | **0.564** | **751** | **0.567** | **737** |
| 5 | GeoCLAP | none | 0.409 | 1428 | 0.421 | 1373 |
| | PSM | none | 0.440 | 1302 | 0.448 | 1279 |
| | Ours | none | **0.564** | **760** | **0.559** | **770** |
| | GeoCLAP | query | 0.581 | 698 | 0.489 | 1036 |
| | PSM | query | 0.753 | 209 | 0.532 | 863 |
| | Ours | query | **0.910** | **78** | **0.567** | **748** |
| | GeoCLAP | audio | 0.482 | 1071 | 0.489 | 1036 |
| | PSM | audio | 0.528 | 881 | 0.532 | 863 |
| | Ours | audio | **0.554** | **764** | **0.567** | **748** |
| *With Metadata* | | | | | | |
| 1 | PSM | query | 0.901 | 113 | 0.943 | 100 |
| | Ours | query | **0.970** | **33** | **0.958** | **64** |
| | PSM | audio | 0.935 | 115 | 0.943 | 100 |
| | Ours | audio | **0.955** | **70** | **0.958** | **64** |
| 3 | PSM | none | 0.827 | 266 | 0.832 | 250 |
| | Ours | none | **0.874** | **163** | **0.879** | **159** |
| | PSM | query | 0.900 | 114 | 0.945 | 102 |
| | Ours | query | **0.972** | **32** | **0.960** | **62** |
| | PSM | audio | 0.936 | 118 | 0.945 | 102 |
| | Ours | audio | **0.957** | **66** | **0.960** | **62** |
| 5 | PSM | none | 0.821 | 281 | 0.826 | 261 |
| | Ours | none | **0.877** | **167** | **0.882** | **167** |
| | PSM | query | 0.896 | 115 | 0.941 | 107 |
| | Ours | query | **0.972** | **32** | **0.963** | **64** |
| | PSM | audio | 0.929 | 124 | 0.941 | 107 |
| | Ours | audio | **0.959** | **68** | **0.963** | **64** |

Table 14. Image-Audio retrieval results for GeoSound with Sentinel imagery at different scales.

| Scale | Method | Composed | I2A-R10% | I2A-MR | A2I-R10% | A2I-MR |
|---|---|---|---|---|---|---|
| *Without Metadata* | | | | | | |
| 1 | GeoCLAP | query | 0.546 | 827 | 0.553 | 804 |
| | PSM | query | 0.803 | 153 | **0.595** | **664** |
| | Ours | query | **0.909** | **79** | 0.566 | 748 |
| | GeoCLAP | audio | 0.542 | 809 | 0.553 | 804 |
| | PSM | audio | **0.586** | **701** | **0.595** | **664** |
| | Ours | audio | 0.555 | 765 | 0.566 | 748 |
| 3 | GeoCLAP | none | 0.454 | 1200 | 0.456 | 1197 |
| | PSM | none | 0.479 | 1086 | 0.487 | 1042 |
| | Ours | none | **0.559** | **776** | **0.561** | **763** |
| | GeoCLAP | query | 0.542 | 840 | 0.555 | 790 |
| | PSM | query | 0.799 | 159 | **0.604** | **657** |
| | Ours | query | **0.910** | **81** | 0.577 | 729 |
| | GeoCLAP | audio | 0.548 | 812 | 0.555 | 790 |
| | PSM | audio | **0.594** | **676** | **0.604** | **657** |
| | Ours | audio | 0.561 | 757 | 0.577 | 729 |
| 5 | GeoCLAP | none | 0.458 | 1194 | 0.457 | 1184 |
| | PSM | none | 0.459 | 1172 | 0.465 | 1138 |
| | Ours | none | **0.545** | **804** | **0.560** | **774** |
| | GeoCLAP | query | 0.542 | 835 | 0.554 | 791 |
| | PSM | query | 0.796 | 158 | **0.584** | **711** |
| | Ours | query | **0.909** | **82** | 0.566 | 751 |
| | GeoCLAP | audio | 0.550 | 812 | 0.554 | 791 |
| | PSM | audio | **0.579** | **720** | **0.584** | **711** |
| | Ours | audio | 0.553 | 784 | 0.566 | 751 |
| *With Metadata* | | | | | | |
| 1 | PSM | query | 0.872 | 142 | 0.940 | 104 |
| | Ours | query | **0.972** | **35** | **0.959** | **70** |
| | PSM | audio | 0.931 | 123 | 0.940 | 104 |
| | Ours | audio | **0.956** | **78** | **0.959** | **70** |
| 3 | PSM | none | 0.795 | 306 | 0.800 | 290 |
| | Ours | none | **0.857** | **208** | **0.858** | **199** |
| | PSM | query | 0.870 | 150 | 0.940 | 104 |
| | Ours | query | **0.970** | **37** | **0.955** | **74** |
| | PSM | audio | 0.929 | 126 | 0.940 | 104 |
| | Ours | audio | **0.949** | **83** | **0.955** | **74** |
| 5 | PSM | none | 0.794 | 316 | 0.794 | 299 |
| | Ours | none | **0.846** | **220** | **0.851** | **216** |
| | PSM | query | 0.868 | 156 | 0.935 | 109 |
| | Ours | query | **0.969** | **37** | **0.954** | **80** |
| | PSM | audio | 0.926 | 131 | 0.935 | 109 |
| | Ours | audio | **0.948** | **88** | **0.954** | **80** |

Table 15. Linear probing evaluation for audio classification performance across different benchmarks.

| Model | BirdCLEF-2022 | BirdCLEF-2023 | BirdCLEF-2024 |
|---|---|---|---|
| CLAP [40] | 42.33 | 32.85 | 39.72 |
| MGA-CLAP [21] | **56.05** | 44.03 | 47.36 |
| ImageBind [12] | 47.11 | 37.46 | 45.04 |
| TaxaBind [32] | 52.60 | 42.19 | 49.31 |
| Ours | 54.84 | **45.69** | **49.59** |

Table 16. Linear probing evaluation for satellite image classification performance across different datasets.

| Model | ViT | EuroSAT | UC-Merced | RESISC-45 |
|---|---|---|---|---|
| SatMAE [9] | Large | 96.43 | 93.81 | 85.00 |
| SatMAE++ [26] | Large | 96.57 | 96.43 | 84.13 |
| Ours | Base | 83.83 | 90.95 | 74.40 |


Figure 6. Some example groups from the GeoSound test set are shown, where each group shares a common set of highly activated codebook concepts, reflecting similar soundscapes of specific geographic areas. The samples in (a) correspond to residential soundscapes, (b) reflect the soundscape of open fields, (c) represent forested area soundscapes, and (d) capture the soundscape of landscapes with water bodies.