Learning Spatially-Aware Language and Audio Embeddings

Bhavika Devnani[1,*]   **Skyler Seto**[2]   **Zakaria Aldeneh**[2]   **Alessandro Toso**[2]
Elena Menyaylenko[2]   **Barry-John Theobald**[2]   **Jonathan Sheaffer**[2]   **Miguel Sarabia**[2]
[1] Georgia Institute of Technology   [2] Apple
bdevnani3@gatech.edu, {sseto, zaldeneh, atoso}@apple.com
{elenam, bjtheobald, sheaffer, miguelsdc}@apple.com

Abstract

Humans can picture a sound scene given an imprecise natural language description. For example, it is easy to imagine an acoustic environment given a phrase like "the lion roar came from right behind me!". For a machine to have the same degree of comprehension, the machine must know what a lion is (semantic attribute), what the concept of "behind" is (spatial attribute) and how these pieces of linguistic information align with the semantic and spatial attributes of the sound (what a roar sounds like when its coming from behind). State-of-the-art audio foundation models, such as CLAP [7, 44], which learn to map between audio scenes and natural textual descriptions, are trained on non-spatial audio and text pairs, and hence lack spatial awareness. In contrast, sound event localization and detection models are limited to recognizing sounds from a fixed number of classes, and they localize the source to absolute position (e.g., 0.2m) rather than a position described using natural language (e.g., "next to me"). To address these gaps, we present ELSA (Embeddings for Language and Spatial Audio), a spatially aware-audio and text embedding model trained using multimodal contrastive learning. ELSA supports non-spatial audio, spatial audio, and open vocabulary text captions describing both the spatial and semantic components of sound. To train ELSA: (a) we spatially augment the audio and captions of three open-source audio datasets totaling 4,738 hours and 890,038 samples of audio comprised from 8,972 simulated spatial configurations, and (b) we design an encoder to capture the semantics of non-spatial audio, and the semantics and spatial attributes of spatial audio using contrastive learning. ELSA is a single model that is competitive with state-of-the-art for both semantic retrieval and 3D source localization. In particular, ELSA achieves $+2.8\%$ mean audio-to-text and text-to-audio R@1 above the LAION-CLAP [44] baseline, and outperforms by $-11.6°$ mean-absolute-error in 3D source localization over the SeldNET [40] baseline on the TUT Sound Events 2018 benchmark [1]. Moreover, we show that the representation-space of ELSA is structured, enabling swapping of direction of audio via vector arithmetic of two directional text embeddings.

1   Introduction

Humans use implicit context when communicating about and comprehending sounds in their environment. For instance, the instruction "Pull over if you hear a siren from behind you" is easily understood by most humans. However, a machine would need to not only recognize the source the sound, i.e., the siren (a semantic cue), but also interpret the spatial reference implied by "behind"

---

*Work done while at Apple.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

relative to its own position (a spatial cue). The machine must then translate these linguistic cues into its understanding of spatial audio to accurately identify, locate, and conditionally respond to the sound. This degree of alignment between spatial audio and natural language is understudied in prior work.

Audio foundation models (AFMs), such as LAION-CLAP [44], have been used for multiple downstream applications, such as language guided audio editing [42, 17], language guided audio and music generation [16, 11, 45], audio representations for image and text [14, 46], setting a precedent for the wide applicability of audio representations aligned with natural language. However, these models, and similar state-of-the-art AFMs, such as Pengi [5], LTU [12], and SALMONN [37], cannot capture the *spatial attributes* as the models are trained only on single-channel/non-spatial audio. Conversely, models such as SELDNet [1], and PILOT [33] are capable of precise spatial attribute classification and regression, but lack capability to generalize to natural language descriptions of spatial and semantic attributes.

To address these challenges, we introduce ELSA, a multimodal foundation model that learns a joint representation space for the spatial attributes and semantics of audio aligned with natural language descriptions of the audio scene. Learning a joint contrastive representation model enables and improves several tasks including: retrieval, multimodal QA, captioning, and generation [44]. Prior work has shown the benefits of learning well-aligned encoders for multimodal tasks within the vision domain [23, 35]. In contrast, mapping between audio and natural language via a large language model, as in [48], can significantly improve language reasoning tasks, especially for zero-shot generalization of pre-trained models, however yields worse performance in classification and QA tasks when fine-tuned in the language domain [41]. ELSA enables similar spatially-aware downstream applications, for instance, one can expand traditional language-guided audio editing to manipulate spatial elements using natural language commands like: "Remove the sound of the plane flying above" or "move the sound of the dog barking from left to right". In this paper we focus, for the first time, on devising and analyzing multimodal task-agnostic representations that capture both the semantics and the spatial attributes of audio aligned with natural language.

Noting a lack of paired spatial audio and language data that can enable training spatially aware audio-language models at scale, to train ELSA, we synthesize a spatial audio corpus consisting of 890,038 samples that span a variety of acoustic room properties, such as size and reverberation, from audio clips of the AudioSet [9] and Freesound [8] corpora. We also synthesize natural language spatial audio captions to match the spatial audio using a large-language model (LLM) to rephrase the initial captions. We demonstrate that ELSA captures spatial attributes and semantics of audio by identifying a set of tasks on which a standard AFM, such as LAION-CLAP, fails. Moreover, we show that ELSA achieves better zero-shot classification of spatial attributes than models trained only for that task. Finally, we show that ELSA maintains the ability to represent non-spatial audio by demonstrating performance competitive with existing state-of-the-art for a number of tasks.

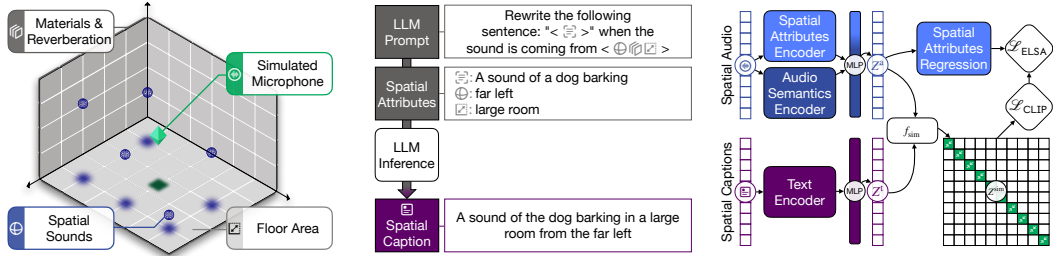
Our key contributions are:

- We present and release a new synthetic dataset of 4738.55 hours, with 890,038 samples and corresponding spatial captions across 8,972 simulated rooms with accurate parametric labels for the room properties and sound source locations. Additionally, we also record a small spatial real-world dataset to verify transfer to the real-world (cf. Section 3).

- We provide ELSA, a multimodal spatial audio-language model that jointly performs semantic classification (sound detection, retrieval), spatial localization, and direction of arrival. ELSA consists of an audio encoder paired with a text encoder that jointly learns semantic and spatial attributes via contrastive learning (cf. Section 4).

- We show that ELSA effectively captures spatial attributes and semantics competitive with baselines. ELSA improves by $-11.6°$ mean-absolute-error on 3D source localization, and by $+2.9\%$ on text-to-audio and audio-to-text mAP@10 scores. (cf. Section 5).

- Further, we show that the representation-space of ELSA is structured, allowing for transposition of spatial sound direction via addition or subtraction of two spatially descriptive text embeddings. (cf. Section 5.4).

2

(a) Our spatial audio pipeline uses simulated rooms with different dimensions, materials, and reverberation, and with sources located at different spatial locations.

(b) We augment the original captions by adding properties from the room simulations and prompt a LLM to rewrite the sentence.

(c) We encode the spatially augmented captions and audio, and then align the representations using a CLIP objective (see Fig. A.F.1 for full architecture).

Figure 1: Our pipeline for learning spatial-audio representations aligned with natural language.

2 Related Work

We provide an overview of the architecture choices and corresponding datasets for training and evaluation of models that can capture the semantics and spatial attributes of audio, including AFMs.

Audio-language Approaches   Prior works (e.g., CLAP [7], LAION-CLAP [44], MULAN [15]) have extended the image-text contrastive pre-training approach introduced by CLIP [27] to link audio representations to textual descriptions. These models use two encoders, one for audio and another for text, to project the representations from the two modalities into a common embedding space. Once trained, the models enable zero-shot prediction and retrieval capabilities on unseen sounds and textual descriptions. Despite their utility, the methods do not capture the spatial attributes of the modeled signals, rather they capture only their semantics. Another line of work (e.g., Pengi [5], LTU [12], and SALMONN [37]) extends LLMs to enable audio understanding in open-vocabulary settings (e.g., audio captioning and audio question answering). Such models learn audio encoders to provide a prefix token to prompt a frozen pre-trained autoregressive LLM, which is then used to generate unconstrained text. These prior methods do not explicitly model the spatial attributes of the audio. Zheng et al. [48] introduced BAT, an audio-based LLM that combines binaural spatial sound perception with natural language understanding with an accompanying question-and-answer dataset that enables model training. BAT focuses on enabling LLMs to reason about binaural spatial audio, which depends on the head-related transfer-function. In contrast our focus is on a task-agnostic and device-agnostic representation of spatial audio aligned with text.

Audio-language Datasets   Learning audio-language models requires access to datasets that link the two modalities. Clotho [6] and AudioCaps [18] are popular audio captioning datasets for which the textual descriptions were collected by annotating sound event datasets (e.g., AudioSet [9], or Freesound [8]) through crowd-sourcing platforms. LAION-Audio-630K [44] is a large-scale audio-text dataset collected by downloading audio and relevant textual descriptions from publicly available websites. All three datasets focus on the semantic attributes of the audio signal and do not have labels for the spatial attributes. SPATIALSOUNDQA [48] is a dataset that consists of simulated binaural audio samples and question-answer pairs, which was used to train BAT [48]. The audio samples were sourced from AudioSet [9], and the question-answer pairs were paraphrased using GPT-4. In contrast with the our environmental description captions, the text in SPATIALSOUNDQA is geared towards question-and-answer tasks. In addition, the dataset employs a binaural representation of spatial audio, rendering the data incompatible with ELSA. STARSS23 [36] is a dataset of real-world multi-channel audio annotated with semantic labels for overlapping sound sources, and the equivalent annotations for the spatial attributes. However, a limitation is that the dataset lacks natural language descriptions of the sound scenes, which are required for aligning the spatial attributes with language descriptions.

3 Paired Spatial Audio and Text Datasets

Multimodal contrastive learning approaches, e.g., CLIP [27] and CLAP [7], use large amounts of multimodal data pairs: 413M and 634k for the LAION versions of both models [32, 44]. Training a model

3

capable of understanding spatial audio as natural language requires a spatial audio dataset annotated with natural language spatial descriptions (e.g., "a dog barking in the far left corner of a room"). To the best of our knowledge, no such dataset is available. Thus, we use a spatial augmentation pipeline composed of two steps: simulating spatial audio in synthetic rooms (cf. Section 3.2 and Fig. 1a), and caption rephrasing using terms that refer to spatial audio attributes (cf. Section 3.2 and Fig. 1b). We use AudioCaps [18], Clotho [6], and Freesound [8] as base datasets for our augmentation pipeline.

The training set ensures at least two spatial augmentations per data point, allowing for the model to see the same audio with at least two different spatial augmentations per epoch. We generate two different sized versions of the evaluation and test sets. The larger version consists, once more, of at least two augmentations per audio sample, whilst the smaller version has no repeated samples and, consequently, is the same size as the original test set. The smaller dataset allows reporting retrieval results on the same sized dataset as the original, as size uniformity is key to consistency in retrieval metrics. The size of the respective datasets is reported in Appendix A.1. For all datasets, we use first-order ambisonics (FOA) as the encoding of spatial audio, which we describe next.

3.1 Spatial Audio Encoding: First Order Ambisonics

Monophonic, non-spatial audio, captures the spectral and temporal nature of sound, which carries a significant portion of context. Spatial audio provides additional context as it contains both spectral-temporal information and directional attributes, characterized by azimuth and elevation $(\theta, \phi) \in \mathbb{S}^2$, and distance. Binaural audio, a common spatial audio distribution format, mimics the signal entering the ear canals. Whilst binaural audio may be a natural choice for playback over headphones, it presents challenges for encoding, storing, and processing spatial information due to the presence of head-related transfer-functions in the signal [2]. To facilitate more processing flexibility, microphone array signals are often encoded as ambisonics [10]. This is accomplished by taking the spherical Fourier transform of the microphone signals and removing their radial component, which is equivalent to representing the spatial signal as a phase-coincident, infinite series in a spherical harmonic basis [30]. In practice, to avoid spatial aliasing, this series is truncated at an order proportional to the number of microphones in the array, with higher orders corresponding to a higher spatial resolution. Ambisonics are linearly mappable into a variety of audio playback formats, including binaural. First-order ambisonics (FOA) can be recorded using readily available four-channel microphone arrays, and have been shown to carry significant spatial information [50]. As such, we develop our models to ingest FOA signals. We leave generalization to higher orders for future work. It is worthwhile noting that once microphone array signals have been encoded into ambisonics, no a-priori knowledge on the structure of the capturing array is needed in order to perform any downstream spatial processing. Thus, ambisonics are agnostic to both recording and playback devices, making any embeddings derived from them equally generalizable.

3.2 Spatial Augmentation of the Audio and Captions

Like TUT Sounds Events 2018 [1] and BAT [48], we use a simulator to spatially augment non-spatial audio. The augmentation pipeline mirrors that of Spatial LibriSpeech [31]. We specify room configurations parameterized by size, shape, and reverberation time, where reverberation time is a function of the room structure and materials with characteristic absorption and scattering coefficients. The simulator further allows specification of the placement and direction of the receiver microphones relative to the source of the sound (see Fig. 1a). For each sample we remove leading and trailing silences, and repeat the audio signal to ensure that samples are at least four seconds long before simulation. A randomly chosen room, placement for the microphone, and placement for the static source is then selected. We ensure that the room augmentations do not overlap between the train, evaluation, and test datasets. The rooms vary in size between 13.3m^2 and 277.4m^2, their full-band T30 reverberance ranges from 114.5ms to 2671.9ms. The full statistics of these synthetic rooms can be found in Appendix A.1.

Our caption augmentation pipeline converts raw numerical values associated with the spatial audio attributes of the room simulator (e.g., distance to the microphone) into natural language descriptors (e.g., "near" or "far"). Our caption augmentation pipeline is shown in Fig. 1b. The full mapping from spatial audio attributes to natural language is given in Appendix A.2.

4

The original caption augmented with the spatial information makes up the input to LLaMA-13B [38], which is prompted to rephrase in the form of a spatially augmented caption. The prompt is:

```
The sound: <original caption> is coming from the <distance> <elevation>
<direction> of a <size> <reverb> room. Rephrase as a short English sentence
describing the sound and all the details of its source.
```

This template prompt overcomes challenges like non-English language in the original caption, missing spatial descriptors in the generated caption, and hallucinations that changed the meaning of the caption. We set the inference temperature of the LLM to 0.9 and the maximum tokens to 1,024. Appendix A.3 contains examples of the obtained spatial captions. We note that the caption re-writes can lead to hallucinations, which is discussed further in Appendix A.4. We leave the quantification and mitigation of hallucinations for future work.

3.3 Spatial Real-World Dataset

Our training data consists of synthetically-augmented audio and captions, so we also recorded a small dataset to verify generalization to real-world data (refer to Sections 5.2 and 5.3 for analysis). Our spatial real-world dataset was recorded using a Zylia 19 microphone spherical array at 48kHz with a bit-depth of 24-bits per sample. The dataset contains environmental sounds typically found in an apartment. In total, we recorded 70 samples of spatial audio in five rooms. Each spatial audio sample in the dataset was captioned with the semantic content (e.g., "sound of a vacuum"), and the direction {"left", "right", "front", "back"}, distance {"far", "near"}, and elevation {"up", "down", "level"}. For privacy, no personally identifiable information was included in the dataset.

4 ELSA Pretraining for Spatial Audio and Language

Our architecture is derived from LAION-CLAP [44], which is composed of an audio encoder and a text encoder that aligns embeddings for similar samples across modalities whilst maintaining the original representational capabilities of the individual modalities.

4.1 Audio Input Features

The audio encoder must capture both the semantics of the audio (e.g., "the sound of a fire alarm") and the spatial attributes (e.g., "the upper right of a reverberant room"). Following LAION-CLAP [7] and BAT [48], we translate the raw audio into the frequency domain. Consider a FOA signal represented by tensor, $\mathbf{A} \in \mathbb{C}^{T \times F \times (N+1)^2}$, where $N = 1$ is the spherical-harmonics order, T the number of time frames and F the number of frequency bins. More information on the derivation of $\mathbf{A}$ can be found in Appendix A.5. The corresponding real-valued log-mel spectrogram feature can be written as:

$$\mathrm{MEL}(t, \nu) = \log\left(|\mathbf{A}(t,f)|^2 \cdot \mathbf{W}_{\mathrm{mel}}(f, \nu)\right), \tag{1}$$

where $\mathbf{W}_{\mathrm{mel}}$ is the corresponding filter, ν is the filter index, t is time, and f is frequency. As summarized in Table 1 of SALSA [25], both mel-spectrograms and intensity vectors (IVs) are effective spatial features for FOAs. We extract the IVs, $\boldsymbol{I}(t, f)$ as follows:

$$\boldsymbol{I}_{\mathrm{active}}(t,f) = \Re\left[A_{0,0}^*(t,f)\begin{pmatrix} A_{1,\text{-}1}(t,f) \\ A_{1,0}(t,f) \\ A_{1,1}(t,f) \end{pmatrix}\right], \quad \boldsymbol{I}_{\mathrm{reactive}}(t,f) = \Im\left[A_{0,0}^*(t,f)\begin{pmatrix} A_{1,\text{-}1}(t,f) \\ A_{1,0}(t,f) \\ A_{1,1}(t,f) \end{pmatrix}\right], \tag{2}$$

where $A_{n,m}$ are the nth and m^{th} order and mode of the ambisonics signal corresponding to its omnidirectional (W) and three dipole (Z, Y, X) components, and $(\cdot)^*$ denotes complex conjugation. Physical normalization constants are omitted here for brevity as IVs are scaled to unit-norm [25].

For ELSA to use semantic features from both non-spatial audio and FOAs, during training we use sample from both the spatially-augmented datasets and the original non-spatial dataset. Since first-order ambisonics has four channels, and non-spatial audio only one, we copy the single-channel non-spatial signal across all channels. Intensity vectors normalize the dipoles by the omni channel, and result in identical IVs for non-spatial audio. We let the model learn this condition. We ablate the effect of using both spatial audio and non-spatial audio in Appendix A.6 and find that using both improves semantic retrieval.

5

4.2 Audio and Text Encoders

Our architecture is composed of an audio encoder and a text encoder. The audio encoder consists of two branches: the *semantic audio branch*, and the *spatial attributes branch*. See Appendix A.7 for a visualization of the full architecture.

For the semantic audio branch, we use HTSAT [3] since it was found to perform best in the LAION-CLAP evaluation [44]. HTSAT is a transformer-based audio encoder with self-attention blocks to achieve high performance in audio classification tasks. We initialize HTSAT with weights provided by LAION-CLAP[2]. For spatial-audio input, we feed only the mel-spectrogram of the omni channel from the first-order ambisonics encoding. The omni channel does not contain spatial characteristics, so its role is equivalent to single channel, non-spatial audio. This branch has 30M parameters.

As far as we are aware, there is no existing established feature encoder for spatial audio. Thus, for our spatial attributes branch we propose a two-branched CNN based on the architecture of [31] that was trained on a multi-task regression loss for azimuth, elevation, distance, and third-octave direct-to-reverberant ratio. The branch was trained for 100 epochs on Spatial LibriSpeech, which uses FOA spatial audio and has enough samples to train the spatial attributes branch. Further details, along with the full training hyper-parameters are discussed in Appendix A.8. This branch is fed the active and reactive intensity vector features described in Eq. (2). This branch has 486k parameters.

The outputs of both the *semantic* (768-dimensional) and the *spatial attributes* (192-dimensional) branches are concatenated to form a 960-dimensional embedding. Using a two-layer multi-layer perceptron (MLP), they are subsequently projected down to a 512-dimensional embedding.

For the text branch, we follow the best performing model in LAION-CLAP [44], and use RoBERTa-base [22]. RoBERTa is a general purpose bidirectional transformer [39], pretrained on a dynamically masked token prediction task, which employs byte-pair encoding [34] for tokenization. We use the same pre-trained model as [44] as the starting point[3]. The text encoder has 125M parameters, and the final embedding has a dimensionality of 712, which also is projected down to 512 by a two-layer MLP, matching the size of the audio encoder output.

4.3 Pretraining Objectives

We learn aligned representations using batched contrastive loss (popularized by CLIP [27]). The loss function rewards the alignment of representations from the same sample but different modalities, and penalizes the alignment of representations from different samples (see Fig. 1c). Our loss (in common with CLIP [27], CLAP [7], and LAION-CLAP [44]) is derived from the InfoNCE loss [26], as we now describe. Given a set of embeddings of any modality $X \in \mathbb{R}^{N \times D}$ where the ith entry, $x_i \in \mathbb{R}^D$ is to be matched with $y \in \mathbb{R}^D$, the following InfoNCE sample loss maximizes the similarity between the pair x_i and y, and minimizes the similarity between all other x and y pairs:

$$\mathcal{L}_{\text{InfoNCE}}(X, x_i, y) = -\log \frac{f_{\text{sim}}(x_i, y)}{\sum_{x_j \in X} f_{\text{sim}}(x_j, y)}, \tag{3}$$

where $f_{\text{sim}}(a, b) = \exp(a \cdot b / \tau)$ is a similarity function with a learnable temperature parameter τ. Taking the average across all audio-text pairs in the batch, where entries at the ith position match each other, we arrive at the CLIP loss:

$$
\begin{aligned}
\mathcal{L}_{\text{CLIP}} &= \frac{1}{2} \left(\frac{1}{N} \sum_{i=0}^{N} \mathcal{L}_{\text{InfoNCE}}(Z^a, z_i^a, z_i^t) + \frac{1}{N} \sum_{i=0}^{N} \mathcal{L}_{\text{InfoNCE}}(Z^t, z_i^t, z_i^a) \right), \\
&= -\frac{1}{2N} \sum_{i=0}^{N} \left(\log \frac{f_{\text{sim}}(z_i^a, z_i^t)}{\sum_{j=0}^{N} f_{\text{sim}}(z_j^a, z_i^t)} + \log \frac{f_{\text{sim}}(z_i^t, z_i^a)}{\sum_{j=0}^{N} f_{\text{sim}}(z_j^t, z_i^a)} \right),
\end{aligned} \tag{4}
$$

Since the rooms we use to spatially-augment the audio are parametric, we have accurate labels associated with spatial features of the audio source. We take advantage of these labels by adding three additional spatial regression objectives. We feed the generated 512 dimension audio embedding into

---

[2]We use `HTSAT-fullset-imagenet-map=0.467.ckpt` from https://github.com/LAION-AI/CLAP
[3]We use the weights for `roberta-base` from: https://dl.fbaipublicfiles.com/fairseq/models/roberta.base.tar.gz

6

Table 1: Comparison of model capabilities and performance for retrieval of semantic captions from AudioCaps, and 3D sound localization for the REAL component TUT Sound Events 2018. ELSA is the only model that allows both open vocabulary language understanding and spatial localization, and performs comparably against the baselines for both tasks.

MODEL	SEMANTIC CAPABILITIES	SPATIAL CAPABILITIES	AUDIOCAPS MAP@10↑	REAL 3D LOCAL.($°$)↓
SeldNET [1]	✗ Limited vocab.	✓	✗	26.6
PILOT [33]	✗ Limited vocab.	✓	✗	4.2
Spatial Librispeech [31]	✗	✓	✗	12.4
LAION-CLAP [44]	✓ Open vocab.	✗	43.8	95.29
ELSA (ours)	✓ Open vocab.	✓	**44.2**	14.97

three 2-layer MLPs of 33k parameters, which respectively regress the direction of arrival (azimuth and elevation) of sound in 3D space, distance of the source to the receiver, and room floor area. These objectives, along with the CLIP loss in Eq. (4), define our final loss:

$$\mathcal{L}_{\text{ELSA}} = \mathcal{L}_{\text{CLIP}} + \mathcal{L}_{\text{dir}} + \mathcal{L}_{\text{dist}} + \mathcal{L}_{\text{area}}, \tag{5}$$

where $\mathcal{L}_{\text{dir}}$ is the cosine similarity between the predicted and target angles, and $\mathcal{L}_{\text{dist}}$ and $\mathcal{L}_{\text{area}}$ is the mean-squared error between the predicted and target distances and room floor area respectively. We ablate the differences between $\mathcal{L}_{\text{ELSA}}$ and $\mathcal{L}_{\text{CLIP}}$ in Appendix A.6 and find that at a negligible cost (0.4%) to semantic retrieval, we get a 15.3% improvement to 3D localization capability and 12.3% improvement in distance estimation when using $\mathcal{L}_{\text{ELSA}}$.

5 Experiments, Results, and Discussion

We demonstrate that ELSA jointly captures the semantics and spatial attributes of sound with either audio or text inputs by answering the following research questions:

RQ1 Does ELSA capture *spatial attributes* in spatial audio (Section 5.2)?

RQ2 Does ELSA capture *semantic information* in both text and audio (Section 5.3)?

RQ3 Does ELSA transfer to our real-world dataset? (Sections 5.2 and 5.3)?

RQ4 Does ELSA provide interpretable multimodal representations (Section 5.4)?

RQ5 Are ELSA embeddings capable of driving automatic captioning (Section 5.5)?

5.1 Training and Evaluation of ELSA

As indicated in Section 4.2, we use pretrained weights for the semantic audio encoder, the spatial attributes encoder, and the text encoder. All components of the model are fine-tuned, which corresponds to 158M trainable parameters, an increase of 0.86% over LAION-CLAP[3].

For our best model, we train for 40 epochs on 12 nodes, each with 8 NVIDIA A100 GPUs and 96 CPU cores with a batch size of 2,304. Training converges within 17 hours. We use the Adam optimizer with a learning rate of 5×10^{-5} and cosine scheduling. We select the checkpoint with the lowest mAP@10 retrieval on the spatially augmented captions.

5.2 Spatial Attributes Evaluation

Table 2: Zero-shot classification accuracy using the cosine similarity between test set audio embeddings and templated probe caption embeddings. The template is "A sound coming from <spatial attribute>" and a value for <spatial attribute> is substituted into the template representing the desired class (e.g., "near" or "far" for distance). A classification is correct if the attribute in the closest test sample matches the attribute in the template. We cannot provide comparisons with baselines since this is a new task.

TASK	S-CLOTHO	S-AC	S-RWD
Distance (2-class)	96.0%	92.9%	67.1%
Direction (4-class)	92.0%	92.8%	35.8%
Elevation (2-class)	100.0%	100.0%	72.1%
Room area (2-class)	76.6%	74.7%	N/A
Reverberation (2-class)	100.0%	83.3%	N/A

We show that ELSA captures the spatial attributes of sound (**RQ1**) by carrying out downstream regression and zero-shot spatial prompt classification. For regression to 3D sound localization, we a train two-layer MLP with 32,768

7

Table 3: Semantic retrieval (R@1, R@5, and R@10) for CLAP and ELSA calculated over the original (non-spatial) versions of Clotho and AudioCaps. Although ELSA is trained using a mixture of non-spatial and spatial audio, it conserves the retrieval performance on non-spatial audio of LAION-AI CLAP, which was trained on only non-spatial data. For the training data, read C as Clotho, AC as AudioCaps, LA as LAION-Audio-630K and FS as Freesound. A superscript ^{S} denotes the spatially-augmented equivalent dataset. We use Freesound, a subset of LAION-Audio-630K due to its more permissive licensing. For a fair comparison, we train a version of CLAP locally with Clotho, Audiocaps and Freesound, which is not reported in the CLAP paper.

| | | AUDIOCAPS | | | | | | CLOTHO | | | | | |
| | | TEXT-TO-AUDIO | | | AUDIO-TO-TEXT | | | TEXT-TO-AUDIO | | | AUDIO-TO-TEXT | | |
MODEL	TRAIN DATA	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
CLAP (paper)	C, AC, LA	34.7	70.5	83.2	45.3	79.5	89.2	16.4	39.0	51.0	21.8	44.6	60.1
CLAP (local)	C, AC, FS	32.7	68.8	81.5	40.7	74.0	84.7	14.4	37.6	50.7	18.3	40.5	55.1
ELSA	$C, AC, FS, C^S, AC^S, FS^S$	**33.2**	68.2	81.0	**40.9**	**74.4**	**86.1**	**15.0**	36.7	**50.8**	**20.1**	**43.2**	**55.4**

parameters using the ELSA audio embeddings generated from the training set. We then evaluate on the REAL component of the TUT Sound Events 2018 dataset [1]. Table 1 confirms that CLAP cannot encode spatial attributes (95.29°), whereas ELSA achieves 14.97° mean-absolute error (MAE) and maintains a higher mAP@10 for semantic retrieval tasks than CLAP. Appendix A.9 shows that there is little variability in the direction-of-arrival error across various spatial attributes. However, we note the errors tend to be higher at the extrema of the dimensions. When compared to methods designed explicitly for 3D sound localization, ELSA performs better than SeldNET (the baseline included with TUT Sound Events 2018) by +11.6°, and that achieves only -2.6° MAE compared to the model in Spatial LibriSpeech[4]. ELSA does not reach the performance of PILOT (4.3°), but this model was specifically-tuned only for 3D sound localization on data derived from TUT Sound Events 2018 [33].

To verify that the spatial attributes are aligned with language, we create new captions using the template: "`A sound coming from <spatial attribute>`", where `<spatial attribute>` can be distance, direction, elevation, room size, and reverberation. For instance, a caption for distance might be "`A sound coming from far away`". The ELSA text embeddings for such captions are extracted from the pre-trained encoder and compared in a zero-shot fashion with ELSA audio embeddings for samples from the test set using cosine similarity. We classify the match as correct if the spatial attribute in the closest audio sample matches the spatial attribute of the query caption, and we report accuracy in Table 2. ELSA achieves >90% correct retrieval for most spatial attributes. For room area, ELSA achieves 74.7% correct retrieval, which we hypothesize is due to the relatively small perceptual differences between small (<50m^2) and large rooms (>100m^2). We observe a transfer gap on retrieval scores when evaluating on our spatial real-world dataset, with ELSA achieving 67.1% (distance), 35.8% (direction) and 72.1% (elevation) correct retrieval. Part of this performance difference is because the spatial attributes were only estimates by the annotators during data capture. On the other hand, performance using pre-trained LAION-CLAP is close to random for all tasks (Appendix A.10), which is expected as CLAP was not trained with spatial audio or captions.

5.3 Semantics Evaluation

Following LAION-CLAP [44], we calculate retrieval results when finding matches from audio-to-text and text-to-audio. To compute retrieval, we encode the test set for each modality, and for every sample we check whether the corresponding sample in the other modality has the closest cosine distance (R@1), is within the five closest samples (R@5), or within the ten closest samples (R@10). The results in Table 3 show that in addition to learning representations of spatial captions and spatial audio, ELSA also performs on par with LAION-CLAP on non-spatial tasks. Table A.T.8 in Appendix A.12 shows the retrieval results when using spatially-augmented versions of AudioCaps and Clotho. We remark that adding Freesound to the training set decreases the retrieval scores in Spatial AudioCaps, but improves retrieval scores in Clotho, due to Clotho being a differently-captioned subset of Freesound. We note that the spatial retrieval performance of ELSA is lower than the non-spatial retrieval performance (for instance, -9.4% and -13.3% on audio-to-text and text-to-audio R@10

---

[4]The Spatial LibriSpeech model can be considered a supervised version of ELSAs contrastive learning, since the authors also train on a synthetically augmented dataset.

8

AudioCaps). This reflects the fact that spatial captions are harder to match, since there is a larger number of attributes and since there are hard-negatives (same semantics, different spatial attributes). Still, ELSA achieves the highest retrieval scores on the spatial real-world dataset (Table A.T.9 in Appendix A.12), showcasing its ability to transfer to the real-world without fine-tuning. Note that we cannot provide comparisons with prior models since using these spatial augmentations is a new task.

5.4 Interpreting the representation structure of ELSA

To confirm that directional characteristics in ELSA spatial caption embeddings are encoded in the same feature space as those in ELSA spatial audio embeddings, we train a direction regressor with a two-layer MLP using the *spatial audio embeddings* in the training split. We subsequently regress the *spatial text embeddings* to azimuth values using our trained regressor and affix direction labels ("left", "right", "front", "back") to each sample based on the azimuth values. We obtain 64.3% accuracy on the four-class problem, indicating alignment between the encoding of the modalities. Similarly, we obtain an accuracy of 76.5% for when classifying over distance labels ("far", "near") and 55.1% over elevation labels ("up", "down").

Besides using regression to confirm that ELSA embeddings capture spatial direction, we verify whether the ELSA embeddings can be clustered by spatial attributes. Fig. 2 shows a UMAP projection of the ELSA embeddings from the test sets of Spatial-AudioCaps and Spatial-Clotho. Note that the UMAP projection was guided with the embeddings and labels of the training sets of both datasets. The figure shows the embeddings cluster well with the direction labels, though there is some degree of confusion between "back" and "front". This is corroborated by the analysis in Appendix A.13, where we compute Wassertein distances directly in the 512-dimensional space. We carried out a similar analysis for spatial distance and found the embeddings cluster clearly between "near" and "far".

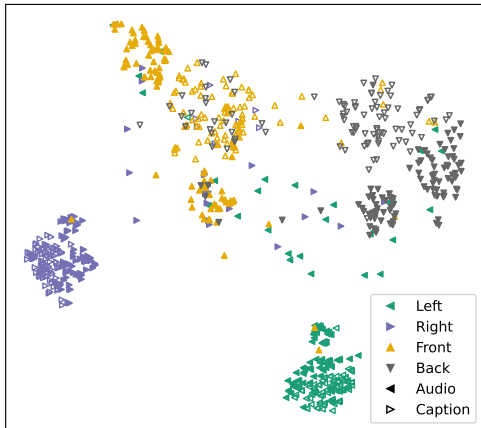



Figure 2: UMAP projection of ELSA embeddings of the test splits of Spatial-Clotho and Spatial-AudioCaps. Filled markers are obtained from spatial audio, and hollow markers are obtained from spatial captions. The UMAP projection was fitted with the train splits of Spatial-Clotho and Spatial-Audio caps, and we made use of supervised dimension reduction to highlight the *direction* differences rather than the semantic differences in the embeddings.

We validate that ELSA audio embeddings capture implicit spatial attributes that are latent in the text-encoder by first training a classifier using the spatial audio in our training data, where the classes are broad directions, such as *above* and *below*. We use LLaMA-13B [38] to generate descriptions of sounds that would typically come from each of these directions, e.g., "the rhythmic drumming of raindrops on a skylight" (for above) and "the faint creaking of an old house settling" (for below). Appendix A.11 lists all generated captions. Finally, we use the classifier trained on audio samples to classify ELSA embeddings for these generated captions with implicit directionality. We find that the classifier can correctly identify 68% of the sounds typically heard from above as being from above, showing that the latent space of the text encoder for ELSA is capturing directionality.

Lastly, we show that we can swap the spatial direction encoded by an ELSA audio embedding with a simple text caption. We first obtain ELSA prototypes for four directions ("left", "right", "front", "back") with the template: "A sound coming from the `direction`". Next, we train a 4-class direction classifier with a two-layer MLP using the spatial audio in the training splits of our spatially-augmented datasets. To swap the direction of the sound, we subtract the text prototype of the original direction and add prototype for the new direction. For evaluation, we swap the spatial direction of every sample in our spatially-augmented test set that was correctly classified by the 4-class direction classifier (96.7% of the audio embeddings). Our results show that 99.7% of the swapped samples are classified correctly with the new spatial direction, which highlights the strong alignment of spatial features across modalities, resulting in the ability to edit spatial attributes of

9

existing spatial audio using text in embedding space. Further details about this experiment are described in Appendix A.14. These results also point to exciting avenues wherein text can condition the manipulation and generation of spatial characteristics of audio. We leave this application for future work.

5.5 Spatial Audio Caption Generation

Decoding multimodal embeddings into natural language can be achieved by prefixing an autoregressive causal language model [24, 13, 19, 4], where the prefix is constructed from a projection of the multimodal embeddings. To facilitate audio captioning using ELSA, we fine-tune a GPT-2 model [28] with 12 attention layers each having 12 heads (with 163M parameters). The ELSA embeddings are projected onto the prefix using a single dense layer (393k parameters). With the ELSA encoder frozen, we train the GPT-2 model on 150k spatial-audio embedding and caption pairs from Spatial-Clotho and Spatial-AudioCaps. We report caption generation metrics in Table 4 and show three generation samples in Appendix A.15. Overall, we find that automatic spatial audio captioning systems are viable though more work is needed to increase the vocabulary size of the generations.

Table 4: Evaluation of Spatial Audio Caption Generation. Metrics were obtained from the Audio Captioning task of the DCASE Challenge[5] by comparing the generated captions produced from spatial audio and the ground-truth captions from the test splits of Spatial-AudioCaps (S-AC) and Spatial-Clotho.

METRIC	RANGE	S-CLOTHO	S-AC
SPIDEr [21]	[0, 5.5]	0.19	0.34
FENSE [49]	[-1.0, +1.0]	0.59	0.68
# Unique words	$[0, \infty)$	1103	1258

6 Conclusions, Limitations, and Further Work

We have presented ELSA, an AFM that aligns representations of spatial audio and equivalent text descriptions. To train such representations we built a pipeline to spatially augmented the audio in existing non-spatial audio-text datasets, such as Clotho [6] and AudioCaps [18], and added spatial information to their respective captions. Our results show that ELSA embeddings capture *both* the semantic contents and the spatial attributes of the audio, with ELSA achieving +2.8% higher scores in audio-to-text and text-to-audio retrieval scores than the state-of-the-art, and obtaining -2.6°MAE in direction of arrival error with respect to an equivalent baseline. Interestingly, by mixing spatial and non-spatial audio and caption pairs, ELSA is able to represent non-spatial audio as well, Finally, we show that the representation space of ELSA is structured in that the directionality of a spatial audio sample can be transposed by simple addition or subtraction of two text representations. Future work will explore acoustic scenarios with overlapping sound sources and sound sources that are moving in the scene. ELSA will also benefit from advances in spatial attributes encoders. In this work, we used the augmented spatial captions as is, but further work should ensure consistency with the semantics before and after augmentation, which will further improve the representational power of ELSA.

Perceiving spatial audio is a fundamental aspect of human nature. As is linking perception with language. Using spatial audio and a contrastive multimodal training approach, ELSA bridges the gap between feature rich spatial audio and language, paving the way for more intuitive and effective human-machine interactions by allowing for richer understanding of the users' environment and generation of immersive sound scenes from natural language.

Broader Impact   Our research has the potential to be used in creation of immersive augmented or virtual reality environments. If not controlled well, these immersive experiences have the potential to become addictive, and thus impact the mental health of individuals or even society as a whole. Another danger is possibility of creating deepfakes of soundscapes, thus making it possible for generated 3D environments to sound very realistic. The proliferation of deepfake soundscapes could lead to misinformation and manipulation, undermining trust in audio media.

---

[5]Available at https://github.com/Labbeti/aac-metrics.

10

Acknowledgements

The authors would like to thank Nicholas Apostoloff, Masha Fedzechkina, Rin Metcalf, Russ Webb, Megan Maher Welsh, and Luca Zappella for their insightful input and discussions on earlier versions of this paper. Moreover, we are thankful to Denise Hui and David Koski for technical support. Names are in alphabetical order by last name within group.

References

[1] Sharath Adavanne, Archontis Politis, Joonas Nikunen, and Tuomas Virtanen. "Sound Event Localization and Detection of Overlapping Sources Using Convolutional Recurrent Neural Networks". In: *Journal of Selected Topics in Signal Processing* 13.1 (2019).

[2] Jens Blauert. "The Technology of Binaural Listening". Springer, 2013.

[3] Ke Chen, Xingjian Du, Bilei Zhu, Zejun Ma, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. "HTS-AT: A Hierarchical Token-semantic Audio Transformer for Sound Classification and Detection". In: *International Conference on Acoustics, Speech and Signal Processing*. IEEE. 2022, pp. 646–650.

[4] Soham Deshmukh, Benjamin Elizalde, Dimitra Emmanouilidou, Bhiksha Raj, Rita Singh, and Huaming Wang. "Training Audio Captioning Models without Audio". In: *International Conference on Acoustics, Speech and Signal Processing*. IEEE. 2024, pp. 371–375.

[5] Soham Deshmukh, Benjamin Elizalde, Rita Singh, and Huaming Wang. "Pengi: An Audio Language Model for Audio Tasks". In: *Advances in Neural Information Processing Systems*. Vol. 36. 2023, pp. 18090–18108.

[6] Konstantinos Drossos, Samuel Lipping, and Tuomas Virtanen. "Clotho: An Audio Captioning Dataset". In: *International Conference on Acoustics, Speech and Signal Processing*. IEEE. 2020, pp. 736–740.

[7] Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. "CLAP: Learning Audio Concepts from Natural Language Supervision". In: *International Conference on Acoustics, Speech and Signal Processing*. IEEE. 2023, pp. 1–5.

[8] Frederic Font, Gerard Roma, and Xavier Serra. "Freesound Technical Demo". In: *International Conference on Multimedia*. ACM. 2013, pp. 411–412.

[9] 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: *International Conference on Acoustics, Speech and Signal Processing*. IEEE. 2017, pp. 776–780.

[10] Michael J. Gerzon. "Periphone (with Height Sound Reproduction)". In: *Journal of the Audio Engineering Society* M07 (1972).

[11] Deepanway Ghosal, Navonil Majumder, Ambuj Mehrish, and Soujanya Poria. *Text-to-Audio Generation using Instruction-Tuned LLM and Latent Diffusion Model*. arXiv: 2304.13731 [eess.AS].

[12] Yuan Gong, Hongyin Luo, Alexander H Liu, Leonid Karlinsky, and James Glass. "Listen, Think, and Understand". In: *International Conference on Learning Representations*. 2024.

[13] Sophia Gu, Christopher Clark, and Aniruddha Kembhavi. "I can't believe there's no images!: Learning Visual Tasks Using Only Language Supervision". In: *International Conference on Computer Vision*. IEEE. 2023, pp. 2672–2683.

[14] Andrey Guzhov, Federico Raue, Jörn Hees, and Andreas Dengel. "AudioCLIP: Extending CLIP to Image, Text and Audio". In: *International Conference on Acoustics, Speech and Signal Processing*. IEEE. 2022, pp. 976–980.

[15] Qingqing Huang, Aren Jansen, Joonseok Lee, Ravi Ganti, Judith Yue Li, and Daniel P W Ellis. "MuLan: A Joint Embedding of Music Audio and Natural Language". In: *International Society for Music Information Retrieval Conference*. ISMIR. 2022, pp. 559–566.

[16] Rongjie Huang, Jiawei Huang, Dongchao Yang, Yi Ren, Luping Liu, Mingze Li, Zhenhui Ye, Jinglin Liu, Xiang Yin, and Zhou Zhao. "Make-An-Audio: Text-To-Audio Generation with Prompt-Enhanced Diffusion Models". In: *International Conference on Machine Learning*. PMLR, 2023, pp. 13916–13932.

11

[17] Xilin Jiang, Cong Han, Yinghao Aaron Li, and Nima Mesgarani. *Listen, Chat, and Edit: Text-Guided Soundscape Modification for Enhanced Auditory Experience*. 2024. arXiv: 2402.03710 [eess.AS].

[18] Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. "Audiocaps: Generating Captions for Audios in the Wild". In: *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics. 2019, pp. 119–132.

[19] Minkyu Kim, Kim Sung-Bin, and Tae-Hyun Oh. "Prefix Tuning for Automated Audio Captioning". In: *International Conference on Acoustics, Speech and Signal Processing*. IEEE. 2023, pp. 1–5.

[20] Rosanne Liu, Joel Lehman, Piero Molino, Felipe Petroski Such, Eric Frank, Alex Sergeev, and Jason Yosinski. "An intriguing failing of convolutional neural networks and the coordconv solution". In: *Advances in Neural information Processing Systems* 31 (2018).

[21] Siqi Liu, Zhenhai Zhu, Ning Ye, Sergio Guadarrama, and Kevin Murphy. "Improved Image Captioning via Policy Gradient optimization of SPIDEr". In: *International Conference on Computer Vision*. IEEE. 2017, pp. 873–881.

[22] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. *RoBERTa: A Robustly Optimized BERT Pretraining Approach*. 2019. arXiv: 1907.11692 [cs.CL].

[23] Jack Merullo, Louis Castricato, Carsten Eickhoff, and Ellie Pavlick. "Linearly Mapping from Image to Text Space". In: *International Conference on Learning Representations*. 2022.

[24] Ron Mokady, Amir Hertz, and Amit H. Bermano. *ClipCap: CLIP Prefix for Image Captioning*. 2021. arXiv: 2111.09734 [cs.CV].

[25] Thi Ngoc Tho Nguyen, Karn N. Watcharasupat, Ngoc Khanh Nguyen, Douglas L. Jones, and Woon-Seng Gan. "SALSA: Spatial Cue-Augmented Log-Spectrogram Features for Polyphonic Sound Event Localization and Detection". In: *Transactions on Audio, Speech, and Language Processing* 30 (2022), pp. 1749–1762.

[26] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. *Representation Learning with Contrastive Predictive Coding*. 2018. arXiv: 1807.03748 [cs.LG].

[27] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. "Learning Transferable Visual Models from Natural Language Supervision". In: *International Conference on Machine Learning*. PMLR. 2021, pp. 8748–8763.

[28] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. *Language Models are Unsupervised Multitask Learners*. 2019.

[29] Boaz Rafaely. "Analysis and Design of Spherical Microphone Arrays". In: *Transactions on Speech and Audio Processing* 13.1 (2004), pp. 135–143.

[30] Boaz Rafaely. "Fundamentals of Spherical Array Processing". Springer, 2019.

[31] Miguel Sarabia, Elena Menyaylenko, Alessandro Toso, Skyler Seto, Zakaria Aldeneh, Shadi Pirhosseinloo, Luca Zappella, Barry-John Theobald, Nicholas Apostoloff, and Jonathan Sheaffer. "Spatial LibriSpeech: An Augmented Dataset for Spatial Audio Learning". In: *Interspeech*. ISCA, 2023, pp. 3724–3728.

[32] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. "LAION-400M: Open Dataset of CLIP-Filtered 400 Million Image-Text Pairs". In: *Data Centric AI NeurIPS Workshop*. 2021.

[33] Christopher Schymura, Benedikt Bönninghoff, Tsubasa Ochiai, Marc Delcroix, Keisuke Kinoshita, Tomohiro Nakatani, Shoko Araki, and Dorothea Kolossa. "PILOT: Introducing Transformers for Probabilistic Sound Event Localization". In: *Interspeech*. ISCA, 2021.

[34] Rico Sennrich, Barry Haddow, and Alexandra Birch. "Neural Machine Translation of Rare Words with Subword Units". In: *Annual Meeting of the Association for Computational Linguistics*. 2016, pp. 1715–1725.

[35] Sheng Shen, Liunian Harold Li, Hao Tan, Mohit Bansal, Anna Rohrbach, Kai-Wei Chang, Zhewei Yao, and Kurt Keutzer. "How Much Can CLIP Benefit Vision-and-Language Tasks?" In: *International Conference on Learning Representations*. 2022.

12

[36] Kazuki Shimada, Archontis Politis, Parthasaarathy Sudarsanam, Daniel A Krause, Kengo Uchida, Sharath Adavanne, Aapo Hakala, Yuichiro Koyama, Naoya Takahashi, Shusuke Takahashi, et al. "STARSS23: An Audio-visual Dataset of Spatial Recordings of Real Scenes with Spatiotemporal Annotations of Sound Events". In: *Advances in Neural Information Processing Systems Datasets and Benchmarks Track*. Vol. 36. 2024, pp. 72931–72957.

[37] Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, and Chao Zhang. "SALMONN: Towards Generic Hearing Abilities for Large Language Models". In: *International Conference on Learning Representations*. 2024.

[38] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. *LLaMA: Open and Efficient Foundation Language Models*. 2023. arXiv: 2302.13971 [cs.CL].

[39] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. "Attention Is All You Need". In: *Advances in Neural Information Processing Systems*. 2017, pp. 6000–6010.

[40] Qing Wang et al. *The NERC-SLIP System for Sound Event Localization and Detection of DCASE2023 Challenge*. Tech. rep. DCASE2023 Challenge, 2023.

[41] Thomas Wang, Adam Roberts, Daniel Hesslow, Teven Le Scao, Hyung Won Chung, Iz Beltagy, Julien Launay, and Colin Raffel. "What Language Model Architecture and Pretraining Objective Works Best for Zero-shot Generalization?" In: *International Conference on Machine Learning*. PMLR. 2022, pp. 22964–22984.

[42] Yuancheng Wang, Zeqian Ju, Xu Tan, Lei He, Zhizheng Wu, Jiang Bian, and Sheng Zhao. "AUDIT: Audio Editing by Following Instructions with Latent Diffusion Models". In: *Advances in Neural Information Processing Systems* 36 (2023), pp. 71340–71357.

[43] Earl G Williams. "Fourier Acoustics: Sound Radiation and Nearfield Acoustical Holography". Academic Press, 1999.

[44] Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. "Large-scale Contrastive Language-audio Pretraining with Feature Fusion and Keyword-to-Caption Augmentation". In: *International Conference on Acoustics, Speech and Signal Processing*. IEEE. 2023, pp. 1–5.

[45] Dongchao Yang, Jianwei Yu, Helin Wang, Wen Wang, Chao Weng, Yuexian Zou, and Dong Yu. "Diffsound: Discrete Diffusion Model for Text-to-sound Generation". In: *Transactions on Audio, Speech, and Language Processing* 31 (2023), pp. 1720–1733.

[46] Guy Yariv, Itai Gat, Lior Wolf, Yossi Adi, and Idan Schwartz. *AudioToken: Adaptation of Text-Conditioned Diffusion Models for Audio-to-Image Generation*. 2023. arXiv: 2305.13050 [cs.SD].

[47] Yang You, Jing Li, Sashank Reddi, Jonathan Hseu, Sanjiv Kumar, Srinadh Bhojanapalli, Xiaodan Song, James Demmel, Kurt Keutzer, and Cho-Jui Hsieh. "Large Batch Optimization for Deep Learning: Training BERT in 76 minutes". In: *International Conference on Learning Representations*. 2020.

[48] Zhisheng Zheng, Puyuan Peng, Ziyang Ma, Xie Chen, Eunsol Choi, and David Harwath. "BAT: Learning to Reason about Spatial Sounds with Large Language Models". In: *International Conference on Machine Learning*. PMLR. 2024, pp. 61454–61469.

[49] Zelin Zhou, Zhiling Zhang, Xuenan Xu, Zeyu Xie, Mengyue Wu, and Kenny Q. Zhu. "Can Audio Captions Be Evaluated With Image Caption Metrics?" In: *International Conference on Acoustics, Speech and Signal Processing*. IEEE. 2022, pp. 981–985.

[50] Franz Zotter and Matthias Frank. "Ambisonics: A Practical 3D Audio Theory for Recording, Studio Production, Sound Reinforcement, and Virtual Reality". Springer, 2019.

13

A Appendix

A.1 Dataset statistics

Table A.T.1 presents a summary of all the paired audio and text datasets we use for training and evaluation.

Table A.T.1: Audio-caption dataset descriptions. The first three rows correspond to the original publicly available datasets, and the subsequent rows correspond to our spatially-augmented variants. For each spatially augmented dataset, there are at least two spatial augmentations per original sample in the train split.

DATASET	SPATIAL AUDIO	SPLITS	NUM. SAMPLES	DURATION (HRS)	CAPTION DESCRIPTION
Clotho	✗	train, val, test	3,839	23.99	5 captions per audio
AudioCaps	✗	train, val, test	49,274	136.87	1–2 captions per audio
FreeSound	✗	train, val, test	414127	2,528.15	1–2 captions per audio, keyword tags
Spatial-Clotho	Synthetic	train, val, test	8,546	55.0	5 spatially augmented captions per audio
Spatial-AudioCaps	Synthetic	train, val, test	98,459	258.12	1–2 spatially augmented captions per audio
Spatial-FreeSound	Synthetic	train, val, test	783,033	4,425.53	1–2 spatially augmented captions per audio
Spatial-RWD	Recorded	test	70	0.25	1–2 human annotated spatial captions per audio

For the spatially-augmented versions of Clotho, AudioCaps, and Freesound, we use 8,972 parametric rooms with the statistic described in Table A.T.2. Note that the parametric rooms in the test set are a subset of the rooms in the training set, however the sources locations on those rooms do not overlap.

Table A.T.2: Spatial attributes of room simulations used to spatially-augmented audio and language datasets

	TRAIN & VALIDATION	TEST
Number of simulated rooms	8,952	4,970
Source azimuth	$[-180.0°, +180.0°]$	$[-180.0°, +180.0°]$
Source elevation	$[-47.5°, +48.7°]$	$[-29.8°, +42.4°]$
Source distance	$[0.5m, 4.0m]$	$[0.9m, 4.0m]$
Room floor area	$[13.3m^2, 277.4m^2]$	$[14.3m^2, 277.4m^2]$
Full-band T30	$[144.5ms, 2671.9ms]$	$[167.8ms, 1254.8ms]$

A.2 Mapping of spatial attributes to natural language

As part of the spatial-augmentation pipeline (described in Section 3.2), we use the mappings in Table A.T.3 to convert spatial attributes to natural language.

Table A.T.3: Mapping between spatial features and natural language descriptors

SPATIAL FEATURE	RANGE	BOUNDS	LANGUAGE DESCRIPTOR
Distance	$[0m, 5m]$	< 1 m > 2 m	*near* *far*
Direction	$[-180°, +180°]$	$[-55°, -125°]$ $[+55°, +125°]$ $[-35°, +35°]$ $[-145°, +45°]$	*left* *right* *front* *back*
Elevation	$[-48.1°, +48.7°]$	$> 40.0°$ $< -40.0°$	*up* *down*
Reverberation	$[144.5ms, 2671.9ms]$	> 1000ms < 200ms	*highly reverberant* *acoustically dampened*
Room Floor size	$[13.3m^2, 277.4m^2]$	$< 50m^2$ $> 100m^2$	*small* *large*

14

A.3   Audio dataset captions

We report the mapping from raw spatial values to spatial captions in Appendix A.2. We select these bounds based on how an audio would be percieved by human ears, e.g. a sound higher than 40 degrees elevation sounds like its coming from a height.

We also present a few random samples of regular to spatial text rewrites by the LLM:

Inputs to LLM

Original caption `A bird is loudly making a lot of noises.`

Distance `far`

Room Size `medium`

Rewritten caption

`In a medium-sized room, a bird is emitting loud sounds from a distant location.`

Inputs to LLM

Original caption `A bunch of people are talking in the background while a man talks to another.`

Room size `medium`

Rewritten caption

`In a medium-sized room, the sound of multiple people conversing in the background can be heard, alongside a man speaking to another individual.`

Inputs to LLM

Original caption `A motor vehicle is running with speed and stopped its engine.`

Distance `far`

Reverb `highly reverberant`

Rewritten caption

`The sound of a motor vehicle running at high speed and then abruptly stopping its engine is emanating from the far end of a highly reverberant room.`

Inputs to LLM

Original caption `A fire crackles in the background while a faint knocking fades away over time.`

Direction `front`

Distance `far`

Room size `small`

Rewritten caption

`The sound of a fire crackling and a faint knocking can be heard emanating from the far end of a small room.`

Inputs to LLM

Original caption `A distant voice and chirping while the wind blows strongly.`

Direction `right`

Distance `near`

Room size `large`

15

Rewritten caption

```
In a large room, a sound of a distant voice and chirping is emanating from the
right side, and it is accompanied by the strong wind blowing nearby.
```


A.4 Hallucinations in LLMs

We note that the caption re-writes can lead to hallucinations in captions. For instance, the phrase ''the purr of a carerra'' describing the sound of the engine of a Porsche Carerra was rephrased as ''the purring cat named Carerra''. Another example, ''the sounds of papers turning'' with direction *below* was rewritten as ''the sound of someone shuffling cards in the basement''. It is not clear what is the overall effect of the hallucinations; as in the first example it changes the semantics of the audio, but in the second example adds plausible and welcome diversity to our caption set. As mentioned in the main section, we leave the quantification and mitigation of sub optimal hallucinations for future work.

A.5 Further background on first-order ambisonics

Consider a continuum of plane-waves impinging on the surface of a sphere, $p(kr, \Omega)$, where p is acoustic pressure, $k = 2\pi f c^{-1}$ is the spatial frequency, r is radial distance and $\Omega \equiv (\theta, \phi) \in \mathbb{S}^2$ is direction in terms of elevation θ and azimuth ϕ. The expansion of this function in a spherical harmonics basis, $p_{nm}(k)$ can be written as,

$$\begin{aligned} p_{nm}(k, r) &= b_n(kr) \int_{\Omega \in S^2} a(k, \Omega)[Y_n^m(\Omega)]^* \mathrm{d}\Omega \\ &= b_n(kr) A_{nm}(k), \end{aligned} \tag{6}$$

Here, $a(k, \Omega)$ denotes the plane-wave density function in the spatial domain, $Y_n^m(\Omega)$ are the spherical-harmonics basis function for order n and mode m, and $b_n(kr)$ is the radial function given for a rigid sphere by [43],

$$b_n(kr) = 4\pi i^n \left[j_n(kr) - \frac{j_n'(kr_0)}{h_n'(kr_0)} h_n(kr) \right] \tag{7}$$

where $j_n(kr)$ and $h_n(kr)$ are the spherical Bessel and Hankel function, respectively, and $(\cdot)'$ denotes their first derivative with respect to the argument. In this work we employ a real-valued spherical harmonics basis and radial functions corresponding to a rigid sphere. We denote $A_{nm}(k)$ as the spherical Fourier transform of the plane-wave density function, $a(k, \Omega)$, and refer to it as an ambisonics signal. We further denote the inverse spherical Fourier transform of the ambisonics signal as,

$$a(k, \Omega) = \sum_{n=0}^{\infty} \sum_{m=-n}^{n} A_{nm}(k) Y_n^m(\Omega), \tag{8}$$

Considering now a microphone array with Q sensors, the integral in (6) becomes a weighted finite summation. In order to avoid spatial aliasing, we conform to $Q = (N + 1)^2$ with optimal spatial sampling of Q [29] sensors, where N is the spherical harmonics order. Accordingly the outer sum in (8) is truncated at order N, and the transformation between the pressure $p(kr, \Omega)$ and ambisonics function $A_{nm}(k)$ is approximated by,

$$p(kr, \Omega) \approx \sum_{n=0}^{N} b_n(kr) \sum_{m=-n}^{n} A_{nm}(k) Y_n^m(\Omega), \tag{9}$$

rewriting (9) in matrix form and solving for A_{nm}, the linear encoding the microphone signals into ambisonics becomes,

$$\mathbf{a}_{nm} = \mathbf{Y}^H \mathrm{diag}(\mathbf{b}_n)^{-1} \mathbf{p} \tag{10}$$

where $\mathbf{Y}$ is a $Q \times (N + 1)^2$ matrix of spherical harmonics,

$$\mathbf{Y} = \begin{bmatrix} Y_0^0(\Omega_1) & Y_1^{-1}(\Omega_1) & \cdots & Y_N^N(\Omega_1) \\ Y_0^0(\Omega_2) & Y_1^{-1}(\Omega_2) & \cdots & Y_N^N(\Omega_2) \\ \vdots & \vdots & \ddots & \vdots \\ Y_0^0(\Omega_Q) & Y_1^{-1}(\Omega_Q) & \cdots & Y_N^N(\Omega_Q) \end{bmatrix} \tag{11}$$

16

$\mathbf{b}_n$ is the radial function vector, and dependency on the spatial frequency k is omitted for brevity. As in this paper we consider signals to be outputs of a short-time Fourier transform, we further denote our ambisonics features as $\mathbf{A} \in \mathbb{C}^{T \times F \times (N+1)^2}$. More specifically, for $N = 1$ and a given time frame t and frequency bin $f = kc/(2\pi)$ our features are contained in the following ambisonics channels:

$$\boldsymbol{A}(t,f) = \begin{bmatrix} A_{0,0}(t,f) \\ A_{1,\text{-}1}(t,f) \\ A_{1,0}(t,f) \\ A_{1,1}(t,f) \end{bmatrix}, \tag{12}$$

these correspond to the W (omnidirectional) and Y, Z, X (dipole) components of the first order ambisonics approximation. It is worthwhile noting that once microphone array signals have been encoded into ambisonics, no a-priori knowledge on the structure of the capturing array is needed in order to perform any downstream spatial processing. Thus, ambisonics are effectively agnostic to both recording and playback devices, making any embeddings derived from them equally generalizable.

A.6 Ablations across model architectures

We ablate over using static intensity vectors in place of the learned encoder, just the learned encoder without spatial regressors, and compare them with our current architecture. Results are shown in Table A.T.4.

Table A.T.4: Comparison of semantic and spatial retrieval performance across data input ablations. mAP@10 refers to the text-to-audio mean average precision @ 10 and audio-to-text mean average precision @ 10. It is a summary metric capturing the model's semantic retrieval capabilities.

	SPATIAL (AUDIOCAPS + CLOTHO)		
MODEL	3D LOCAL. (°)	DIST. (cm)	MAP@10
Static Intensity Vectors	27.86	60.2	23.43
Pretrained Spatial Branch & only CLIP Loss	27.4	54.31	24.93
Pretrained Spatial Branch & Spatial Losses	23.2	47.71	24.81

Table A.T.5: The retrieval metrics here demonstrate why mixing both mono and spatial audio is required to achieve the best possible retrieval performance. We see that the model needs to see both the regular and spatially augmented versions of the audio and captions to achieve best performance.

| | TRAIN DATA | | AUDIOCAPS | | | | | | CLOTHO | | | | | |
| | | | TEXT-TO-AUDIO | | | AUDIO-TO-TEXT | | | TEXT-TO-AUDIO | | | AUDIO-TO-TEXT | | |
MODEL	AUDIO	CAPTIONS	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
CLAP	Non-spatial	Non-spatial	32.7	68.8	81.5	40.7	74.0	84.7	14.4	37.6	50.7	18.3	40.5	55.1
ELSA	Spatial	Non-spatial	27.1	62.7	76.1	36.6	68.7	78.4	11.3	32.6	44.4	12.4	28.3	50.0
ELSA	Spatial	Spatial	25.3	59.3	72.5	34.8	64.5	75.2	9.9	31.0	39.8	12.1	35.3	47.3
ELSA	Mixed	Mixed	33.2	68.2	81.0	40.9	74.4	86.1	15.0	36.7	50.8	20.1	43.2	55.4

A.7 ELSA architecture

The full architecture of the ELSA model is shown in Fig. A.F.1.

A.8 Spatial attributes branch of audio encoder details

The full architecture for the spatial attributes branch of the audio encoder is shown below in Fig. A.F.2.

The spatial attributes branch has 485,828 parameters, and was pre-trained with a learning rate of 10^{-3} on the LAMB optimizer [47] with weight decay factor of 0.01 and without scheduling the learning rate. The batch size was 1024 and the model was trained for 100 epochs on a single node with 8 NVIDIA V100 GPUs and 80CPUs. Training took 12h 20min. The training set was composed of 134,712 10-second segments from the first-order ambisonics samples of the Spatial LibriSpeech train set. We used a multi-task regression loss, $\mathcal{L}_{\text{pre}}$, to jointly learn the azimuth, elevation, distance, room volume, and 20-third octave bins (between 100Hz and 8kHz) for direct-to-reverberant ratio and T30.

17

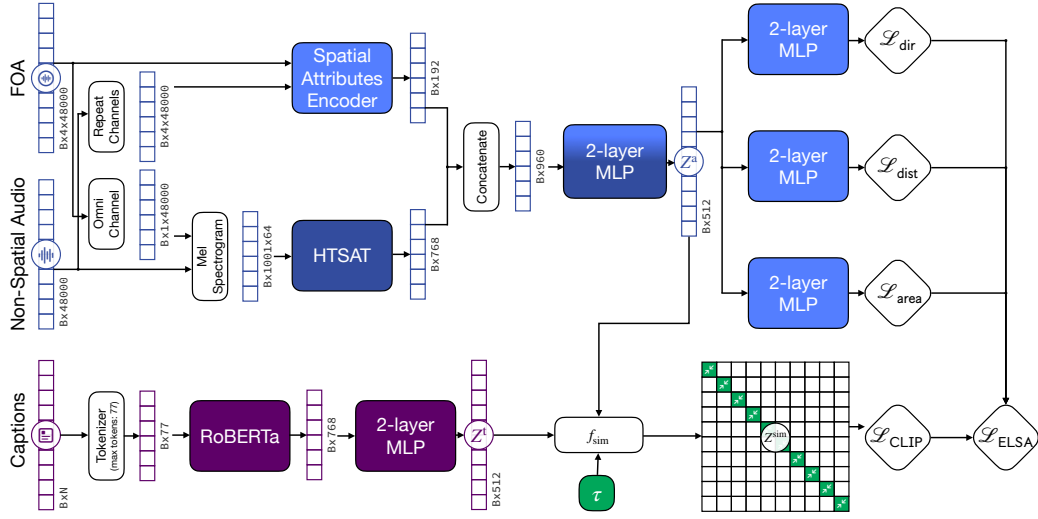



Figure A.F.1: Full architecture diagram for ELSA. Filled blocks include trainable parameters.

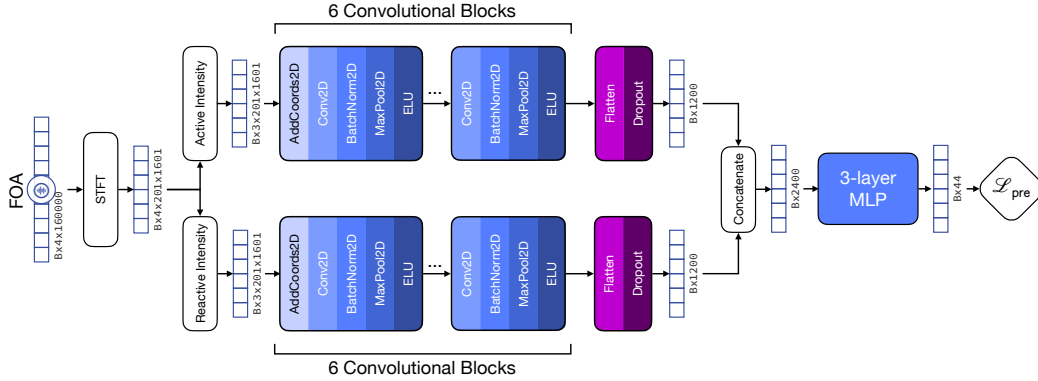



Figure A.F.2: Architecture diagram for Spatial Attributes Branch. Filled blocks include trainable parameters. The AddCoords2D block is described in [20].

$\mathcal{L}_{\text{pre}}$ is the sum of the cosine loss for azimuth and elevation and the mean squared error for all other predictions.

A.9 Fine-grained of direction-of-arrival error analysis

We analyze the errors of a two-layer MLP trained to regress the direction-of-arrival (same setting as the last column in Table 1). We observe how the errors vary along the following dimensions: source azimuth, source elevation, source distance, room floor area, room mean T30, and TUT Sound Events 2018 semantic classes. Results are rendered as boxplots in Fig. A.F.3 below.

Table A.T.6 further report mean, standard deviation, and number of samples per bin about the direction-of-arrival errors across the previously analysed dimensions.

A.10 Spatial attributes retrieval for LAION-CLAP

Section 5.2 showed that ELSA embeddings can be classified in a zero-shot fashion by using a templated probe caption. Table A.T.7 shows the same experiment applied to LAION-CLAP (with the ELSA results retained for context) by feeding LAION-CLAP the omni-channel of the spatial datasets. As expected the performance of LAION-CLAP in this setting is close to random, for instance LAION-CLAP achieves 48% accuracy on the two class distance classification task and 28.2% on the four-class direction classification task.

18

(a) Azimuth (rad)

(b) Elevation (rad)

(c) Distance (m)

(d) Floor area (m^2)

(e) Mean T30 (ms)

(f) TUT Sound Events 2018 Semantic classes

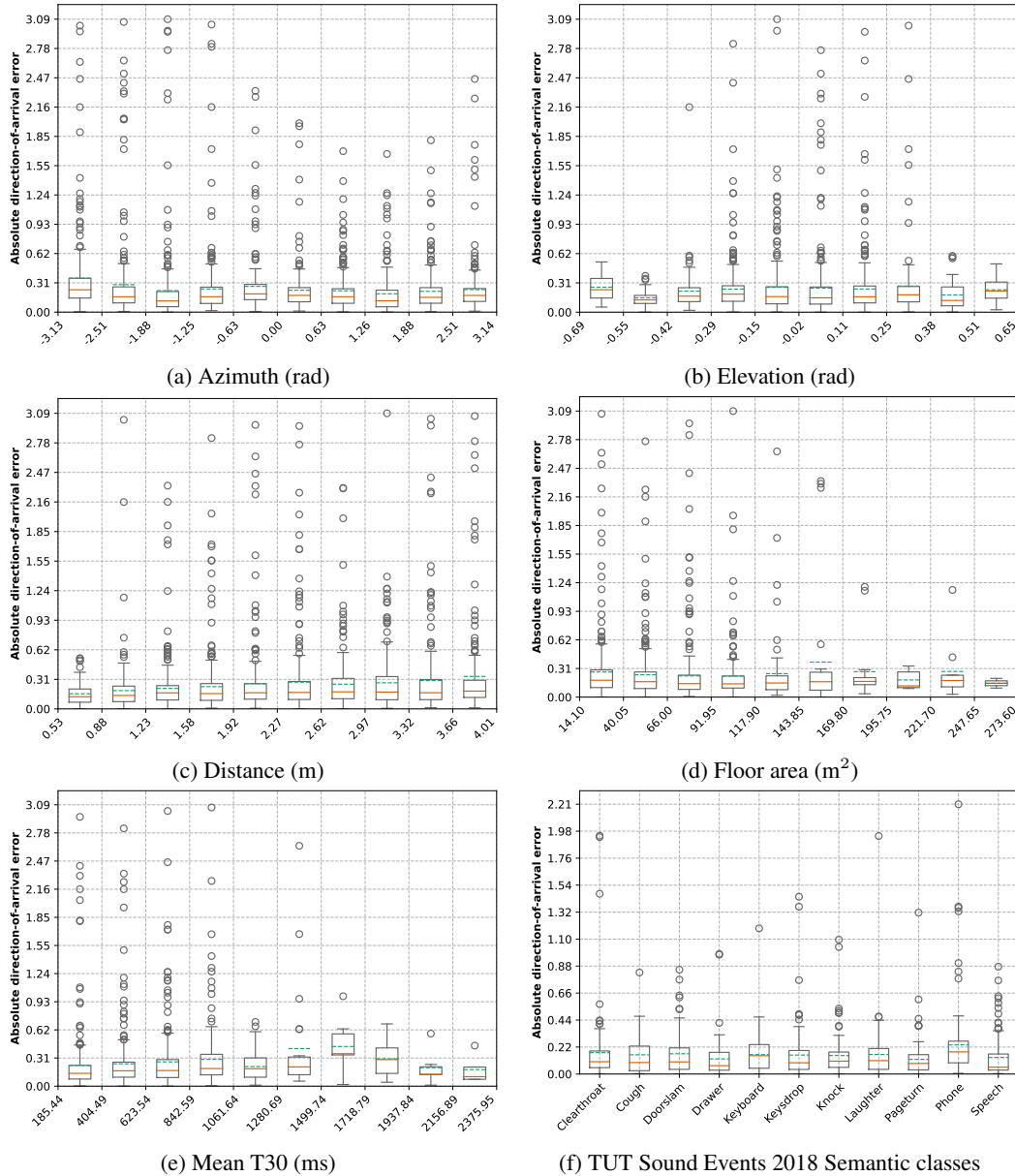



Figure A.F.3: Boxplots of absolute direction-of arrival errors predicted by 2-layer MLP. Figs. (a)–(e) show the Spatial Audiocaps and Spatial Clotho test sets errors by different categories. Fig. (f) shows the predictions of the test set of TUT Sounds 2018 by different semantic classes. For all figures, boxes represent the interquartile range, solid orange lines are the median, and dashed green lines are the mean.

19

Table A.T.6: Mean and standard deviation of absolute direction-of arrival errors (in radians) predicted by 2-layer MLP. Tables (a)–(e) show the Spatial Audiocaps and Spatial Clotho test sets errors by different dimensions. Table (f) shows the predictions of the test set of TUT Sounds 2018 by different semantic classes.

(a) DOA error by azimuth

AZIMUTH RANGES (rad)	DOA ERR. MEAN	DOA ERR. STD. DEV	# BIN SAMPLES
$[-3.13, -2.51)$	0.36	0.44	272
$[-2.51, -1.88)$	0.29	0.46	251
$[-1.88, -1.25)$	0.23	0.43	258
$[-1.25, -0.63)$	0.24	0.37	245
$[-0.63, +0.00)$	0.27	0.31	215
$[+0.00, +0.63)$	0.23	0.25	224
$[+0.63, +1.26)$	0.23	0.24	270
$[+1.26, +1.88)$	0.19	0.23	220
$[+1.88, +2.51)$	0.22	0.22	259
$[+2.51, +3.14)$	0.24	0.29	244

(b) DOA error by elevation

ELEVATION RANGES (rad)	DOA ERR. MEAN	DOA ERR. STD. DEV	# BIN SAMPLES
$[-0.69, -0.55)$	0.26	0.15	12
$[-0.55, -0.42)$	0.15	0.11	20
$[-0.42, -0.29)$	0.22	0.25	88
$[-0.29, -0.15)$	0.24	0.28	312
$[-0.15, -0.02)$	0.26	0.35	660
$[-0.02, +0.11)$	0.25	0.38	735
$[+0.11, +0.25)$	0.24	0.33	456
$[+0.25, +0.38)$	0.28	0.40	126
$[+0.38, +0.51)$	0.18	0.16	35
$[+0.51, +0.65)$	0.24	0.13	14

(c) DOA error by distance

DISTANCE RANGES (m)	DOA ERR. MEAN	DOA ERR. STD. DEV	# BIN SAMPLES
$[0.53, 0.88)$	0.16	0.12	100
$[0.88, 1.23)$	0.19	0.26	252
$[1.23, 1.58)$	0.21	0.26	340
$[1.58, 1.92)$	0.23	0.30	322
$[1.92, 2.27)$	0.26	0.39	314
$[2.27, 2.62)$	0.28	0.38	287
$[2.62, 2.97)$	0.26	0.29	268
$[2.97, 3.32)$	0.27	0.33	208
$[3.32, 3.66)$	0.30	0.45	203
$[3.66, 4.01)$	0.34	0.51	164

(d) DOA error by room floor area

FLOOR AREA RANGES (m^2)	DOA ERR. MEAN	DOA ERR. STD. DEV	# BIN SAMPLES
$[14.10, 40.05)$	0.27	0.36	820
$[40.05, 66.00)$	0.24	0.31	983
$[66.00, 91.95)$	0.23	0.33	346
$[91.95, 117.90)$	0.23	0.33	173
$[117.90, 143.85)$	0.25	0.40	69
$[143.85, 169.80)$	0.38	0.65	30
$[169.80, 195.75)$	0.28	0.32	18
$[195.75, 221.70)$	0.19	0.10	7
$[221.70, 247.65)$	0.28	0.31	10
$[247.65, 273.60)$	0.15	0.05	2

(e) DOA error by T30

T30 RANGES (ms)	DOA ERR. MEAN	DOA ERR. STD. DEV	# BIN SAMPLES
$[185.44, 404.49)$	0.23	0.35	585
$[404.49, 623.54)$	0.25	0.34	1030
$[623.54, 842.59)$	0.27	0.35	521
$[842.59, 1061.64)$	0.29	0.35	191
$[1061.64, 1280.69)$	0.22	0.15	78
$[1280.69, 1499.74)$	0.41	0.57	25
$[1499.74, 1718.79)$	0.44	0.26	9
$[1718.79, 1937.84)$	0.30	0.19	9
$[1937.84, 2156.89)$	0.20	0.18	6
$[2156.89, 2375.95)$	0.18	0.15	4

(f) DOA error by TUT Sound Events 2018 semantic class

SEMANTIC CLASS	DOA ERR. MEAN	DOA ERR. STD. DEV	# BIN SAMPLES
Drawer	0.12	0.16	97
Laughter	0.16	0.22	95
Cough	0.16	0.16	87
Clearthroat	0.17	0.29	115
Keyboard	0.16	0.15	97
Speech	0.14	0.17	105
Phone	0.24	0.30	117
Pageturn	0.12	0.15	115
Knock	0.15	0.17	112
Doorslam	0.17	0.17	101
Keysdrop	0.15	0.21	111

Table A.T.7: Complete version of Table 2 with LAION-CLAP results. LAION-CLAP results are obtained by passing omni channel of spatial dataset through pre-trained model.

TASK	ELSA			LAION-CLAP		
	S-CLOTHO	S-AC	S-RWD	S-CLOTHO	S-AC	S-RWD
Distance (2-class)	96.0%	92.9%	67.1%	48.0%	54.3%	53.0%
Direction (4-class)	92.0%	92.8%	35.8%	28.2%	29.3%	27.3%
Elevation (2-class)	100.0%	100.0%	72.1%	56.7%	51.3%	59.4%
Room area (2-class)	76.6%	74.7%	N/A	46.3%	66.5%	N/A
Reverberation (2-class)	100.0%	83.3%	N/A	57.3%	52.5%	N/A

20

A.11 Text corpus used when testing ELSA's implicitly learned spatial attributes

As mentioned in main text, we validate that ELSA's audio embeddings can capture implicit spatial attributes, which are latent in the text-encoder. We use LLaMA-13B [38] to generate descriptions of 50 sounds that typically come from each of above and below. The sentences do not necessarily include explicit spatial descriptions (e.g., "A fire alarm going off", a sound typically from above but not explicitly stated). The original prompt and generated sentences can be found below. We then train a two-class classifier (top vs bottom) using the spatial audio in the train sets of our spatial-augmented datasets, and classify each of the sentences. Results show an above-random classification accuracy of 68.75% for *above* sentences, 58.82% for *below* sentences. This experiment shows that, to a degree, the text-encoder (RoBERTa) is able to leverage its semantic bias of placement of objects in the real world and encode it as spatial features that the spatial encoder understands. This task is made quite hard by the observation that our simulated dataset does not reflect the natural world (an aeroplane sound could be simulated from below) and the fact that `RoBERTa-base` has significantly fewer parameters and a smaller training set more recent large-language models such as LLaMA-13B. This experiment is purely qualitative, and we leave a more in-depth exploration of knowledge sharing between the pretrained spatial and text encoders to future work.

Sentences for sounds from above

- Sound from the upper part of a room.
- The sound of an aeroplane flying.
- The sound of a bird chirping.
- The sound of a helicopter in the sky.
- A low rumble of distant thunder.
- A sharp crack of lightning striking.
- The pitter-patter of rain on a roof.
- The rhythmic whoosh of wind through trees.
- A distant siren wailing high in the air.
- The muffled thud of footsteps on a floor upstairs.
- A child's laughter echoing from an upstairs room.
- A playful meow or bark from a pet overhead.
- The frantic buzzing of a trapped fly.
- The soft hum of a ceiling fan rotating.
- The rhythmic beeping of a smoke alarm.
- The distant hooting of an owl.
- The rustling of leaves as a squirrel scurries across a roof.
- The muffled boom of fireworks exploding high above.
- The melodic ringing of church bells.
- A fighter jet screaming across the heavens.
- The rhythmic thump of a basketball bouncing overhead.
- The scraping of furniture being moved on a floor above.
- The high-pitched whine of a mosquito circling your head.
- The rhythmic drumming of raindrops on a skylight.
- A loud thump followed by a startled yelp (someone tripped upstairs).
- The rhythmic tap-tap-tap of a woodpecker on a tree trunk.
- The frantic buzzing of a swarm of bees overhead.
- The mournful cry of a seagull circling above the beach.
- The rhythmic clatter of hail bouncing off a roof.
- The melodic singing of a bird outside your window.
- The gentle whoosh of a hot air balloon floating overhead.
- The rhythmic thump-thump-thump of helicopter blades.
- Someone walking on the roof.
- The soft cooing of pigeons perched on a building ledge.
- The gentle pitter-patter of rain on a tent roof.
- The high-pitched shriek of a child on a roller coaster.
- The rhythmic chirping of crickets in a field at night.
- The muffled thump of a heavy object being dropped from above.
- The rhythmic whir of a helicopter hovering nearby.
- The faint melody of a lullaby drifting down a stairwell.
- A sudden gust of wind whistling through the trees.
- The rhythmic clicking of a computer mouse from upstairs.
- The melodic ringing of a wind chime swaying in the breeze.

21

- The rhythmic pounding of rain on a metal roof.
- The rhythmic chirping of birds waking you up at dawn.
- The muffled conversation of people walking on a floor above.
- The muffled snoring of someone sleeping upstairs.
- The chirping and squawking of a flock of birds taking flight.
- The rhythmic click-clack of tap shoes dancing on a floor above.
- The rhythmic tapping of a woodpecker searching for insects.

Sentences for sounds from below

- The sound of an underground subway train.
- The sound of feet walking on a pavement.
- The muffled boom of a distant explosion.
- The rhythmic dripping of a leaky pipe in the basement.
- The gurgling of water in a drainpipe.
- The muffled thump of something heavy being dropped downstairs.
- The low hum of a refrigerator running.
- The rhythmic squeak of floorboards underfoot.
- The muffled clinking of glasses from below.
- The faint laughter echoing from downstairs.
- The rhythmic drumming of a clothes dryer in the basement.
- The rhythmic beeping of a malfunctioning appliance in the basement.
- The scratching sound of a pet exploring under furniture.
- The scurrying of mice in the floorboards.
- The rhythmic dripping of a bath faucet you forgot to turn off completely.
- The muffled roar of a furnace kicking on.
- The rhythmic tick-tock of a grandfather clock.
- The faint vibration of a subwoofer from a downstairs stereo.
- The muffled chatter of people in a room below.
- The rhythmic click-clack of a keyboard from below.
- The muffled thump of a door closing downstairs.
- The rhythmic pinging of a pinball machine in a basement arcade.
- The dripping sound of melting ice from a refrigerator.
- The rhythmic whoosh of a basement exhaust fan.
- The faint hum of electrical wiring in the walls.
- The rhythmic drumming of rain on a basement window.
- The low rumble of distant traffic filtering through the floorboards.
- The faint creaking of an old house settling.
- The rhythmic thump of a bouncing ball from downstairs.
- The muffled clinking of coins in a jar breaking on the floor.
- The rhythmic whoosh of a vacuum cleaner downstairs.
- The faint strains of music seeping up from a basement party.
- The rhythmic beeping of a smoke alarm in the basement (hopefully a false alarm).
- The muffled shouts of children playing downstairs.
- The rhythmic tapping of a sewer pipe being repaired.
- The rhythmic dripping of condensation on a cold water pipe.
- The rhythmic click of a deadbolt lock being secured downstairs.
- The muffled roar of a lawnmower outside.
- The rhythmic scratching of a pet trying to dig a hole in the carpet.
- The muffled clinking of silverware being dropped in the sink.
- The rhythmic thump of a basketball bouncing on the floor below.
- The rhythmic ping-pong of a table tennis match in progress downstairs.
- The rhythmic whirring of a washing machine agitating clothes downstairs.
- The muffled snoring of someone sleeping downstairs.
- The rhythmic drumming of rain on a basement windowpane at night.
- The faint, high-pitched whine of a mosquito buzzing around your ankles.
- The muffled clatter of pots and pans being moved around in the kitchen.
- The rhythmic whoosh of a sprinkler system watering the lawn.

22

A.12 Semantic retrieval for spatial data

We report ELSA's semantic retrieval on Spatial-Clotho, and Spatial-Audiocaps in Table A.T.8. Likewise, Table A.T.9 shows the retrieval scores of ELSA on our spatial real-world dataset.

Table A.T.8: Semantic Retrieval Metrics calculated over spatially augmented version of Clotho and AudioCaps eval sets – identical in size as the non-spatial sets.

| | | SPATIAL-AUDIOCAPS | | | | | | SPATIAL-CLOTHO | | | | | |
| | | TEXT-TO-AUDIO | | | AUDIO-TO-TEXT | | | TEXT-TO-AUDIO | | | AUDIO-TO-TEXT | | |
MODEL	TRAIN DATA	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
ELSA	Sp(Clotho + AC)	25.3	58.9	73.5	32.6	61.5	73.0	9.4	25.9	38.0	10.5	27.8	40.2
ELSA	Sp(Clotho + AC + FS)	24.2	56.7	71.82	30.5	59.42	72.8	11.26	30.72	43.07	12.6	32.0	44.08

Table A.T.9: Semantic Retrieval Metrics calculated over our spatial real-world dataset.

| | | SPATIAL-RWD | | | | | |
| | | TEXT-TO-AUDIO | | | AUDIO-TO-TEXT | | |
MODEL	TRAIN DATA	R@1	R@5	R@10	R@1	R@5	R@10
ELSA	Sp(Clotho + AC)	18.6	54.3	75.7	25.7	55.7	71.4
ELSA	Sp(Clotho + AC + FS)	41.4	68.6	88.6	35.7	67.14	80.0

A.13 Further analysis on embedding clusters

In Section 5.4 we analysed the UMAP projection of the ELSA embeddings of the test set of Spatial AudioCaps and Spatial Clotho. Table A.T.10 (a) shows the Wasserstein distances computed directly in the 512-dimensional space, where we see the data clusters by direction with lower Wasserstein distances between "front" and "back". Similarly, Fig. A.F.4 and Table A.T.10 (b) show the ELSA embeddings can be clustered according to spatial distance characteristics.

Table A.T.10: Wasserstein distances of 512-dimensional ELSA embeddings, clustered by either (a) direction or (b) distance.

(a) Direction clustering distances

	LEFT	RIGHT	FRONT	BACK
LEFT	0.00	1.04	0.94	0.98
RIGHT	1.04	0.00	0.92	0.97
FRONT	0.94	0.92	0.00	0.81
BACK	0.98	0.97	0.81	0.00

(b) Distance clustering distances

	NEAR	FAR
NEAR	0.00	1.10
FAR	1.10	0.00

23

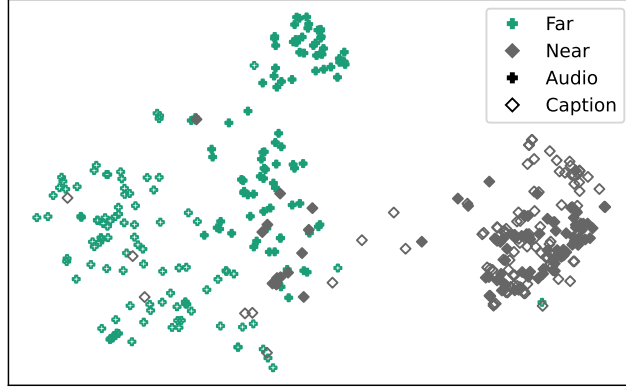



Figure A.F.4: UMAP projection of ELSA embeddings of the test splits of Spatial-Clotho and Spatial-AudioCaps. Filled markers are obtained from spatial audio, and hollow markers are obtained from spatial captions. The UMAP projection was fitted with the train splits of Spatial-Clotho and Spatial-Audio caps, and we made use of supervised dimension reduction to highlight the *distance* differences rather than the semantic differences in the embeddings.

A.14 Swapping of Spatial Direction Experiments

As already introduced in Section 5.4. Our spatial direction swapping pipeline consists of the following steps:

1. We obtain the ELSA embeddings four directions ("left", "right", "front", "back") with the template: "`A sound coming from the direction`". These are our direction prototypes.

2. Train a 4-class direction classifier with a 2-layer MLP (33k parameters) on the training set of Spatial-AudioCaps and Spatial-Clotho. We obtained a 96.7% classification accuracy of the test set of Spatial-AudioCaps and Spatial-Clotho.

3. For every correctly-classified sample in the test sets, we obtain their ELSA embedding, subtract the prototype of the original direction, and add a prototype for the new direction.

Additionally, we measure any changes in sound *semantics* by computing the difference in recall@10 between the ELSA audio embedding and the ELSA embedding of the sample's non-spatial description.

Detailed results are shown in table Table A.T.11. Overall, we find an average 99.7% of the samples are correctly classified with the new direction after transposition, and an average change of -0.2% in recall@10. These results show ELSA directional attributes can be linearly swapped without affecting the semantics of sound.

Table A.T.11: Direction swapping of ELSA embeddings. See Appendix A.14 for a detailed explanation of how we swapped the embedding directions. $\notin$ is the number of test samples misclassified by our direction classifier, and subsequently excluded. N is the number of samples that were used for direction transposition. R@10 is the recall@10 computed over the corresponding non-spatial captions. θ is the classification accuracy of the transposed sample. $\Delta_{R@10}$ is the change in recall@10 after performing the change of direction.

| | | | | | NEW DIRECTION | | | | | | | |
| | | | | LEFT | | FRONT | | RIGHT | | BACK | |
		$\notin$	N	R@10	θ	$\Delta_{R@10}$	θ	$\Delta_{R@10}$	θ	$\Delta_{R@10}$	θ	$\Delta_{R@10}$
ORIGINAL DIR.	LEFT	14	156	94.9%	–	–	100.0%	+1.3%	100.0%	+0.0%	100.0%	-1.3%
	FRONT	13	486	81.1%	100.0%	+1.9%	–	–	100.0%	+0.6%	100.0%	-0.4%
	RIGHT	12	196	92.3%	100.0%	-0.5%	100.0%	+0.5%	–	–	100.0%	-0.5%
	BACK	9	564	81.6%	97.3%	-1.4%	100.0%	-0.5%	100.0%	-0.9%	–	–

24

Furthermore, we wanted to verify what happened if we removed the original direction but did not add back a new direction. Table A.T.12 shows this ablation. Interestingly, the classification does not result in random classification accuracy but rather 0% accuracy for all four original directions.

Table A.T.12: Direction removal of ELSA embeddings. See Appendix A.14 for a detailed explanation of how we swapped the embedding directions. $\notin$ is the number of test samples misclassified by our direction classifier, and subsequently excluded. N is the number of samples that were used for direction transposition. R@10 is the recall@10 computed over the corresponding non-spatial captions. θ is the classification accuracy of the transposed sample. $\Delta_{R@10}$ is the change in recall@10 after performing the change of direction.

| | $\notin$ | N | R@10 | Dir. Removed | |
				θ	$\Delta_{R@10}$
LEFT	14	156	94.9%	0.0%	-0.6%
FRONT	13	486	81.1%	0.0%	-0.6%
RIGHT	12	196	92.3%	0.0%	-5.1%
BACK	9	564	81.6%	0.0%	-2.3%

A.15 Further details on Spatial Audio Caption Generation

Section 5.5 introduced a spatial audio caption generation system. In what follows, we illustrate some of the generations produced by the system as well as the corresponding ground-truth annotation. Below that, in Fig. A.F.5, we include an architecture diagram for the spatial audio caption generation system.

Generated caption

```
In a medium-sized room located at the far back, an electric motor is emitting a
high-pitched whine, accompanied by a whirring noise.  In the background, adult
male voice can be heard speaking.
```

Ground-truth caption from test set

```
From deep within a medium-sized room, the noise of a robust industrial engine
can be heard whirring loudly.
```

Generated caption

```
The sound of water flowing and splashing is emanating from the front of a
room.
```

Ground-truth caption from test set

```
The sound of gentle rowing and paddling in the water is emanating from the
vicinity of a medium-sized room.
```

Generated caption

```
The sound of cheering coming from a crowd is heard near the medium-sized
room.
```

Ground-truth caption from test set

25

The sound of applause, indicating that people are praising the musicians after their performance, is emanating from the medium-sized room.

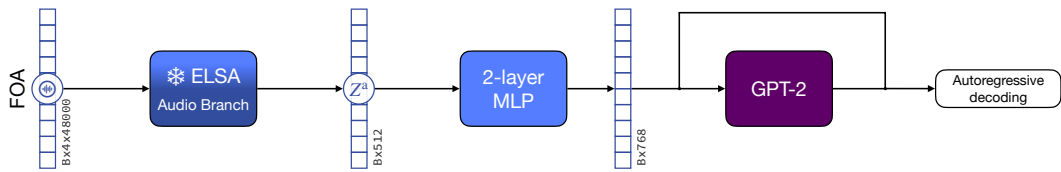



Figure A.F.5: Architecture diagram for spatial audio caption generation.

26

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We propose a new dataset, discuss the dataset in Section 3, and demonstrate results of ELSA trained on this dataset in Section 5.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the limitations of our work in terms of some pitfalls of using synthetically augmented captions in the conclusion — Section 6 of the main paper as well as Section A.9 of the appendix.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

27

Answer: [NA]

Justification: The paper does not contain any novel theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Dataset creation details are included in Section 3 including details of prompts, and dataset creation. Hyperparameter details are included in Section 5.1 and the appendix. Implementation details of our architecture are included in Section 4. Furthermore, we will release our code and models to aid reproducibility.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

28

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Code, datasets, and models will be made publicly available at https://github.com/apple/ml-spatial-audio-elsa.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We highlight training hyperparameters in Section 5.1. Details of the dataset creation are included in Section 3.1, and details of our dataset evaluations are included in Sections 5.2-5.4.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: The original datasets provide only a single partition for training/testing/evaluation splits. We base our experiments on these splits so that results are directly comparable with other state of the art methods.

   Guidelines:

   - The answer NA means that the paper does not include experiments.

29

- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The training dataset took ~2 weeks to generate, and utilized 96 CPUs and 16T of disk space. Each full training run of the model takes roughly 1600 A100 GPUhs. This was tested on multi-node GPU machines with up to 12 nodes. There were a total of 40 estimated runs to completion. For the version of our model only trained on the smaller datasets Clotho and AudioCaps, convergence takes 4 hours otherwise takes 17 hours.

   Guidelines:
   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: For training we use publicly available datasets, for which we augment with only spatial information. The real-world dataset that we captured for our experiments was created without capturing personally identifiable information. Furthermore, the dataset contains sounds of everyday items, e.g., a coffee grinder or paper rustling, and we know of no concerns regarding demographics of end-users.

   Guidelines:
   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

30

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Positive impact is detailed in the introduction(Section 1) and conclusion(Section 6), potential negative impacts are in the broader impacts paragraph (Section 6).

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: Our societal risks only apply for generation of soundscapes, which is out of the scope for the paper.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

31

Justification: All datasets and model architectures are appropriately and explicitly cited. Github links are also included for code used.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: New assets including models and datasets are included in Sections 3 and 4 and plan to be released for reproducibility.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: We do not conduct any experiments involving human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

32

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: We do not conduct any experiments involving human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

33