In-the-wild Audio Spatialization with Flexible Text-guided Localization

Tianrui Pan, Jie Liu*, Zewen Huang, Jie Tang, Gangshan Wu
State Key Laboratory for Novel Software Technology, Nanjing University
a24164839@163.com, liujie@nju.edu.cn,
502023370017@smail.nju.edu.cn, {tangjie,gswu}@nju.edu.cn

Abstract

To enhance immersive experiences, binaural audio offers spatial awareness of sounding objects in AR, VR, and embodied AI applications. While existing audio spatialization methods can generally map any available monaural audio to binaural audio signals, they often lack the flexible and interactive control needed in complex multi-object user-interactive environments. To address this, we propose a Text-guided Audio Spatialization (TAS) framework that utilizes flexible text prompts and evaluates our model from unified generation and comprehension perspectives. Due to the limited availability of premium and large-scale stereo data, we construct the SpatialTAS dataset, which encompasses 376,000 simulated binaural audio samples to facilitate the training of our model. Our model learns binaural differences guided by 3D spatial location and relative position prompts, augmented by flipped-channel audio. It outperforms existing methods on both simulated and real-recorded datasets, demonstrating superior generalization and accuracy. Besides, we develop an assessment model based on Llama-3.1-8B, which evaluates the spatial semantic coherence between our generated binaural audio and text prompts through a spatial reasoning task. Results demonstrate that text prompts provide flexible and interactive control to generate binaural audio with excellent quality and semantic consistency in spatial locations. Dataset is available at https://github.com/Alice01010101/TASU.

1 Introduction

Humans can identify the location of objects by processing auditory differences between their ears, even when they cannot see or are not physically present in the scene. Binaural audio contains spatial information for each sound source, it is essential for applications in Virtual Reality (VR) or Augmented Reality (AR) (Li et al., 2018; Kim et al., 2019b; Xu et al., 2024), and embodied AI (Liu et al., 2024c). The audio spatialization task (Gao and Grauman, 2019; Zhou et al., 2020; Rachavarapu et al., 2021; Parida et al., 2022; Garg et al., 2023; Dagli et al., 2024) continues to be a vibrant area of research. This task involves mapping monaural audio signals to binaural audio signals, allowing users to experience immersive surroundings as if they were physically present in the scenes. Most existing methods are visually guided (Gao and Grauman, 2019; Zhou et al., 2020; Garg et al., 2023), performing mono-to-binaural mapping using visual frames captured by cameras of different Field Of Views (FOV). However, accurate mapping between sound sources in binaural audio and visible objects in frames is impeded by sound sources located outside the camera's view and complex environments with extraneous noise.

To address these challenges, we propose a Text-guided Audio Spatialization (TAS) framework that incorporates flexible text prompts and evaluates our model innovatively from the perspective of unified generation and understanding. To the best of our knowledge, the only relevant study in this area Li et al. (2024b) manually labeled text prompts for the FAIR-Play dataset (Gao and Grauman, 2019) from extracted visual frames, resulting in suboptimal performance due to its simplistic approach and limited dataset scale. To mitigate the lack of corresponding datasets, we propose sampling from a large-scale simulated binaural dataset (Zheng et al., 2024) and refining it with more detailed text descriptions. This results in the SpatialTAS dataset, which contains approximately 376K training samples. Since providing precise azimuth or elevation information is not always feasible in practical scenarios, we generate two primary types of descriptions, as illustrated in Figure 1. The first type categorizes eight spatial directions based on the Cartesian product of spherical coordinates: *(left, right)*,

*Corresponding author (liujie@nju.edu.cn).

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1989–2001

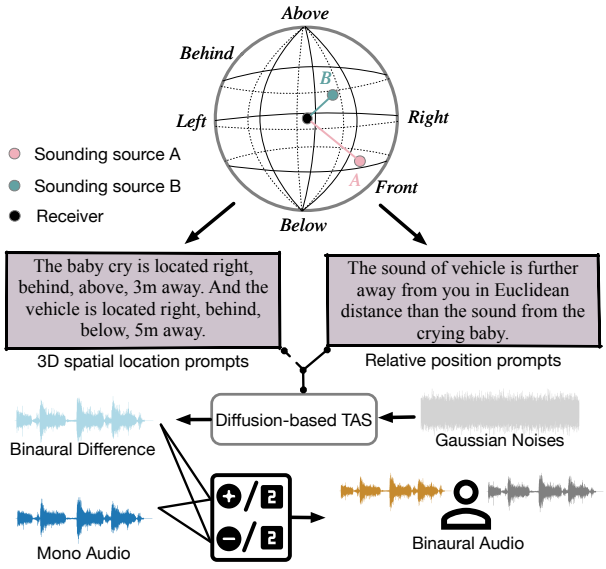

Figure 1: We propose the Text-guided Audio Spatialization (TAS) framework. It utilizes diverse text descriptions to specify the 3D spatial information of multiple sound sources, serving as prompts to transform monaural audio into binaural audio in complex environments.

(front, behind), *(above, below)*, along with their distances to the receiver. In certain real-time interaction scenarios, humans can only make subjective judgments about the relative spatial relationships between two concurrently active sound events. For the second type, we offer descriptions of the relative positions between any two sound sources. These text descriptions enable selective location instructions for specific target objects, thereby enhancing user-friendliness and adaptability to various contexts. This is in contrast to most previous methods (Gao and Grauman, 2019; Zhou et al., 2020; Rachavarapu et al., 2021; Parida et al., 2022; Garg et al., 2023; Dagli et al., 2024) that require guidance for all sound sources within an audio mixture to avoid obvious performance drop. Inspired by PseudoBinaural (Xu et al., 2021), we aim to train our model on the constructed large-scale simulated SpatialTAS dataset, which can transfer freely to in-the-wild monaural audios (Section 5.2).

Recent works (Li et al., 2024a,b) have achieved impressive performance using diffusion models in the audio spatialization task. However, these approaches employ a diffusion model directly in the waveform space, utilizing a cross-attention module to interact with audio and text embeddings. In contrast to this approach, we leverage a latent diffusion model (Rombach et al., 2022) that is directly conditioned on text embeddings to learn the binaural difference between the left and right audio channels, as illustrated in Figure 1. By learning the latent representations of audio signals without modeling the cross-modal relationship, our model improves both generation quality and computational efficiency. Furthermore, recognizing the absence of spatial audio alignment in the pretrained text encoder during training, we introduce a text-audio coherence module. This module employs flipped-channel audio to finetune the encoder, thereby enriching the spatial representation of text embeddings.

While numerous metrics exist for evaluating monaural audio, specific metrics for generated binaural audio remain lacking. In this work, we first establish an assessment model by finetuning Llama-3.1-8B (Dubey et al., 2024) on the SpatialTAS with the spatial audio reasoning task. Then we utilize the assessment model to assess the spatial semantic coherence between our generated audio and text prompts. Experimental results on the SpatialTAS dataset demonstrate that our generated binaural audio not only exhibits high audio quality but also captures distinct and interpretable spatial characteristics for spatial audio understanding. Furthermore, it shows strong generalization ability when tested on the FAIR-Play (Gao and Grauman, 2019) and 360° YouTube-Binaural (Garg et al., 2023) datasets, which consist of real-world binaural recordings, including various audio types such as music, speech, and natural sounds.

2 Related Work

2.1 Audio Spatialization

Some studies utilize video frames for self-supervision to infer the positions of sound-emitting objects (Morgado et al., 2018; Garg et al., 2023; Gao and Grauman, 2019; Zhou et al., 2020). Morgado et al. (2018) introduced two datasets for audio spatialization using 360° videos: REC-STREET and YT-ALL. Garg et al. (2023) enhanced the YT-Clean dataset by converting ambisonic audio to binaural audio with Normal Field-Of-View (NFOV) video clips, creating the YouTube-Binaural dataset, which we use alongside the original 360° videos. Gao and Grauman (2019) proposed the FAIR-Play dataset, focusing on NFOV video and binaural audio with multiple music tracks. Other studies improved alignment between binaural audio and visual features(Garg et al., 2023; Liu et al., 2024b; Li et al., 2024a). Recently, Li et al. (2024b) labeled the FAIR-Play dataset with object location descrip-

tions and suggested guiding audio spatialization with text.

2.2 Binaural Audio Generation

Recently, several text-to-audio generation methods have been proposed (Liu et al., 2023, 2024a; Vyas et al., 2023; Evans et al., 2024a,b; Lee et al., 2023; Yang et al., 2023), with some focusing specifically on text-to-binaural audio generation (Singh Kushwaha et al., 2024; Sun et al., 2024). Singh Kushwaha et al. (2024) utilized text as the sole input, introducing a multi-conditional encoder to unify spatial and semantic information for context-aligned binaural audio generation. Similarly, Sun et al. (2024) proposed the BEWO-1M dataset, demonstrating a novel approach with promising results. Since large-scale monaural datasets are readily available in the real world, we focus on text-guided audio spatialization, leveraging text prompts to provide flexible and interactive control that better aligns with real-world application needs.

3 Method

3.1 Generating Prompts for Training

Our object is to establish a text-guided audio spatialization framework that uses positional text descriptions T_{prompts} to transform a monaural audio A_{mono} into binaural audio, along with a unified evaluation for generation and understanding.

Given the limited scale of most real-recorded binaural audio datasets and the absence of text prompts for sound source locations, we introduce SpatialTAS, a large-scale simulated dataset meticulously crafted by sampling and refining data from the SpatialSoundQA (Zheng et al., 2024) dataset by GPT-4o (Hurst et al., 2024). Notably, SpatialTAS incorporates more fine-grained text prompts tailored for binaural audio generation. As detailed in Table 1, our dataset encompasses approximately 256K samples with text descriptions for Direction Of Arrival (DOA) and Distance Estimation (DE), complemented by an additional 120K samples featuring descriptions of relative relationships between sound sources. The SpatialTAS dataset provides comprehensive 3D spatial location prompts that convey the direction and distance of each sound source, along with versatile relative position prompts that facilitate flexible specification of spatial relationships between any two sound sources. In Table 1, Example **A** and Example **B** exemplify detailed spatial location prompts. Example **A** represents a scenario with a single sounding source, while Example **B** depicts a situation with two sound sources in an audio mixture. Regarding the versatile relative position prompts, Example **C** conveys information about the relative distance between the two sources, whereas Examples **D**, **E**, **F**, and **G** describe their relative spatial locations. The dataset comprises hundreds of diverse audio events carefully selected from 10-second audio clips in AudioSet (Gemmeke et al., 2017). We aim to train a model on this large-scale simulated dataset, enabling seamless transfer to in-the-wild audios.

3.2 Audio Spatialization Framework

During training, we train a diffusion model to learn the channel difference A_{lr} from Gaussian noise, which signifies the distinction between the left and right channels. Given the simulated binaural audio $A_b = (A_l, A_r)$, the monaural audio A_{mono} is obtained by mixing the left and right channels, while the target channel difference audio A_{lr} is obtained by subtracting the channels:

$$A_{\text{mono}} = A_l + A_r, A_{lr} = A_l - A_r, \quad (1)$$

where A_{mono} is utilized as model input during training. We train the latent diffusion model to learn the channel difference A_{lr} from the Gaussian distribution. During inference, we compute generated binaural audio $\hat{A}_b = (\hat{A}_l, \hat{A}_r)$ as follows:

$$\hat{A}_l = \frac{A_{\text{mono}} + A_{lr}}{2}, \hat{A}_r = \frac{A_{\text{mono}} - A_{lr}}{2}. \quad (2)$$

The generated binaural audio $\hat{A}_b$ retains the same spatial positional information for each sound source as in A_b. Furthermore, the model demonstrates strong generalization capabilities for real-world binaural audio generation, encompassing various audio types, including music, speech, and diverse sound effects. We train a latent diffusion model F_θ to learn the binaural difference A_{lr} conditioned on text embeddings T_e and embedded monaural audio A_e, together with a spatial coherence module.

Conditional latent diffusion model. As illustrated in the lower left of Figure 2, we employ a Variational AutoEncoder (VAE) (Kingma, 2013) latent encoder $\text{Enc}(\cdot)$ to compress the mel-spectrogram of A_{lr}, which has a shape of $\mathbb{R}^{T \times F}$, into a compact continuous representation $z \in \mathbb{R}^{\frac{T}{r} \times \frac{F}{r} \times C}$. Here, T and F represent the time length and frequency dimensions, respectively. C denotes the number of

Table 1: **Overview of text condition types.** The SpatialTAS dataset includes about 256,000 samples with 3D spatial location and distance prompts for each sound source, along with approximately 120,000 samples for relative spatial relationships among multiple sound sources. **Sources** indicates the number of sound sources present in each sample.

Text Type	Sources	Example
DOA & DE (256K, 68%)	1	**A:** The emergency vehicle is located right, behind, below, 5m away.
	2	**B:** The music is located left, behind, below, 8.5m away. And the whip is located right, behind, below, 5m away.
Relative Relationships (120K, 32%)	2	**C:** The distance between the sound of the animal and the sound of the spray is 3m away. **D:** The sound from the music on the back is located further away, while the sound from the telephone dialing with DTMF is closer to the front. **E:** The sound from the scratching originates on the left, and the sound from the children playing originates on the right. **F:** The sound from the music is above and the sound from the boat, water vehicle is below. **G:** The sound from speech is further away from you in Euclidean distance than the sound from a mechanical fan.

channels in the latent representation, and r is the downsampling ratio that determines the compression level of the latent space. After the diffusion model, we use the VAE latent decoder $\mathrm{Dec}(\cdot)$ to reconstruct the latent representation z back into the mel-spectrogram format of A_{lr}. Additionally, we incorporate a HiFi-GAN vocoder (Kong et al., 2020a) to convert the mel-spectrogram into a high-quality waveform. Both the latent encoder $\mathrm{Enc}(\cdot)$ and the latent decoder $\mathrm{Dec}(\cdot)$ consist of stacked convolutional modules.

Given the encoded latent representation of the input audio $z_0 = \mathrm{Enc}(A_{lr})$, we apply a forward process during training to obtain the noised representation z_t at each time step t. This is done by injecting noise ϵ according to the equation $z_t = \alpha z_0 + \beta\epsilon$, following the noise schedule (Song et al., 2020). Here, ϵ is random noise drawn from an isotropic Gaussian distribution $\mathcal{N}(0, I)$. We define the training loss $\mathcal{L}_\theta$ as the objective to predict the noise ϵ added to the noisy latent representation, guided by the text embedding T_e and the embedding of the monaural audio $A_e = E_a(A_{\text{mono}})$. This is achieved by minimizing the following loss function:

$$\mathcal{L}_\theta = \mathbb{E}_{\epsilon\in\mathcal{N}(0,I),t}||\epsilon - F_\theta(z_t, t, T_e, A_e)||_2^2. \quad (3)$$

The Classifier-Free Guidance (CFG) (Ho and Salimans, 2022) is crucial for generating audio that semantically matches and temporally aligns with the text instructions, while preserving the model's generative diversity and enhancing its generalization capability. Therefore, during training, we randomly replace the condition pair (T_e, A_e) with a zero tensor with a probability of 0.1. And during sampling, we modify the vector field using the formula as follows:

$$\hat{F}_\theta(z_t, t, T_e, A_e) = \gamma F_\theta(z_t, t, T_e, A_e) + (1-\gamma)F_\theta(z_t, t, \varnothing, \varnothing), \quad (4)$$

where γ is the guidance scale trading off the sample diversity and generation quality, and $\hat{F}_\theta$ degenerates into the original vector field F_θ when $\gamma = 1$.

Text and audio embeddings. CLAP (Elizalde et al., 2023) and T5 (Raffel et al., 2020) are commonly used models for extracting text embeddings. While CLAP captures global features, it lacks temporal sensitivity (Elizalde et al., 2023). An ablation study by Sun et al. (2024) shows that CLAP accelerates convergence compared to T5 but performs worse in spatial tasks. To improve text embeddings with better temporal cues and spatial information, we utilize the pretrained FLAN-T5 language model (Chung et al., 2024). This enhanced version of T5 has been fine-tuned on a variety of tasks, enabling us to extract text embeddings T_e from $T_{prompts}$.

Text spatial coherence augmentation. Most audio-language models, such as CLAP (Elizalde et al., 2023) and FLAN-T5 (Chung et al., 2024), lack specialized training on datasets that provide detailed text spatial coherence for sound localization. To address this, we propose a module that enhances the spatial expressive capacity of the text embeddings. We generate misalignment samples between $A_{lr} := A_l - A_r$ and the flipped $A_{rl} := A_r - A_l$ to capture spatial localization differences. As shown in the upper left of Figure 2, the classifier P integrates the selected features with the text representation T_e to assess whether the audio differences align with the text descriptions. This encourages the text features to reason about the relative positions of sound sources and identify cues indicat-

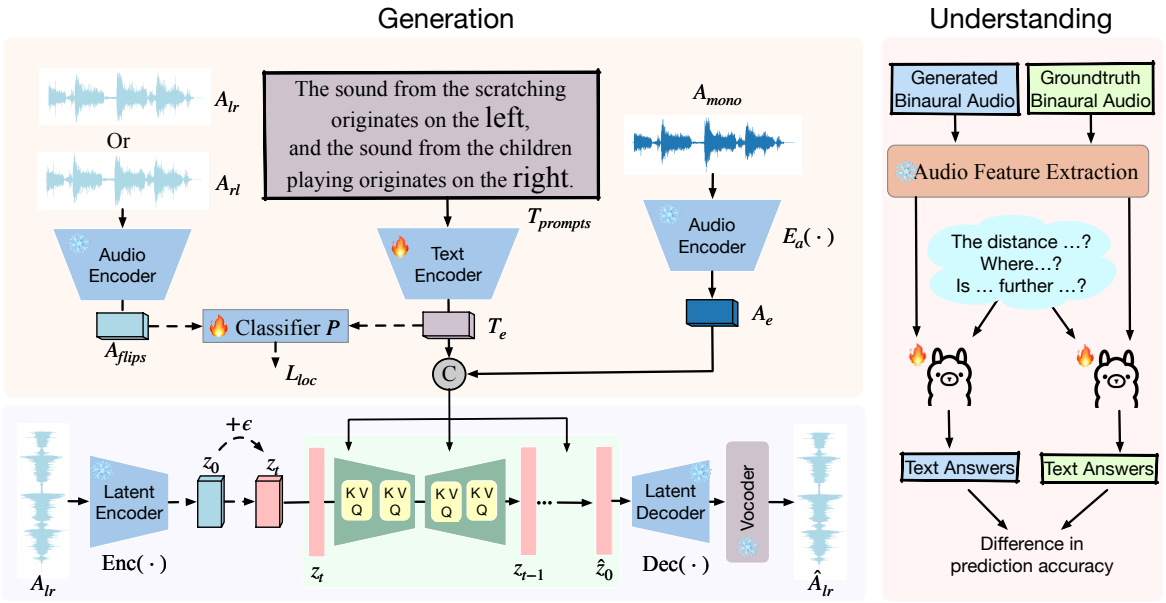

Figure 2: **The detailed structure for the text-guided audio spatialization model.** The dashed lines indicate processes that occur only during training. We train a latent diffusion model that adds noise to the monaural audio A_{mono} based on the concatenation of the encoded text embedding T_e and audio embedding A_e. During inference, the model predicts the binaural difference A_{lr} from the Gaussian noise. Additionally, we finetune a LLM to perform spatial reasoning, verifying the accuracy of the spatial semantic information in our generated binaural audio.

ing the perceived direction of sound. To evaluate the classifier's performance in predicting audio flipping, we calculate the Binaural Cross-Entropy (BCE) loss, represented as ground truth indicator $g = P(A_{lr}|A_{rl}, T_e)$, where $|$ denotes the logical OR operation. The indicator g indicates the ground truth of whether the audio is flipped, leading to the computation of the BCE loss for spatial coherence as follows:

$$\mathcal{L}_{loc} = \text{BCE}\left(P(A_{lr}|A_{rl}, T_e), g\right). \quad (5)$$

The total loss is the combination of the diffusion loss $\mathcal{L}_\theta$ and the spatial coherence loss $\mathcal{L}_{loc}$. The $\mathcal{L}_\theta$ is aimed at optimizing the parameters of the diffusion model, while $\mathcal{L}_{loc}$ is mainly designed to finetune the text encoder.

3.3 Spatial Understanding Metrics

In addition to evaluating audio quality through generation metrics, we assess the spatial semantic coherence between our generated binaural audio and text prompts using a spatial audio reasoning task. This evaluation is detailed in the understanding part of Figure 2. Firstly, we follow Zheng et al. (2024) to develop an assessment model for spatial audio question answering. We fine-tune the Llama-3.1-8B model (Dubey et al., 2024) on SpatialTAS, integrating the pretrained SpatialAudioEncoder (Zheng et al., 2024) to extract spatial audio features. Secondly, we send the ground-truth binaural audio and our generated binaural audio to the assessment model along with the corresponding spatial questions, obtaining the prediction accuracy discrepancy between them. A lower discrepancy indicates superior spatial fidelity in our generated binaural audio. Spatial question types are detailed in Appendix C.

4 Experiments

4.1 Model Implementation Details

For our experiments, we employ the pretrained VAE and HiFi-GAN vocoder (Kong et al., 2020a) from Liu et al. (2024a), with both modules frozen during training. It is trained on the combination of AudioSet (Gemmeke et al., 2017), AudioCaps (Kim et al., 2019a), BBC Sound Effects and Freesound (Fonseca et al., 2021) datasets. Our model utilizes a U-Net backbone for the diffusion process, consisting of four encoder and decoder blocks that incorporate downsampling and upsampling operations, with a bottleneck layer positioned between them. Multi-head attention is employed in the last three encoder blocks and the first three decoder blocks, featuring 64 head dimensions and 8 heads per layer. The Variational Autoencoder (VAE) is configured with a

compression level r of 4 and a latent dimension d of 8. During the forward process, we implement N=1000 steps with a linear noise schedule that ranges from β_1=0.0015 to β_N=0.0195 for noise generation. Additionally, we leverage the DDIM sampling method (Song et al., 2020) with 200 sampling steps. For classifier-free guidance, we set the guidance scale λ to 2.5, as detailed in Equation Equation (4). Training is conducted using the AdamW optimizer (Loshchilov, 2017) with a learning rate of 10^{-4}, β_1=0.95, β_2=0.999, ϵ=10^{-6}, and a weight decay of 10^{-3} training for 500,000 steps.

4.2 Dataset

SpatialTAS Dataset. The SpatialTAS dataset, derived from the SpatialSoundQA (Zheng et al., 2024) dataset, contains large-scale simulated binaural audio with detailed and flexible text descriptions of sound source locations. We split the dataset into 376,104 training samples, 732 validation samples, and 4,000 testing samples. The training samples consist of 138,338 single-object DOA and DP samples, 117,519 two-object DOA and DP samples, 50,501 relative direction samples, and 52,747 relative distance samples. The testing samples are evenly distributed, with 1,000 samples for each category.

Revisiting FAIR-Play and YouTube-Binaural Dataset. The FAIR-Play Dataset (Gao and Grauman, 2019) contains 1,871 ten-second video clips accompanied by binaural audio recordings, totaling 5.2 hours of content, primarily focused on musical instrument sounds. To evaluate our model further, we also use the audio from the YouTube-Binaural Dataset (Garg et al., 2023), which includes 426 corresponding 360° videos. This dataset is sourced from the YT-Clean dataset (Morgado et al., 2018), featuring in-the-wild 360° YouTube videos collected using spatial audio-related queries, with limited superimposed sources like room conversations and individuals playing instruments. For fair comparison, we extract one frame from each video and generate text prompts describing the locations of each sound source. Using GPT-4o (Hurst et al., 2024), we set task parameters related to the Field Of View (FOV) and the receiver's position, instructing it to generate captions based on the frames. More details about the caption generation process can be found in Appendix A.

4.3 Evaluation Metrics

During evaluation, we use both generation metrics and understanding metrics to assess the generated binaural audio. The generation quality metrics include *Fréchet Distance* (FD), *Fréchet Audio Distance* (FAD), *Kullback-Leibler Divergence* (KL), and *Inception Score* (IS). We also compare our model with previous non-generative models using *STFT Distance* (STFT) and *Envelope Distance* (ENV). The understanding metrics comprise *Direction of Arrival* (DOA) and *Distance Estimation* (DE) for perception-related questions, as well as *Direction* and *Distance* for reasoning questions. More details about these metrics are provided in Appendix B.

5 Results

We first present the experimental results on the test set of the proposed SpatialTAS dataset, using both generation and understanding metrics. Next, we report results from the real-recorded FAIR-Play dataset (Gao and Grauman, 2019) and the 360° Youtube-Binaural dataset (Garg et al., 2023), with the corresponding image-to-caption text descriptions. We then conduct ablation studies focused on the effects of separately modifying the direction, distance, or relative position components of the text prompts. Finally, we visualize several generated results alongside their spectrograms, using different kind of text prompts.

5.1 SpatialTAS Evaluation Results

The performance of our model is comprehensively evaluated on the testing set of SpatialTAS. The detailed results are presented in Table 2, where we compare our approach with two baselines: *Mono-Mono* and PseudoBinaural (Xu et al., 2021). *Mono-Mono* serves as a baseline to verify whether our model can effectively distinguish between the two channels, achieved by duplicating the same monaural audio to create a two-channel input. PseudoBinaural (Xu et al., 2021) shares a similar concept with our method in leveraging large-scale pseudo-generated binaural audio for training and demonstrating generalization to real audio. Originally proposed with a U-Net structure and cross-attention mechanism utilizing extracted visual features, we re-train PseudoBinaural on SpatialTAS with the corresponding text descriptions to ensure a fair comparison.

As detailed in Table 2, we conduct an exten-

Method	Generation Metrics				Understanding Metrics			
					Perception		Reasoning	
	FD↓	FAD↓	KL↓	IS↑	DOA↓	DE↓	Direction↓	Distance↓
Mono-Mono	9.03	3.67	0.99	1.61	19.66	18.12	12.79	15.33
PseudoBinaural (2021)*	7.23	2.81	0.65	1.85	6.39	4.00	10.36	12.91
Ours	**4.93**	**1.44**	**0.58**	**2.23**	**3.07**	**2.45**	**6.99**	**8.16**
Ours w/o text	6.77	2.54	0.63	2.00	5.87	4.03	9.25	11.40
Ours w/o Flipper	5.08	1.72	0.61	2.15	4.14	2.89	8.63	10.03

* indicates that we re-train it on the SpatialTAS dataset.

Table 2: **Results on the testing set of SpatialTAS.** Mono-Mono refers to duplicating the mono audio. Our model demonstrates strong performance in both Generation Metrics for audio quality and Understanding Metrics for spatial semantic correctness. Additionally, we present ablation results without text conditions and the flipped-channel audio augmentation module.

sive comparison of the models on a range of quality metrics that evaluate the overall quality of the generated audio, as well as spatialization metrics that specifically assess the accuracy of spatialization achieved through text-based spatial question-answering. Our model consistently demonstrates superior performance across multiple metrics, particularly in the spatial perception and reasoning tasks, which involve evaluating the generated audio based on questions focusing on "the relative positions between any two sounding sources" and "estimating the relative distance between any two sounding sources". Notably, in the reasoning part of the understanding metrics, we observe a significant performance improvement of 5.80% and 7.17% compared to the Mono-Mono baseline. In contrast, the PseudoBinaural approach achieves improvements of only 2.43% and 2.42% over Mono-Mono. This observation suggests that PseudoBinaural may lack the necessary capabilities to effectively generate corresponding binaural audio guided by relative position text descriptions. To further analyze the impact of different components in our model, we conduct ablation studies by evaluating models without text-guided descriptions and models trained without the text spatial coherence augmentation (i.e., without the binaural channel flippers). The results clearly demonstrate the significance of both text conditions and the spatial coherence module in achieving superior performance.

5.2 Real-recorded Binaural Audio Evaluation

We extend our evaluation to the FAIR-Play Dataset and the 360° Youtube-Binaural Dataset, which encompass in-the-wild binaural audio recordings of music and life-like sounds. Since these datasets

Method	FAIR-Play Dataset			
	STFT↓	ENV↓	WAV↓	SNR↑
Mono-Mono	1.155	0.153	7.666	5.735
L2BNet (2021)	1.028	0.148	-	-
Mono2Binaural (2019)	0.959	0.141	6.496	6.232
APNet (2020)	0.889	0.136	5.758	6.972
Sep-binaural (2020)	0.879	0.135	6.526	6.422
Main-net (2021)	0.867	0.135	5.750	6.985
Complete-net (2021)	0.856	0.134	5.787	6.959
AVSN (2024b)	0.849	0.133	-	-
Cyclic (2024a)	0.787	0.128	5.244	7.546
TAS (2024b)	0.914	0.137	6.092	6.771
Ours	**0.773**	**0.126**	**5.019**	**7.966**

Table 3: **Results on the FAIR-Play Dataset.** Our model performs well in real-world scenarios with diverse musical sound sources and outperforms visually guided models, underscoring the importance of text prompts.

are originally video-based, we generate text descriptions for the locations of each sounding source based on the videos using GPT-4o (Hurst et al., 2024). Notably, we generate different spatial position descriptions according to the extracted frames with varying Field of View (FOV), considering that the extracted frames in the FAIR-Play Dataset are not 360° views, while those in the 360° Youtube-Binaural Dataset are omnidirectional views. This approach ensures that our model is evaluated on a diverse set of real-world binaural audio recordings.

As comprehensively presented in Table 3, we conduct an extensive comparison of our model with other visual-guided and text-guided methods. Our model consistently outperforms the others across almost all metrics. It is noticing that TAS (Li et al.,

Method	360° Youtube-Binaural STFT↓	ENV↓
Mono-Mono	4.715	0.261
Audio-only	3.129	0.213
Mono2Binaural (2019)	2.892	0.208
APNet (2024b)	2.733	0.204
SimBinaural (2023)	2.544	0.196
Ours	**2.471**	**0.188**

Table 4: **Results on the 360° Youtube-Binaural Dataset.** The results indicate that our model easily extends to various types of real-recorded sounds, including speech and diverse natural sounds.

2024b) exhibits inferior performance compared to previous visual-guided methods. In contrast, our method surpasses these visual-guided methods. This observation suggests that utilizing more flexible text descriptions for the location of sounding sources, encompassing both 3D spatial position descriptions and relative position descriptions, provides the model with more generalized guidance for audio spatialization. Furthermore, as illustrated in Table 4, we demonstrate performance improvements on the 360° Youtube-Binaural Dataset, showcasing the generalization capabilities of our model to in-the-wild scenarios.

5.3 Ablations for Text Prompts

As illustrated in Figure 3, we present the results of our ablation studies, focusing on how changes in text prompts related to direction, distance, and multi-source relative positions affect sound localization. Figure 3(a) demonstrates the impact of changing the directional component of the text prompt from "right" to "left". This adjustment enables us to evaluate the Interaural Time Difference (ITD), which measures the time delay for sound to reach each ear. The goal of estimating the ITD is to ascertain the difference in arrival times of a sound at two microphones. Our results indicate that modifying the directional aspect effectively localizes the sound to the specified direction. Figure 3(b) illustrates the Interaural Level Difference (ILD) when the distance of the sound source is changed from "4m away" to "9m away". The ILD refers to the difference in sound pressure levels reaching each ear. We observe that altering the distance results in a lower ILD, demonstrating how distance affects perceived loudness. Figure 3(c) represents the differences in spectrograms when changing the relative position from "is nearer to" to "is farther away from". Given the significant frequency differences between the sounds of a baby crying and dance music, we can analyze the changes in the directional spectrogram by examining variations in energy levels. The sound of a baby crying primarily occupies the lower left section of the spectrogram, while dance music predominantly occupies the upper region. This results in a noticeable change in energy: the baby cry exhibits a transition from high to low energy, whereas the dance music shows a shift from low to high power. Overall, these findings demonstrate that text prompts can provide more detailed and flexible control over the localization of sound sources.

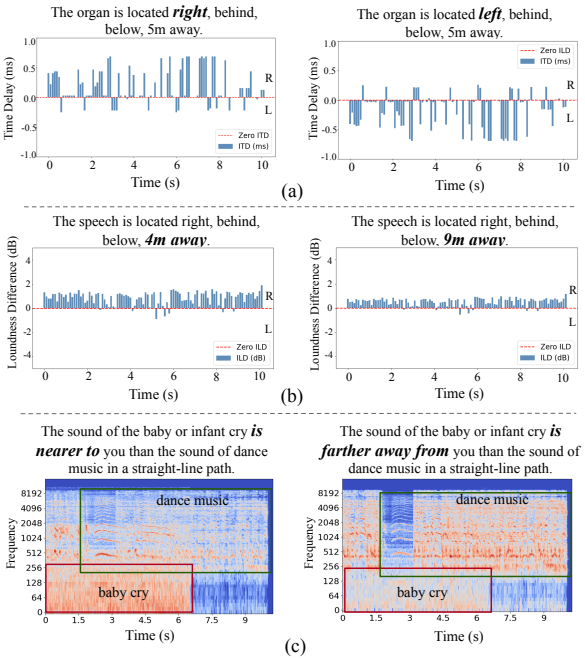

Figure 3: **Ablations for text prompts.** We systematically alter the direction, distance, and relative position in the text prompts, and present the differences observed before and after these changes.

5.4 Visualization Results

As illustrated in Figure 4, we present several generation results from the test set of the SpatialTAS dataset. First, we display text prompts that provide detailed descriptions of single-object locations for music and speech. Next, we showcase flexible text prompts designed for multi-object relative location descriptions for a broader range of natural sounds. The results indicate that our method generates audio with a more natural distribution that closely aligns with the ground truth compared to PseudoBinaural.

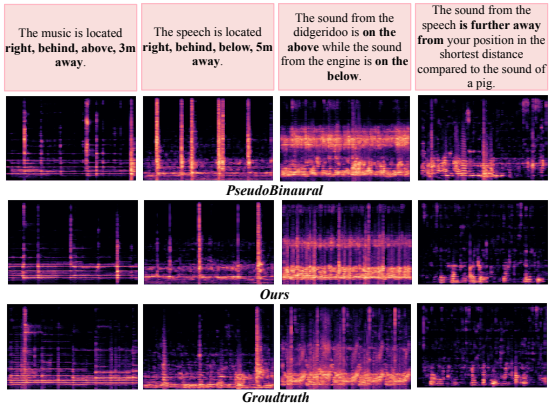

Figure 4: **Visualization for binaural difference prediction.** We present the binaural difference results using various spatial text prompts, including 3D sound source descriptions and relative position descriptions for music, speech, and natural sounds.

6 Conclusion

We propose a Text-guided Audio Spatialization (TAS) framework, providing a more flexible and interactive control to map monaural audios to binaural ones. We especially train the latent diffusion model on large-scale simulated datasets and can perform well on real-recorded datasets. We evaluate the binaural audio quality from generation metrics and spatial coherence through spatial audio reasoning with LLM. Results show that we can generate binaural audios with both high-quality and semantic consistency in spatial locations.

Limitations

Our model does not account for changes in the location of each sounding object. For instance, a car approaching the listener would produce a change in perceived distance from far to near. Additionally, due to data limitations, our model currently relies solely on text modality to guide audio spatialization. We do not incorporate both text and image modalities, or even motion cues from videos as conditioning factors.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant No. 62402211) and the Natural Science Foundation of Jiangsu Province (Grant No. BK20241248). Jie Liu is the corresponding author.

References

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.

Rishit Dagli, Shivesh Prakash, Robert Wu, and Houman Khosravani. 2024. See-2-sound: Zero-shot spatial environment-to-spatial sound. *arXiv preprint arXiv:2406.06612*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. 2023. Clap learning audio concepts from natural language supervision. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Zach Evans, Julian D Parker, CJ Carr, Zack Zukowski, Josiah Taylor, and Jordi Pons. 2024a. Long-form music generation with latent diffusion. *arXiv preprint arXiv:2404.10301*.

Zach Evans, Julian D Parker, CJ Carr, Zack Zukowski, Josiah Taylor, and Jordi Pons. 2024b. Stable audio open. *arXiv preprint arXiv:2407.14358*.

Eduardo Fonseca, Xavier Favory, Jordi Pons, Frederic Font, and Xavier Serra. 2021. Fsd50k: an open dataset of human-labeled sound events. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:829–852.

Ruohan Gao and Kristen Grauman. 2019. 2.5 d visual sound. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 324–333.

Rishabh Garg, Ruohan Gao, and Kristen Grauman. 2023. Visually-guided audio spatialization in video with geometry-aware multi-task learning. *International Journal of Computer Vision*, 131(10):2723–2737.

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

Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al. 2017. Cnn architectures for large-scale audio classification. In *2017 ieee international conference on acoustics, speech and signal processing (icassp)*, pages 131–135. IEEE.

Jonathan Ho and Tim Salimans. 2022. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. 2019a. Audiocaps: Generating captions for audios in the wild. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 119–132.

Hansung Kim, Luca Remaggi, Philip JB Jackson, and Adrian Hilton. 2019b. Immersive spatial audio reproduction for vr/ar using room acoustic modelling from 360 images. In *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pages 120–126. IEEE.

Diederik P Kingma. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. 2020a. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. *Advances in neural information processing systems*, 33:17022–17033.

Qiuqiang Kong, Yin Cao, Turab Iqbal, Yuxuan Wang, Wenwu Wang, and Mark D Plumbley. 2020b. Panns: Large-scale pretrained audio neural networks for audio pattern recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:2880–2894.

Yeonghyeon Lee, Inmo Yeon, Juhan Nam, and Joon Son Chung. 2023. Voiceldm: Text-to-speech with environmental context. *Preprint*, arXiv:2309.13664.

Dingzeyu Li, Timothy R Langlois, and Changxi Zheng. 2018. Scene-aware audio for 360 videos. *ACM Transactions on Graphics (TOG)*, 37(4):1–12.

Zhaojian Li, Bin Zhao, and Yuan Yuan. 2024a. Cyclic learning for binaural audio generation and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26669–26678.

Zhaojian Li, Bin Zhao, and Yuan Yuan. 2024b. Tas: Personalized text-guided audio spatialization. In *ACM Multimedia 2024*.

Haohe Liu, Zehua Chen, Yi Yuan, Xinhao Mei, Xubo Liu, Danilo Mandic, Wenwu Wang, and Mark D Plumbley. 2023. Audioldm: Text-to-audio generation with latent diffusion models. *arXiv preprint arXiv:2301.12503*.

Haohe Liu, Yi Yuan, Xubo Liu, Xinhao Mei, Qiuqiang Kong, Qiao Tian, Yuping Wang, Wenwu Wang, Yuxuan Wang, and Mark D Plumbley. 2024a. Audioldm 2: Learning holistic audio generation with self-supervised pretraining. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Miao Liu, Jing Wang, Xinyuan Qian, and Xiang Xie. 2024b. Visually guided binaural audio generation with cross-modal consistency. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7980–7984. IEEE.

Yang Liu, Weixing Chen, Yongjie Bai, Xiaodan Liang, Guanbin Li, Wen Gao, and Liang Lin. 2024c. Aligning cyber space with physical world: A comprehensive survey on embodied ai. *arXiv preprint arXiv:2407.06886*.

I Loshchilov. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Pedro Morgado, Nuno Nvasconcelos, Timothy Langlois, and Oliver Wang. 2018. Self-supervised generation of spatial audio for 360 video. *Advances in neural information processing systems*, 31.

Kranti Kumar Parida, Siddharth Srivastava, and Gaurav Sharma. 2022. Beyond mono to binaural: Generating binaural audio from mono audio with depth and cross modal attention. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3347–3356.

Kranthi Kumar Rachavarapu, Vignesh Sundaresha, AN Rajagopalan, et al. 2021. Localize to binauralize: Audio spatialization from visual sound source localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1930–1939.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

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

Saksham Singh Kushwaha, Jianbo Ma, Mark RP Thomas, Yapeng Tian, and Avery Bruni. 2024. Diffsage: End-to-end spatial audio generation using diffusion models. *arXiv e-prints*, pages arXiv–2410.

Jiaming Song, Chenlin Meng, and Stefano Ermon. 2020. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*.

Peiwen Sun, Sitong Cheng, Xiangtai Li, Zhen Ye, Huadai Liu, Honggang Zhang, Wei Xue, and Yike Guo. 2024. Both ears wide open: Towards language-driven spatial audio generation. *arXiv preprint arXiv:2410.10676*.

Apoorv Vyas, Bowen Shi, Matthew Le, Andros Tjandra, Yi-Chiao Wu, Baishan Guo, Jiemin Zhang, Xinyue Zhang, Robert Adkins, William Ngan, et al. 2023. Audiobox: Unified audio generation with natural language prompts. *arXiv preprint arXiv:2312.15821*.

Xudong Xu, Dejan Markovic, Jacob Sandakly, Todd Keebler, Steven Krenn, and Alexander Richard. 2024. Sounding bodies: modeling 3d spatial sound of humans using body pose and audio. *Advances in Neural Information Processing Systems*, 36.

Xudong Xu, Hang Zhou, Ziwei Liu, Bo Dai, Xiaogang Wang, and Dahua Lin. 2021. Visually informed binaural audio generation without binaural audios. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15485–15494.

Dongchao Yang, Jinchuan Tian, Xu Tan, Rongjie Huang, Songxiang Liu, Xuankai Chang, Jiatong Shi, Sheng Zhao, Jiang Bian, Xixin Wu, Zhou Zhao, Shinji Watanabe, and Helen Meng. 2023. Uniaudio: An audio foundation model toward universal audio generation. *Preprint*, arXiv:2310.00704.

Wen Zhang and Jie Shao. 2021. Multi-attention audio-visual fusion network for audio spatialization. In *Proceedings of the 2021 International Conference on Multimedia Retrieval*, pages 394–401.

Zhisheng Zheng, Puyuan Peng, Ziyang Ma, Xie Chen, Eunsol Choi, and David Harwath. 2024. Bat: Learning to reason about spatial sounds with large language models. *arXiv preprint arXiv:2402.01591*.

Hang Zhou, Xudong Xu, Dahua Lin, Xiaogang Wang, and Ziwei Liu. 2020. Sep-stereo: Visually guided stereophonic audio generation by associating source separation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16*, pages 52–69. Springer.

A Image Caption Engineering

We extract the required sound sources of the video frame and its corresponding description of the orientation and distance, which is summarized into a caption using large language model(LLM). Prompting allows a pre-trained model to adapt to different tasks via different prompts without modifying any parameters. LLMs like GPT-4o have shown strong zero-shot and few-shot ability via prompting. Prompting has been successful for a variety of natural language tasks, hence we design prompt for GPT-4o for sound source detection and attribute inference in images. We provide an image of video and a list of detected sound sources. Then we require sound objects with attributes (relative orientation and distance from the lens). The prompt-guided caption complies with (1) accurately detect the sound source object (2) describe the required attributes as general captions do, and (3) provide auxiliary information in the caption if necessary. Figure 5 illustrates the GPT-4o prompt we use for image caption engineering.

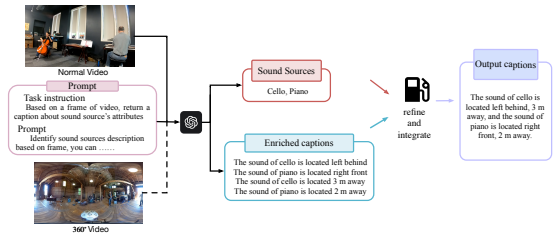

Figure 5: The caption engineering and an example.

B Evaluation Details

FD and FAD assess the distribution similarity between real and generated audio using different classifiers, with FAD employing VGGish (Hershey et al., 2017) and FD using PANNs (Kong et al., 2020b). KL quantifies distribution similarity, while IS evaluates the quality and diversity of the generated audio. Additionally, we compare our model with previous non-generative models using *STFT Distance* (STFT) and *Envelope Distance* (ENV). STFT is calculated as the Euclidean distance between the ground-truth and predicted complex spectrograms, scaled to represent raw audio energy levels. ENV involves computing the envelope of both ground-truth and predicted waveforms, as direct waveform comparisons may not capture perceptual similarity effectively.

C QA Pairs for Spatial Audio Reasoning

As shown in Appendix C, the questions can be categorized into spatial perception and spatial reasoning types. The perception questions primarily focus on Direction of Arrival (DOA) and Distance Estimation (DE), addressing the direction and distance descriptions for each sound source. In contrast, the reasoning questions involve the relative direction and distance between any two sound sources.

D Discussion about failure cases

Case 1 When two sources have similar characteristics with similar energy distributions in spectrogram, the generated results lead to distortion for the text embeddings may map to the same spectrogram part. In the future, we will specifically apply spectrogram-similar sound sources and spectrogram-different sound sources for targeted analysis.

Case 2 Incorrect text-embedding to audio mapping can result in unwanted sounds, especially in speech, which presents more stringent requirements compared to music and natural sounds. To address this issue, we will curate a diverse and representative dataset, employ advanced embedding techniques to capture nuanced differences, implement regularization methods to mitigate overfitting, and apply domain adaptation tailored to specific audio types.

Spatialization	QA Type	Example
Perception	DOA & DER	**Q:** (In single sound source.) How would you describe the location of the music's sound in terms of direction and distance? **A:** right, behind, below; 4m **Q:** (In double sound source.) At what distance and in which direction, is the writing's sound originating? **A:** left, behind, above; 2.5m
Reasoning	Direction & Distance	**Q:** Measuring the shortest path in a straight line, is the sound of camera more distant from you than the sound of music? / **A:** Yes **Q:** Is the sound of bird flight, flapping wings further from you than music when considering the direct paths? / **A:** Yes **Q:** Can you estimate the distance from the sound of the speech to the sound of the drawer open or close? **A:** 1.5m **Q:** What is the sound on the below side of the sound of the wind instrument, woodwind instrument? **A:** slap, smack **Q:** Could you determine whether the breaking's sound is to the left or right of the music's sound? **A:** left

Table 5: **QA pairs used the spatial llm reasoning task.** The first four types focus on perception, while the last emphasizes reasoning. DP: Distance Prediction; DOA: Direction-of-Arrival. Numbers (e.g., 139K, 15.9%) indicate the QA sample count and their percentages in the dataset.