

MACS: Multi-source Audio-to-image Generation with Contextual Significance and Semantic Alignment

Hao Zhou^{1†} Xiaobao Guo^{1†*} Yuzhe Zhu¹ Adams Wai-Kin Kong¹

¹Nanyang Technological University, Singapore

zhou0552@e.ntu.edu.sg, xiaobao.guo@ntu.edu.sg, g240005@e.ntu.edu.sg, adamskong@ntu.edu.sg

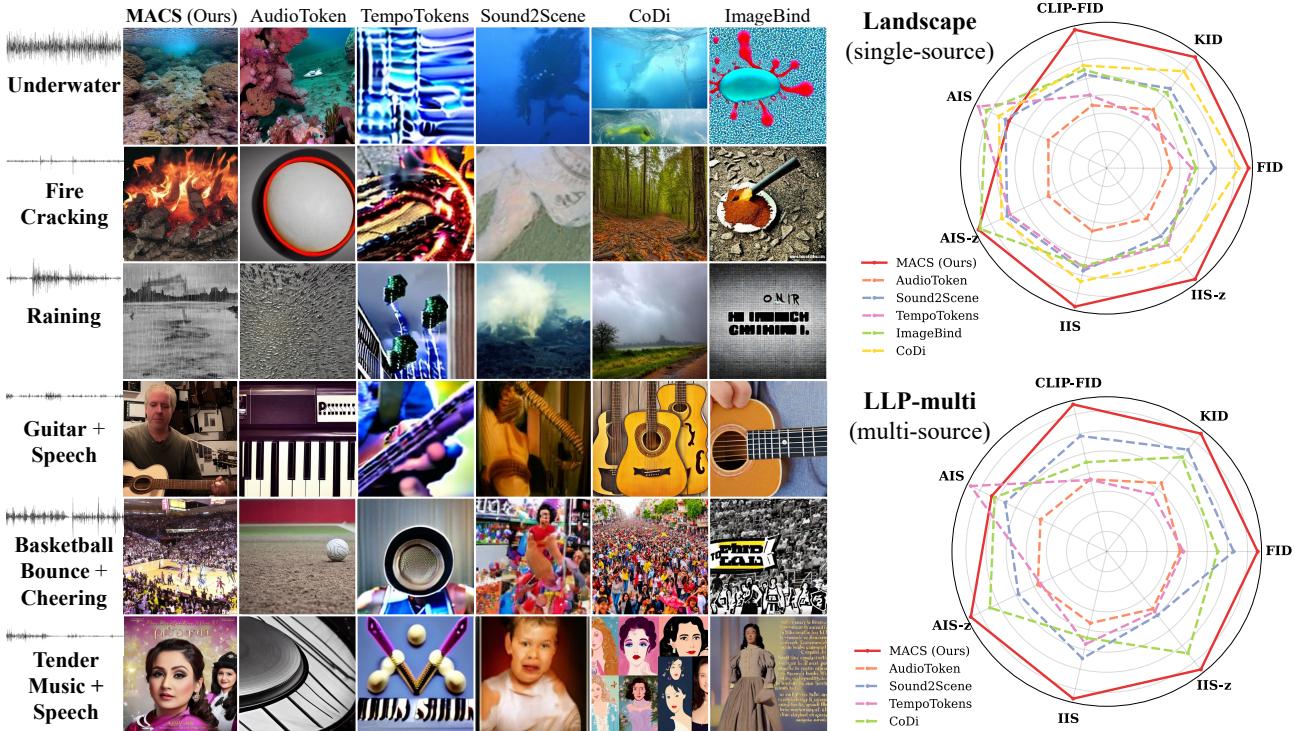


Figure 1. **Qualitative and Quantitative Comparison of MACS and other SOTA Methods.** **Left:** Generated images from single-source (line 1-3) and multi-source (line 4-6) audio datasets. **Right:** Radar maps illustrating performance on single-source (top) and multi-source (bottom) datasets. Values are normalized, and the lower-is-better metrics (FID, CLIP-FID, and KID) are inverted for consistency.

Abstract

Propelled by the breakthrough in deep generative models, audio-to-image generation has emerged as a pivotal cross-model task that converts complex auditory signals into rich visual representations. However, previous works only focus on single-source audio inputs for image generation, ignoring the multi-source characteristic in natural auditory scenes, thus limiting the performance in gener-

ating comprehensive visual content. To bridge this gap, a method called MACS is proposed to conduct multi-source audio-to-image generation. This is the first work that explicitly separates multi-source audio to capture the rich audio components before image generation. MACS is a two-stage method. In the first stage, multi-source audio inputs are separated by a weakly supervised method, where the audio and text labels are semantically aligned by casting into a common space using the large pre-trained CLAP model. We introduce a ranking loss to consider the contextual significance of the separated audio signals. In the second stage, efficient image generation is achieved by mapping

[†]Equal contribution.

*Corresponding author.

the separated audio signals to the generation condition using only a trainable adapter and a MLP layer. We preprocess the LLP dataset as the first full multi-source audio-to-image generation benchmark. The experiments are conducted on multi-source, mixed-source, and single-source audio-to-image generation tasks. The proposed MACS outperforms the current state-of-the-art methods in 17 of the 21 evaluation indexes on all tasks and delivers superior visual quality. The code will be publicly available.

1. Introduction

Audio-to-image generation has emerged as a cross-modal task that transforms rich and dynamic audio signals into semantically coherent visual representations. Early works in this area have demonstrated that audio cues, often rich with temporal dynamics and nuanced semantic information, can guide the synthesis of images [3, 4, 7, 56]. Recently, inspired by the success of diffusion models [15] and multimodal learning [17, 35, 51], more research works have demonstrated that models originally designed for text-to-image synthesis can be successfully adapted for audio inputs [34, 52], making it easier to develop an audio-to-image generation model. Audio-to-image generation is useful in many applications such as creative arts [23], multi-media content generation [11], and enhanced teaching and learning experiences by generating immersive visuals from sound by VR or AR systems [8].

Despite the recent success in conditioning image synthesis on audio, most of the existing literature focuses on single-source audio inputs [34, 42, 52]. Instead of a single, isolated sound, natural auditory scenes are typically multi-source, with overlapping and mixed components, where each source contributes distinct semantic cues to the overall auditory scene [57, 58]. Since these natural audio signals are mixed by the physical environment, they are not ideal for directly training an image-generation model. Consequently, previous methods fail to capture the full richness of real-world audio, limiting their ability to generate contextually comprehensive images (see Fig. 1; other methods could not combine the “Basketball Bounce + Cheering” scene effectively, while MACS generated a coherent one).

Although multi-source audio-to-image generation is a potential approach to address the limits, it introduces some additional challenges: 1) **audio mixture separation**—Overlapping audio signals need to be disentangled to enable accurate image generation. This requires robust audio source separation techniques that isolate distinct components while preserving their unique characteristics; 2) **contextual significance and semantic alignment**—The individual contribution and semantic cues of the separated audio stream must be preserved before merging them into coherent visual content. The model must learn to balance the contextual significance of the individual audios so that the

generated images do not overlook the contributions and semantics of some individual audios or overemphasize others; 3) **diffusion generation with multiple audio signals**—The system needs to map multiple concurrent audio representations to a single visual output using diffusion models, where the overall scene can be generated effectively. Overcoming these challenges is crucial for developing a multi-source audio-to-image framework that disentangles individual sounds and generates high-quality images reflecting the complex interplay of real-world audio.

To bridge the gap between multi-source audio and image generation, we propose **MACS**, which enables image generation from complex multi-source audio signals. To the best of our knowledge, this is the first method that aims to explicitly generate images from multi-source audios. We present a “*separation before generation*” approach. Specifically, we propose a two-stage framework that first learns to separate an audio mixture into its constituent single audio signals, where the contextual significance and semantics are preserved. We propose a multi-source audio separation network based on UNet [6, 30, 39]. For semantic alignment, we project individual audio signals and their corresponding labels into the CLAP space using a contrastive loss [51]. Leveraging the pre-trained model *provides additional prior knowledge and enriches the semantic representation of the audio*. We also introduce a ranking loss to *capture the contextual significance of each audio signal*. By disentangling the mixed audio from the physical environment, our approach enables the model to learn how to *combine* these signals *more effectively* for image generation. In the second stage, the individual audio signals are transformed and mapped to a visual output by the diffusion process [15], where we use the trainable decoupled cross-attention module [54] and an MLP layer for efficient mapping while keeping the rest of the model frozen. As shown in Fig. 1, MACS can effectively address the aforementioned challenges and achieve desirable performance.

To sum up, our contributions are as follows:

- We propose MACS, the first audio-to-image framework that explicitly separates multi-source audio inputs.
- We propose to preserve the contextual significance and semantics of the separated audio signals by introducing a ranking loss and a contrastive loss in the CLAP space.
- We propose a scheme based on the decoupled cross-attention module to efficiently merge multiple audio signals into a single image in the diffusion process.
- MACS outperforms the compared SOTA models on multi-source, mixed-source, and single-source audio-to-image generation tasks with significant margins.

2. Related Works

2.1. Sound Source Separation

Sound source separation aims to decompose a mixed audio signal into its constituent sound sources. Recently, researchers have pursued different training schemes, including supervised, unsupervised, and weakly-supervised methods to separate sound sources. Supervised methods typically train on large amounts of mixtures with isolated ground-truth sources [29, 47]. However, these methods mainly focus on specific domains like speech and music, and it is challenging to get isolated sounds in real-world scenarios. In contrast, unsupervised techniques leverage massive amounts of unlabeled mixtures to learn rich representations that can be fine-tuned for separation. For example, PIT [55] and MixIT [49] learn to output multiple sub-audios directly from multi-source audio. However, they require a post-selection process, such as a trained sound classifier, to tell what sounds are in each prediction result. MixPIT [18] directly outputs the predicted mixture in the Mixture of Mixtures (MoM) setting, but the number of separations is limited. To bridge the gap between supervised and unsupervised methods, the weakly-supervised approaches explore large-scale mixture datasets and rely on high-level semantic information rather than exact source ground truth for guidance [33]. However, most prior research works emphasize vision- or text-conditioned audio separation [6, 30]. In this work, we follow the paradigm of the weakly-supervised methods and focus on leveraging versatile semantic information on unconditional sound separation.

2.2. Contrastive Multimodal Pre-training

Multimodal contrast pre-training has proven versatile and effective in rich data representations and multimodal alignment [48]. Contrastive learning usually trains the dual or multi-stream encoders to pull the positive pairs closer while pushing away the negative ones with a contrastive objective function. CILP [35] is the seminal work for joint visual-text learning that enables powerful zero-shot capabilities, which is trained on around 400M image–text pairs. Later, other works are proposed for joint visual-text representations using simple transformer architectures [5, 19], or aiming for robust representations and better alignment, such as ALIGN [17] and BLIP-2 [25]. To introduce audio to CLIP, researchers often replace one modality or add more modalities. For example, AudioCLIP [12] extends CLIP to the audio domain by training a joint embedding for audio, visual, and textual modalities via contrastive learning. Wav2CLIP [50] distills the knowledge from the image encoder to the audio encoder from CLIP for audio-visual alignment. An aligned multimodal space is beneficial for many downstream tasks, such as audio-visual classification [12, 50] and sound-guided image manipulation [21, 22]. For audio-text contrastive learning, Wu *et*

al. [51] propose CLAP, which is trained on over 600K audio-caption pairs to learn a joint audio-text embedding space. In this work, we leverage the strong alignment ability of CLAP for audio separation.

2.3. Audio-conditioned Image Generation

Recently, text-to-image generation networks has developed rapidly [24, 37, 38]. By going beyond the text as a sole conditioning signal, researchers are motivated to bridge the gap between auditory and visual modalities in the generation task. Audio offers a rich source of conditional information that can lead to a more comprehensive and contextually relevant image generation. Previous models mainly focus on specific audio domains such as music [3] and speech [32]. Some works rely on Generative Adversarial Networks (GAN) to synthesize images conditioned on general audio inputs [46]. Although it demonstrates the feasibility of generating images from the audio domain, it suffers limited diversity and quality. More recent works shift the network toward diffusion models due to the higher inherent stability and sample quality. AudioToken [52] trains an audio embedding model to transfer audios into text tokens and uses these tokens as the condition prompt of Stable Diffusion [38]. ImageBind [11] aims at versatile generation that learns the embedding space of multiple modalities, such as text, image, audio, and depth. It develops the audio-to-image generation task with unCLIP [36] for better multimodal alignment. Moreover, diffusion models, enhanced with specialized adapter modules, emerge as the state-of-the-art approach [31, 54] for conditional image generation and editing. They are lightweight and flexible to be integrated into a pre-training framework. Unlike previous works that use single-source audio for image generation, we tackle the more complex task of generating images from multi-source audio by explicitly separating the mixtures.

3. Methods

The proposed method, MACS, is a two-stage architecture that transforms multi-source audio into semantically rich images (see Fig. 2). In the first stage, the model separates the mixed audio into individual sub-audio signals, which are then embedded using a pre-trained CLAP model for semantic alignment. This process leverages additional prior knowledge to enrich the semantic representation, and a ranking loss is adopted to rank the importance of each signal. In the second stage, these embeddings condition the image generation pipeline through a trainable, decoupled cross-attention module that effectively fuses multiple audio cues. Since natural audio is inherently mixed by the physical world and not directly suitable for image generation, separating the signals allows the model to learn how to optimally recombine them for improved synthesis. Therefore, “separation before generation” is necessary and effective.

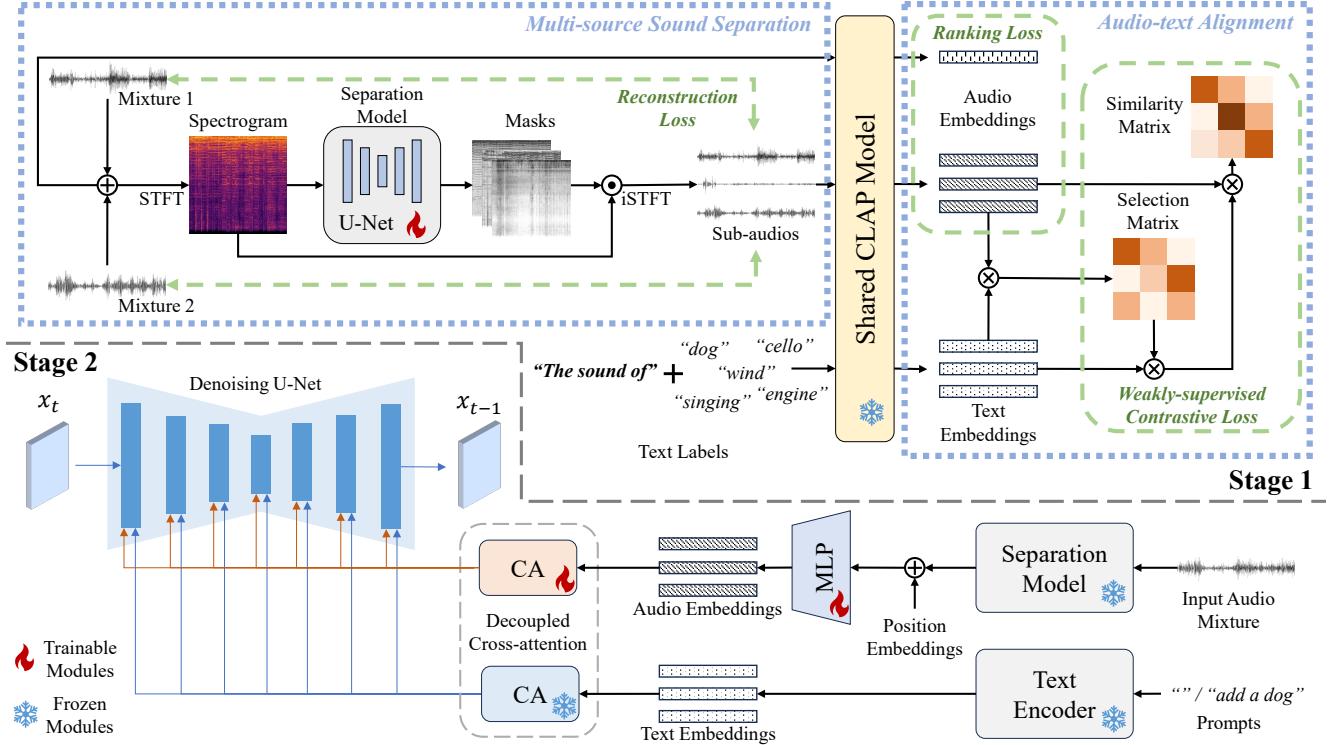


Figure 2. An overview of the proposed two-stage MACS architecture. *Stage 1:* A multi-source Sound Separation (MSS) model decomposes audio mixtures into sub-audios by reconstruction loss. The separated audios are then embedded using the CLAP model, with the contrastive loss and the ranking loss to ensure audio-text semantic alignment and contextual significance. *Stage 2:* A diffusion-based generation model integrates audio embeddings by a decoupled cross-attention module to produce high-quality, semantically accurate images. This two-stage design facilitates scalable pre-training of MSS, and it is efficient, requiring only a few trainable layers for generation.

3.1. Multi-source Audio Separation

Problem Formulation. Given a multi-source audio \mathbf{m} , the multi-source audio separation problem aims to find M binary masks by a separation model \mathcal{G}_θ such that each mask can be applied to the original mixture to get the separated audio component. Specifically, \mathbf{m} is first transformed into a spectrogram using the Short Time Fourier Transform (STFT), followed by a UNet-like model, \mathcal{U}_θ , to generate M masks. These masks are applied to the original spectrogram by multiplying them with the magnitude component to generate the masked spectrograms. Then, the resulting masked spectrograms are converted back into the waveform domain using inverse STFT (iSTFT). The forward process of the separation model \mathcal{G}_θ can be formulated as:

$$\mathcal{G}_\theta(\mathbf{m}) = \mathbf{T}^{-1}(|\mathbf{T}(\mathbf{m})| \odot \mathcal{U}_\theta(|\mathbf{T}(\mathbf{m})|), \phi(\mathbf{T}(\mathbf{m}))), \quad (1)$$

where $\mathbf{T}(\cdot)$ denotes the STFT function, and $|\cdot|$ and $\phi(\cdot)$ represent the magnitude and phase components involved in the STFT process, respectively. The UNet model only takes the magnitude as input as the phase information is not crucial to many audio-related tasks [30] and is only used for waveform reconstruction.

Mixed Audio Separation. Conventional audio separation methods mix multiple single-source signals to create

a composite input, then use the original signals as ground truth to train the model to recover each component. However, these methods are limited in fully capturing the complexity of real-world scenarios and struggle to generalize well to real recordings. Following the unsupervised method MixIT [49], we also form a **Mixture of Mixtures (MoM)** as input and train the model to separate its constituent sources by optimizing a reconstruction loss. Similar to previous works [6, 30], we use spectrograms and UNet models for separations rather than the waveforms used in MixIT. However, the difference is that our method is unconditional, *i.e.*, our UNet does not have conditioning inputs.

Formally, considering two audio mixtures, \mathbf{m}_1 and \mathbf{m}_2 , a new mixture of mixtures $\mathbf{m} = \mathbf{m}_1 + \mathbf{m}_2$ is formed. The output of $\mathcal{G}_\theta(\mathbf{m})$ is a set of separated audio signals $\mathcal{S} = \{s_1, \dots, s_M\}$, where their sum reconstructs \mathbf{m} . The reconstruction is conducted between \mathcal{S} and the original mixtures \mathbf{m}_1 and \mathbf{m}_2 . Typically, M is chosen to be greater than 2. To avoid the permutation variance, the reconstruction loss iterates over all possible bipartitions of the M separated signals and selects the combination that yields the

minimum reconstruction error [49]. It can be defined as

$$\begin{aligned}\mathcal{L}_{Rec} = \min_{(\Lambda_1, \Lambda_2) \in \Lambda} & [\mathcal{L}_{SISDR}(\mathbf{m}_1, \sum_{i \in \Lambda_1} \mathbf{s}_i) \\ & + \mathcal{L}_{SISDR}(\mathbf{m}_2, \sum_{i \in \Lambda_2} \mathbf{s}_i)],\end{aligned}\quad (2)$$

where Λ represents the set of all bipartitions of the index set $A = \{1, 2, \dots, M\}$, formally defined as:

$$\Lambda = \{(\Lambda_1, \Lambda_2) \mid \Lambda_1 \cup \Lambda_2 = A, \Lambda_1 \cap \Lambda_2 = \emptyset, \Lambda_1, \Lambda_2 \neq \emptyset\}. \quad (3)$$

To quantify the discrepancy between the estimated audio signal $\hat{\mathbf{s}}_j = \sum_{i \in \Lambda_j} \mathbf{s}_i, j = \{1, 2\}$ and its original audio \mathbf{m}_j , the negative scale-invariant signal-to-distortion ratio (SI-SDR) [20] is employed as the measurement for the reconstruction error,

$$\mathcal{L}_{SISDR}(\mathbf{m}_j, \hat{\mathbf{s}}_j) = -10 \log_{10} \frac{\|\alpha \mathbf{m}_j\|_2^2}{\|\alpha \mathbf{m}_j - \hat{\mathbf{s}}_j\|_2^2}, \quad (4)$$

where the scaling factor α is computed as:

$$\alpha = \frac{\hat{\mathbf{s}}_j^\top \mathbf{m}_j}{\|\mathbf{m}_j\|_2^2}. \quad (5)$$

The model learns to reconstruct the mixtures by minimizing the minimal reconstruction loss computed across all possible bipartitions of the separated outputs, without being influenced by suboptimal selections of Λ_1 and Λ_2 .

3.2. Audio-text Alignment

Although the multi-source audio separation model introduced in Section 3.1 can learn rich audio representations, it lacks high-level semantic alignment due to the unsupervised training scheme, which may not be desirable for audio-image generation. To address this issue, we introduce audio-text alignment, where a pre-trained audio-text model, CLAP [51], is employed to project the separated audio signals, the mixed audio, and the text labels into a shared semantic space using audio encoder \mathcal{P}_A and language encoder \mathcal{P}_L . We consider two aspects of the audio-text alignment: 1) the contextual significance of the separated audio signals in their original mixture and 2) the semantic alignment between the separated audio signals and their text labels.

Specifically, we project all the separated audio signals and their labels into the CLAP model. For every text label, like “violin”, we add a prefix of “The sound of” to form a complete phrase. For audio mixtures with fewer than M text labels, a placeholder label, “Noise”, is introduced to maintain consistency. The separated audios $\mathcal{S} = [\mathbf{s}_1, \dots, \mathbf{s}_M]$ and text labels $\mathcal{T} = [\mathbf{t}_1, \dots, \mathbf{t}_{M'}]$ are then embedded using the CLAP audio encoder \mathcal{P}_A and language encoder \mathcal{P}_L , respectively, resulting in audio embeddings $\mathcal{E}^A = \mathcal{P}_A(\mathcal{S}) \in \mathbb{R}^{M \times D}$ and text embeddings

$\mathcal{E}^T = \mathcal{P}_T(\mathcal{T}) \in \mathbb{R}^{M' \times D}$, where D represents the embedding dimension.

Ranking Loss. We adopt ranking loss that accounts for the contextual significance of the separated audio sources within a mixture, which is beneficial for the model to learn the significance of different separated audios. In the real world, specific aspects of audio—such as certain periods or timbres—may hold greater importance. For instance, background noise may be semantically less significant, whereas distinctive sounds like a dog barking or music could carry more meaningful information. However, the outputs from the multi-source separation audio are randomly ordered without prioritizing the important source.

To enable our audio separation model to identify and prioritize more significant audio separations, we introduce a *ranking loss*, formulated as:

$$\mathcal{L}_{Rank} = 1 - r_s(\mathbf{S}, \text{Sorted}(\mathbf{S})), \quad (6)$$

where $r_s(\cdot, \cdot)$ represents the Spearman’s rank correlation coefficient [41] quantifying the degree of discrepancy in the ranking of data between two arrays, and $\mathbf{S} \in \mathbb{R}^M$ consists of the cosine similarities between the CLAP embedding of the original audio mixture, $\mathcal{P}_A(\mathbf{m}) \in \mathbb{R}^D$, and the separated audio embeddings, $\mathcal{E}_i^A \in \mathbb{R}^{M \times D}$. Function $\text{Sorted}(\cdot)$ outputs the sorted array in descending order. In fact, there is no strict restriction to the selection of sorting function. To ensure differentiability during training, we adopt the ranking optimization method proposed by [2], which enables deep learning models to directly optimize ranking functions. By incorporating this ranking loss, the model is encouraged to discern and rank the significance of different audio separations, reinforcing its ability to preserve and emphasize key semantic information.

Contrastive Loss. Our second focus is on semantic alignment between the separated audio signals and their text labels, which is critical for image generation. To address the lack of explicit semantic cues in unsupervised methods, we employ a contrastive loss to align each audio’s semantic information with its corresponding text label.

Each separated audio \mathbf{s}_i is expected to semantically correspond to one text label from its associated labels. However, at this stage, we lack explicit information about the content of each separate audio. To address this, we leverage the joint embedding space and perform a *soft assignment* to align the separated audio embeddings with the text embeddings. This is achieved as follows:

$$\mathcal{E}'^T = \text{Softmax} \left(\frac{\langle \mathcal{E}^A \rangle \langle \mathcal{E}^T \rangle^\top}{\tau} \right) \mathcal{E}^T, \quad \mathcal{E}'^T \in \mathbb{R}^{M \times D}, \quad (7)$$

where $\langle \cdot \rangle$ denotes L_2 normalization, and τ is a fixed temperature parameter set to 1e-2. This soft assignment ensures that each separated audio embedding is aligned with

its most relevant text embedding, avoiding the rigidness of hard assignment and facilitating better semantic preservation during training.

To further align the embeddings of the two modalities, we compute the *contrastive loss* [35], formulated as:

$$\begin{aligned} \mathcal{L}_{CL} = & -\frac{1}{2M} \sum_{i=1}^M \log \frac{\exp(W_{ii})}{\sum_{j=1}^M \exp(W_{ij})} \\ & -\frac{1}{2M} \sum_{i=1}^M \log \frac{\exp(W_{ii})}{\sum_{j=1}^M \exp(W_{ji})}, \end{aligned} \quad (8)$$

where W is the similarity matrix defined as

$$W = \frac{\langle \mathcal{E}^A \rangle \langle \mathcal{E}'^T \rangle^\top}{\tau'} \in \mathbb{R}^{M \times M}, \quad (9)$$

with τ' being a trainable temperature parameter. We compute the cosine similarity between the soft-assigned text embeddings and the separated audio embeddings. The contrastive loss pulls matching pairs closer and pushes non-corresponding pairs apart, thereby improving semantic alignment.

Overall, the model is trained using the following loss function in the first stage:

$$\mathcal{L}_1 = \lambda \mathcal{L}_{Rec} + \mu \mathcal{L}_{CL} + \gamma \mathcal{L}_{Rank}, \quad (10)$$

where λ , μ and γ are weights of the losses. At this stage, the model is pre-trained solely on audio data and text labels, leveraging large-scale annotated datasets. Empirical evaluations show that it produces reliable audio embeddings that generalize well to image generation tasks.

3.3. Multi-source Audio-to-image Generation

To leverage the power of text-to-image models like Stable Diffusion, we concurrently transform multiple audio embeddings into conditioning inputs using a decoupled cross-attention module [54]. Originally designed as an *adapter* for image and text inputs, this module flexibly fuses multi-modal embeddings with minimal trainable parameters. *We extend it to process multiple audio inputs simultaneously, demonstrating its versatility for audio-to-image generation.*

Specifically, before passing the multiple audio embeddings into the adapter, we integrate them with trainable position embeddings, $\mathcal{E}^{Pos} \in \mathbb{R}^{M \times D}$, and transform the combined embeddings into the required dimensionality using a Multi-Layer Perceptron (MLP) equipped with layer normalization (refer to Fig. 2):

$$\mathcal{E}'^A = \text{MLP}(\mathcal{E}^A + \mathcal{E}^{Pos}) \in \mathbb{R}^{M \times D'}, \quad (11)$$

where D' represents the dimensionality required for conditioning. For a given set of query features \mathbf{H} from x_t in the

UNet, the audio embeddings are processed through a cross-attention layer:

$$\mathbf{H}_A = \text{Softmax} \left(\frac{(\mathbf{H}\mathbf{W}_q)(\mathcal{E}'^A\mathbf{W}_k)^\top}{\sqrt{D'}} \right) (\mathcal{E}'^A\mathbf{W}_v), \quad (12)$$

where \mathbf{W}_q , \mathbf{W}_k , and \mathbf{W}_v denote the query, key, and value weight matrices in the attention mechanism. \mathbf{W}_k and \mathbf{W}_v are newly initialized key and value weights, while the query \mathbf{W}_q is shared across modalities.

The decoupled cross-attention module can also take text modality as input. It is worth noting that the text prompts are *optional* in the MACS framework. The text embeddings can be set to *empty* if the conditions are solely audio inputs. Integrating a meaningful text embedding can further extend the MACS for *Audio-text Joint Image Generation*. Specifically, in the Stable Diffusion model, text prompts are embedded by the CLIP text encoder, which is denoted as \mathcal{E}^P . For a given set of query features \mathbf{H} , the cross-attention mechanism for text computes:

$$\mathbf{H}_T = \text{Softmax} \left(\frac{(\mathbf{H}\mathbf{W}_q)(\mathcal{E}^P\mathbf{W}'_k)^\top}{\sqrt{D'}} \right) (\mathcal{E}^P\mathbf{W}'_v), \quad (13)$$

where \mathbf{W}_q , \mathbf{W}'_k , and \mathbf{W}'_v denote the query, key, and value weight matrices.

The final output of the decoupled cross-attention layer is given by:

$$\mathbf{H}' = \mathbf{H}_T + \mathbf{H}_A. \quad (14)$$

In the second stage of MACS, during training, only \mathbf{W}_k , \mathbf{W}_v , \mathcal{E}^{Pos} , and the parameters of the MLP are trainable while the model in the first stage is frozen. The training objective follows the same loss function as Stable Diffusion (refer to Appendix I):

$$\mathcal{L}_2 = \mathcal{L}_{SD} = \mathbb{E}_{\mathbf{z}, \epsilon, t} \|\epsilon - \epsilon_\theta(\mathbf{z}_t, t, c)\|_2^2, \quad (15)$$

where c is the combined conditioning input consisting of \mathcal{E}'^A and \mathcal{E}^P . In our work, we extend the decoupled cross-attention module to audio inputs, enabling the mapping of multiple audio signals to a single image. Furthermore, the module's inherent flexibility allows for the seamless integration of text modality for efficient joint generation.

4. Experiments

4.1. Datasets

LLP-multi. Previous works focus on single-source, audio-conditioned image generation, leaving a gap in multi-source benchmarks. To tackle this, we preprocess the LLP dataset [45], a subset of AudioSet [10] originally designed for audio-visual video parsing, and form a dataset named “LLP-multi.” In the pre-processing, we select videos with more than one label and extract 6,595 frames with high

Table 1. Performance Comparison with the baselines on LLP-multi (multi-source). The best results are **bold**, and the second-best results are underlined. The method with a star* is excluded for comparison but reference only. The results are averaged over 5-fold cross-validation.

Method	FID \downarrow	CLIP-FID \downarrow	KID \downarrow	AIS \uparrow	AIS-z \uparrow	IIS \uparrow	IIS-z \uparrow
ImageBind* [11]	76.81	21.17	0.0088	0.0885	1.4219	0.6127	2.0361
AudioToken [52]	143.62	52.21	0.0431	0.0591	0.6201	0.4914	0.6799
Sound2Scene [42]	<u>105.14</u>	<u>33.79</u>	<u>0.0240</u>	0.0711	0.8176	<u>0.5545</u>	0.7877
TempoTokens [53]	141.37	52.45	0.0494	0.0828	0.5932	0.5288	0.7259
CoDi [44]	116.67	44.96	0.0283	0.0747	<u>1.1068</u>	0.5179	<u>1.4429</u>
AudioToken [52] (w/ MSS)	130.77	47.03	0.0396	0.0633	0.6621	0.5173	0.6940
MACS	87.09	20.47	0.0157	0.0754	1.3038	0.6269	1.7231

Table 2. Performance Comparison on AudioSet-Eval (mixed-source). The best results are **bold**, and the second-best results are underlined. The method with a star* is excluded for comparison but reference only. The results are averaged over 5-fold cross-validation.

Method	FID \downarrow	CLIP-FID \downarrow	KID \downarrow	AIS \uparrow	AIS-z \uparrow	IIS \uparrow	IIS-z \uparrow
ImageBind* [11]	41.69	14.76	0.0083	0.0892	1.1498	0.5808	1.7928
AudioToken [52]	102.85	40.68	0.0397	0.0663	0.5664	0.5426	0.6829
Sound2Scene [42]	<u>63.94</u>	<u>26.61</u>	0.0207	0.0725	0.7310	0.5445	0.6837
TempoTokens [53]	108.73	45.37	0.0510	0.0879	0.3863	0.5335	0.5153
CoDi [44]	70.20	31.49	0.0206	0.0789	1.0128	0.4920	1.0869
AudioToken [52] (w/ MSS)	96.93	37.67	0.0305	0.0702	0.6218	<u>0.5479</u>	0.7924
MACS	62.40	19.65	0.0142	0.0724	0.8736	0.5532	1.1328

audio-visual coexistence (6,314 with 2 labels, 242 with 3, and 35 with 4). LLP-multi is ideal for multi-source tasks as it contains multiple concurrent audio-visual events with corresponding annotations.

AudioSet-Eval. AudioSet [10] comprises over 2 million videos featuring diverse sounds such as human voices, animal noises, music, and environmental sounds. For mixed-source evaluation, we use its evaluation set. After manually filtering out about 20 extremely dark and low-resolution videos, we obtained 15,712 clips, 20.7% (3,255) with a single event label and 79.3% (12,457) with multiple event labels. With 610 distinct event labels, this dataset exhibits strong class diversity and is referred to as “AudioSet-Eval.”

Landscape. Landscape [21] is a high-quality dataset with natural scenes that contain one audio event for each video. There are 9 distinct labels for 1,000 video clips. It is widely adopted in audio-to-image generation tasks, and we also use it for the single-source evaluation dataset with 90%-10% train-test split [40].

FSD50K. FSD50K [9] is an audio dataset containing over 51,000 text-labeled audio clips totaling over 100 hours. The audio content is manually labeled using 200 classes drawn from the AudioSet ontology. *FSD50K is used to pre-train the audio separation model in the first stage of MACS.*

4.2. Evaluation Metrics

For a comprehensive performance evaluation, we gathered **seven metrics** for quantitative evaluation. (a) the overall quality of the generated images (*Fréchet Inception Distance (FID)* [14], *CLIP-FID*, and *Kernel Inception Distance (KID)* [1]), (b) pairwise similarity between generated

images and ground truth images (*Image-Image Similarity (IIS)* [52] and *IIS-z*, where “z” means using z-score), and (c) pairwise semantic similarity between audio and images (*Audio-Image Similarity (AIS)* [52] and *AIS-z*, “z” for z-score). More details are provided in the supplement Appendix F.

4.3. Quantitative Analysis

To ensure a comprehensive evaluation, we conducted experiments using *multi-source*, *mixed-source*, and *single-source* audio datasets. We compared MACS with five state-of-the-art (SOTA) methods: AudioToken [52], Sound2Scene [42], TempoTokens [53], ImageBind [11], and CoDi [44]. Note that ImageBind and CoDi are competitive *foundation models*, and we also tested these two on the evaluation datasets. Besides, since LLP-multi and AudioSet-Eval are included in ImageBind’s pre-training dataset [11], its performances in Tab. 1 and 2 are presented in gray for *reference only*.

- **Multi-source Audio.** We benchmarked MACS against five SOTA methods on **LLP-multi** (see Table 1), a *fully multi-source* audio dataset. Results show that MACS significantly improves image quality, as evidenced by the three left metrics, and it also enhances content fidelity and maintains semantic consistency across multiple sources, as shown by other metrics. We contend that the “*separation before generation*” strategy is key to its performance compared with methods that condition directly on mixed audio. Moreover, MACS achieves competitive results and outperforms ImageBind on two metrics, underscoring its strong capability in multi-source audio-to-image generation.

- **Mixed-source Audio.** We further compare MACS

Table 3. Performance Comparison on Landscape (single-source). The best results are **bold**, and the second-best results are underlined. The evaluation is conducted on the standard train-test split [40].

Method	FID \downarrow	CLIP-FID \downarrow	KID \downarrow	AIS \uparrow	AIS-z \uparrow	IIS \uparrow	IIS-z \uparrow
AudioToken [52]	236.63	54.42	0.0402	0.0708	0.3527	0.6030	0.2900
Sound2Scene [42]	186.12	43.25	0.0280	0.1042	0.6519	0.6762	0.6368
TempoTokens [53]	212.69	50.70	0.0450	0.1251	0.6307	0.6703	0.8057
ImageBind [11]	207.93	41.49	0.0304	<u>0.1189</u>	<u>0.8483</u>	0.6673	0.7681
CoDi [44]	<u>158.31</u>	39.97	<u>0.0180</u>	0.1094	0.6912	<u>0.6961</u>	<u>1.0942</u>
AudioToken [52] (w/ MSS)	202.54	50.57	0.0289	0.0817	0.4351	0.6161	0.3725
MACS	147.23	26.91	0.0098	0.1015	0.8602	0.7422	1.4805

with other methods on the mixed-source audio dataset. **AudioSet-Eval** is more complex than LLP-multi, featuring over 600 event classes. Nevertheless, MACS still achieves desirable performance in generating high-quality images. One observation from Tab. 1 and 2 is that the overall image quality on mixed-source audio datasets is superior. Compared with baselines that ignore the mixed nature of audio, MACS is more flexible in handling both single-source and multi-source inputs. Consequently, MACS offers significant advantages and can be easily extended to generate images from an arbitrary number of audio sources.

- **Single-source Audio.** MACS also exhibits strong performance on **Landscape**. As illustrated in Tab. 3, MACS achieves SOTA performance on most metrics, outperforming the second-best results by a substantial margin. In comparison to other leading baselines and two strong foundation models, the audio separation process in MACS enhances single audio quality by minimizing noise; thereby, the generated images can be both higher in quality and semantic relevance.

- **Multi-source Sound Separation (MSS) is Adaptable.** To demonstrate the effectiveness of MSS, we integrated MACS stage 1 outputs into AudioToken [52], a model well-suited for adaptation. AudioToken converts audio clips into embeddings that are concatenated with the tokenized prompt “A *photo of a ...*” for Stable Diffusion. We modified it to generate M audio tokens from the M separated signals and refer to this version as “AudioToken (w/ MSS)” (see Tabs. 1–3). Compared with the original AudioToken, MSS significantly improves performance, demonstrating both the effectiveness and adaptability of the proposed MACS. More analysis regarding MSS can be found in the supplement Appendix C.

4.4. Qualitative Results

- **MACS produces higher quality visuals.** The left section of Fig. 1 displays images generated from single- and multi-source audio using MACS and the baseline methods. MACS produces more realistic images that are contextually and semantically aligned with the audio. For example, generating detailed underwater scenes and vivid flames in the “Underwater” and “Fire Crackling” categories, while

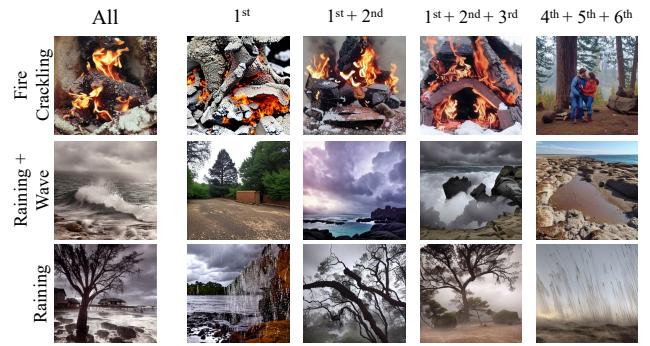


Figure 3. **Ranking loss helps sort the contextual significance.** Embeddings in higher rank positions contain more important semantic information for accurate image generation.

the baselines often yield abstract visuals. Similarly, for multi-source audio, MACS consistently delivers images that match the expected scenes, whereas methods like CoDi and ImageBind frequently fail to align with the audio content (e.g., in the “Basketball Bounce + Cheering” category). More qualitative results can be found in the supplement Appendix D.

- **Ranking Loss Helps Contextual Importance Learning.** The ranking loss in Eq. (6) trains the model to capture the contextual importance and priorities of audio signals, which is beneficial for image generation. With $M = 6$, we evaluated five configurations: *all embeddings, only the first, the first two, the first three, and the last three*. Qualitative results in Fig. 3 show that higher-ranked embeddings capture more important audio events, with the first three carrying most of the semantic information. In single-audio cases, the first embedding alone is sufficient to generate semantically correct images (see the first and last rows).

4.5. Ablation Studies

Due to space limitations, we report LLP-multi ablation results in the supplement Appendix B. We removed each of the three components one at a time, ranking loss (RL), contrastive loss (CL), and decoupled cross-attention (DC), while keeping the rest of MACS unchanged. In every case, performance degraded across all metrics compared with the

complete MACS model, underscoring the critical role of audio–text alignment in generating high-quality, semantically accurate images for multi-source audio-image generation.

5. Limitations

MACS effectively generates images from multiple mixed audio signals. However, failure cases may arise when the audio is ambiguous or lacks clear semantic cues. That is challenging for the model to produce images that accurately reflect the intended content. The model may also struggle with complex or abstract audios that lack direct visual counterparts, resulting in vague or mismatched images. We plan to address these limitations in future work.

6. Conclusion

In this work, we introduce MACS, the first two-stage architecture that explicitly separates mixed audio signals for audio-to-image generation. MACS preserves the contextual and semantic relationships between separated audio and text labels and uses a decoupled cross-attention module to integrate multiple audio signals efficiently. Extensive experiments demonstrate that the “separation before generation” strategy is effective, and MACS achieves state-of-the-art performance across both mixed-source and single-source audio-to-image generation tasks.

References

- [1] Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans. *arXiv preprint arXiv:1801.01401*, 2018. [7](#), [3](#)
- [2] Mathieu Blondel, Olivier Teboul, Quentin Berthet, and Josip Djolonga. Fast differentiable sorting and ranking. In *International Conference on Machine Learning*, pages 950–959. PMLR, 2020. [5](#)
- [3] Moitreyra Chatterjee and Anoop Cherian. Sound2sight: Generating visual dynamics from sound and context. In *Proceedings of the European Conference on Computer Vision*, pages 701–719. Springer, 2020. [2](#), [3](#)
- [4] Lele Chen, Sudhanshu Srivastava, Zhiyao Duan, and Chen-liang Xu. Deep cross-modal audio-visual generation. In *Proceedings of the on Thematic Workshops of ACM Multimedia 2017*, pages 349–357, 2017. [2](#)
- [5] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *Proceedings of the European Conference on Computer Vision*, pages 104–120. Springer, 2020. [3](#)
- [6] Hao-Wen Dong, Naoya Takahashi, Yuki Mitsufuji, Julian McAuley, and Taylor Berg-Kirkpatrick. Clipsep: Learning text-queried sound separation with noisy unlabeled videos. In *Proceedings of International Conference on Learning Representations (ICLR)*, 2023. [2](#), [3](#), [4](#)
- [7] Amanda Cardoso Duarte, Francisco Roldan, Miquel Tubau, Janna Escur, Santiago Pascual, Amaia Salvador, Eva Moredano, Kevin McGuinness, Jordi Torres, and Xavier Giro-i-Nieto. Wav2pix: Speech-conditioned face generation using generative adversarial networks. In *ICASSP*, pages 8633–8637, 2019. [2](#)
- [8] Tira Nur Fitria. Augmented reality (ar) and virtual reality (vr) technology in education: Media of teaching and learning: A review. *International Journal of Computer and Information System (IJCIS)*, 4(1):14–25, 2023. [2](#)
- [9] Eduardo Fonseca, Xavier Favory, Jordi Pons, Frederic Font, and Xavier Serra. Fsd50k: an open dataset of human-labeled sound events. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:829–852, 2021. [7](#)
- [10] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 776–780. IEEE, 2017. [6](#), [7](#)
- [11] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15180–15190, 2023. [2](#), [3](#), [7](#), [8](#), [4](#)
- [12] Andrey Guzhov, Federico Raue, Jörn Hees, and Andreas Dengel. Audioclip: Extending clip to image, text and audio. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 976–980. IEEE, 2022. [3](#)
- [13] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016. [1](#)
- [14] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in Neural Information Processing Systems*, 30, 2017. [7](#), [3](#)
- [15] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. [2](#), [5](#)
- [16] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7132–7141, 2018. [1](#)
- [17] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pages 4904–4916. PMLR, 2021. [2](#), [3](#)
- [18] Ertug Karamath and Serap Kirbuz. Mixcycle: unsupervised speech separation via cyclic mixture permutation invariant training. *IEEE Signal Processing Letters*, 29:2637–2641, 2022. [3](#)
- [19] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*, pages 5583–5594. PMLR, 2021. [3](#)
- [20] Jonathan Le Roux, Scott Wisdom, Hakan Erdogan, and John R Hershey. Sdr-half-baked or well done? In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 626–630. IEEE, 2019. [5](#)

- [21] Seung Hyun Lee, Wonseok Roh, Wonmin Byeon, Sang Ho Yoon, Chanyoung Kim, Jinkyu Kim, and Sangpil Kim. Sound-guided semantic image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3377–3386, 2022. 3, 7
- [22] Seung Hyun Lee, Hyung-gun Chi, Gyeongrok Oh, Wonmin Byeon, Sang Ho Yoon, Hyunje Park, Wonjun Cho, Jinkyu Kim, and Sangpil Kim. Robust sound-guided image manipulation. *Neural Networks*, 175:106271, 2024. 3
- [23] Taegyeong Lee, Jeonghun Kang, Hyeonyu Kim, and Tae-hwan Kim. Generating realistic images from in-the-wild sounds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7160–7170, 2023. 2
- [24] Bowen Li, Xiaojuan Qi, Thomas Lukasiewicz, and Philip Torr. Controllable text-to-image generation. *Advances in Neural Information Processing Systems*, 32, 2019. 3
- [25] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning*, pages 19730–19742. PMLR, 2023. 3
- [26] Zhaoyang Liu, Yinan He, Wenhui Wang, Weiyun Wang, Yi Wang, Shoufa Chen, Qinglong Zhang, Zeqiang Lai, Yang Yang, Qingyun Li, Jiashuo Yu, et al. Interngpt: Solving vision-centric tasks by interacting with chatgpt beyond language. *arXiv preprint arXiv:2305.05662*, 2023. 4
- [27] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *Proceedings of International Conference on Learning Representations (ICLR)*, 2019. 5
- [28] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. *Advances in Neural Information Processing Systems*, 35:5775–5787, 2022. 5
- [29] Yi Luo and Jianwei Yu. Music source separation with band-split rnn. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:1893–1901, 2023. 3
- [30] Tanvir Mahmud, Saeed Amizadeh, Kazuhito Koishida, and Diana Marculescu. Weakly-supervised audio separation via bi-modal semantic similarity. *arXiv preprint arXiv:2404.01740*, 2024. 2, 3, 4
- [31] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4296–4304, 2024. 3
- [32] Tae-Hyun Oh, Tali Dekel, Changil Kim, Inbar Mosseri, William T Freeman, Michael Rubinstein, and Wojciech Matusik. Speech2face: Learning the face behind a voice. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7539–7548, 2019. 3
- [33] Fatemeh Pishdadian, Gordon Wichern, and Jonathan Le Roux. Finding strength in weakness: Learning to separate sounds with weak supervision. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:2386–2399, 2020. 3
- [34] Can Qin, Ning Yu, Chen Xing, Shu Zhang, Zeyuan Chen, Stefano Ermon, Yun Fu, Caiming Xiong, and Ran Xu. Gluegen: Plug and play multi-modal encoders for x-to-image generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23085–23096, 2023. 2
- [35] 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*, pages 8748–8763. PMLR, 2021. 2, 3, 6
- [36] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1 (2):3, 2022. 3
- [37] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In *International Conference on Machine Learning*, pages 1060–1069. PMLR, 2016. 3
- [38] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 3, 5
- [39] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention—MICCAI 2015: 18th international conference, Munich, Germany, October 5–9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015. 2, 5
- [40] Ludan Ruan, Yiyang Ma, Huan Yang, Huiguo He, Bei Liu, Jianlong Fu, Nicholas Jing Yuan, Qin Jin, and Baining Guo. Mm-diffusion: Learning multi-modal diffusion models for joint audio and video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10219–10228, 2023. 7, 8, 5
- [41] Charles Spearman. The proof and measurement of association between two things. *The American Journal of Psychology*, 100(3/4):441–471, 1987. 5
- [42] Kim Sung-Bin, Arda Senocak, Hyunwoo Ha, Andrew Owens, and Tae-Hyun Oh. Sound to visual scene generation by audio-to-visual latent alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6430–6440, 2023. 2, 7, 8, 4
- [43] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2818–2826, 2016. 3
- [44] Zineng Tang, Ziyi Yang, Chenguang Zhu, Michael Zeng, and Mohit Bansal. Any-to-any generation via composable diffusion. *Advances in Neural Information Processing Systems*, 36:16083–16099, 2023. 7, 8, 4
- [45] Yapeng Tian, Dingzeyu Li, and Chenliang Xu. Unified multisensory perception: Weakly-supervised audio-visual video parsing. In *Proceedings of the European Conference on Computer Vision*, pages 436–454. Springer, 2020. 6, 3

- [46] Chia-Hung Wan, Shun-Po Chuang, and Hung-Yi Lee. Towards audio to scene image synthesis using generative adversarial network. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 496–500. IEEE, 2019. 3
- [47] DeLiang Wang and Jitong Chen. Supervised speech separation based on deep learning: An overview. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(10):1702–1726, 2018. 3
- [48] Xiao Wang, Guangyao Chen, Guangwu Qian, Pengcheng Gao, Xiao-Yong Wei, Yaowei Wang, Yonghong Tian, and Wen Gao. Large-scale multi-modal pre-trained models: A comprehensive survey. *Machine Intelligence Research*, 20(4):447–482, 2023. 3
- [49] Scott Wisdom, Efthymios Tzinis, Hakan Erdogan, Ron Weiss, Kevin Wilson, and John Hershey. Unsupervised sound separation using mixture invariant training. *Advances in Neural Information Processing Systems*, 33:3846–3857, 2020. 3, 4, 5
- [50] Ho-Hsiang Wu, Prem Seetharaman, Kundan Kumar, and Juan Pablo Bello. Wav2clip: Learning robust audio representations from clip. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4563–4567. IEEE, 2022. 3, 4
- [51] 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 *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023. 2, 3, 5
- [52] Guy Yariv, Itai Gat, Lior Wolf, Yossi Adi, and Idan Schwartz. Audiotoken: Adaptation of text-conditioned diffusion models for audio-to-image generation. *arXiv preprint arXiv:2305.13050*, 2023. 2, 3, 7, 8, 4, 5
- [53] Guy Yariv, Itai Gat, Sagie Benaim, Lior Wolf, Idan Schwartz, and Yossi Adi. Diverse and aligned audio-to-video generation via text-to-video model adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 6639–6647, 2024. 7, 8, 4
- [54] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ipadapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023. 2, 3, 6
- [55] Dong Yu, Morten Kolbæk, Zheng-Hua Tan, and Jesper Jensen. Permutation invariant training of deep models for speaker-independent multi-talker speech separation. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 241–245. IEEE, 2017. 3
- [56] Hang Zhou, Yu Liu, Ziwei Liu, Ping Luo, and Xiaogang Wang. Talking face generation by adversarially disentangled audio-visual representation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 9299–9306, 2019. 2
- [57] Jinxing Zhou, Xuyang Shen, Jianyuan Wang, Jiayi Zhang, Weixuan Sun, Jing Zhang, Stan Birchfield, Dan Guo, Lingpeng Kong, Meng Wang, et al. Audio-visual segmentation with semantics. *International Journal of Computer Vision*, pages 1–21, 2024. 2
- [58] Xiatian Zhu, Chen Change Loy, and Shaogang Gong. Learning from multiple sources for video summarisation. *International Journal of Computer Vision*, 117:247–268, 2016. 2

MACS: Multi-source Audio-to-image Generation with Contextual Significance and Semantic Alignment

Supplementary Material

A. Architecture of Multi-source Sound Separation Model in MACS

The overall architecture of the UNet separation model is shown in Fig. 4, where the input is a spectrogram and the output comprises source-specific masks. The architecture features symmetric encoder and decoder paths connected by skip connections, ensuring the preservation of spatial details and hierarchical representations across feature maps.

Encoder: The encoder progressively downsamples the input spectrogram while increasing the channel depth to capture hierarchical audio features. Beginning with the input spectrogram of dimensions $1 \times H \times W$, the encoder reduces the spatial resolution through a sequence of downsampling operations, halving the resolution at each stage until reaching $H/16 \times W/16$. The channel depth follows a structured progression from C to $16C$. Each stage of the encoder employs a downsample block, implemented with a 4×4 convolution with a stride of 2, followed by a ResBlock. The ResBlock combines two 3×3 convolutional layers with batch normalization (BN) and Gaussian Error Linear Unit (GELU) [13] activations, integrated through a residual connection to improve feature learning. The sum is handled by a Squeeze-and-Excitation Net (SENet) [16] to improve the model’s ability to capture interdependencies across different feature maps.

Bottleneck: We utilize a SENet block as the bottleneck of UNet model. This mechanism refines the feature representation by adaptively reweighting the feature channels based on global context, allowing the model to emphasize the most informative aspects of the audio representation.

Decoder: The decoder mirrors the encoder in structure, progressively upsampling the feature maps to restore the original resolution. Each stage in the decoder consists of an upsample block, implemented using transposed 4×4 convolutions with a stride of 2, followed by a ResBlock to refine the high-resolution feature maps. Skip connections between encoder and decoder layers at matching resolutions are employed to fuse high-frequency spatial details from the encoder into the decoder. This channel-wise concatenation at each resolution level ensures the preservation of critical spatial and spectral information throughout the decoding process.

Output Layers: The final output layer employs a 1×1 convolution with sigmoid activation to reduce the channel depth to M , corresponding to the number of output source masks. The resulting masks retain the same spatial dimen-

sions as the input spectrogram, with a shape of $M \times H \times W$. This design effectively captures both global and local audio features, enabling robust source separation and reconstruction.

B. Ablation Studies

We performed ablation studies in Tab. 4 on LLP-multi to assess the contribution of key components in the MACS framework. We removed three components one at a time: the ranking loss (RL), contrastive loss (CL), and decoupled cross-attention (DC), while keeping the remaining parts of MACS unchanged. For MACS w/o DC, we directly fed the MLP layer’s embeddings into the frozen UNet of the Stable Diffusion model. The results show that removing any component degrades performance across all metrics. Notably, MACS without the contrastive loss performs the worst in all metrics except IIS, highlighting the crucial role of audio–text alignment in generating high-quality, semantically accurate images in multi-source audio-to-image generation.

C. Effects of Multi-source Sound Separation

We pre-trained the Multi-source Sound Separation (MSS) model on FSD50K in stage 1 to enhance its audio representation capabilities. To investigate the effect of pre-training on downstream separation performance, we evaluated three configurations as shown in Tab. 5: (1) **Vanilla**, where only a reconstruction loss is applied; (2) **Pre-trained**, our default model that uses both reconstruction and audio–text alignment losses (ranking and contrastive loss); and (3) **Fine-tuned**, which builds on the pre-trained model with an additional 10 epochs of fine-tuning on the target dataset. We measure semantic alignment using the cosine similarity between the M' text label embeddings of the mixture and the M separated audio embeddings, and we report the average standard deviation of these scores on the target test set. Our results show that incorporating semantic alignment loss during pre-training significantly improves the semantic quality of the separated audios and that fine-tuning on the target dataset yields a comparable effect. These findings suggest that effective MSS pre-training is both powerful and sufficient for audio-to-image generation, potentially eliminating the need for additional fine-tuning on the target dataset.

D. More Qualitative Results

We present more qualitative results of MACS. Given a single audio or mixed audio, MACS generates images from

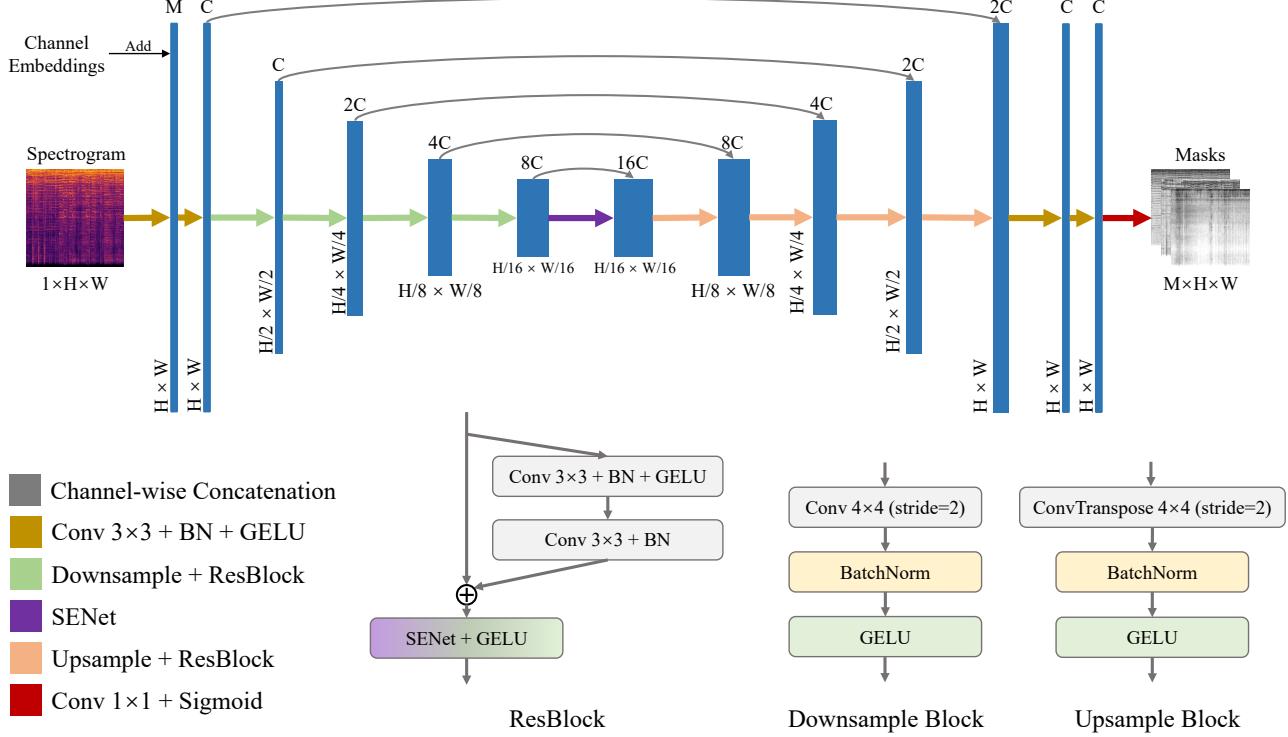


Figure 4. The overall architecture of UNet sound separation model of MACS.

Table 4. Ablation studies on LLP-multi comparing the impact of different components on the performance of the MACS model. The best results are **bold**, and the second best results are underlined.

Config	FID	CLIP-FID	KID	AIS	AIS-z	IIS	IIS-z
MACS (w/o RL)	<u>89.85</u>	<u>23.45</u>	<u>0.0179</u>	<u>0.0731</u>	1.1772	0.6044	1.5593
MACS (w/o CL)	95.33	25.02	0.0183	0.0704	1.1648	0.6067	1.4935
MACS (w/o DC)	92.24	23.85	0.0186	0.0719	<u>1.2026</u>	<u>0.6140</u>	<u>1.6610</u>
MACS	87.09	20.47	0.0157	0.0754	1.3038	0.6269	1.7231

different initial noises of the Stable Diffusion model in each row in the figures.

Single-source Audio. Results are presented in Fig. 5, Fig. 6. We can observe that the generated images accurately capture the semantics of single audio inputs, whether from natural sounds or object sounds. This demonstrates that MACS remains effective in producing high-quality single-source audio-to-image results.

Multi-source Audio. Results are presented in Fig. 7, Fig. 8 and Fig. 9. The images reveal that for double-source audio inputs, MACS produces high-quality images that accurately reflect the audio, with clear details (e.g., “*Taping guitar + speech*”). In more complex scenarios—such as those with three sound sources or crowd sounds—MACS captures the overall semantics accurately; however, details like faces and hands are less well-rendered (e.g., “*Brass Instrument + French Horn + Orchestra*”). These findings suggest that further improvements are needed for complex

scenarios. Enhancing multi-source audio-to-image generation based on MACS will likely require additional data and training power. Overall, MACS generates visually coherent and plausible images for multi-source audio inputs.

Audio-text Joint Image Generation. As mentioned in the method section, text input is optional, but it can further enrich the representations and enhance image quality. We set the text prompt to a non-empty value to evaluate MACS’s ability to generate images conditioned on both audio and text (see Fig. 10, Fig. 11, and Fig. 12). The results demonstrate that MACS effectively integrates textual information while preserving the semantic content of the audio, highlighting its flexibility in handling multimodal conditioning to produce visually coherent images that reflect both modalities. Additionally, the framework supports straightforward style transfer via text prompts, allowing users to modify the visual style of generated images while retaining their core semantic elements.

Table 5. The average standard deviation of the similarity scores between the text labels of each audio mixture and its separations across three datasets.

Config	Dataset		
	LLP-Multi	AudioSet-Eval	Landscape
Vanilla	0.0412	0.0410	0.0466
Pre-trained	0.0867	0.0810	0.0654
Finetuned	0.0903	0.0865	0.0689

E. Details of Pre-processing LLP-multi

For the preprocessing of the LLP dataset, we utilize the pre-trained model proposed by [45], which segments videos and parses them into distinct temporal audio, visual, and audio-visual events associated with semantic labels. Our objective is to obtain a frame that semantically aligns with the corresponding multi-source audio, enhancing the consistency between modalities.

Given a T -second video, the model generates audio event predictions $E_A \in [0, 1]^{T \times C}$ and visual event predictions $E_V \in [0, 1]^{T \times C}$, where C represents the number of classes in the dataset. Specifically, the video is split into one-second segments, and the model detects whether the object of each class is present in the audio or visual modality (1 for presence and 0 for absence), enabling a structured representation of multimodal event occurrences. To determine the co-existence of audio and visual events, we compute the element-wise multiplication of E_A and E_V , resulting in $E_{AV} = E_A \odot E_V$. Next, we select the frame from the seconds where the highest number of audio-visual co-existing events occurs. This step ensures that the selection process focuses on frames where both audio and visual events co-occur, narrowing the scope to identify the most representative frame in a video. Therefore, the selected frames and their multi-source audio are highly aligned.

F. Details of Selected Metrics

Fréchet Inception Distance (FID), CLIP-FID, and Kernel Inception Distance (KID) are quantitative measures that evaluate the quality of generated images by comparing the distributions of high-level features extracted from real and generated images. They differ in their mathematical formulations and the choice of embedding space. FID uses features from the Inception model, CLIP-FID leverages features from the CLIP model, and KID employs a kernel-based approach. We use all three metrics in this work to provide a comprehensive comparison.

For pairwise evaluation metrics in this work, we use Image-Image Similarity (IIS) and IIS-z for image pairs. We use Audio-Image Similarity (AIS) and AIS-z for image-audio pairs. These metrics aim to measure the similarities of the paired data using transformed representations, often by projecting to the same embedding space.

Note that these metrics are only comparable within the same dataset. Cross-comparison is not meaningful. For example, the difference between LLP-multi and AudioSet-Eval is much smaller than LLP-multi and Landscape because LLP-multi and AudioSet-Eval are both from AudioSet, while Landscape is different from them. We introduce the details of the metrics below.

- **Fréchet Inception Distance (FID):** FID [14] is a widely used metric for assessing the perceptual quality and diversity of generated images. It measures the realism of generated images by computing the difference in feature distributions between the generated and real images. These features are extracted from an Inception v3 [43] model, which captures high-level representations of image content. Specifically, FID assumes that the extracted features follow a multivariate normal distribution and computes the Fréchet distance between the feature distributions of real and generated images. Given the mean and covariance of the real image features, (μ_r, Σ_r) , and those of the generated image features, (μ_g, Σ_g) , FID is defined as:

$$\text{FID} = \|\mu_r - \mu_g\|^2 + \text{Tr}(\Sigma_r + \Sigma_g - 2(\Sigma_r \Sigma_g)^{\frac{1}{2}}), \quad (16)$$

where $\|\mu_r - \mu_g\|^2$ captures the difference in mean values, and the covariance term accounts for variations in feature distribution. Lower FID scores indicate greater similarity between real and generated images, signifying higher-quality synthesis.

- **CLIP-FID:** To provide an alternative evaluation aligned with multi-modal representations, we also employ the CLIP-FID metric, which replaces the Inception v3 model [43] with the CLIP image encoder. This modification ensures that the perceptual quality is assessed based on CLIP’s feature space, which is trained for semantic alignment between images and text.

- **Kernel Inception Distance (KID):** Similar to FID, KID [1] also quantifies the discrepancy between real and generated image distributions using features extracted from Inception v3. However, unlike FID, KID does not assume a normal distribution of features. Instead, it estimates the squared Maximum Mean Discrepancy (MMD) between feature distributions using a polynomial kernel function $k(x, y)$. KID is computed as:

$$\text{KID} = \mathbb{E}[k(x, x')] + \mathbb{E}[k(y, y')] - 2\mathbb{E}[k(x, y)], \quad (17)$$

where x, x' are feature vectors from real images, and y, y' are feature vectors from generated images. KID is advantageous in scenarios where the normality assumption does not hold, making it a more flexible measure of distributional discrepancy. Similar to FID, lower KID values indicate better alignment between real and generated image distributions, signifying higher visual fidelity.

- **Image-Image Similarity (IIS) and IIS-z:** Additionally, we adopt IIS [52], which measures the similarity between pairs of images by computing the CLIP-based similarity of their respective embedding vectors. This metric provides a direct evaluation of how close a generated image resembles its corresponding ground truth counterpart in the learned feature space.

To account for variations in overall similarity across different models, we also compute the average z-score of IIS, denoted as **IIS-z**. This normalization helps assess the relative prominence of a generated image’s similarity to its ground truth compared to other samples in the dataset. The z-score formulation ensures that the metric is not biased by the absolute similarity values, which may differ significantly depending on the model architecture or training approach. In practice, for a generated image $a_i \in \mathcal{G}$ and its corresponding ground truth $b_i \in \mathcal{T}$, we first compute the mean μ_i and standard deviation σ_i of similarity scores between a_i and all other ground truth images b_j , where $j \neq i$. The z-score for the generated image is then calculated as:

$$z_i = \frac{\text{SIM}(a_i, b_i) - \mu_i}{\sigma_i}, \quad (18)$$

where $\text{SIM}(a_i, b_i)$ represents the CLIP similarity between the generated image and its corresponding ground truth. By incorporating IIS-z, we ensure a more robust evaluation of model performance, as it highlights cases where a generated image exhibits significantly higher semantic alignment with its ground truth compared to random pairings within the dataset.

- **Audio-Image Similarity (AIS) and AIS-z:** AIS [52] is designed to evaluate the semantic alignment between input audio and generated images by measuring their similarity in a shared feature space. In our experiments, we utilize the Wav2CLIP model [50] as the audio encoder to extract feature representations from input audio signals. These embeddings are then compared with the corresponding image features obtained from a CLIP image encoder, providing a quantitative measure of how well the generated images preserve the semantic content of the input audio.

To ensure a more robust and interpretable evaluation, we also report the average z-score of AIS, denoted as **AIS-z**. This auxiliary metric does a similar calculation compared with IIS-z. By incorporating AIS-z, we provide a more relative measure of semantic alignment, enabling fairer comparisons between different approaches.

G. Details of Baselines

To evaluate the performance of audio-conditioned image generation, we compare our model with five state-of-the-art baselines.

AudioToken [52] leverages audio–image semantic alignment to generate images from audio inputs; however, it is primarily designed for single-source audio and struggles to disambiguate overlapping events in multi-source scenarios, which leads to ambiguity in the generated images.

Sound2Scene [42] enhances audio features by incorporating visual information, learning to align audio to a shared visual latent space. While this approach improves the semantic alignment between audio and visual modalities, the generation quality is constrained by the limitations of the GAN architecture. Specifically, Sound2Scene produces images with a resolution of 128×128 , which restricts the level of detail and fidelity in the generated images.

TempoTokens [53], originally an audio-to-video generation model, synthesizes dynamic visual content by maintaining temporal coherence. When adapted for audio-to-image tasks by extracting a representative mid-frame from generated videos, it fails to capture the full richness of simultaneous audio signals, thereby compromising performance in multi-source settings.

ImageBind [11] is a multimodal embedding model that cannot generate images directly, so we take advantage of InternGPT [26] combining it with Stable Diffusion 2.1 unCLIP. Specifically, ImageBind encodes the audio data into the shared embedding space, and the embeddings are subsequently fed into unCLIP as the condition.

CoDi [44] can generate corresponding images by conditioning on audio inputs, leveraging its ability to process and align multiple modalities. However, in multi-source audio scenarios, CoDi’s performance may be influenced by factors such as the potential challenges in maintaining coherence across generated modalities. Moreover, CoDi simply selects the mid-frames of videos as the training set, which may not always correspond to the most semantically relevant or representative frames for a given audio input.

We trained AudioToken and Sound2Scene on the selected datasets following the training protocols specified in their respective papers. For the remaining models, we utilized publicly released pre-trained weights. Specifically, we finetuned TempoTokens on AudioSet-Eval and LLP-Multi for 10 epochs with a learning rate of 1×10^{-5} . For the Landscape dataset, we directly used the provided pre-trained weights, as the model trained on this dataset is publicly available. Additionally, we employed the pre-trained weights of ImageBind and CoDi directly for inference without further finetuning.

H. Implementation Details

Input: We pre-process the raw audio by padding or truncating each clip to 8 seconds at a 16,000 Hz sample rate.

Stage 1: The UNet in the multi-source audio separation model comprises 5 layers with input channel sizes of (64, 64, 128, 256, 512) and corresponding output channel sizes of (1024, 512, 256, 128, 128), as shown in Fig. 4. We set $M = 6$ because most mixed audio clips contain no more than 6 sources.

The audio separation model is pre-trained on the FSD50K dataset for 10 epochs with a learning rate of 1e-3. To balance the different loss terms in Eq. (10), we employ a linear scheduling strategy for the loss weights:

$$\lambda = 1 - \frac{\text{#Current Epoch} - 1}{\text{#Total Epochs} - 1}, \quad \mu = \gamma = 1 - \lambda. \quad (19)$$

This scheduling approach ensures that the model initially prioritizes audio separation by emphasizing λ during the early epochs. As training progresses, the focus gradually shifts toward learning contextual importance and semantic alignment by increasing μ and γ .

Stage 2: To make a fair comparison with previous works [52], we use pre-trained Stable Diffusion 1.4 as the backbone generation model. During sampling, we use DPM-Solver [28] as the sampling method with steps of 25. The classifier-free guidance scale is set to 7.5 for our evaluation setup. We set the output embeddings of the audio separation model to zeros as the unconditional input.

We freeze the backbone weights and train the remaining components of our model on each dataset for 15 epochs using a learning rate of 1e-4. Specifically, the text prompt is left empty during both training and inference when generating images based solely on audio.

Optimization and GPUs: In both stages, AdamW [27] optimizer is adopted with β_1 of 0.9, β_2 of 0.999 and weight decay of 1e-2. We use an equivalent batch size of 16 with gradient accumulation. All experiments are conducted on a single NVIDIA RTX 4090D GPU.

Protocols: For LLP-Multi and AudioSet-Eval, we use 5-fold cross-validation and report the average results. For the Landscape dataset, we follow the train-test split from [40], where 90% of the data is for training and 10% is for testing.

I. Preliminary on Latent Diffusion Model

Diffusion models [15] are generative models that learn a data distribution $p(\mathbf{x})$ through a forward process of gradually corrupting data samples $\mathbf{x} \sim p(\mathbf{x})$ with Gaussian noise and a learned reverse denoising process. While effective for high-quality image synthesis, their computational cost scales with the iterative refinement required in pixel space.

Latent Diffusion Models (LDMs) [38] address this inefficiency by operating in a lower-dimensional latent space.

By training an autoencoder with encoder \mathcal{E} and decoder \mathcal{D} , LDMs map images \mathbf{x} to compressed latents $\mathbf{z} = \mathcal{E}(\mathbf{x})$, where semantically meaningful features are preserved. The forward process gradually adds noise to \mathbf{z} over T steps via a variance schedule β_1, \dots, β_T :

$$q(\mathbf{z}_t | \mathbf{z}_{t-1}) = \mathcal{N}(\mathbf{z}_t; \sqrt{1 - \beta_t} \mathbf{z}_{t-1}, \beta_t \mathbf{I}). \quad (20)$$

To be expressed in closed form [15], this becomes:

$$q(\mathbf{z}_t | \mathbf{z}) = \mathcal{N}(\mathbf{z}_t; \sqrt{\bar{\alpha}_t} \mathbf{z}_{t-1}, (1 - \bar{\alpha}_t) \mathbf{I}), \quad (21)$$

where $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{i=1}^t \alpha_i$.

During training, a denoising network, typically a UNet [39], is optimized to predict the noise ϵ added to \mathbf{z}_t :

$$\mathcal{L} = \mathbb{E}_{\mathbf{z}, \epsilon, t} \|\epsilon - \epsilon_\theta(\mathbf{z}_t, t, c)\|_2^2, \quad (22)$$

where ϵ_θ represents the noise prediction model parameterized by θ , and c is an optional conditioning variable. While LDMs like Stable Diffusion are widely used with text prompts, their conditioning mechanism is inherently multimodal – c can be represent diverse inputs such as class labels, text embeddings, or even audio signals.

The reverse process leverages the learned ϵ_θ to iteratively denoise \mathbf{z}_T :

$$q(\mathbf{z}_{t-1} | \mathbf{z}_t) = \mathcal{N}(\mathbf{z}_{t-1}; \mu_\theta(\mathbf{z}_t, t, c), \Sigma_\theta(\mathbf{z}_t, t, c)), \quad (23)$$

where μ_θ and Σ_θ are derived from ϵ_θ . Crucially, conditioning ϵ_θ on audio features allows LDMs to synthesize images semantically aligned with acoustic input. In the second stage of our proposed MACS, audio signals are processed into c , enabling the model to generate visual content that reflects the acoustic context. After T denoising steps, the refined latent $\hat{\mathbf{z}}$ is decoded to pixel space via $\hat{\mathbf{x}} = \mathcal{D}(\hat{\mathbf{z}})$.

This extension underscores the flexibility of LDMs in integrating diverse modalities. By integrating audio as a conditioning input within the diffusion process, we demonstrate that LDMs can effectively bridge auditory and visual domains, enabling robust audio-to-image synthesis tasks.

J. Ethical Considerations

Like other image-generation frameworks, audio-to-image methods raise ethical concerns that warrant careful evaluation. Potential issues include privacy and security risks, as the technology could be misused to produce misleading or deceptive images from audio inputs, potentially contributing to misinformation or unintended identity reconstruction. Additionally, biases in the generated images—stemming from the training data—should be addressed to prevent discrimination. In our work, the datasets do not contain identifiable or private information, and the generated images do not pose security risks. However, misuse of MACS could still lead to privacy and security issues, and we urge users to mitigate these risks and refrain from misuse.

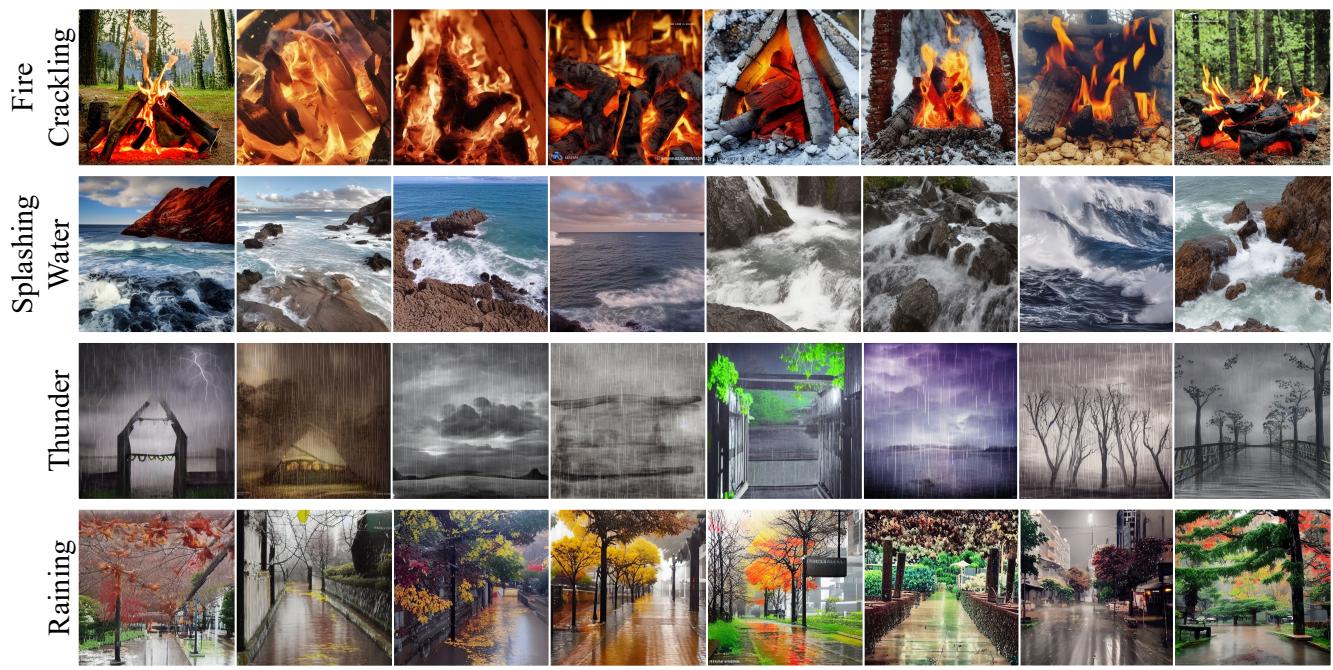


Figure 5. Qualitative results generated from single-source audios. (1/2)

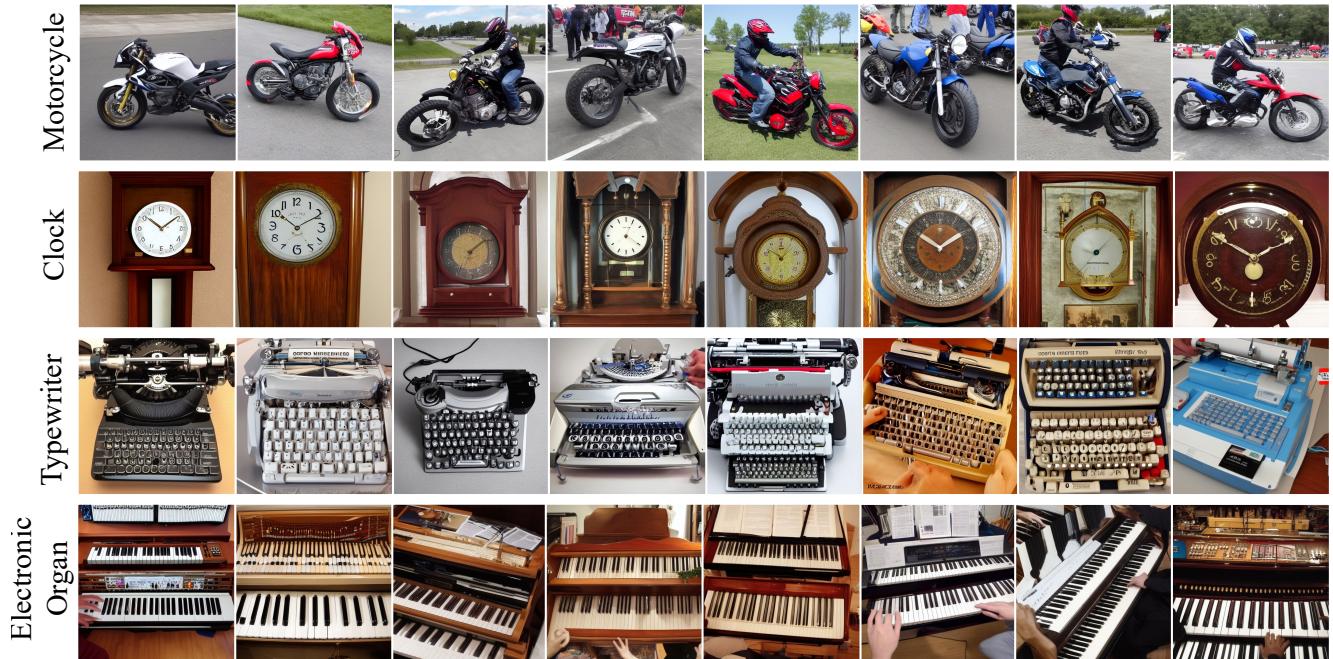


Figure 6. Qualitative results generated from single-source audios. (2/2)

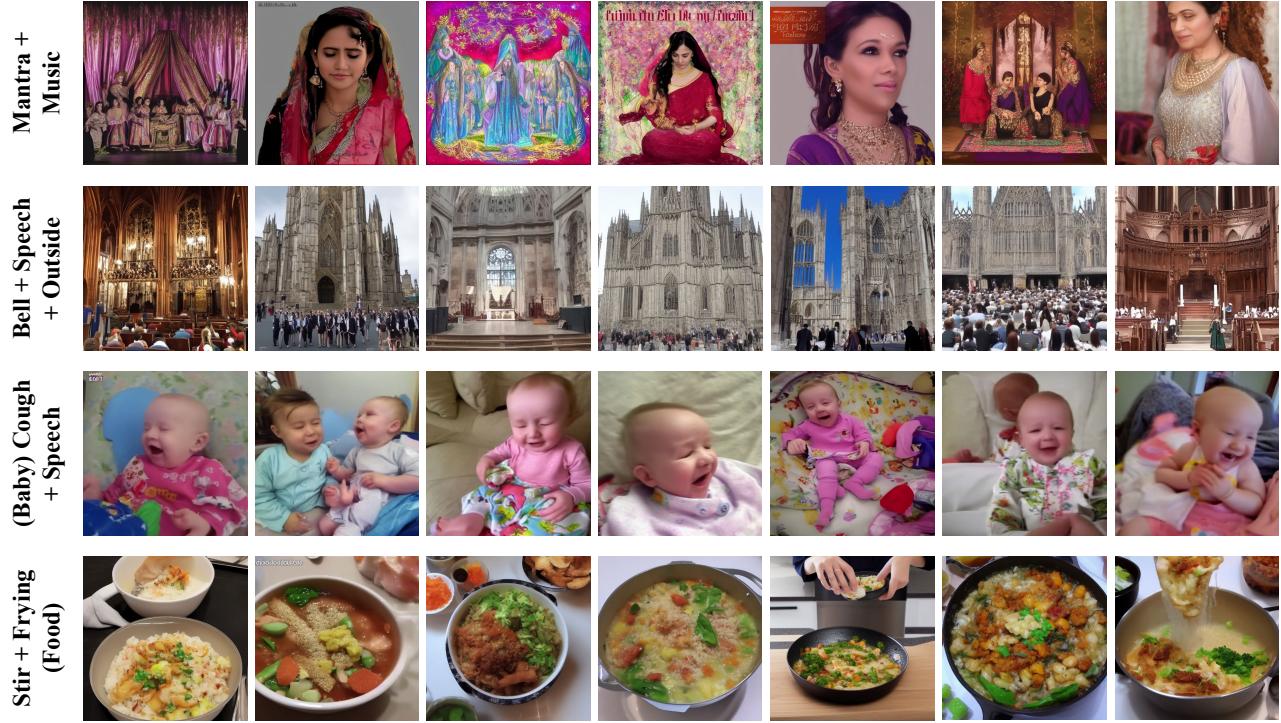


Figure 7. Qualitative results generated from multi-source audios. (1/3)

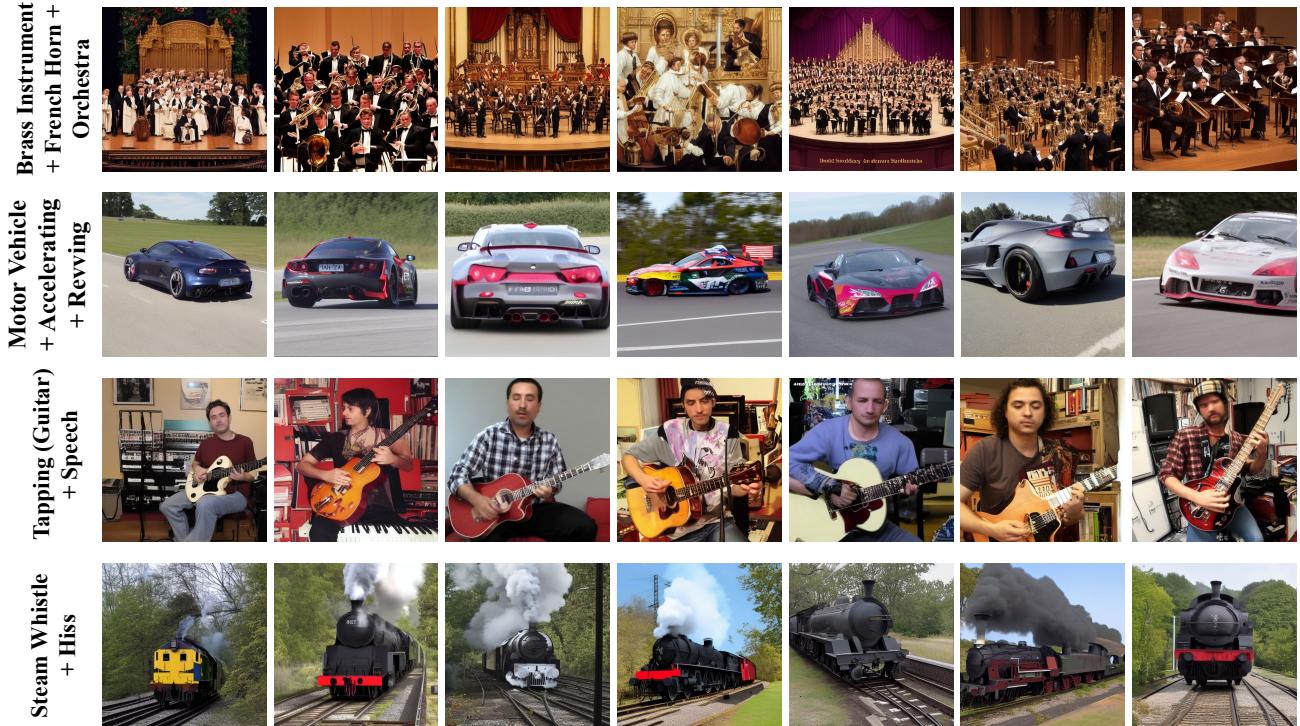


Figure 8. Qualitative results generated from multi-source audios. (2/3)

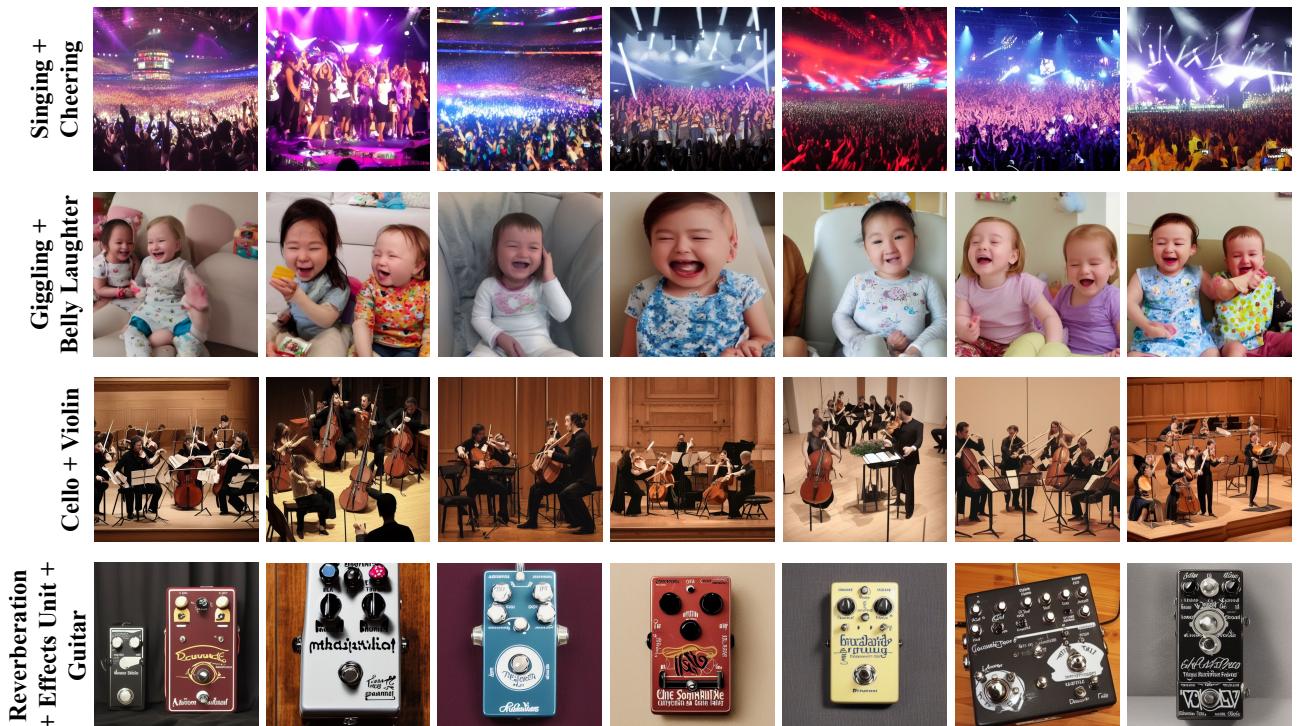


Figure 9. Qualitative results generated from multi-source audios. (3/3)

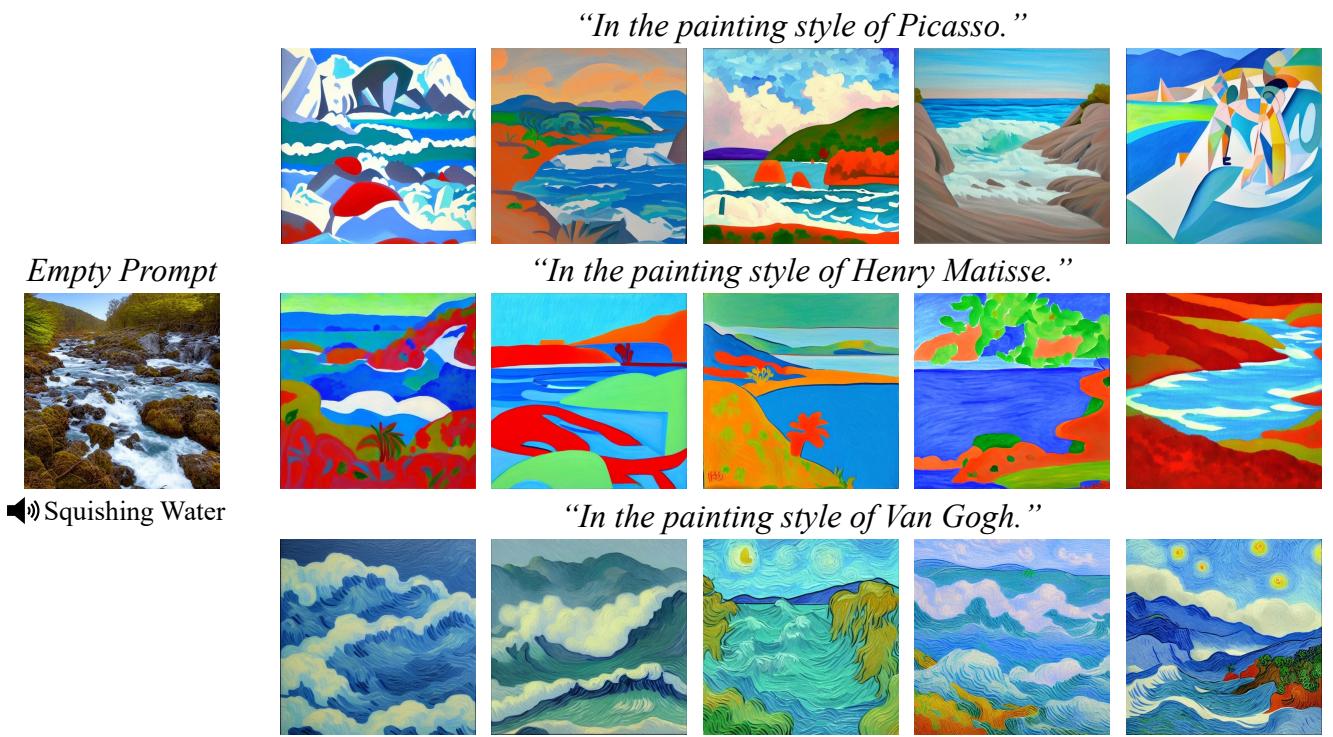


Figure 10. Qualitative results jointly generated from audios and texts. (1/3)

“Seen from the window.”

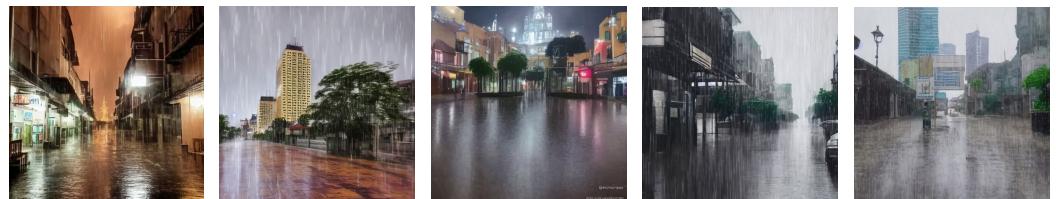


Empty Prompt



🔊 Raining

“Downtown area.”

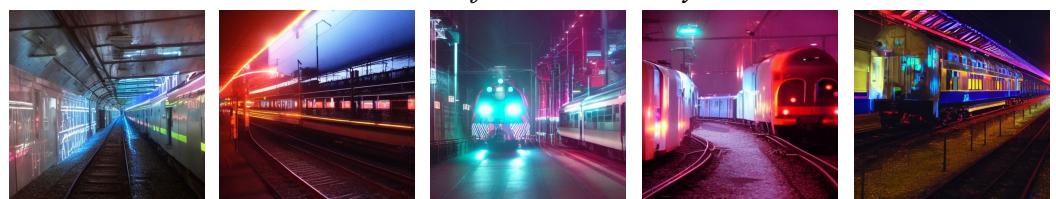


“A running dog.”



Figure 11. Qualitative results jointly generated from audios and texts. (2/3)

“Neon-lit, futuristic, moody.”



Empty Prompt



🔊 Steam Whistle
🔊 Hiss

“In the style of LEGO.”



“Raining heavily.”



Figure 12. Qualitative results jointly generated from audios and texts. (3/3)