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

Hao Zhou[1]*, **Xiaobao Guo**[1]*, **Yuzhe Zhu**[1], **Adams Wai-Kin Kong**[1]

[1] Nanyang Technological University

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

Abstract

Propelled by the breakthrough in deep generative models, audio-to-image generation has emerged as a pivotal cross-modal 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 generating comprehensive visual content. To bridge this gap, we propose a method called MACS to conduct multi-source audio-to-image generation. To our best knowledge, 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, effective image generation is achieved by mapping 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 out of the 21 evaluation indexes on all tasks and delivers superior visual quality. Code available at https://github.com/alxzzhou/MACS.

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 (Zhou et al. 2019; Chen et al. 2017; Duarte et al. 2019; Chatterjee and Cherian 2020). Recently, inspired by the success of diffusion models (Ho, Jain, and Abbeel 2020) and multimodal learning (Radford et al. 2021; Wu et al. 2023; Jia et al. 2021), more research works have demonstrated that models originally designed for text-to-image synthesis can be successfully adapted for audio inputs (Yariv et al. 2023; Qin et al. 2023), making it easier to develop an audio-to-image generation model. Audio-to-image generation is useful in many applications such as creative arts (Lee et al. 2023), multimedia content generation (Girdhar et al. 2023), and enhanced teaching and learning experiences by generating immersive visuals from sound by VR or AR systems (Fitria 2023).

Despite the recent success in conditioning image synthesis on audio, most of the existing literature focuses on single-source audio inputs (Yariv et al. 2023; Sung-Bin et al. 2023; Qin et al. 2023). In contrast, natural soundscapes are complex, often composed of multiple overlapping audio sources. Previous methods fall short on mixed audio signals; for example, they fail to effectively combine sounds like "guitar and speech," resulting in incoherent images (see Fig. 1). These methods struggle to disentangle and utilize the full in-

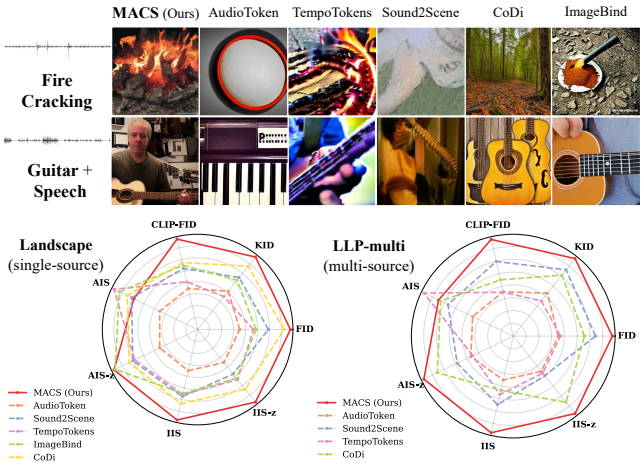

Figure 1: **Qualitative and Quantitative Comparison of MACS and Other SOTA Methods. Upper:** Generated images from single-source and multi-source audio datasets. **Lower:** Radar maps illustrating performance on single-source (left) and multi-source (right) datasets. Values are normalized, and the lower-is-better metrics (FID, CLIP-FID, and KID) are inverted for consistency.

*These authors contributed equally.

formation embedded in real-world auditory scenes, limiting their ability to generate contextually comprehensive images.

To bridge the gap between multi-source audio and image generation, we propose **MACS**, the first framework that explicitly tackles multi-source audio-to-image generation. Our core idea is simple yet effective: "separation before generation." Rather than directly mapping complex mixtures to images, MACS first disentangles mixed audio into individual sources and then synthesizes an image that captures their combined semantics. MACS is a two-stage framework that addresses three main challenges: **1)** Mixed audio separation. Overlapping audio sources must be disentangled using robust separation techniques that preserve each source's unique characteristics; **2)** Contextual significance and semantic alignment. The semantic content and relative importance of each separated source must be maintained and appropriately balanced to ensure coherent visual synthesis; **3)** Multi-source conditioning in diffusion. The system needs to map multiple concurrent audio representations to a single visual output using a diffusion model, where the overall scene can be generated effectively.

Specifically, we propose a multi-source audio separation model based on UNet (Ronneberger, Fischer, and Brox 2015). For semantic alignment, we project individual audio signals and their corresponding labels into the CLAP (Wu et al. 2023) space using a contrastive loss. 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. Analysis in Fig. 6 shows that separated audio embeddings can produce more localized and semantically aligned attention maps. In the second stage, the individual audio signals are transformed and mapped to a visual output through the diffusion process, where we use the trainable decoupled cross-attention module (Ye et al. 2023) and an MLP layer for effective mapping while keeping the rest of the model frozen.

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 effectively 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 models rely on large datasets with isolated ground-truth sources (Wang and Chen 2018; Luo and Yu 2023), but are often limited to domains like speech and music, where clean source data is available. Unsupervised methods, such as PIT (Yu et al. 2017) and MixIT (Wisdom et al. 2020), use unlabeled mixtures to learn representations, but require post-selection (e.g., a trained classifier) to identify separated sources. MixPIT (Karamatlı and Kırbız 2022) handles mixture-of-mixtures inputs directly but is limited in the number of separable sources. To overcome the limitations above, the weakly-supervised approaches explore large-scale mixture datasets and rely on high-level semantic information rather than exact source ground truth for guidance (Pishdadian, Wichern, and Le Roux 2020). However, most prior research works study vision- or text-conditioned audio separation (Dong et al. 2023; Mahmud et al. 2024). 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 Multimodal Contrastive Pretraining and Audio-conditioned Image Generation

Contrastive multimodal pretraining has become a cornerstone for learning aligned representations across text, image, and audio modalities (Wang et al. 2023; Radford et al. 2021; Chen et al. 2020; Kim, Son, and Kim 2021; Jia et al. 2021; Li et al. 2023). Pioneering works like CLIP (Radford et al. 2021) train dual encoders with a contrastive loss on large-scale image-text pairs, enabling powerful cross-modal understanding. This framework has since been extended to other modalities. AudioCLIP (Guzhov et al. 2022) and Wav2CLIP (Wu et al. 2022) introduce audio into the CLIP architecture, while CLAP (Wu et al. 2023) learns a joint embedding from over 600K audio-caption pairs for audio-text tasks. These aligned spaces facilitate diverse downstream applications, such as classification (Wu et al. 2022; Guzhov et al. 2022) and sound-guided image manipulation (Lee et al. 2022, 2024).

Building on these embeddings, beyond the popular text-to-image generation (Li et al. 2019; Reed et al. 2016; Rombach et al. 2022), recent works explore audio-conditioned image generation (Chatterjee and Cherian 2020; Oh et al. 2019). Early attempts using GANs (Wan, Chuang, and Lee 2019) faced limitations in diversity and fidelity. More recent methods leverage diffusion models (Rombach et al. 2022) due to their superior stability and quality. For instance, AudioToken (Yariv et al. 2023) maps audio to discrete tokens compatible with Stable Diffusion, while ImageBind (Girdhar et al. 2023) learns a unified embedding space for multiple modalities including audio, enabling generation via unCLIP (Ramesh et al. 2022). Adapter-based diffusion models (Mou et al. 2024; Ye et al. 2023) further enhance conditional generation with modular and effective conditioning.

While prior works in audio-to-image generation focus on single-source audio, we tackle the more complex and realistic setting of mixed audio inputs. Unlike methods that

directly map entire audio scenes to images, we first separate audio sources before generation, allowing finer alignment between sound events and visual content. Our approach leverages the strong alignment capabilities of CLAP and diffusion models to enable image generation from rich, multi-source acoustic scenes.

3 Methods

The proposed MACS framework introduces a two-stage architecture for generating semantically meaningful images from multi-source audio. By adopting a "*separation before generation*" strategy, MACS effectively disentangles and recombines complex audio signals, leading to more accurate and expressive image generation. The brief description of MACS can be found in the caption of Fig. 2.

3.1 Multi-source Audio Separation

Problem Formulation Given a multi-source audio mixture $\mathbf{m}$, the separation model $\mathcal{G}_\theta$ estimates M binary masks using a UNet $\mathcal{U}_\theta$. First, $\mathbf{m}$ is mapped to its spectrogram $\mathcal{T}(\mathbf{m})$ via Short Time Fourier Transform (STFT), and the UNet predicts masks from the magnitude $|\mathcal{T}(\mathbf{m})|$. Each mask is applied element-wise to the magnitude, and the masked spectrograms are converted back to waveforms using the inverse STFT (iSTFT) with the original phase $\phi(\mathcal{T}(\mathbf{m}))$. Formally,

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

The UNet operates on magnitude only, as phase is often unnecessary for many audio tasks (Mahmud et al. 2024), and is used solely for waveform reconstruction.

Mixed Audio Separation Conventional audio separation trains on synthetic mixtures with single-source ground truth, but often fails on real-world audio. Inspired by MixIT (Wisdom et al. 2020), we use an unsupervised Mixture of Mixtures (MoM) input and optimize a reconstruction loss to recover constituent sources. Like (Dong et al. 2023; Mahmud et al. 2024), we apply UNet-based separation on spectrograms; however, our method is fully *unconditional*, requiring no auxiliary inputs to the UNet.

Formally, given two mixtures $\boldsymbol{m}_1$ and $\boldsymbol{m}_2$, a new mixture of mixtures $\boldsymbol{m} = \boldsymbol{m}_1 + \boldsymbol{m}_2$ is formed. The separation model $\mathcal{G}_{\boldsymbol{\theta}}(\boldsymbol{m})$ produces M separated signals $\boldsymbol{S} = \{\boldsymbol{s}_1, \ldots, \boldsymbol{s}_M\}$, whose sum reconstructs $\boldsymbol{m}$. To train the model, a reconstruction loss compares $\boldsymbol{S}$ and the original mixtures $\boldsymbol{m}_1$ and $\boldsymbol{m}_2$. Since $M > 2$ and source order is ambiguous, all possible bipartitions of $\boldsymbol{S}$ are evaluated. The loss selects the partition (Λ_1, Λ_2) that minimizes the total reconstruction error:

$$\begin{aligned}\mathcal{L}_{Rec} = \min_{(\Lambda_1,\Lambda_2)\in\Lambda} [&\mathcal{L}_{SISDR}(\boldsymbol{m}_1, \sum_{i\in\Lambda_1} \boldsymbol{s}_i) \\ &+\mathcal{L}_{SISDR}(\boldsymbol{m}_2, \sum_{i\in\Lambda_2} \boldsymbol{s}_i)],\end{aligned} \tag{1}$$

where Λ represents the set of all non-empty, disjoint bipartitions of the index set $A = \{1, 2, \ldots, M\}$:

$$\Lambda = \{(\Lambda_1, \Lambda_2) \mid \Lambda_1 \cup \Lambda_2 = A,\ \Lambda_1 \cap \Lambda_2 = \emptyset,\ \Lambda_1, \Lambda_2 \neq \emptyset\}. \tag{2}$$

The reconstruction loss $\mathcal{L}_{SISDR}$ is measured by the negative scale-invariant signal-to-distortion ratio (SI-SDR) (Le Roux et al. 2019):

$$\mathcal{L}_{SISDR}(\boldsymbol{m}_j, \hat{\boldsymbol{s}}_j) = -10\log_{10}\frac{\|\alpha\boldsymbol{m}_j\|_2^2}{\|\alpha\boldsymbol{m}_j - \hat{\boldsymbol{s}}_j\|_2^2}, \tag{3}$$

where $\hat{\boldsymbol{s}}_j = \sum_{i\in\Lambda_j} \boldsymbol{s}_i, j = \{1, 2\}$, and the scaling factor α is $\frac{\hat{\boldsymbol{s}}_j^\top \boldsymbol{m}_j}{\|\boldsymbol{m}_j\|_2^2}$. By minimizing the best-matching bipartition loss, the model learns to reconstruct the original mixtures without relying on fixed source assignments.

3.2 Audio-text Alignment

While the separation model in section 3.1 captures rich audio features, its unsupervised training lacks the high-level semantic alignment required for audio-to-image generation. To address this, we employ CLAP (Wu et al. 2023) to project separated signals, mixed audio, and text labels into a shared embedding space via an audio encoder $\mathcal{P}_A$ and a language encoder $\mathcal{P}_L$. We target two alignment objectives: (1) the contextual significance of each separated signal within the mixture, and (2) its semantic correspondence to its text label.

Specifically, let $\mathcal{S} = [\mathbf{s}_1, \ldots, \mathbf{s}_M]$ be the separated signals and $\mathcal{T} = [t_1, \ldots, t_{M'}]$ their labels. Each label t_i is prefixed with *"The sound of"*; if $M' < M$, we pad $\mathcal{T}$ with *"Noise"* to length M. We then obtain

$$\mathcal{E}^A = \mathcal{P}_A(\mathcal{S}) \in \mathbb{R}^{M\times D}, \quad \mathcal{E}^T = \mathcal{P}_L(\mathcal{T}) \in \mathbb{R}^{M\times D},$$

where D denotes the embedding dimension. This procedure ensures both contextual and semantic alignment between audio signals and their textual descriptions.

Ranking Loss We adopt a ranking loss to capture the contextual significance of separated audio sources within a mixture, helping the model identify which components are more semantically important. In real-world audio, some elements in a mixture carry greater importance. For example, background noise is often less relevant than distinct sounds like a dog bark or music. However, separation outputs are unordered, lacking inherent prioritization and making the ranking loss essential for guiding the model's focus.

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

$$\mathcal{L}_{Rank} = 1 - r_s(\mathbf{S}, \mathrm{Sorted}(\mathbf{S})), \tag{4}$$

where $r_s(\cdot,\cdot)$ represents the Spearman's rank correlation coefficient (Spearman 1987) 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, $\mathcal{P}_A(\boldsymbol{m}) \in \mathbb{R}^D$, and the separated audio embeddings, $\mathcal{E}^A \in \mathbb{R}^{M\times D}$. Function $\mathrm{Sorted}(\cdot)$ outputs the sorted array in descending order. There is no strict requirement for the choice of sorting function. To ensure differentiability during training, we adopt the ranking optimization method from (Blondel et al. 2020), which allows direct optimization of ranking functions in deep models. This ranking loss guides the model to identify and prioritize important audio sources, enhancing its ability to preserve key semantic information.

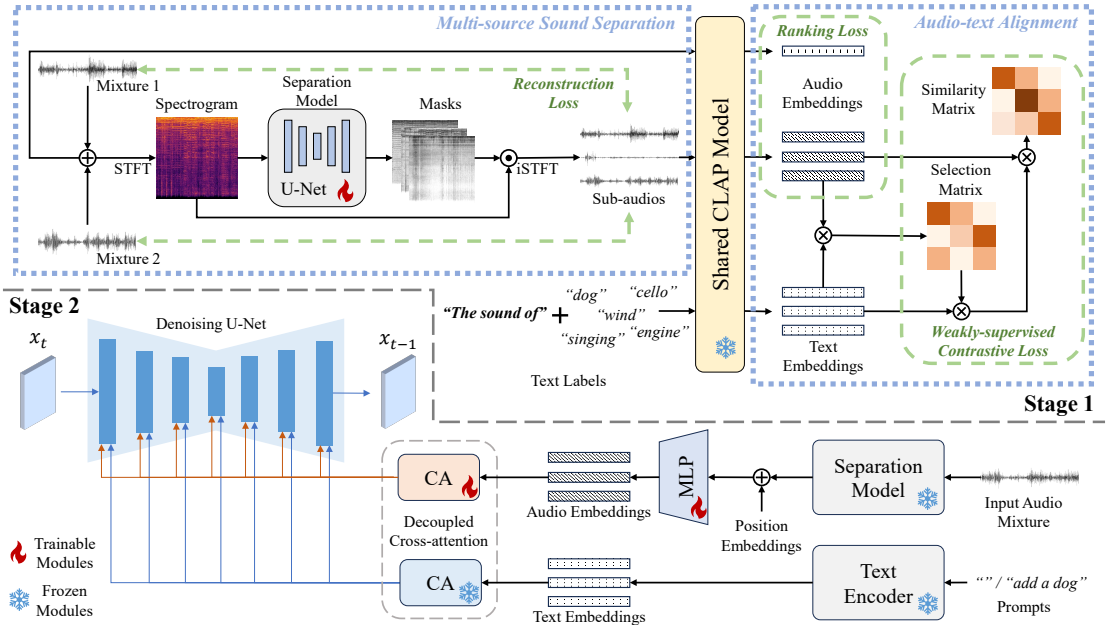

Figure 2: **An overview of the proposed two-stage MACS architecture.** *Stage 1*: A Multi-source Sound Separation (MSS) model decomposes audio mixtures into sub-audios using reconstruction loss. The separated audios are embedded with CLAP, guided by contrastive and ranking losses to ensure audio-text semantic alignment and contextual significance. *Stage 2*: A diffusion-based generator uses a decoupled cross-attention module to integrate audio embeddings and produce high-quality, semantically accurate images. MACS enables scalable MSS pre-training and image generation with fewer trainable layers.

Contrastive Loss Our second focus is aligning separated audio signals with their text labels, which is crucial for image generation. To compensate for missing semantic cues in unsupervised learning, we apply a contrastive loss to align each audio with its corresponding text label.

Each separated audio s_i is expected to semantically align with one of its associated text labels. However, explicit label information is not available at this stage, as the model receives only audio data as input. To address this, we perform a *soft assignment* in a joint embedding space by aligning audio and text embeddings as follows:

$$\mathcal{E}'^{T} = \text{Softmax}\left(\frac{\langle \mathcal{E}^{A} \rangle \langle \mathcal{E}^{T} \rangle^{\top}}{\tau}\right)\mathcal{E}^{T}, \quad \mathcal{E}'^{T} \in \mathbb{R}^{M\times D}, \tag{5}$$

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, enabling semantic consistency without enforcing rigid one-to-one assignments.

To align audio and text embeddings, we employ the contrastive loss (Radford et al. 2021):

$$\begin{aligned}\mathcal{L}_{CL} = &-\frac{1}{2M}\sum_{i=1}^{M}\log\frac{\exp(W_{ii})}{\sum_{j=1}^{M}\exp(W_{ij})} \\ &-\frac{1}{2M}\sum_{i=1}^{M}\log\frac{\exp(W_{ii})}{\sum_{j=1}^{M}\exp(W_{ji})},\end{aligned} \tag{6}$$

where the similarity matrix

$$W = \frac{\langle \mathcal{E}^{A} \rangle \langle \mathcal{E}'^{T} \rangle^{\top}}{\tau'} \in \mathbb{R}^{M\times M}, \tag{7}$$

computes cosine similarities, and τ' is a learnable temperature. This loss pulls matched audio–text pairs together and pushes mismatches apart, enhancing semantic alignment.

The overall training objective for the first stage is:

$$\mathcal{L}_1 = \lambda\mathcal{L}_{Rec} + \mu\mathcal{L}_{CL} + \gamma\mathcal{L}_{Rank}, \tag{8}$$

where λ, μ and γ are weights of the losses. In this stage, the model is pre-trained on audio and text labels from large-scale annotated datasets, learning reliable audio embeddings that generalize effectively to the image generation task.

3.3 Multi-source Audio-to-image Generation

To leverage text-to-image models such as Stable Diffusion, we adopt a decoupled cross-attention module (Ye et al. 2023), originally designed as a lightweight adapter for image–text fusion, and extend it to handle multiple audio inputs. Given M audio embeddings $\mathcal{E}^{A} \in \mathbb{R}^{M\times D}$, we first add trainable positional embeddings $\mathcal{E}^{Pos} \in \mathbb{R}^{M\times D}$ and project the sum into the conditioning dimensionality D' via a multi-layer perceptron (with layer normalization):

$$\mathcal{E}'^{A} = \text{MLP}\left(\mathcal{E}^{A} + \mathcal{E}^{Pos}\right) \in \mathbb{R}^{M\times D'}. \tag{9}$$

For UNet query features $\mathbf{H}$, the audio cross-attention is

$$\mathbf{H}_A = \text{Softmax}\left(\frac{(\mathbf{H}\mathbf{W}_q)(\mathcal{E}'^{A}\mathbf{W}_k)^{\top}}{\sqrt{D'}}\right)(\mathcal{E}'^{A}\mathbf{W}_v), \tag{10}$$

where $\mathbf{W}_q$ is shared across modalities, and $\mathbf{W}_k, \mathbf{W}_v$ are newly initialized.

The module also supports optional text conditioning. When a prompt is provided, CLIP encodes it as $\mathcal{E}^{P}$, yielding

$$\mathbf{H}_T = \text{Softmax}\left(\frac{(\mathbf{H}\mathbf{W}_q)(\mathcal{E}^{P}\mathbf{W}'_k)^{\top}}{\sqrt{D'}}\right)(\mathcal{E}^{P}\mathbf{W}'_v). \tag{11}$$

The combined output is

$$\mathbf{H}' = \mathbf{H}_A + \mathbf{H}_T. \tag{12}$$

In the second stage of training, only $\mathbf{W}_k$, $\mathbf{W}_v$, $\mathcal{E}^{Pos}$, and the MLP parameters are updated, while the base model remains frozen. We optimize using the Stable Diffusion loss:

$$\mathcal{L}_2 = \mathbb{E}_{\mathbf{z},\epsilon,t}\left\|\epsilon - \epsilon_\theta(\mathbf{z}_t, t, c)\right\|_2^2, \tag{13}$$

where $c = \{\mathcal{E}'^A, \mathcal{E}^P\}$. In our work, we extend the decoupled cross-attention module to audio inputs, enabling mapping multiple audio inputs—and optionally text—into a single generated image.

4 Experiments

4.1 Datasets

LLP-multi. To address the lack of multi-source benchmarks, we construct LLP-multi from the LLP dataset (Tian, Li, and Xu 2020), a subset of AudioSet (Gemmeke et al. 2017). We select videos with multiple labels and extract 6,595 frames with high audio-visual coexistence (6,314 with 2 labels, 242 with 3, and 35 with 4). LLP-multi captures concurrent audio-visual events with corresponding annotations, making it well-suited for *multi-source tasks*.

AudioSet-Eval. AudioSet (Gemmeke et al. 2017) contains over 2 million videos across diverse sound categories such as human voices, animal noises, and music. We use its evaluation split, filtering out 20 poor-quality clips, resulting in 15,712 samples, 20.7% with a single label and 79.3% with multiple labels. With 610 distinct classes, this diverse set is used for *mixed-source evaluation*.

Landscape. The Landscape dataset (Lee et al. 2022), widely adopted in audio-to-image generation tasks, features 1,000 natural scene videos, each with one audio event from 9 classes. Following prior work (Ruan et al. 2023), we use a 90/10 train-test split for *single-source evaluation*.

FSD50K. FSD50K (Fonseca et al. 2021) includes over 51,000 manually labeled audio clips across 200 classes from the AudioSet ontology. *FSD50K is used to pre-train the audio separation model in the first stage of MACS.*

4.2 Training Setup

We use AdamW (Loshchilov and Hutter 2019) in both stages with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and a weight decay of 1e-2. Training uses a batch size of 16 (with gradient accumulation) on one RTX 4090D GPU. For LLP-multi and AudioSet-Eval, we report average results from 5-fold cross-validation. For Landscape, we follow the 90/10 train-test split in (Ruan et al. 2023). See Appendix G for more details.

4.3 Evaluation Metrics

For a comprehensive performance evaluation, we gathered **seven metrics** for quantitative evaluation. **(a)** the overall quality of the generated images including *Fréchet Inception Distance (FID)* (Heusel et al. 2017), *CLIP-FID*, and *Kernel Inception Distance (KID)* (Bińkowski et al. 2018), **(b)** pairwise similarity between generated images and ground truth images including *Image-Image Similarity (IIS)* (Yariv et al. 2023) and *IIS-z*, where "z" means using z-score, and **(c)** pairwise semantic similarity between audio and images including *Audio-Image Similarity (AIS)* (Yariv et al. 2023) and *AIS-z*, "z" for z-score. More details are in Appendix E.

4.4 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 methods: AudioToken (Yariv et al. 2023), Sound2Scene (Sung-Bin et al. 2023), TempoTokens (Yariv et al. 2024), ImageBind (Girdhar et al. 2023), and CoDi (Tang et al. 2023). 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 (Girdhar et al. 2023), its performances in Tab. 1 and Tab. 2 are presented in gray for *reference only*. The results are averaged over 5-fold cross-validation.

Multi-source Audio. We benchmarked MACS against five SOTA methods on **LLP-multi** (see Tab. 1), a *fully* multi-source audio dataset. MACS significantly improves image quality (left three metrics) and further enhances content fidelity and semantic consistency across multiple sources (right four 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 evaluate MACS on the mixed-source AudioSet-Eval dataset, which is more challenging than LLP-multi with over 600 event classes. Despite the increased complexity, MACS consistently generates high-quality images. As shown in Tab.1 and Tab.2, overall image quality improves on mixed-source datasets. Unlike baselines that overlook audio mixing, MACS handles both the single- and multi-source inputs, making it versatile and scalable to an arbitrary number of audio sources.

Single-source Audio. MACS also performs strongly on Landscape, achieving state-of-the-art results on most metrics (Tab. 3) and outperforming the next-best method by a substantial margin. Compared to 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 evaluate MSS, we integrated MACS stage 1 outputs into AudioToken (Yariv et al. 2023), which transforms audio clips into embeddings concatenated with the prompt "**A photo of a...**" for Stable Diffusion. We extended it to produce M audio tokens from the M separated signals, denoting this variant as "AudioToken (w/MSS)" (see Tab. 1–3). Compared to the original, MSS markedly enhances performance, showing MACS's effectiveness and adaptability.

Audio Separation Model is Flexible. We evaluated model performance using an alternative separation mechanism by replacing our module with MixIT and assess-

Method	FID↓	CLIP-FID↓	KID↓	AIS↑	AIS-z↑	IIS↑	IIS-z↑
ImageBind* (Girdhar et al. 2023)	76.81	21.17	0.0088	0.0885	1.4219	0.6127	2.0361
AudioToken (Yariv et al. 2023)	143.62	52.21	0.0431	0.0591	0.6201	0.4914	0.6799
Sound2Scene (Sung-Bin et al. 2023)	105.14	33.79	0.0240	0.0711	0.8176	0.5545	0.7877
TempoTokens (Yariv et al. 2024)	141.37	52.45	0.0494	**0.0828**	0.5932	0.5288	0.7259
CoDi (Tang et al. 2023)	116.67	44.96	0.0283	0.0747	1.1068	0.5179	1.4429
AudioToken (Yariv et al. 2023) (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 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.

Method	FID↓	CLIP-FID↓	KID↓	AIS↑	AIS-z↑	IIS↑	IIS-z↑
ImageBind* (Girdhar et al. 2023)	41.69	14.76	0.0083	0.0892	1.1498	0.5808	1.7928
AudioToken (Yariv et al. 2023)	102.85	40.68	0.0397	0.0663	0.5664	0.5426	0.6829
Sound2Scene (Sung-Bin et al. 2023)	63.94	26.61	0.0207	0.0725	0.7310	0.5445	0.6837
TempoTokens (Yariv et al. 2024)	108.73	45.37	0.0510	**0.0879**	0.3863	0.5335	0.5153
CoDi (Tang et al. 2023)	70.20	31.49	0.0206	0.0789	**1.0128**	0.4920	1.0869
AudioToken (Yariv et al. 2023) (w/ MSS)	96.93	37.67	0.0305	0.0702	0.6218	0.5479	0.7924
MACS	**62.40**	**19.65**	**0.0142**	0.0724	0.8736	**0.5532**	**1.1328**

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.

ing generation quality on the LLP-multi dataset (Tab. 4). Although MixIT operates at the waveform level, our spectrogram-based separator outperforms it on every metric, underscoring its advantages. Nevertheless, MixIT remains competitive, demonstrating the audio-separation component's versatility within the MACS framework.

Pre-training Facilitates Multi-source Sound Separation. We pre-trained the Multi-source Sound Separation (MSS) model on FSD50K in stage 1 to enrich its audio representations. To assess the impact of pre-training on separation performance, we evaluated three configurations (see Fig. 3): (1) **Vanilla**, using only reconstruction loss; (2) **Pre-trained**, our default model with reconstruction and audio–text alignment losses (ranking and contrastive); and (3) **Fine-tuned**, the pre-trained model further trained for 10 epochs on the target dataset. We measured semantic alignment by computing the cosine similarity between M' (mixture text-label embeddings) and M (separated-audio embeddings), and reported the average standard deviation on the test set. Results demonstrate that adding semantic alignment loss during pre-training substantially enhances separation quality, with fine-tuning providing comparable gains. These findings indicate that robust MSS pre-training alone can suffice for audio-to-image generation, potentially eliminating the need for further fine-tuning.

4.5 Qualitative Results

MACS produces higher quality visuals. The upper part of Fig. 1 presents images generated from single- and multi-source audio using MACS and baseline methods. MACS consistently produces more realistic and semantically aligned images. For instance, generating vivid flames in the *"Fire Crackling"* category, whereas baselines often produce abstract visuals. For multi-source audio, MACS reliably captures the expected scenes, while methods such as CoDi and ImageBind frequently fail to reflect the audio content. See more qualitative examples in Appendix C.

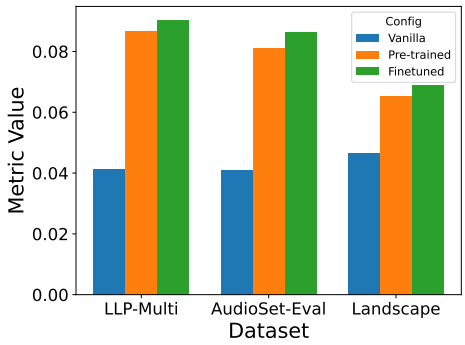

Figure 3: The average standard deviation of the similarity scores between the text labels of each audio mixture and its separations across three datasets.

Ranking Loss Helps Contextual Importance Learning. The ranking loss in Eq. 4 trains the model to capture the contextual importance of audio signals 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. As shown in Fig. 4, higher-ranked embeddings represent more salient audio events, with the first three encoding most of the semantic information. In single-source cases, the first embedding alone suffices to generate semantically accurate images (see the first and last rows).

Audios are Interpolable under MACS. To evaluate the model's ability to transition between sounds, we interpolate between the audio clips X (dog bark) and Y (motor vehicle,

Method	FID↓	CLIP-FID↓	KID↓	AIS↑	AIS-z↑	IIS↑	IIS-z↑
AudioToken (Yariv et al. 2023)	236.63	54.42	0.0402	0.0708	0.3527	0.6030	0.2900
Sound2Scene (Sung-Bin et al. 2023)	186.12	43.25	0.0280	0.1042	0.6519	0.6762	0.6368
TempoTokens (Yariv et al. 2024)	212.69	50.70	0.0450	**0.1251**	0.6307	0.6703	0.8057
ImageBind (Girdhar et al. 2023)	207.93	41.49	0.0304	0.1189	0.8483	0.6673	0.7681
CoDi (Tang et al. 2023)	158.31	39.97	0.0180	0.1094	0.6912	0.6961	1.0942
AudioToken (Yariv et al. 2023) (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**

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 (Ruan et al. 2023).

Sep. Model	FID↓	CLIP-FID↓	KID↓	AIS↑	AIS-z↑	IIS↑	IIS-z↑
MixIT [49] (Waveform)	98.73	28.42	0.0201	0.0688	1.0471	0.5782	1.3819
MACS (Spectrogram)	**87.09**	**20.47**	**0.0157**	**0.0754**	**1.3038**	**0.6269**	**1.7231**

Table 4: MACS benchmarked with MixIT on LLP-multi.

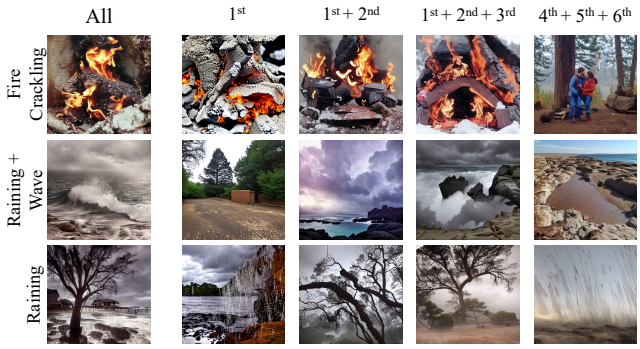

Figure 4: **Ranking loss helps sort the contextual significance.** Embeddings in higher ranking contain more important semantic information for accurate image generation.

Figure 5: Generated images from interpolations between two audio clips (dog bark and motor vehicle).

engine, revving):

$$Z(\alpha) = \alpha X + (1-\alpha)Y, \quad \alpha \in \{0, 0.25, 0.5, 0.75, 1\},$$

where $\alpha = 0$ and 1 correspond to pure engine and bark. At $\alpha = 0.5$, both dog and car appear. As shown in Fig. 5, MACS successfully blends and disentangles semantic features from the mixed audio inputs.

Association of Separated Audios with Distinct Image Regions. We use Grad-CAM (Selvaraju et al. 2017) to visualize how individual audio embeddings correspond to specific regions in the images generated by MACS. As shown in Fig. 6, mixed audio embeddings result in diffuse attention maps, while disentangled embeddings produce more focused, object-aligned regions, indicating stronger semantic alignment.

4.6 Ablation Studies

Due to space constraints, LLP-multi ablation results are reported in Appendix B. We ablated each of the three components, ranking loss (RL), contrastive loss (CL), and decoupled cross-attention (DC), individually, keeping the rest of MACS unchanged. In all cases, performance declined across all metrics, highlighting the importance of audio–text alignment for generating high-quality, semantically accurate images in multi-source audio-image generation.

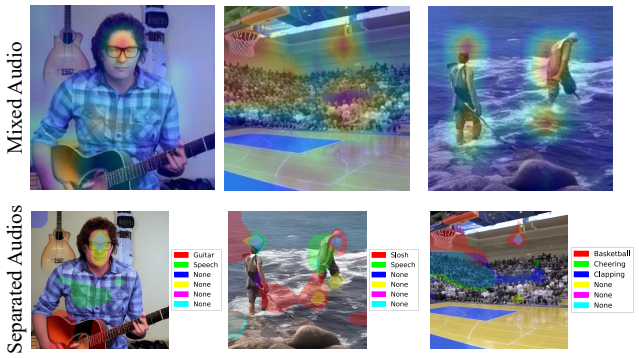

Figure 6: Attention maps of generated images w/ or w/o audio separation (Audio-text similarity score thresholded 0.5).

5 Conclusion

We present MACS, the first two-stage architecture that explicitly separates mixed audio signals for audio-to-image generation. MACS preserves the contextual and semantic alignment between separated audio and text labels, and employs a decoupled cross-attention module to effectively integrate multiple audio inputs. Extensive experiments show that the "separation before generation" strategy is effective, with MACS achieving state-of-the-art performance on both mixed- and single-source audio-to-image generation tasks.

References

Bińkowski, M.; Sutherland, D. J.; Arbel, M.; and Gretton, A. 2018. Demystifying mmd gans. *arXiv preprint arXiv:1801.01401*.

Blondel, M.; Teboul, O.; Berthet, Q.; and Djolonga, J. 2020. Fast differentiable sorting and ranking. In *International Conference on Machine Learning*, 950–959. PMLR.

Chatterjee, M.; and Cherian, A. 2020. Sound2sight: Generating visual dynamics from sound and context. In *Proceedings of the European Conference on Computer Vision*, 701–719. Springer.

Chen, L.; Srivastava, S.; Duan, Z.; and Xu, C. 2017. Deep cross-modal audio-visual generation. In *Proceedings of the on Thematic Workshops of ACM Multimedia 2017*, 349–357.

Chen, Y.-C.; Li, L.; Yu, L.; El Kholy, A.; Ahmed, F.; Gan, Z.; Cheng, Y.; and Liu, J. 2020. Uniter: Universal image-text representation learning. In *Proceedings of the European Conference on Computer Vision*, 104–120. Springer.

Dong, H.-W.; Takahashi, N.; Mitsufuji, Y.; McAuley, J.; and Berg-Kirkpatrick, T. 2023. CLIPSep: Learning Text-queried Sound Separation with Noisy Unlabeled Videos. In *Proceedings of International Conference on Learning Representations (ICLR)*.

Duarte, A. C.; Roldan, F.; Tubau, M.; Escur, J.; Pascual, S.; Salvador, A.; Mohedano, E.; McGuinness, K.; Torres, J.; and Giro-i Nieto, X. 2019. WAV2PIX: Speech-conditioned Face Generation using Generative Adversarial Networks. In *ICASSP*, volume 2019, 8633–8637.

Fitria, T. N. 2023. 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.

Fonseca, E.; Favory, X.; Pons, J.; Font, F.; and Serra, X. 2021. Fsd50k: an open dataset of human-labeled sound events. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30: 829–852.

Gemmeke, J. F.; Ellis, D. P.; Freedman, D.; Jansen, A.; Lawrence, W.; Moore, R. C.; Plakal, M.; and Ritter, M. 2017. Audio set: An ontology and human-labeled dataset for audio events. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 776–780. IEEE.

Girdhar, R.; El-Nouby, A.; Liu, Z.; Singh, M.; Alwala, K. V.; Joulin, A.; and Misra, I. 2023. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15180–15190.

Guzhov, A.; Raue, F.; Hees, J.; and Dengel, A. 2022. Audioclip: Extending clip to image, text and audio. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 976–980. IEEE.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in Neural Information Processing Systems*, 30.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33: 6840–6851.

Jia, C.; Yang, Y.; Xia, Y.; Chen, Y.-T.; Parekh, Z.; Pham, H.; Le, Q.; Sung, Y.-H.; Li, Z.; and Duerig, T. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, 4904–4916. PMLR.

Karamatlı, E.; and Kırbız, S. 2022. Mixcycle: unsupervised speech separation via cyclic mixture permutation invariant training. *IEEE Signal Processing Letters*, 29: 2637–2641.

Kim, W.; Son, B.; and Kim, I. 2021. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*, 5583–5594. PMLR.

Le Roux, J.; Wisdom, S.; Erdogan, H.; and Hershey, J. R. 2019. SDR–half-baked or well done? In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 626–630. IEEE.

Lee, S. H.; Chi, H.-g.; Oh, G.; Byeon, W.; Yoon, S. H.; Park, H.; Cho, W.; Kim, J.; and Kim, S. 2024. Robust sound-guided image manipulation. *Neural Networks*, 175: 106271.

Lee, S. H.; Oh, G.; Byeon, W.; Kim, C.; Ryoo, W. J.; Yoon, S. H.; Cho, H.; Bae, J.; Kim, J.; and Kim, S. 2022. Sound-guided semantic video generation. In *European Conference on Computer Vision*, 34–50. Springer.

Lee, T.; Kang, J.; Kim, H.; and Kim, T. 2023. Generating Realistic Images from In-the-wild Sounds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 7160–7170.

Li, B.; Qi, X.; Lukasiewicz, T.; and Torr, P. 2019. Controllable text-to-image generation. *Advances in Neural Information Processing Systems*, 32.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning*, 19730–19742. PMLR.

Loshchilov, I.; and Hutter, F. 2019. Decoupled Weight Decay Regularization. In *Proceedings of International Conference on Learning Representations (ICLR)*.

Luo, Y.; and Yu, J. 2023. Music source separation with band-split RNN. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31: 1893–1901.

Mahmud, T.; Amizadeh, S.; Koishida, K.; and Marculescu, D. 2024. Weakly-supervised Audio Separation via Bi-modal Semantic Similarity. *arXiv preprint arXiv:2404.01740*.

Mou, C.; Wang, X.; Xie, L.; Wu, Y.; Zhang, J.; Qi, Z.; and Shan, Y. 2024. 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*, volume 38, 4296–4304.

Oh, T.-H.; Dekel, T.; Kim, C.; Mosseri, I.; Freeman, W. T.; Rubinstein, M.; and Matusik, W. 2019. Speech2face: Learning the face behind a voice. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7539–7548.

Pishdadian, F.; Wichern, G.; and Le Roux, J. 2020. Finding strength in weakness: Learning to separate sounds with weak supervision. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28: 2386–2399.

Qin, C.; Yu, N.; Xing, C.; Zhang, S.; Chen, Z.; Ermon, S.; Fu, Y.; Xiong, C.; and Xu, R. 2023. Gluegen: Plug and play multi-modal encoders for x-to-image generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 23085–23096.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 8748–8763. PMLR.

Ramesh, A.; Dhariwal, P.; Nichol, A.; Chu, C.; and Chen, M. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2): 3.

Reed, S.; Akata, Z.; Yan, X.; Logeswaran, L.; Schiele, B.; and Lee, H. 2016. Generative adversarial text to image synthesis. In *International Conference on Machine Learning*, 1060–1069. PMLR.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10684–10695.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. 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*, 234–241. Springer.

Ruan, L.; Ma, Y.; Yang, H.; He, H.; Liu, B.; Fu, J.; Yuan, N. J.; Jin, Q.; and Guo, B. 2023. 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*, 10219–10228.

Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; and Batra, D. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, 618–626.

Spearman, C. 1987. The proof and measurement of association between two things. *The American Journal of Psychology*, 100(3/4): 441–471.

Sung-Bin, K.; Senocak, A.; Ha, H.; Owens, A.; and Oh, T.-H. 2023. Sound to visual scene generation by audio-to-visual latent alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6430–6440.

Tang, Z.; Yang, Z.; Zhu, C.; Zeng, M.; and Bansal, M. 2023. Any-to-any generation via composable diffusion. *Advances in Neural Information Processing Systems*, 36: 16083–16099.

Tian, Y.; Li, D.; and Xu, C. 2020. Unified multisensory perception: Weakly-supervised audio-visual video parsing. In *Proceedings of the European Conference on Computer Vision*, 436–454. Springer.

Wan, C.-H.; Chuang, S.-P.; and Lee, H.-Y. 2019. Towards audio to scene image synthesis using generative adversarial network. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 496–500. IEEE.

Wang, D.; and Chen, J. 2018. Supervised speech separation based on deep learning: An overview. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(10): 1702–1726.

Wang, X.; Chen, G.; Qian, G.; Gao, P.; Wei, X.-Y.; Wang, Y.; Tian, Y.; and Gao, W. 2023. Large-scale multi-modal pre-trained models: A comprehensive survey. *Machine Intelligence Research*, 20(4): 447–482.

Wisdom, S.; Tzinis, E.; Erdogan, H.; Weiss, R.; Wilson, K.; and Hershey, J. 2020. Unsupervised sound separation using mixture invariant training. *Advances in Neural Information Processing Systems*, 33: 3846–3857.

Wu, H.-H.; Seetharaman, P.; Kumar, K.; and Bello, J. P. 2022. Wav2clip: Learning robust audio representations from clip. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 4563–4567. IEEE.

Wu, Y.; Chen, K.; Zhang, T.; Hui, Y.; Berg-Kirkpatrick, T.; and Dubnov, S. 2023. 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)*, 1–5. IEEE.

Yariv, G.; Gat, I.; Benaim, S.; Wolf, L.; Schwartz, I.; and Adi, Y. 2024. Diverse and aligned audio-to-video generation via text-to-video model adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 6639–6647.

Yariv, G.; Gat, I.; Wolf, L.; Adi, Y.; and Schwartz, I. 2023. Audiotoken: Adaptation of text-conditioned diffusion models for audio-to-image generation. *arXiv preprint arXiv:2305.13050*.

Ye, H.; Zhang, J.; Liu, S.; Han, X.; and Yang, W. 2023. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*.

Yu, D.; Kolbæk, M.; Tan, Z.-H.; and Jensen, J. 2017. Permutation invariant training of deep models for speaker-independent multi-talker speech separation. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 241–245. IEEE.

Zhou, H.; Liu, Y.; Liu, Z.; Luo, P.; and Wang, X. 2019. Talking face generation by adversarially disentangled audio-visual representation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 9299–9306.