Bridging Audio-Visual Semantics with Language-Guided Synthesis

Zihao Wei Zixuan Pan Yidong Huang Ziqiao Ma
Ziyang Chen Joyce Chai Andrew Owens
University of Michigan

Abstract

One of the underlying assumptions behind audio-visual learning models is that the two modalities convey overlapping information. However, this assumption is widely violated in practice, which results in degraded performance. To address this problem, we propose to replace mismatched audio-visual signals using cross-modal generative models. Our approach uses language-based supervision to perform this generation.We show that data synthetically generated through this process is well-suited for a variety of representation learning methods. The features that we learn this way outperform those trained solely on real data for a range of downstream tasks, including audio classification, audio-visual retrieval, and visual sound localization.

1. Introduction

A core assumption behind existing audio-visual representation learning methods is that the two modalities tend to convey the same underlying scene structures — that, more often than not, when we hear an engine noise we also see the car that produced it. In practice, this assumption is frequently violated. In popular internet video datasets [5, 14, 30, 38], for example, sound sources are often distant, occluded, or lie outside the camera's field of view (Fig. 1). While recent representation learning methods are designed to tolerate *some* misalignment between modalities, these misaligned examples collectively degrade the performance of the learned models.

In this paper, we identify examples where audio and visual signals are most relevant using multimodal embedding models. We keep these examples and use cross-modal prediction methods to replace one of the signals of the rest audio visual pairs. This results in signals that convey the same scene structures, and thus it is more useful for audio-visual representation learning. To perform this cross-modal prediction, we use text-conditioned generation models, treating language as a bridge between modalities. We take advantage of the fact that large-scale text-conditioned generation and captioning models are widely available. We generate a caption for a given image (or audio track) and then pass this caption into a cross-modal prediction model. Our use of language as an intermediate representation also crucially allows us to steer the generation process by prompting, which we use to ensure the generated images are not homogeneous.

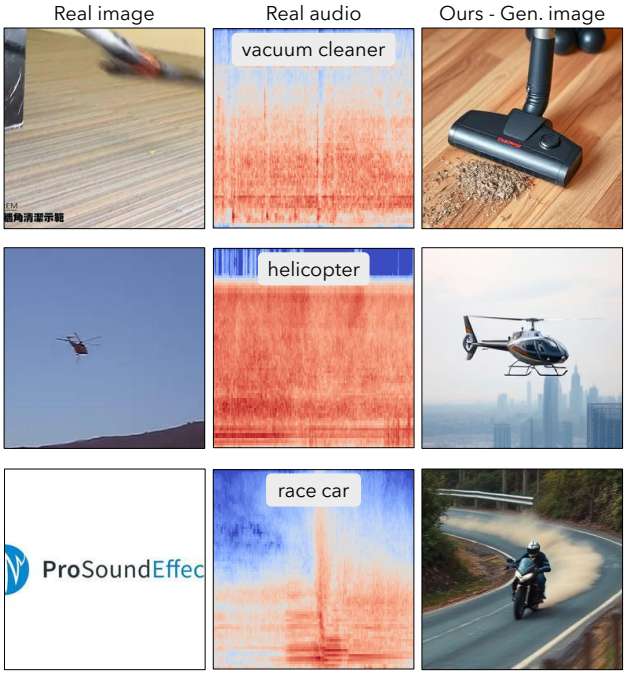

Figure 1. **Misalignment between images and audio.** We show selected videos from VGGSound [5], a popular internet video dataset. The audio and visual signals either do not convey the same events, or the visual signal is impoverished. We identify cases like these and use cross-modal prediction to replace one of the two signals, improving alignment across modalities and the suitability of the data for representation learning. Here we show images that have been generated from the audio.

Our approach is inspired by an emerging line of work that has successfully used synthetically generated data for a variety of visual learning tasks [3, 22, 61, 62], particularly work that uses generated images for representation learning [62]. However, in contrast to these approaches, we

use synthetic data to impute missing signals in paired multimodal datasets using conditional generation, rather than generating the entire dataset "from scratch."

We evaluate our pipeline by generating a synthetic version of VGGSound dataset [5] and comparing it with the real one. Through human ratings and automated evaluation metrics, we find that the generated audio-visual pairs are more semantically aligned with one other. We evaluate the quality of the generated data using a variety of different representation learning methods, including audio-visual contrastive learning [20, 34, 65] and masked autoencoders [16, 19, 21]. We find that the resulting representations outperform those trained on the original dataset downstream tasks, including audio and visual classification, cross-modal retrieval, and sound localization. These experiments suggest:

- Perhaps surprisingly, we find that simply replacing all images in VGGSound with generated images leads to improved representation learning performance (over using the real images).
- Models that use a mix of real and synthetic data outperform those trained on real or synthetic data alone.
- The representation learned with our generated data works consistently well in a variety of downstream tasks.

2. Related Work

Audio-visual representation learning. Self-supervised methods for audio-visual representation learning have been widely studied to establish connections between audio and visual [1, 12, 13, 25, 27]. Recently more and more attention has been drawn to contrastive learning-based methods. CAVP, AVID and its variants [34, 35, 40, 50] pull the embeddings of paired audio-visual samples together while pushing apart non-paired samples. Approaches like EZ-VSL and its variants [39, 55, 57] apply contrastive learning at the pixel level, effectively localizing the sounding object in images. Masking-based methods, such as AV-MAE [16] and MAViL [24], enable cross-modal learning through a joint decoder, while CAV-MAE [19] further improves performance by combining contrastive learning with masked modeling. Cacophony [69] extends these methods by incorporating text via an autoregressive captioner. These approaches rely on the implicit assumption that the audio-visual pairs are well aligned, though this is often not the case with collected data. We hypothesize that a higher-quality dataset will yield better representations, ultimately improving performance across downstream tasks. One direct approach is generating images from audio or vice versa. However, as shown in Fig. 4, this is challenging due to the limited availability of high-quality paired data for training audio-visual generative models [59, 68]. In our work, we address this by using a language-guided data synthesis pipeline to create aligned data, enabling us to leverage these various representation learning methods for evaluation.

Language supervision for vision and audio representation learning. The study of cross-modal semantic mapping between language and visual/audio perception has a longstanding history [26, 29, 36]. Natural language has been widely used as a form of supervision to support transferable vision and audio representation learning [31, 41, 45]. Much recent excitement stems from CLIP [44], which learns joint embeddings of language and visuals from large-scale paired data. Subsequent models, such as CLAP, extend this approach to audio and text representation learning [7, 65]. VatLM [70] and AudioCLIP [20] further broadened this by learning cross-modal representations across all three modalities—text, audio, and visuals—through mutual alignment. Most similar to our approach is ImageBind [18], which uses RGB images as the hub modality, aligning embeddings from all other modalities to the image embedding. However, instead of directly learning coherent representations, we explore language as an intermediary to generate semantically aligned counterparts for audio and visual data.

Representation learning with synthetic data. To address data scarcity and quality issues, synthetic data has been employed for learning representations [3]. Recent advances in large language models have enabled text rewriting for multimodal data augmentation [10, 69]. Similarly, in the vision domain, images can be augmented by replacing them with those generated through diffusion models [3, 22, 61, 62]. Synthetic data has also been used to improve sound localization methods in the audio-visual field by leveraging alignment between modalities [54, 55]. Rather than focusing on task-specific applications, this work aims to explore the general audio-visual representation learning capabilities enabled by synthetic data. One of the primary challenges with synthetic data is its limited distribution [2, 17, 51]. Efforts to address this have focused on improving the diversity of generated datasets through techniques like prompt engineering [61]. Retrieval can also be incorporated into our pipeline for dataset preparation [15]. In addition to these innovative prompt engineering methods, we apply a filtering pipeline to ensure alignment with real-world data, mitigating the collapse issue and enhancing the representation learning for our target modalities.

3. Method

Our method uses a data generation and filtering pipeline to address distribution difference between visual and audio data, as shown in Fig. 2.

3.1. Data Generation Pipeline

We apply a bidirectional data generation pipeline, so that we can generate both image and audio by using its counterpart as a reference. For both sides, we use language as a middle

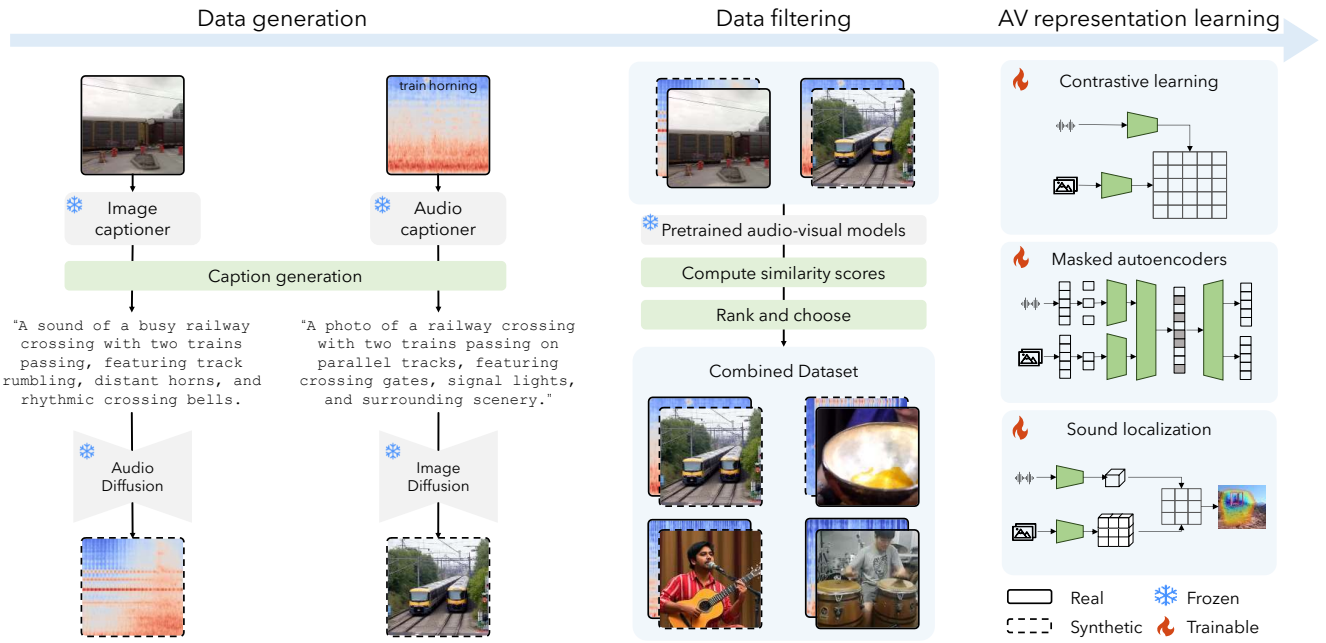

Figure 2. **Pipeline.** Our methods consist of two parts: dataset generation and data filtering. In the bidirectional data generation process, we generate captions for either audio or images, refine these captions, and use them to synthesize corresponding images or audio. In the data filtering pipeline, we calculate similarity scores to identify and retain image-audio pairs with high similarity, ensuring that the most semantically aligned pairs in the original dataset are preserved. The combined dataset can further be used in the representation learning method for downstream tasks.

stage, as it is easy to evaluate, control, and modify.

For synthetic image generation, we first use an audio captioner to generate a description of the audio, containing various sounding objects or events, such as cars, trains, and people speaking. Compared to the labels in datasets like VGGSound [5] or AudioSet [14], the generated captions are often more complete, including multiple sounding events that may be missing from the labels due to faint sounds or labeling errors. Also, by using an audio captioner, we can deal with the label-free dataset that is directly collected from the Internet. However, these captions are not immediately suitable for image generation. We apply prompt engineering and utilize large language models to reformat and diversify the captions. The reorganized prompts are then input into an off-the-shelf image generator to create synthetic images. For audio, we use a similar reversed process with an image captioner and audio generator.

3.2. Caption Synthesis

A key challenge in synthetic data generation is the issue of data distribution. Simple prompts like "A photo of a car" often lead to a narrow distribution of images that look similar, even when the guidance rate is low. This narrow distribution can significantly hinder pretraining performance, especially when scaling the dataset's size. Previous works like SynCLR [3] address this by utilizing context learning-based methods to introduce diversity to the generated caption. These methods will usually have a large bank of templates and environment lists and modify the caption to be highly vivid. However, these methods usually only deal with two modalities, so no matter how they change the prompt the generated images are still aligned with the text. In our cases, the caption is only a middle stage, so too much modification would lead to the LLM's hallucination problem [61, 66] and make the generated caption no longer align with the audio, which will harm the further downstream tasks. Thus, we trade-off between the caption's vividness and correctness. Instead of introducing too many entities as [61], we design a set of prompts that leverages the generalization ability of LLM to place the given objects in the foreground, and then put them in a reasonable environment. Additionally, we use a list of negative prompts to prevent image distortions and further enhance the quality of the generated images.

3.3. Combining Synthetic and Real Data

One of our concerns is that the real data itself has the underlying alignment, so we should not completely discard them. Therefore, we developed a filtering pipeline to automatically select well-aligned audio-visual pairs in the original dataset and maintain them. To measure the alignment, we follow the clip score [23], where we apply large pretrained models like ImageBind [18] to extract features from

the visual and audio pairs and compute their similarity. To overcome the potential problem that the data distribution we tested may differ from the data distribution they pre-trained on, we normalize the get similarity and acquire a ranked score, where 1 represents the highest similarity and 0 is the lowest. We retain the best-aligned audio-visual pairs. To deal with the test-time distribution shift, we find that for vision-related tasks, using a high ratio of synthetic audio with real images is very effective, while for audio-related tasks, incorporating a large portion of synthetic images with real audio further boosts model performance. Additionally, the ratio of real to synthetic data needs to be carefully chosen, as there exists a distribution gap between real and synthetic data, which can lead to poor downstream performance when they combined.

4. Experiments

In this section, we thoroughly examine our pipeline. We first conduct a human study and automatic evaluation to evaluate the dataset quality. Then we further evaluate the generated data by pre-training with different self-supervised representation learning methods and testing their performance on different downstream tasks.

4.1. Implementation Details

Data generation pipeline. For image generation, we use SALMONN [56, 58] to caption the audio and LLAMA2-7b [63] to generate the caption for image generation. The image generation model is Flux.1-schnell [32], which is a distilled text to image diffusion model. To ensure maximum diversity, the guidance scale is set to 0. For audio generation, LLAVA-7b [33] is used to generate image captions, and Stable-Audio [9] converts text to audio. For filtering, we typically retain the top 5% of the most-aligned audio-visual pairs, with the remaining dataset filled according to task requirements. For audio-oriented tasks, we pair real audio with synthetic images, while for visual-oriented tasks, we pair real images with synthetic audio. Additionally, we can also replace the lowest 5% of real data with pure synthetic data to eliminate potentially low-quality problems. Without specially mentioned, default hyper-parameters are used in most cases for generation tasks.

Evaluation details. To benchmark the performance gains achieved by our data generation pipeline, we compare the downstream task capabilities of various representation learning algorithms when trained on the original dataset versus our synthesized versions. Specifically, we prepare four versions of the synthesized datasets: (1) real audio with synthetic images, (2) synthetic audio with real images, (3) a combined dataset oriented at audio tasks, and (4) a combined dataset targeted at visual tasks. In these combined versions, the composition of real and synthetic data is determined by the previously filtering strategy. For audio-based tasks, we use the real audio and audio-oriented combined datasets, while for image-based tasks, we use the remaining two versions. To ensure fair comparisons, all versions of the synthesized dataset, as well as the original real dataset, contain the same number of audio-visual pairs.

For representation learning methods, we select several widely used techniques, including contrastive learning methods like CAVP [34], CLAP [65], and AudioClip [20], as well as masked autoencoders like CAV-MAE [19]. The pretrained models are evaluated on downstream tasks such as classification, along with audio-visual applications like audio-visual retrieval and visual sound localization. If a dataset is well aligned, it should enable representation models to learn more effectively, leading to better performance in downstream tasks.

Datasets and training details. In line with prior work [19], we apply our pipeline to the VGGSound dataset [5] for evaluation. We follow AudioCLIP's procedure to preprocess the dataset [20], extracting the middle frame of each video as the image input and resampling audio to 16kHz. Unless otherwise noted, all models are trained from scratch. We use the default hyperparameter configurations provided by LAION-CLAP [65] and CAV-MAE [19].

For audio-domain downstream tasks, we evaluate audio classification on ESC-50 [42], FSD-50k [11], Urban-8k [49], and VGGSound test sets [5]. To assess visual representation capabilities, we conduct linear probing to the pretrained model on CIFAR-10, CIFAR-100 [28], ImageNet-100 [60], and ImageNet-1k [48]. In sound localization, we use EZ-VSL as the backbone for localization [39], trained on the VGGSound training set and tested on Flickr-SoundNet [52, 53] and VGG-SS [6]. All models are trained on 4 NVIDIA A40 GPUs, with pretraining for 30 epochs and downstream tasks for 10 epochs.

Table 1. **Human Evaluation Results.** We collect human preference data to evaluate both the audio to image and image to audio pipelines

Task	Prefer real	Prefer syn	Both good	Both bad
Audio → Image	18.9	**46.7**	25.5	8.9
Image → Audio	21.1	**55.6**	15.5	7.8

Table 2. **Automatic evaluation of the dataset.** We evaluate the semantic distance between the visual and audio pairs by captioning both pairs and calculating the language embeddings' similarity between two captions using SentenceBert [46].

Visual	Audio	SentenceBert [46]
Real	Real	38.72
Real	Syn	41.08
Syn	Real	**52.12**

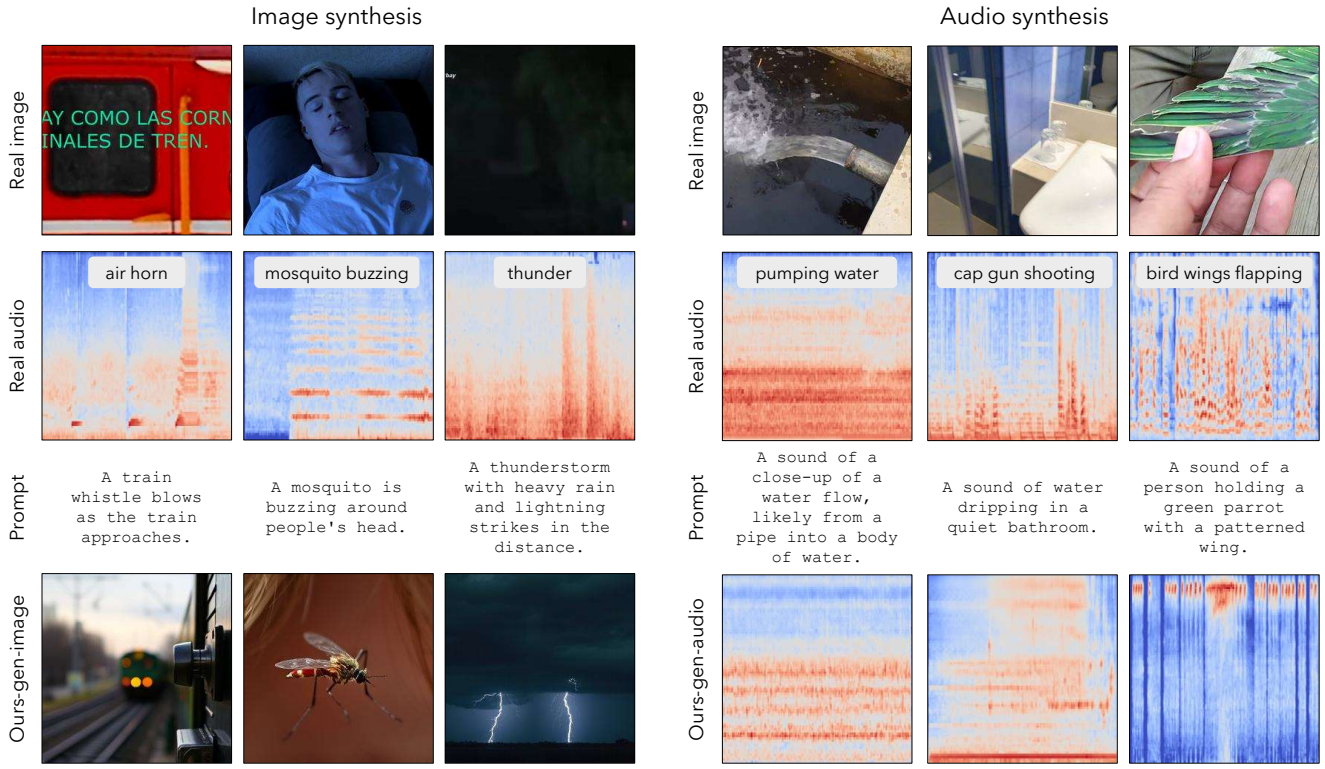

Figure 3. **Qualitative result.** We selected several samples from the generated dataset, showcasing image synthesis from the audio on the left and audio synthesis based on images on the right. For both cases, the first two rows display the original paired data in VGGSound [5], with the ground truth text label shown on the mel-spectrogram. We then provide the prompt used for image or audio generation, followed by the final generated result.

4.2. Evaluation of the Synthetic Dataset

Human evaluation. We randomly selected 500 image-audio pairs from the VggSound dataset [5]. Users are presented with a reference set containing an original real image and audio pair, along with a synthesized version (either a synthesized image or audio). They are asked to choose which is more aligned with the reference or indicate if both are equally good or poor. Each user is assigned 10 image generation samples and 10 audio generation samples, both random;y picked from the whole set. Order of real and generated samples are also randomly shuffled to prevent users falling into a certain paradigm.

We demonstrate in Tab. 1 that users show a preference for synthetic data.. It can be found that in over 72.2% cases the users consider the generated images to represent the reference audio well and in 71.1% cases, the generated audio matches the given image. Compared to the real images, the synthetic data was preferred by 27.8% and 34.5% more for images and audio, respectively.

Automatic evaluation. We choose to evaluate in the language domain instead of using the similarity of the embeddings from the popular audio-visual foundation models like imagebind [18]. Since these models are trained on datasets such as AudioSet [14], which primarily contain music and speech, they tend to overscore data similar to their training distribution. Thus, audio-visual pairs with music or speech are scored high, even if the target sound events are absent.

For our evaluation, we use SALMONN [58] and LLAVA [33] to caption the real audio and image and its synthetic counterparts. Then we compute the semantic distance between these captions can be calculated by comparing the cosine similarity between their embeddings with Sentence-Bert [46]. The result is shown in Tab. 2, and we find the semantic distance is improved for both the generative images and audio. We also compute the Spearman correlation between the automatic results and the human evaluation giving a correlation of 0.21, demonstrating the automatic evaluation pipeline is correlated with the human evaluation.

Qualitative comparison. In Fig. 3, we select several image-audio pairs from VGGSound to compare the real images and audios with those generated by our synthetic pipeline qualitatively. We also include the original real data and the prompt for generation for better comparison. It can be found that the generated samples are of higher quality than the ones in the original dataset. For image synthesis, the original image may suffer from problems like distortion,

Table 3. **Linear probing results on audio classification.** Audio classification accuracy using various representation learning methods pretrained on different variations of VGGSound [5]. In the data type section, $\mathcal{A}$, $\mathcal{V}$, $\mathcal{T}$ mean the data type for audio, visual, and text, where 'Real' indicates that this modality in this variation is composed of real data, 'Synthetic' denotes that the modality is entirely generated, and 'Combined' represents a mixture of both real and synthetic data. For the text modality, 'Label' signifies the use of original text labels as model input, while 'Caption' refers to using the audio's caption as input.

Method	Data Type			ESC-50 [42]	FSD-50k [11]	Urban-8k [49]
	$\mathcal{A}$	$\mathcal{V}$	$\mathcal{T}$			
CAV-MAE [19]	Real	Real	-	77.8	35.1	82.5
	Real	Synthetic	-	86.5	**38.3**	83.6
	Combined	Combined	-	**87.0**	36.2	**84.0**
CAVP [34]	Real	Real	-	72.8	40.1	74.3
	Real	Synthetic	-	82.0	45.7	81.4
	Combined	Combined	-	**83.8**	**45.8**	**82.8**
CLAP [65]	Real	-	Label	83.3	46.1	**81.5**
	Real	-	Caption	**84.3**	**46.3**	80.1
AudioCLIP [20]	Real	Real	Label	83.0	45.7	77.5
	Real	Real	Caption	85.8	46.2	78.9
	Real	Synthetic	Caption	86.5	46.8	**82.8**
	Combined	Combined	Caption	**87.0**	**47.1**	82.0

non-realistic, and poor light quality, while the generated can overcome these problems by adding negative prompts during generation. For the audio side, the real audio always contains much noise like background music (e.g. pumping water example) or people's speech in the distance (e.g. cap gun shooting example), while the synthetic audio is usually more clear and distinguishable.

Also for both cases, the generated image or audio is semantically more aligned with its paired audio or image. One of the reasons is that the generated captions are more detailed than the given labels, consistently capturing the elements provided by the audio or visual. For example, the thunder pair's label does not contain the rain sound. By fully obtaining the sounding objects, the generated images will be better aligned with the corresponding audio. This also maintains the truth in the audio generation pipeline. The synthetic audio will only contain the objects that are shown in the scene without the out-of-scene sound here serves as noise.

4.3. Evaluation on Downstream Tasks

Classification. We present linear probing results for various pretraining methods [19, 20, 34] on popular audio classification datasets in Tab. 3 and on image classification datasets in Tab. 4. We also provide the linear probing result for audio-visual classification for methods supporting this task in Tab. 5. Our results demonstrate a significant performance boost in all classification tasks when models are pretrained on our synthetic dataset compared to those trained solely on real data. This underscores the ability of our synthetic data pipeline to enhance downstream performance across a range of tasks. The results also show that the synthetic data is effective across diverse representation training algorithms, including contrastive methods like CAVP [34] and reconstruction-based methods like CAV-MAE [19].

Notably, we observe that using triple-modality aligned data—audio, visual, and text—yields superior audio classification results. Initially, AudioCLIP and CLAP achieved similar performance or even dropped when trained on real data, which may be because the uncorrelated pairs in the dataset harm the learning process. However, when trained with synthesized data, AudioCLIP outperforms CLAP significantly, demonstrating the benefit of leveraging synthesized multi-modal aligned data in representation learning.

Cross-modal retrieval. We evaluated the retrieval performance of AudioClip [20] trained on both synthetic and real images in Tab. 6 on VGGSound's test set. Besides the full test set, we also build a clean subset of it, since the original one, like its training set, contains many misaligned samples. Evaluating models on the cleaned one is more indicative of the model's actual ability. We create the subset by calculating CLIP score [23] between VGGSound labels and their corresponding images. Samples with scores below 25 were removed. We also include the numbers tested on the entire test set for reference. Additionally, we assessed retrieval performance by category, defining a match as instances where the retrieved audio or image belonged to the same category as the reference.

In zero-shot retrieval tasks, models trained on synthetic data performed worse than those trained exclusively on real data, particularly when using the full real dataset. Apart from the data quality issue, the disparity is mainly due to the

Table 4. **Linear probing results on image classification.** Image classification accuracy of different representation learning methods pretrained on variations of VGGSound [5]. Results are reported on CIFAR 10 [28], CIFAR 100 [28], ImageNet-100 [60], and ImageNet-1k [48]. Data modalities $\mathcal{V}$ (visual), $\mathcal{A}$ (audio), and $\mathcal{T}$ (text) follow the definitions provided earlier.

Method	Data Type			CIFAR 10 [28]	CIFAR 100 [28]	IN-100 [60]	IN-1k [48]
	$\mathcal{V}$	$\mathcal{A}$	$\mathcal{T}$				
CAVP [34]	Real	Real	-	57.6	32.2	31.7	15.1
	Real	Synthetic	-	62.8	**37.1**	34.1	**18.7**
	Combined	Combined	-	**63.4**	36.7	**34.3**	18.2
AudioCLIP [20]	Real	Real	Label	61.2	34.8	30.1	15.7
	Real	Synthetic	Caption	65.2	38.6	**37.1**	19.8
	Combined	Combined	Caption	**65.5**	**38.7**	35.8	**19.9**

Table 5. **Linear probing results on audio-visual classification.** Audio-visual classification accuracy with CAV-MAE [19]. Results are reported on test split of VGGSound [5]

Method	Data Type		VGGSound [5]
	$\mathcal{A}$	$\mathcal{V}$	
CAV-MAE [19]	Real	Real	50.9
	Real	Synthetic	51.5
	Combined	Combined	**52.7**

substantial distribution gap between synthetic and real images, causing the visual branch trained on synthetic data to struggle in mapping real images into the same latent space as synthetic ones. As a result, the embeddings no longer align with the audio, leading to lower retrieval scores.

However, after fine-tuning the model pre-trained on synthetic data with real VGGSound images for an additional 5 epochs, we observed significant performance improvements. To ensure a fair comparison, the model trained on real data alone was also fine-tuned for the same number of epochs. We also hypothesize that the audio and text branches have already learned effective representations, and the main challenge is adapting the visual branch to real images. We verified this by freezing the audio and text branches and fine-tuning only the visual branch achieving similar improvements as the fully tuned model. This approach also yielded a speedup of over five times compared to fully fine-tuning the model.

Visual sound localization. We further evaluated visual sound localization with EZ-VSL [39] as the backbone on Flickr-SoundNet [52, 53] and VGG-SS[6]. In this task, the model is given an audio clip and an image and tasked with identifying the region in the image producing the sound. Thus, the self-supervised method relies heavily on alignment, which pulls together features from corresponding audio-visual pairs while pushing away unrelated examples in the batch. From the result shown in Tab. 7, we find the model benefits a lot from our augmented dataset, as when trained with our synthetic data, the cIoU and AuC scores improved consistently on both test sets, especially the Flickr-SoundNet.

Table 6. **Audio-visual retrieval results.** Category-based audio-visual retrieval accuracy using AudioCLIP [20] as the backbone. We evaluate both zero-shot retrieval and fine-tuned retrieval. Here, $\mathcal{A}$, $\mathcal{V}$, and $\mathcal{T}$ denote the audio, visual, and text branches, respectively, while ✓ and ✗ indicate whether the encoder for the corresponding modality is frozen or unfrozen during fine-tuning.

Image type	Branches			VGGSound-clean		VGGSound-test	
	$\mathcal{A}$	$\mathcal{V}$	$\mathcal{T}$	$\mathcal{A} \rightarrow \mathcal{V}$	$\mathcal{V} \rightarrow \mathcal{A}$	$\mathcal{A} \rightarrow \mathcal{V}$	$\mathcal{V} \rightarrow \mathcal{A}$
Real	✗	✗	✗	25.1	31.0	10.4	14.5
Synthetic	✗	✗	✗	13.1	19.7	1.9	1.1
Real	✓	✓	✓	25.4	31.1	10.8	14.7
Synthetic	✓	✓	✓	26.1	**34.9**	**11.8**	**18.1**
Synthetic	✗	✓	✗	**26.3**	33.1	10.7	17.3

Table 7. **Visual sound localization result.** Performance evaluation of sound localization using EZ-VSL [39], trained with either real or synthetized audio. Results are reported as CIoU (Center Intersection over Union) and AuC (Area under the Curve) on the Flickr [52] and VGG-SS [6] datasets.

method	Visual	Audio	Flickr [52]		VGG-SS [6]	
			CIoU	AuC	CIoU	AuC
EZ-VSL [39]	Real	Real	78.3	61.0	36.3	38.4
	Real	Syn	**81.9**	**62.1**	**36.5**	**38.6**
EZ-VSL$_{\text{obj}}$ [39]	Real	Real	81.5	63.0	40.8	40.4
	Real	Syn	**82.7**	**63.5**	**41.1**	**40.5**

4.4. Ablation Study

Synthesizing strategies. We analyze different image generation strategies using audio-to-image generation as a case: direct generation using audio-conditioned models [59] and three variants of our generation pipeline: direct generation with captions, SynCLR's caption engineering [61], and our caption engineering strategy, which placing the objects into a reasonable environment.

For evaluation, we first visualize some samples in Fig. 4.

Table 8. **Ablation on different generation strategies.** Performance comparison between different image generation strategies. Results are reported for audio classification accuracy (pretrained with CAVP [34]) and sound localization performance (pretrained with EZ-VSL [39]). In all cases, the audio data are maintained real.

Method	Audio CLS.		Sound Loc.	
	ESC-50	FSD-50k	CIoU	AuC
Any2Any [59]	74.0	44.0	18.1	33.3
Direct	82.0	45.1	73.4	57.8
SynCLR [61]	69.0	40.1	**83.1**	61.5
Ours	**82.0**	**45.7**	81.9	**62.1**

The images generated by audio-to-image models and the no caption edition version shared a common issue of limited diversity; for example, the generated drum images appear similar, sharing the same view, background, and lighting conditions. Audio-to-visual image generation also leads to more errors, such as failure in generating the guitar, because the audio cannot accurately capture all the visual information. We observe that prompt-engineered generation mitigates the issue of limited diversity. However, SynCLR's approach, which often makes extensive changes to the caption and introduces excessive elements, sometimes results in unreasonable images.

To further assess the quality of image generation, we use two performance metrics: audio classification and sound localization. For audio classification, we train a CAVP model and perform linear probing on ESC50 and FSD50k. For sound localization, we train an EZ-VSL model and test it on Flickr-SoundNet. The results are shown in Tab. 8. We find that models prioritizing the alignment of the generated images with the audio perform well in contrastive learning cases, as they are simpler to learn. The SynCLR model, however, nearly collapses, supporting our hypothesis that excessive caption modifications misalign text and audio, leading to uncorrelated images that hinder audio-visual learning. Nonetheless, the diversity introduced by SynCLR can greatly aid in sound localization. One potential reason is that it can generate objects in a variety of scales, which captures more possibilities in the real world. Compared to all the baselines, our method balances correctness and diversity, resulting in consistent improvements.

Frame numbers. We extend our experiment to train with multiple frames. Here we use CAV-MAE [19] as the backbone model for evaluation. For the real data, we uniformly sampled five frames for each video in the dataset. For the synthetic dataset, we sampled five times by using the same prompt. During the training process, the model will randomly choose one frame as the image input. The result is shown in Tab. 9. By adding samples, the real data's result improves quite a lot as more sample numbers reduce the possibility of choosing the bad frames. However, we still find that our model can still benefit from multi-samples and we outperform the ones trained with entirely real datasets.

Table 9. **Train with multiple samples.** Performance comparison on audio-visual classification between different frame sampling strategies. CAV-MAE [19] is used as backbone.

Visual	Audio	VGGSound
Single-real frame	Real	50.9
Multi-real frames	Real	51.1
Single-syn frame	Real	51.5
Multi-syn frames	Real	**51.9**

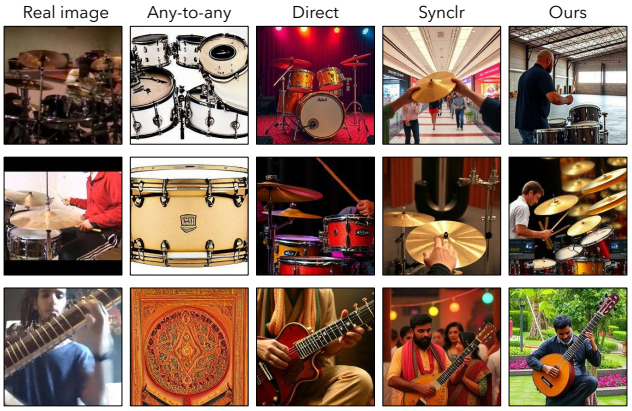

Figure 4. **Qualitative comparison between different generation strategies.** We sample images generated by different methods conditioned on the same audio. Real images are also included as a reference.

5. Conclusion

We propose a data generation pipeline that generates well-aligned audio images paired with text as a bridge, improving the training self-supervised audio-visual correspondence learning. Additionally, we introduce a data filtering method to combine real and synthetic data, further improving data quality. By training with our synthetic dataset, the models outperform those trained with real data on a wide range of downstream tasks. These findings highlight the potential of using synthetic data for scaling up self-supervised models and in the meantime making training more efficient.

Limitation and broader impact. All of our approaches use the popular transformer-based backbones and are trained only on the VGGSound dataset [5]. Expanding the scope of the experiments on various datasets could offer additional insights. Furthermore, our model's performance may be influenced by inherent biases and hallucinations present in generative models.

References

[1] Humam Alwassel, Dhruv Mahajan, Bruno Korbar, Lorenzo Torresani, Bernard Ghanem, and Du Tran. Self-supervised learning by cross-modal audio-video clustering, 2020. 2

[2] Anonymous. Beyond model collapse: Scaling up with synthesized data requires verification. In *Submitted to The Thirteenth International Conference on Learning Representations*, 2024. under review. 2

[3] Shekoofeh Azizi, Simon Kornblith, Chitwan Saharia, Mohammad Norouzi, and David J. Fleet. Synthetic data from diffusion models improves imagenet classification. *Transactions on Machine Learning Research*, 2023. 1, 2, 3

[4] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023. 12

[5] Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. Vggsound: A large-scale audio-visual dataset. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 721–725, 2020. 1, 2, 3, 4, 5, 6, 7, 8

[6] Honglie Chen, Weidi Xie, Triantafyllos Afouras, Arsha Nagrani, Andrea Vedaldi, and Andrew Zisserman. Localizing visual sounds the hard way. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16867–16876, 2021. 4, 7

[7] Ke Chen, Xingjian Du, Bilei Zhu, Zejun Ma, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. Hts-at: A hierarchical token-semantic audio transformer for sound classification and detection. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, 2022. 2, 12

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. 12

[9] Zach Evans, Julian D. Parker, CJ Carr, Zack Zukowski, Josiah Taylor, and Jordi Pons. Stable audio open, 2024. 4

[10] Lijie Fan, Dilip Krishnan, Phillip Isola, Dina Katabi, and Yonglong Tian. Improving CLIP training with language rewrites. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 2

[11] 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, 2022. 4, 6

[12] Chuang Gan, Deng Huang, Hang Zhao, Joshua B. Tenenbaum, and Antonio Torralba. Music gesture for visual sound separation, 2020. 2

[13] Ruohan Gao and Kristen Grauman. Co-separating sounds of visual objects, 2019. 2

[14] Jort F. Gemmeke, Daniel P. W. 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 *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 776–780, 2017. 1, 3, 5

[15] Scott Geng, Cheng-Yu Hsieh, Vivek Ramanujan, Matthew Wallingford, Chun-Liang Li, Pang Wei Koh, and Ranjay Krishna. The unmet promise of synthetic training images: Using retrieved real images performs better, 2024. 2

[16] Mariana-Iuliana Georgescu, Eduardo Fonseca, Radu Tudor Ionescu, Mario Lucic, Cordelia Schmid, and Anurag Arnab. Audiovisual masked autoencoders. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 16098–16108, 2023. 2

[17] Matthias Gerstgrasser, Rylan Schaeffer, Apratim Dey, Rafael Rafailov, Tomasz Korbak, Henry Sleight, Rajashree Agrawal, John Hughes, Dhruv Bhandarkar Pai, Andrey Gromov, Dan Roberts, Diyi Yang, David L. Donoho, and Sanmi Koyejo. Is model collapse inevitable? breaking the curse of recursion by accumulating real and synthetic data. In *First Conference on Language Modeling*, 2024. 2

[18] 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 *CVPR*, 2023. 2, 3, 5, 12

[19] Yuan Gong, Andrew Rouditchenko, Alexander H. Liu, David Harwath, Leonid Karlinsky, Hilde Kuehne, and James R. Glass. Contrastive audio-visual masked autoencoder. In *The Eleventh International Conference on Learning Representations*, 2023. 2, 4, 6, 7, 8, 12

[20] Andrey Guzhov, Federico Raue, Jörn Hees, and Andreas Dengel. Audioclip: Extending clip to image, text and audio, 2021. 2, 4, 6, 7, 12

[21] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners, 2021. 2

[22] Ruifei He, Shuyang Sun, Xin Yu, Chuhui Xue, Wenqing Zhang, Philip Torr, Song Bai, and XIAOJUAN QI. IS SYNTHETIC DATA FROM GENERATIVE MODELS READY FOR IMAGE RECOGNITION? In *The Eleventh International Conference on Learning Representations*, 2023. 1, 2

[23] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning, 2022. 3, 6, 12

[24] Po-Yao Huang, Vasu Sharma, Hu Xu, Chaitanya Ryali, haoqi fan, Yanghao Li, Shang-Wen Li, Gargi Ghosh, Jitendra Malik, and Christoph Feichtenhofer. Mavil: Masked audio-video learners. In *Advances in Neural Information Processing Systems*, pages 20371–20393. Curran Associates, Inc., 2023. 2

[25] E. Kidron, Y.Y. Schechner, and M. Elad. Pixels that sound. In *2005 IEEE Computer Society Conference on Computer*

Vision and Pattern Recognition (CVPR'05), pages 88–95 vol. 1, 2005. 2

[26] Douwe Kiela and Stephen Clark. Multi-and cross-modal semantics beyond vision: Grounding in auditory perception. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 2461–2470, 2015. 2

[27] Bruno Korbar, Du Tran, and Lorenzo Torresani. Cooperative learning of audio and video models from self-supervised synchronization, 2018. 2

[28] A. Krizhevsky and G. Hinton. Learning multiple layers of features from tiny images. *Master's thesis, Department of Computer Science, University of Toronto*, 2009. 4, 7

[29] Angeliki Lazaridou, Elia Bruni, and Marco Baroni. Is this a wampimuk? cross-modal mapping between distributional semantics and the visual world. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1403–1414, 2014. 2

[30] Sangho Lee, Jiwan Chung, Youngjae Yu, Gunhee Kim, Thomas Breuel, Gal Chechik, and Yale Song. Acav100m: Automatic curation of large-scale datasets for audio-visual video representation learning. In *ICCV*, 2021. 1

[31] Weixin Liang, James Zou, and Zhou Yu. Alice: Active learning with contrastive natural language explanations, 2020. 2

[32] Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling, 2023. 4, 13

[33] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023. 4, 5, 12

[34] Simian Luo, Chuanhao Yan, Chenxu Hu, and Hang Zhao. Diff-foley: Synchronized video-to-audio synthesis with latent diffusion models. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 2, 4, 6, 7, 8, 12

[35] Shuang Ma, Zhaoyang Zeng, Daniel McDuff, and Yale Song. Active contrastive learning of audio-visual video representations, 2021. 2

[36] Ziqiao Ma, Jiayi Pan, and Joyce Chai. World-to-words: Grounded open vocabulary acquisition through fast mapping in vision-language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 524–544, 2023. 2

[37] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. SDEdit: Guided image synthesis and editing with stochastic differential equations. In *International Conference on Learning Representations*, 2022. 12, 13

[38] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. HowTo100M: Learning a Text-Video Embedding by Watching Hundred Million Narrated Video Clips. In *ICCV*, 2019. 1

[39] Shentong Mo and Pedro Morgado. Localizing visual sounds the easy way. *arXiv preprint arXiv:2203.09324*, 2022. 2, 4, 7, 8

[40] Pedro Morgado, Ishan Misra, and Nuno Vasconcelos. Robust audio-visual instance discrimination, 2021. 2

[41] Yasuhide Mori, Hironobu Takahashi, and Ryuichi Oka. Image-to-word transformation based on dividing and vector quantizing images with words. In *First international workshop on multimedia intelligent storage and retrieval management*. Citeseer, 1999. 2

[42] Karol J Piczak. Esc: Dataset for environmental sound classification. In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 1015–1018, 2015. 4, 6

[43] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis, 2023. 13

[44] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 2

[45] Vignesh Ramanathan, Percy Liang, and Li Fei-Fei. Video event understanding using natural language descriptions. In *2013 IEEE International Conference on Computer Vision*, pages 905–912, 2013. 2

[46] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks, 2019. 4, 5

[47] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2022. 13

[48] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115 (3):211–252, 2015. 4, 7

[49] J. Salamon, C. Jacoby, and J. P. Bello. A dataset and taxonomy for urban sound research. In *22nd ACM International Conference on Multimedia (ACM-MM'14)*, pages 1041–1044, Orlando, FL, USA, 2014. 4, 6

[50] Pritam Sarkar and Ali Etemad. Self-supervised audio-visual representation learning with relaxed cross-modal synchronicity, 2022. 2

[51] Mohamed El Amine Seddik, Suei-Wen Chen, Soufiane Hayou, Pierre Youssef, and Merouane Abdelkader DEBBAH. How bad is training on synthetic data? a statistical analysis of language model collapse. In *First Conference on Language Modeling*, 2024. 2

[52] Arda Senocak, Tae-Hyun Oh, Junsik Kim, Ming-Hsuan Yang, and In So Kweon. Learning to localize sound source in visual scenes. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 4, 7

[53] Arda Senocak, Tae-Hyun Oh, Junsik Kim, Ming-Hsuan Yang, and In So Kweon. Learning to localize sound source in visual scenes: Analysis and applications. *TPAMI*, 2020. 4, 7

[54] Arda Senocak, Hyeonggon Ryu, Junsik Kim, Tae-Hyun Oh, Hanspeter Pfister, and Joon Son Chung. Sound source localization is all about cross-modal alignment. In *Proceedings*

of the IEEE/CVF International Conference on Computer Vision, pages 7777–7787, 2023. 2

[55] Arda Senocak, Hyeonggon Ryu, Junsik Kim, Tae-Hyun Oh, Hanspeter Pfister, and Joon Son Chung. Aligning sight and sound: Advanced sound source localization through audio-visual alignment. *arXiv preprint arXiv:2407.13676*, 2024. 2

[56] Guangzhi Sun, Wenyi Yu, Changli Tang, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun MA, Yuxuan Wang, and Chao Zhang. video-SALMONN: Speech-enhanced audio-visual large language models. In *Forty-first International Conference on Machine Learning*, 2024. 4

[57] Weixuan Sun, Jiayi Zhang, Jianyuan Wang, Zheyuan Liu, Yiran Zhong, Tianpeng Feng, Yandong Guo, Yanhao Zhang, and Nick Barnes. Learning audio-visual source localization via false negative aware contrastive learning, 2023. 2

[58] Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun MA, and Chao Zhang. SALMONN: Towards generic hearing abilities for large language models. In *The Twelfth International Conference on Learning Representations*, 2024. 4, 5, 12

[59] Zineng Tang, Ziyi Yang, Chenguang Zhu, Michael Zeng, and Mohit Bansal. Any-to-any generation via composable diffusion. *arXiv preprint arXiv:2305.11846*, 2023. 2, 7, 8

[60] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, pages 776–794. Springer, 2020. 4, 7

[61] Yonglong Tian, Lijie Fan, Kaifeng Chen, Dina Katabi, Dilip Krishnan, and Phillip Isola. Learning vision from models rivals learning vision from data, 2023. 1, 2, 3, 7, 8

[62] Yonglong Tian, Lijie Fan, Phillip Isola, Huiwen Chang, and Dilip Krishnan. Stablerep: Synthetic images from text-to-image models make strong visual representation learners. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 1, 2

[63] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models, 2023. 4

[64] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023. 12

[65] 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*, 2023. 2, 4, 6, 12

[66] Ziwei Xu, Sanjay Jain, and Mohan Kankanhalli. Hallucination is inevitable: An innate limitation of large language models, 2024. 3

[67] An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zhihao Fan. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024. 12

[68] Guy Yariv, Itai Gat, Lior Wolf, Yossi Adi, and Idan Schwartz. Audiotoken: Adaptation of text-conditioned diffusion models for audio-to-image generation. *Interspeech*, 2023. 2

[69] Ge Zhu and Zhiyao Duan. Cacophony: An improved contrastive audio-text model. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, abs/2402.06986, 2024. 2

[70] Qiushi Zhu, Long Zhou, Ziqiang Zhang, Shujie Liu, Binxing Jiao, Jie Zhang, Lirong Dai, Daxin Jiang, Jinyu Li, and Furu Wei. Vatlm: Visual-audio-text pre-training with unified masked prediction for speech representation learning. *IEEE Transactions on Multimedia*, 26:1055–1064, 2024. 2

A. Implementation details

A.1. Hyperparameter

For contrastive learning, we reproduce the CAVP [34], CLAP [65], and AudioCLIP [20], with a ViT-B/16 [8] model for visual, a transformer [64] for text and a HTSAT-tiny [7] for audio. For a fair comparison, we developed code for all three methods based on CLAP's code base [65] and used the same set of hyperparameters. For masked autoencoders, we follow CAV-MAE's implementation [19] where two ViT-B/16 encoders are trained for both modalities separately except the final block whose parameter is shared between the two modalities. The detailed parameter is shown in Tab. 10 and Tab. 11. For the parameters not mentioned, we use the default parameters contained in the original code base.

Table 10. **Hyper-parameter used to train CLAP, CAVP and AudioCLIP.** For finetune learning rate, three learning rates are searched to find the model with best performance.

config	Pretrain	Finetune
optimizer	Adam	Adam
base lr	1e-4	{1e-4, 2.5e-4, 5e-4}
weight decay	0	0
optimizer	β_1, β_2=0.9, 0.98	β_1, β_2=0.9, 0.98
batch size	64	32
lr schedule	Cosine	constant
epochs	30	10

Table 11. **Hyper-parameter used to train CAV-MAE [19].**

config	Pretrain	Finetune
optimizer	Adam	Adam
base lr	1e-4	1e-4
weight decay	5e-7	5e-7
optimizer	β_1, β_2=0.95, 0.999	β_1, β_2=0.95, 0.999
batch size	128	40
lr schedule	ReduceLROnPlateau	MultiStepLR
warmup epochs	10	0
full epochs	25	10
masking ratio	75%	N/A
contrast loss weight	0.01	N/A
mae loss weight	1.0	N/A
label smootht	0.0	0.1

A.2. Pipeline Model Selection

The choice of pretrained backbone models can significantly impact the performance of our pipeline. We experimented with multiple generators and captioners, as well as cross-modal matching models.

Cross-Modal Matching Models. We evaluated the performance of various cross-modal matching models for filtering, where these models compute similarity scores between audio and image modalities (Tab. 12). The datasets filtered using these similarity scores were subsequently used to pretrain the CAVP model, and the pretrained model's performance was assessed on audio classification tasks. For methods like LLAVA [33] and CLIP score [23] which only acquire visual and text as input, we directly use the label for filtering as this maximizes the correctness. For ImageBind [18], we use audio and visual as input. We found that large-language-model-based models like LLAVA encounter hallucination issues, which leads to a low score. CLIP also performs worse than ImageBind, as ImageBind can directly calculate the similarity between audio and visual. Thus, we chose ImageBind for filtering's backbone model.

Table 12. **Performance of different Cross-Modal Matching Models for filtering** Experiments are conducted on FSD-50k audio classification task with pretrained model using CAVP as the backbone.

Filter Method	modalities	FSD-50k
Pure synthetic	N/A	45.7
CLIP score	Visual-text	44.7
LLAVA	Visual-text	40.9
ImageBind	Audio-visual	**45.8**

Audio Captioning model. We evaluated the performance of several audio captioning models, including Salmonn [58], Qwen [4], and Qwen-v2 [67], using the downstream task performance of CLAP as the evaluation metric (Tab. 13). Based on these evaluations, we selected Salmonn as the final model due to its optimal balance of computational efficiency and classification performance.

Table 13. **Comparing audio captioner with CLAP.** Salmonn gives the best performance on FSD-50k audio classification.

Text	FSD-50k
Real label	46.1
Salmonn [58]	**46.3**
Qwen [4]	45.1
Qwenv2 [67]	46.2

Synthetic Image Generation and Editing. We explored the most effective methods for generating synthetic images by evaluating their impact on downstream audio classification tasks, with results summarized in Tab. 14. The performance of each method was measured using CAVP as the evaluation metric. In addition to generating images from scratch, we tested image editing methods, such as SDEdit [37], which attempts to add the sounding object directly to existing images. However, our experiments revealed that these editing methods consistently underperformed compared to images generated from scratch.

Among the tested methods, the Flux-schnell model has

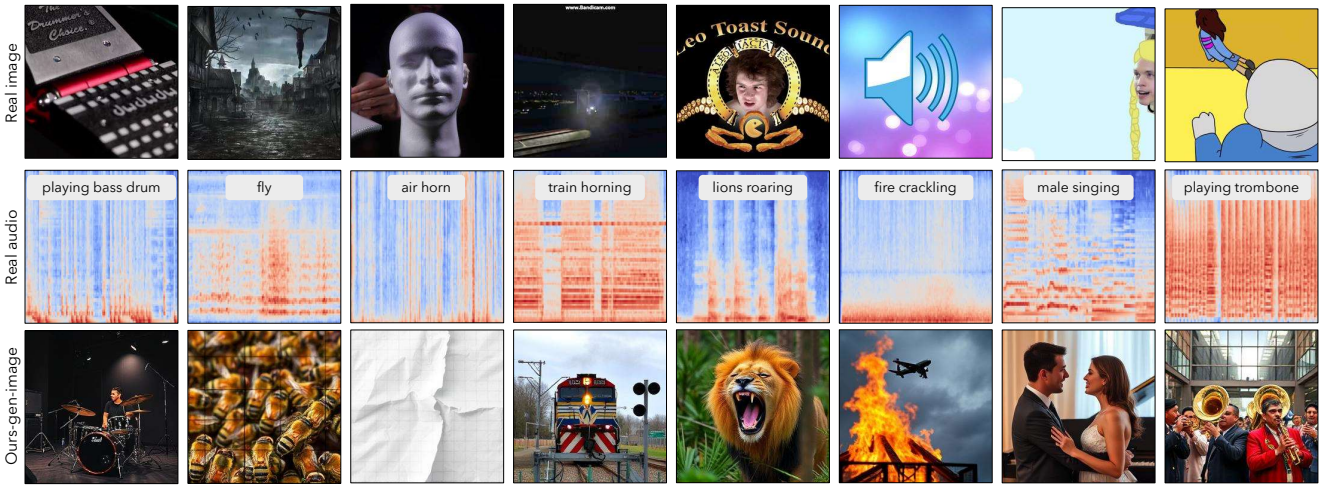

Figure 5. **More generated image examples.**

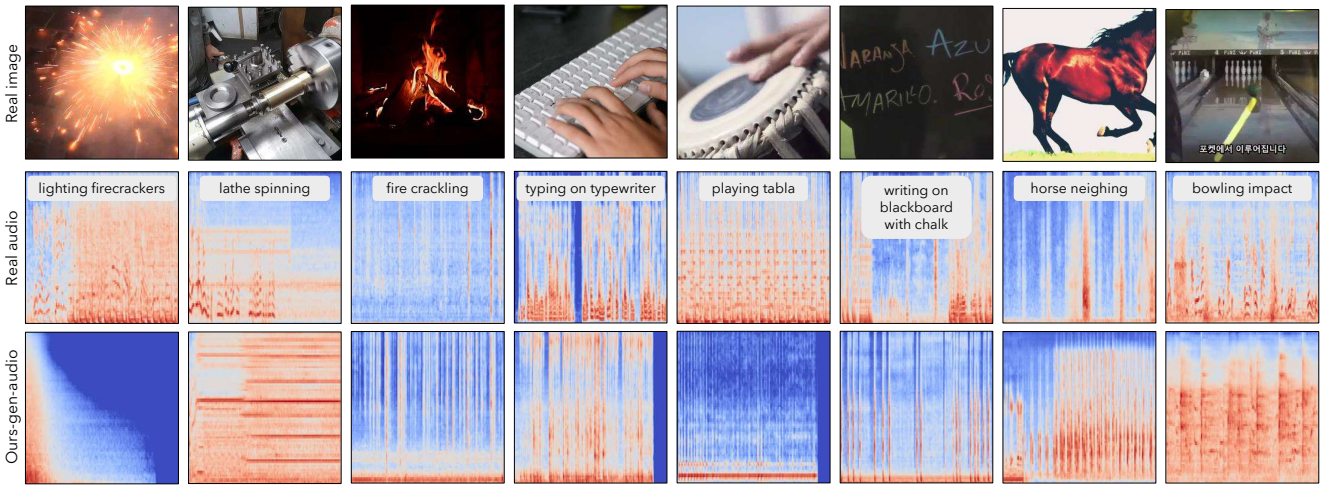

Figure 6. **More generated audio examples.**

the best performance, producing the highest accuracy on audio classification tasks while maintaining the fastest image generation speeds. This balance of speed and quality makes the distilled Flux model the most suitable choice for our synthetic image generation pipeline.

B. Filtering threshold

We perform a hyperparameter search on the best ratio of real data to keep. To improve the real image ratio, we always add the real image with the highest audio-visual score from the ImageBind. To increase the synthetic audio ratio, we substitute the audio with the lowest audio-visual score. We pre-train the model on VGGSound-variants and perform audio classification on ESC50. Results show that having 5% real image and 95% real audio gives the best accuracy.

Table 14. **Comparison of Image Generation and Editing Methods Using CAVP.** For SDEdit, Stable Diffusion 1.5 was used as the backbone. Generation speed was measured using a single NVIDIA A40 GPU with default hyperparameters. The distilled Flux model achieves the best performance on FSD-50k audio classification while offering the fastest generation speed.

Method	Speed (s/image)	FSD-50k
Real	-	40.1
SDEdit [37]	-	40.3
Stable-1.5 [47]	4	43.5
Stable-2.1 [47]	5	43.8
Stable-XL [43]	5	45.2
Flux-schnell [32]	1	**45.7**

Table 15. Performance with varying proportions of synthetic audio and image ratio.

	90% Syn Image	95% Syn Image	100% Syn Image
0% Syn Audio	82.0	83.3	82.0
5% Syn Audio	82.8	**83.8**	82.3
10% Syn Audio	83.3	81.8	82.0

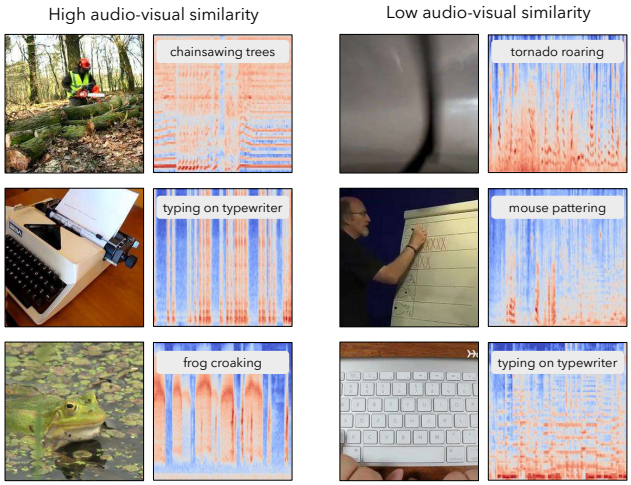

Figure 7. **Samples with high and low audio-visual similarity score.** Left are audio-visual samples with the highest ImageBind similarity, and right are samples with the lowest similarity.

C. Filtering samples

We demonstrate the filtered results with the top three highest scores and the lowest scores in Fig. 7. For the low-similarity samples, the first two incorrectly match the label, and the third one has significant background noise.

D. More examples

We provide more examples of our generated audio-visual pairs in Fig. 5 and Fig. 6. Example audio clips are also included in the video.

E. Human study

To elaborate on our human study, we present the website UI in Fig. 8 and provide a video demonstrating how it works. Essentially, users choose between synthetic samples and real samples to indicate which one is better or if both are good or bad. Samples given to each user are different and randomly chosen from the dataset. The order of real and synthetic samples is also randomly shuffled.

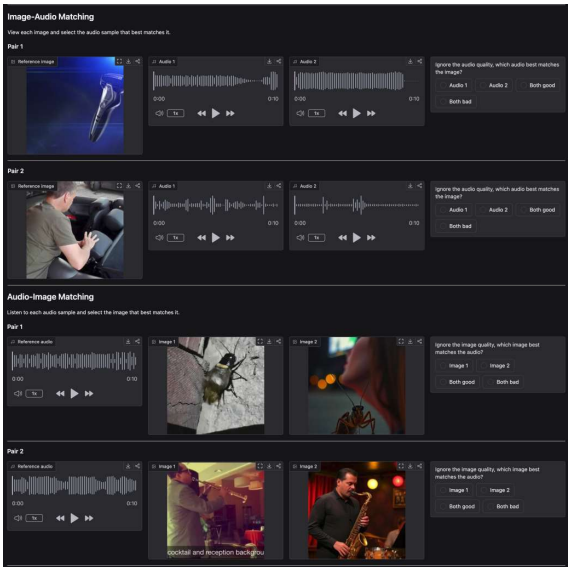

Figure 8. **User study UI.**