V2A-Mapper: A Lightweight Solution for Vision-to-Audio Generation by Connecting Foundation Models

Heng Wang[1]*, **Jianbo Ma**[2], **Santiago Pascual**[2], **Richard Cartwright**[2], **Weidong Cai**[1]

[1]University of Sydney
[2]Dolby Laboratories
{heng.wang, tom.cai}@sydney.edu.au, {jianbo.ma, santiago.pascual, richard.cartwright}@dolby.com

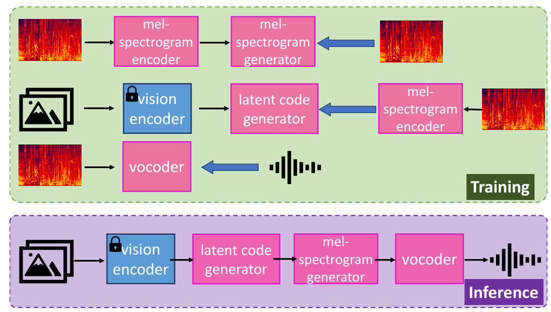
Abstract

Building artificial intelligence (AI) systems on top of a set of foundation models (FMs) is becoming a new paradigm in AI research. Their representative and generative abilities learnt from vast amounts of data can be easily adapted and transferred to a wide range of downstream tasks without extra training from scratch. However, leveraging FMs in cross-modal generation remains under-researched when audio modality is involved. On the other hand, automatically generating semantically-relevant sound from visual input is an important problem in cross-modal generation studies. To solve this vision-to-audio (V2A) generation problem, existing methods tend to design and build complex systems from scratch using modestly sized datasets. In this paper, we propose a lightweight solution to this problem by leveraging foundation models, specifically CLIP, CLAP, and AudioLDM. We first investigate the domain gap between the latent space of the visual CLIP and the auditory CLAP models. Then we propose a simple yet effective mapper mechanism (V2A-Mapper) to bridge the domain gap by translating the visual input between CLIP and CLAP spaces. Conditioned on the translated CLAP embedding, pretrained audio generative FM AudioLDM is adopted to produce high-fidelity and visually-aligned sound. Compared to previous approaches, our method only requires a quick training of the V2A-Mapper. We further analyze and conduct extensive experiments on the choice of the V2A-Mapper and show that a generative mapper is better at fidelity and variability (FD) while a regression mapper is slightly better at relevance (CS). Both objective and subjective evaluation on two V2A datasets demonstrate the superiority of our proposed method compared to current state-of-the-art approaches - trained with 86% fewer parameters but achieving 53% and 19% improvement in FD and CS, respectively. Supplementary materials such as audio samples are provided at our demo website: https://v2a-mapper.github.io/.

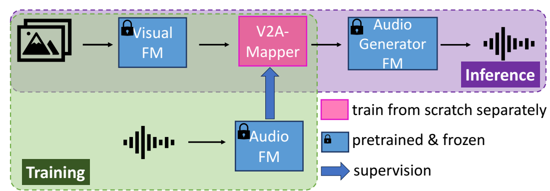
Introduction

Foundation models (FMs), trained on large-scale data and often making use of self-supervised learning, offer problem-agnostic representative or generative capabilities to downstream tasks via adaptation (Bommasani et al. 2021). They have demonstrated robust generalization and knowledge transfer ability across a broad spectrum of tasks in recent AI research (Zhou et al. 2023; Cao et al. 2023; Yin et al. 2023). Despite success in many uni-modal tasks spanning language (Paaß and Giesselbach 2023), vision (Awais et al. 2023), and audio (Li et al. 2023), the adaptation of FMs in problems involving multiple modalities such as cross-modal generation is greatly dominated by vision-language research (Du et al. 2022). Although attempts (Ao et al. 2022; Kreuk et al. 2023; Yang et al. 2023; Huang et al. 2023; Liu et al. 2023a,b; Yuan et al. 2023; Ghosal et al. 2023) have been made lately to bring FMs into text-to-audio generation and achieved remarkable performance, the viability of adopting FMs in vision-to-audio generation is still unclear.

Vision and audio are two essential and correlated sources through which people perceive the world. Humans have the



(a) Previous V2A generation approaches.



(b) Our lightweight solution to V2A generation.

Figure 1: Schematic illustrations of training and inference pipelines from previous V2A algorithms and our lightweight solution, respectively. Leveraging foundation models (FMs), we only require the training of a single V2A-Mapper while current works involve multiple modules to train.

*Work done during an internship at Dolby.

ability to imagine the corresponding sound when just observing a visual event (Owens and Efros 2018). Mimicry of this human-like cross-modal generation ability is applicable to various scenarios such as enhancing the experience of immersion in virtual reality, automating video editing for content creators, and assisting people with visual impairment (Ghose and Prevost 2020; Luo et al. 2023). Such rich visual-audio consistency and wide application have drawn constant interest in vision-to-audio (V2A) generation (Wei et al. 2022). Not restricted to a specific in-domain sound type (e.g., background music (Di et al. 2021), dance music (Zhu et al. 2023b), speech (Prajwal et al. 2020)), in this paper, we aim to generate natural sound from visual input in more diverse real-world scenarios, a V2A task that poses a markedly elevated level of difficulty (Zhou et al. 2018).

To solve this open-domain V2A generation problem, current methods (Iashin and Rahtu 2021; Sheffer and Adi 2023; Dong et al. 2023) often involve a complex system of separately optimized submodules trained with limited size of datasets as illustrated in Fig. 1(a). It could be cumbersome and resource-intensive to train each module individually and the generalization capability of each module could be restricted due to the lack of sufficient training data.

In this work, we explore the feasibility of adopting foundation models in open-domain vision-to-audio generation task. As shown in Fig. 1(b), our lightweight method only requires the training of a **V2A-Mapper** to bridge the domain gap between the vision representative FM CLIP (Radford et al. 2021) and the audio generative FM AudioLDM (Liu et al. 2023a). The V2A-Mapper is supervised by the audio representative FM CLAP (Wu et al. 2023) to learn the translation from visual space to auditory space. Leveraging the generalization and knowledge transfer ability of foundation models, the V2A-Mapper is trained with the same modestly sized dataset but the overall system can achieve much better performance. Our contribution includes: **1)** investigating the potential of bringing FMs into the field of vision-to-audio generation; **2)** proposing a simple but effective V2A-Mapper to connect visual and auditory FMs; **3)** investigating both generative and regression strategies of the V2A-Mapper; **4)** both subjective and objective evaluation on two V2A datasets demonstrate the efficiency and effectiveness of our method - it is trained with 86% fewer parameters but can achieve up to 53% and 19% improvement in fidelity (FD) and relevance (CS).

Related Works

Vision-to-Audio Generation. Earlier V2A works (Owens et al. 2016; Chen et al. 2017; Hao, Zhang, and Guan 2018) deal with limited sound in controlled environments. VEGAS (Zhou et al. 2018) for the first time introduced open-domain sound generation from in-the-wild visual input. But VEGAS and later works (Chen et al. 2018, 2020b) had to train a separate model for each sound type which is hard to scale up. To solve this issue, SpecVQGAN (Iashin and Rahtu 2021) designed the first label-free approach where a single model can produce diverse sound types. SpecVQGAN used a pretrained image classifier network to extract visual features from which a Transformer-based (Vaswani et al. 2017) autoregressive model synthesizes the mel-spectrogram. Upgrading this label-free approach, Im2Wav (Sheffer and Adi 2023) used the vision foundation model CLIP (Radford et al. 2021) to get visual features of multimodal semantic information. Instead of predicting the mel-spectrogram directly, Im2Wav autoregressively generates its latent code based on the visual prompt and a VQ-VAE (Van Den Oord, Vinyals et al. 2017) is trained to encode and decode between the mel-spectrogram and the latent space as shown in Fig. 1(a). Similar to Im2Wav, CLIPSonic-IQ (Dong et al. 2023) also adopted CLIP but they trained a diffusion model (Nichol and Dhariwal 2021) to directly generate the mel-spectrogram as in SpecVQGAN. All of these attempts train multiple modules with limited amount of data from scratch. In this paper, we propose to utilize FMs to inherit their generalization ability obtained from large-scale training. Optimizing a V2A-Mapper to connect FMs with the same modestly sized dataset, our method is lightweight in the training phase and effective in the generalization capability.

Foundation Model Adaptation. Adapting FMs to downstream tasks has been actively explored in NLP for uni-modal tasks, which can be categorized into prompt-based (Le Scao and Rush 2021), fine-tune-based (Zaken, Goldberg, and Ravfogel 2022; Hu et al. 2021), and lightweight adapter-based methods (Houlsby et al. 2019). When introducing this new paradigm into multimodal domain, pioneering works in the vision-language (VL) field follow the third strategy and freeze FMs to avoid catastrophic forgetting (McCloskey and Cohen 1989). PICa (Yang et al. 2022) considered language FM GPT3 as a knowledge base for visual question answering tasks while ClipCap (Mokady, Hertz, and Bermano 2021) and Flamingo (Alayrac et al. 2022) learnt auxiliary modules (i.e., interleaving new layers or tokens) to utilize vision and language FMs for image captioning. Compared to VL field, there is much less research on FM adaptation in vision-audio domain. In this paper, we propose a simple yet effective V2A-Mapper to connect visual and auditory FMs for open-domain V2A generation task. In line with VL works, we keep our FMs frozen but, unlike previous attempts, we do not change the inner architecture of FMs. Our method keeps FMs completely intact and only adds a mapper, which guarantees easy deployment and updating.

Method

Our lightweight solution includes a visual encoder FM (CLIP), an audio encoder FM (CLAP), an audio generator FM (AudioLDM), and a trainable V2A-Mapper. Fig. 2 presents how we train the V2A-Mapper with frozen CLIP and CLAP models and how we incorporate it with frozen CLIP and AudioLDM models to produce high-fidelity and visually-aligned sound. In this section, we first revisit the adopted foundation models. We then analyze the domain gap between visual and auditory spaces and introduce how we train the V2A-Mapper to bridge the gap. Lastly, we present the details of our generative diffusion-based V2A-Mapper.

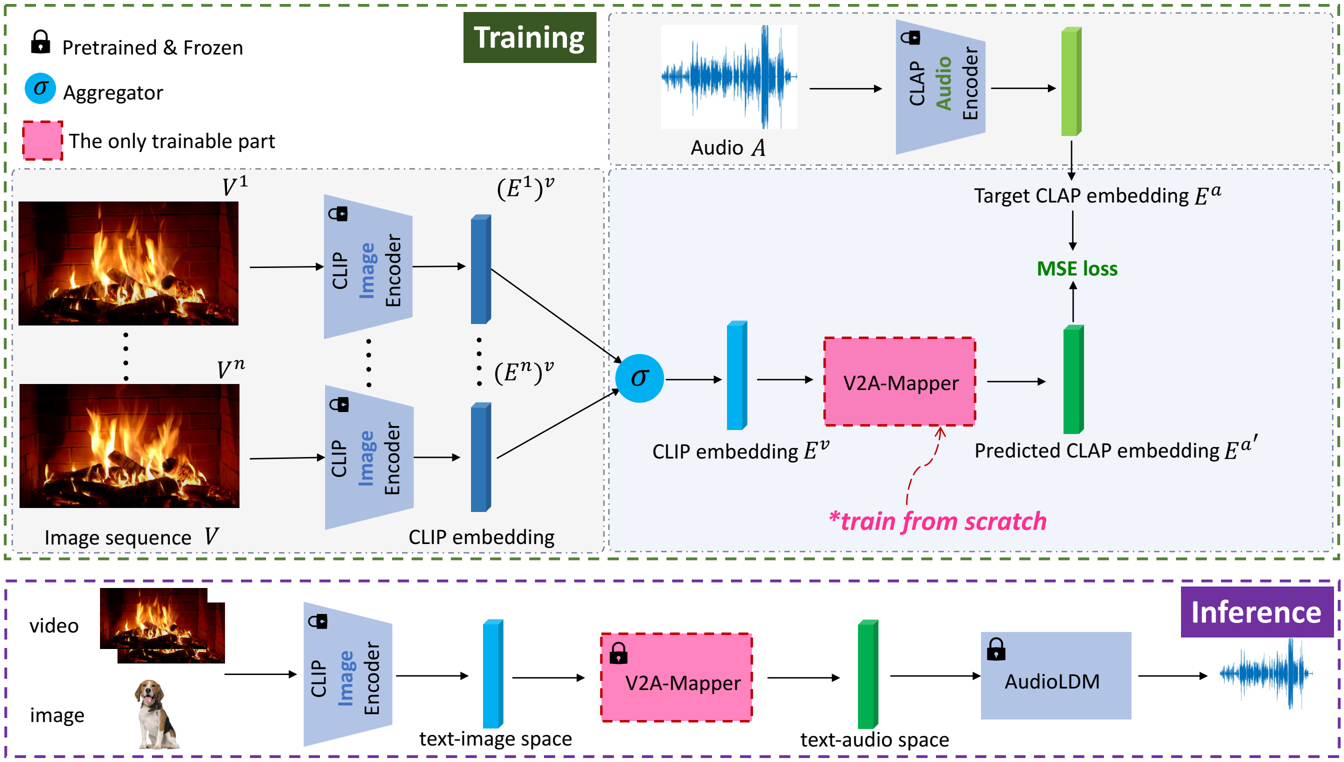


Figure 2: Top: Our lightweight V2A-Mapper training process. We first extract the source visual embedding E^v and the target audio embedding E^a with frozen pretrained foundation models CLIP and CLAP. We explore different aggregators σ to project the video data into a single feature vector. We then train the proposed V2A-Mapper using the audio-visual pair $\{E^v, E^a\}$ with MSE loss. Bottom: The compact inference pipeline of our method for vision-to-audio generation. We first adopt pretrained CLIP image encoder to project video/image into text-image space (the aggregation process for video input is omitted for brevity) and then use the trained V2A-Mapper to translate the visual embedding into CLAP text-audio space. Conditioned on the pseudo CLAP audio embedding, AudioLDM can be utilized to produce the sound waveform.

Selected Foundation Models

We choose the following foundation models because they are currently the state-of-the-art FMs for vision representation, audio representation, and audio generation, respectively. They can be replaced given better alternatives.

CLIP. As our V2A generation task spans across two modalities, adapting multimodal FMs is a natural way of utilizing their semantic features for tasks involving multiple domains (Lu et al. 2021). CLIP (Radford et al. 2021) is a text-image representation model which is trained to maximize the similarity between 400M paired text and image data via contrastive learning. Since the vision space learnt by CLIP is guided by language supervision which is of high-level semantic meaning, the visual feature is rich in semantic information. Therefore, we use a pretrained CLIP model to extract the features of visual prompts.

CLAP. CLAP (Wu et al. 2023) is currently the largest audio representation FM trained with 2.5M text-audio paired data. Similar to CLIP, CLAP learns a joint text-audio embedding space via contrastive learning under the language supervision. A critical reason we choose CLIP and CLAP is that they both share the text modality as a common domain during their training. We assume text could serve as a bridge which makes the translation from vision to audio easier.

AudioLDM. AudioLDM (Liu et al. 2023a) is a continuous latent diffusion model (LDM) trained in a self-supervised way with 3.3M 10-second audio clips. Conditioned on CLAP audio embedding, it generates the latent code of audio mel-spectrogram which can be decoded and converted into audio waveform. The original work only explores the usage of the LDM part in text-to-audio (T2A) generation task. Since CLAP represents text and audio jointly, AudioLDM can directly take text as input when being adapted to T2A task. We note that despite being proposed specifically for T2A generation task, AudioLDM is expected to adapt more naturally to audio features. This inspires us to ponder if we could translate a vision feature into its corresponding audio embedding in CLAP space, then we could keep AudioLDM completely intact and utilize it as an off-the-shelf audio generator FM.

Bridge the Domain Gap Between Vision and Audio

We first investigate if there exists a domain gap between the vision and audio spaces learnt by CLIP and CLAP respectively. Following (Liang et al. 2022), we measure it by ran-

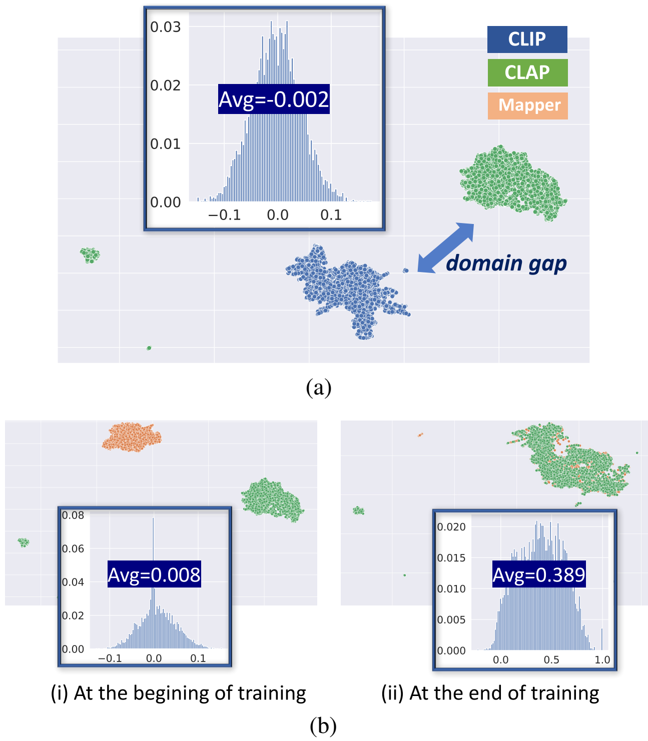


Figure 3: We visualize the domain gap between CLIP image space and CLAP audio space in (a). In (b), we present the process of closing the domain gap during the training of the V2A-Mapper. The accompanying histograms display the cosine similarity between paired embeddings from two domains. The larger the value is, the closer two domains are.

domly selecting 5000 samples from the video dataset VGGSound (Chen et al. 2020a) to get an estimation of both the visual and auditory feature distributions. Specifically, we encode the video frames into 512-d feature vectors with pretrained CLIP image encoder and average them along the time axis to get one single embedding for each video. For audio data, we project each audio sample into 512-d feature vector with pretrained CLAP audio encoder. We then use UMAP visualization (Sainburg, McInnes, and Gentner 2021) to project 5000 CLIP image embeddings and 5000 CLAP audio embeddings into the same 2-d space. As shown in Fig. 3(a), the average cosine similarity between paired visual (CLIP) and audio (CLAP) features is near 0 and there indeed exists a considerable gap between CLIP image domain and CLAP audio domain.

To bridge the domain gap, we propose to train a mapper, namely V2A-Mapper, between CLIP and CLAP so that the visual embedding could be translated into the CLAP space. The upper part of Fig. 2 shows the training pipeline. A video V_i is a sequence of n images $\left\{V_i^1, ..., V_i^n\right\}$. To get the visual embedding for a video, we use frozen CLIP model to encode each frame into 512-d feature vector to get a set of frame features $\left\{(E_i^1)^v, ..., (E_i^n)^v\right\}$. We then use an aggregator function σ to get a single vector E_i^v as the visual feature for the video input. The aggregator function could be: 1) randomly picking one vector; 2) picking the vector of the middle frame; 3) averaging along time axis. According to the experiments, the third option obtains the best performance in both fidelity and relevance. Similarly, for the paired audio data A_i, we encode it into 512-d feature vector E_i^a with frozen CLAP model. Once we have paired visual features E_i^v and auditory features E_i^a, we can train the mapper to convert the CLIP embedding E_i^v into a pseudo CLAP embedding $E_i^{a'}$. We use Mean Square Error loss to guide the training. The training process can be formulated as below:

$$L = \mathbb{E}_{i \sim [1,K]} \left[\| \ E_i^a - E_i^{a'} \ \|^2 \right], \tag{1}$$

where K is the batch size and $E_i^{a'}$ is from $mapper(E_i^v)$. Fig. 3(b) visualizes the domain shift after the training. Since the mapper is randomly initialized at the beginning, the translated embedding cluster is still far from the target CLAP space as displayed in Fig. 3(b)(i). When training finishes, the translated space and the target CLAP space become overlapped as suggested in Fig. 3(b)(ii) indicating the mapper is optimized successfully.

Diffusion-based V2A-Mapper

Since the mapper is expected to project the embedding from visual space to audio space, a natural way to implement the mapper is a stack of multilayer perceptrons (MLPs) as a one-to-one regression task. Inspired by DALLE2's prior model (Ramesh et al. 2022), we consider the projection process as a conditional generation task, which models a one-to-many mapping ensuring the diversity and generalization of the target audio distribution. Specifically, we train the mapper as a diffusion model (Ho, Jain, and Abbeel 2020; Song et al. 2020). It includes a forward process where Gaussian noises are gradually added to the target audio embedding $E_{i,0}^a$ until it approaches to a standard Gaussian distribution $E_{i,T}^a$ (i.e., completely random) for T timesteps and a reverse process where the target is gradually recovered from the noisy distribution by canceling the added noises with a network in a recursive manner. Following DALLE2, instead of predicting the intermediate noises added at each step (Ho, Jain, and Abbeel 2020), we directly predict the target audio embedding. Therefore, we train the mapper network f_θ to predict audio embedding $E_{i,0}^a$ based on the timestep t, the noisy audio embedding $E_{i,t}^a$ at timestep t, and the condition visual embedding E_i^v. Hence, the training objective in Eq. 1 can be formulated as:

$$L = \mathbb{E}_{i \sim [1,K], t \sim [1,T]} \left[\| \ E_{i,0}^a - f_\theta \left(t, E_{i,t}^a, E_i^v \right) \ \|^2 \right]. \tag{2}$$

We experiment with two different architectures for the mapper network f_θ - simple MLPs and Transformer. For the Transformer variant, we craft a learnable token of 512-d whose output from the Transformer is considered as the recovered audio embedding. We then take the time embedding, noisy audio embedding, as well as the visual condition as three other tokens of the same shape (i.e., 512-d) to the Transformer encoder to obtain the recovered audio embedding. For the simple MLP variant, we concatenate all the three tokens as input and output the final 512-d vector as the

Method	VGGSound			ImageHear	Infer. Time (s) ↓	#Trainable Param. (M) ↓
	FD ↓	FAD ↓	CS ↑	CS ↑		
Reference	0	0	8.925	-	-	-
Im2Wav	51.500	6.005	7.827	9.843	864.12	360.40
CLIPSonic-IQ	27.124	3.495	7.251	11.392	53.94	142.58
Ours	**24.168**	**0.841**	**9.720**	**11.950**	**35.45**	**48.83**

Table 1: Objective comparison with SOTA methods on VGGSound (video-to-sound generation) and ImageHear (image-to-sound generation). The inference time is measured as the average time spent for 100 samples through the whole pipeline from input visual prompts to output waveforms on one NVIDIA RTX A6000 GPU. Our method achieves the best on all the objective metrics.

Method	VGGSound		ImageHear	
	Fidelity ↑	Relevance ↑	Fidelity ↑	Relevance ↑
Reference	3.580±0.455	4.178±0.533	-	-
Im2Wav	1.838±0.511	2.415±0.645	1.840±0.502	2.705±0.398
CLIPSonic-IQ	2.533±0.522	2.140±0.551	2.888±0.502	3.215±0.291
Ours	**2.845±0.491**	**2.808±0.651**	**3.425±0.459**	**3.310±0.295**

Table 2: Subjective comparison with SOTA methods on VGGSound (video-to-sound generation) and ImageHear (image-to-sound generation). Our lightweight solution outperforms previous methods in both sound quality and the relevance to visual prompt from a human perception perspective.

predicted audio embedding via fully-connected network. We find the Transformer is a better way to incorporate the condition compared to simple concatenation in MLPs.

Experiments

Experimental Setup

Datasets. We train our V2A-Mapper and all the variants on VGGSound video dataset (Chen et al. 2020a). VGGSound contains 199,176 10-second video clips extracted from videos uploaded to YouTube with audio-visual correspondence. Note that VGGSound has never been used as training data for the foundation models we adapt. Following the original train/test splits, we train on 183,730 videos and evaluate on 15,446 videos. To testify the generalization ability of our V2A-Mapper, we also test on out-of-distribution dataset ImageHear (Sheffer and Adi 2023) which contains 101 images from 30 visual classes (2-8 images per class). We generate 10-second audio samples for all the evaluations.

Metrics. We measure the performance on two aspects, fidelity and the relevance to the visual prompt. Specifically, we use Fréchet Distance (FD) to measure the overall quality and variability of generated audio clips. FD computes the distance of embedding distributions between the synthesized and the real samples. To compare with previous methods, we also compute the Fréchet Audio Distance (FAD) (Kilgour et al. 2019). FD and FAD differ at the embedding extractor - FD uses PANNs (Kong et al. 2020) while FAD adopts VGGish (Hershey et al. 2017). Similar to (Liu et al. 2023a), we choose FD as our main evaluation metric regarding the sound quality since PANNs is superior to VGGish by considering long distance temporal change. For the relevance evaluation, we use CLIP-Score (CS) (Sheffer and Adi 2023) to get the cosine similarity between the CLIP embedding of the visual input and the Wav2CLIP (Wu et al. 2022) embedding of the generated sound. As Wav2CLIP learns an audio encoder via contrastive loss on VGGSound with the guidance of frozen CLIP image encoder, if the generated sound matches the visual input, the Wav2CLIP embedding is expected to be similar to its paired CLIP embedding.

Subjective Testing. To complement the objective metrics, we also conduct a listening test to measure the fidelity of the generated audio clips and their relevance to visual prompts from a human perception perspective. We ask 20 listeners to rate audio clips of 20 randomly selected visual samples on a discrete 5-point scale in terms of fidelity and relevance, respectively. The average rating across all listeners for each algorithm is computed as Mean Opinion Score (MOS) (International Telecommunication Union 1996). We also re-code the responses into paired comparisons and infer the relative standings via indirect scaling (Agresti 1992). We calculate the degree by which other approaches exceed our method as the Just Meaningful Difference (JMD) score (e.g. a negative value indicates inferiority of other algorithms compared to ours). More details of our human evaluation are provided in the supplementary.

Implementation Details. We use "ViT-B/32" version for CLIP model[1]. For CLAP model and audio generator, we use pretrained models from AudioLDM[2]. For the diffusion-

[1]https://github.com/openai/CLIP

[2]https://github.com/haoheliu/AudioLDM

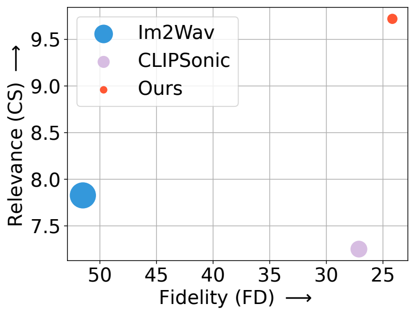


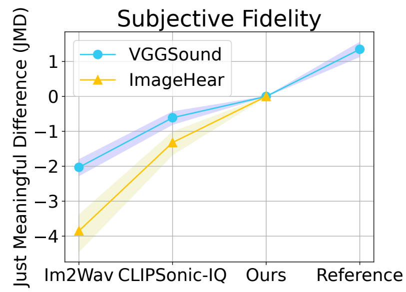
Figure 4: Our method achieves better results on both metrics and contains fewer trainable parameters (smaller circle size).

based V2A-Mapper, we use a cosine noise schedule with 1000 diffusion steps during training and 200 steps at inference time. We use AdamW with a learning rate of 1.1e-4, a batch size of 448 visual-audio embedding pairs, and a dropout rate of 0.1 in classifier-free guidance. We provide more implementation details including datasets used in adopted FMs, the full experiments of architecture hyperparameter tuning, and the guidance scale tuning in the supplementary.

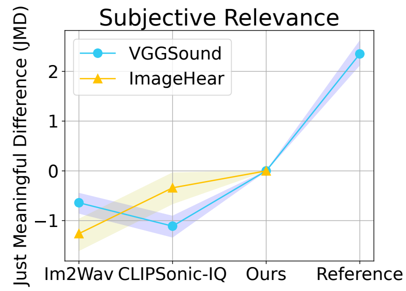
Compare with SOTA

Im2Wav (Sheffer and Adi 2023) is the current state-of-the-art method in open-domain vision-to-audio generation. It involves training of two Transformer decoders of different scales for latent code generation, a VQ-VAE for audio mel-spectrogram encoding and decoding, and a vocoder for waveform conversion. CLIPSonic-IQ (Dong et al. 2023) is a concurrent work to ours and they train a diffusion model to directly generate mel-spectrogram conditioned on visual representation. They also require the training of a BigVGAN (Lee et al. 2022) to convert the generated mel-spectrogram into audio waveform. Compared to these methods, our approach only requires the training of a single V2A-Mapper. Trained with the same modestly sized VGGSound data, our method achieves better performance as a result of the knowledge transfer from foundation models.

Objective Results. Tab. 1 shows that the proposed method achieves superior performance in all objective metrics. We also plot the comparison on VGGSound in Fig. 4 to showcase our method achieves better results on both relevance and fidelity and contains fewer trainable parameters. Compared to Im2Wav, our method trains with 86% fewer parameters but achieves 53% and 19% improvement in FD and CS, respectively. It is also noticeable that our method is significantly faster than Im2Wav (x24 faster) during inference. Our method also outperforms CLIPSonic-IQ in all the metrics with fewer parameters and faster inference speed. Note that our method exceeds even the reference for the relevance metric (CS). We conjecture that this is because VGGSound contains noisy data whose audio and visual streams might not be highly-relevant, which could suggest the proposed method is robust to noisy training data. We recommend readers to watch the sports live video in our demo website to observe this phenomenon.



(a) Fidelity.



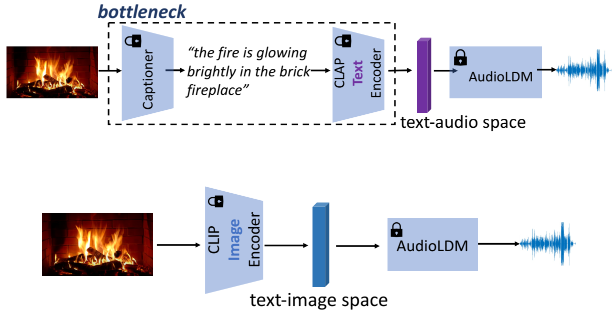
(b) Relevance.

Figure 5: The Just Meaningful Difference steps of current methods relative to our algorithm with 95% bootstrap confidence intervals.

Subjective Results. As shown in Tab. 2 and Fig. 5, our method exceeds previous works in both fidelity and relevance. We notice that the usage of diffusion model could especially boost the audio quality as indicated by the improvement achieved by both CLIPSonic-IQ and our method. While CLIPSonic-IQ fails at the relevance aspect when taking videos as input, our method consistently outperforms the SOTA method on both videos and images. However, we note that there is still a gap between our performance and the ground truth. Empirically, we find temporal alignment to be a major issue that leads to unsatisfactory relevance rating, which we will attempt to address in our future work.

Ablation Study

Different Ways of Utilizing FMs. Since AudioLDM is proposed for text-to-audio generation, the naive way of utilizing it for V2A synthesis is interleaving a captioning model to generate text input as shown in Fig. 6(a). To verify this vision-txt-audio idea, we adopt SOTA captioner BLIP (Li et al. 2022) to generate descriptions for images. For video-to-audio generation, we use the tag information provided in VGGSound. As reported in Tab. 3, although using text as bridge could mitigate the gap to some extent, it is still inferior to our method with the V2A-Mapper in both fidelity and relevance. This result indicates that the captioner is actually a bottleneck whose performance would directly affect what audio is to be generated from AudioLDM. We provide two examples where the captioner fails at predicting correct



(b) w/o mapper.

Figure 6: Different ways of using FMs. (a) adopts a captioner to generate text description as a bridge to use AudioLDM while (b) directly puts the visual features from CLIP image encoder as the condition to AudioLDM.

Method	VGGSound		ImageHear
	FD ↓	CS ↑	CS ↑
vision-txt-audio	56.397	6.672	7.310
w/o mapper	72.527	5.258	4.026
w/ mapper	**24.168**	**9.720**	**11.950**

Table 3: Ablation study with different ways of using FMs for vision-to-audio synthesis.

object category in "Why Vision-Text-Audio is Bottlenecked by the Captioner" section of our demo website. We suggest readers to check the audio results to examine the bottleneck challenge. Instead of decoding the visual condition into text format, our V2A-Mapper keeps the visual information as its latent code form and explicitly translates it from CLIP's visual space to CLAP's audio space, which could avoid information loss occurred during vision-txt-audio conversion. If the V2A-Mapper is skipped as illustrated in Fig. 6(b), the domain gap between vision and audio space prevents AudioLDM from generating high-fidelity and visually-relevant sound. Audio examples showcasing the difference are presented in "Domain Gap Bridging Process" section of our demo website. The recent text-to-audio generation work Make-An-Audio (Huang et al. 2023) trained their audio generator with CLIP text embedding and adopted CLIP image embedding as input to handle vision-to-audio generation. Similar to the "w/o mapper" strategy, the domain gap between the visual condition and the target embedding space which their audio generator works on is not addressed. We refer readers to our demo website to observe the comparison with Make-An-Audio.

Inside the Mapper: Generative vs. Regression. The V2A-Mapper can be implemented in a generative or a regression strategy. A generative V2A-Mapper learns a one-to-many mapping while a regression one builds a one-to-one projection. As displayed in Tab. 4, although regression model could learn a slightly better relevance due to the one-to-one mapping, the generated sound lacks diversity and fidelity as suggested by much worse FD scores. A generative mapper is critical to ensure the variability as also observed in text-to-image synthesis (Ramesh et al. 2022). To showcase the diversity of our method, we provide three samples for each visual input in the "Variability of Our V2A Generation Model" section of our demo website. And compared to linear projections, the attention mechanism used in Transformer could integrate the visual condition in a better way.

Arch. of the V2A-Mapper		VGGSound		ImageHear
		FD ↓	CS ↑	CS ↑
Regression	MLPs	35.059	9.927	12.048
	Transformer	**29.378**	**10.076**	**12.317**
Generative	diff. w/ MLPs	28.803	8.685	10.449
	diff. w/ Transformer	**24.168**	**9.720**	**11.950**

Table 4: Ablation study with different V2A-Mapper strategies (regression vs. generative) and architectures (MLPs vs. Transformer).

Aggregation	VGGSound		ImageHear
	FD ↓	CS ↑	CS ↑
random	24.826	9.200	11.465
middle	25.569	9.192	11.901
average	**24.168**	**9.720**	**11.950**

Table 5: Ablation study with different ways of aggregation for video feature representation during training.

Method		VGGSound		ImageHear
		FD ↓	CS ↑	CS ↑
w/o mapper	BLIP	**53.621**	4.948	**4.314**
	CLIP	72.527	**5.258**	4.026
w/ mapper	BLIP	24.788	9.402	10.836
	CLIP	**24.168**	**9.720**	**11.950**

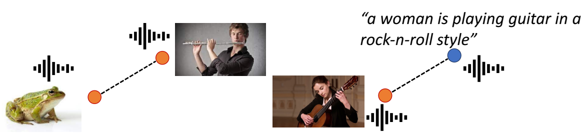
Table 6: Ablation study with different vision-language models.

Different Aggregator Methods σ. We explore three different ways of aggregating visual information of videos: 1) randomly select one frame as the key frame to represent the video; 2) instead of using random frame, choose the middle one; 3) average the CLIP features of all the frames along the time axis. Tab. 5 shows the performance of models trained with different aggregation methods. Since the task is to generate a large time-span (10 seconds) of a highly dynamic signal (audio), having time-related information in the condition could help. The average of abstract frame embeddings with rich semantic contents throughout the temporal dynamics is a better summary of the video than a single frame.

Different Pretrained Vision FMs. Our V2A-Mapper can be generalized to other vision-language models such as BLIP (Li et al. 2022). As shown in Tab. 6, the proposed

Method	VGGSound		ImageHear	Time (s) ↓	#Param. (M) ↓
	FD ↓	CS ↑	CS ↑		
audioldm-s	25.635	9.547	11.586	9.33	185.04
audioldm-s-v2	**24.168**	**9.720**	11.950	**9.33**	**185.04**
audioldm-l	25.130	9.531	**12.016**	11.58	739.14

Table 7: Ablation study with different pretrained AudioLDM models.



(a) Guided by image. (b) Guided by text.

Figure 7: Our V2A-Mapper enables interpolation guided by both image and text. Audios are provided in demo website.

V2A-Mapper can boost the performance of both CLIP- and BLIP-based systems. We also note that no matter what vision-language model is used and how big the domain gap between the vision and audio spaces is, the proposed V2A-Mapper can bridge the gap and translate visual information into audio space - two systems achieve similar performance with the proposed V2A-Mapper.

Different Pretrained Audio FMs. We ablate with different pretrained audio generators from AudioLDM: 1) audioldm-s is the base model; 2) audioldm-s-v2 is the base model but trained with more steps; 3) audioldm-l is the model with larger architecture. As shown in Tab. 7, either scaling the model up or optimizing its training for longer steps can help enhance the performance to some extent. Therefore, we hypothesize that a better audio generator FM could further improve the quality and relevance in the future.

Latent Space Interpolation

As the visual condition is translated into the CLAP latent space, we could interpolate audio embeddings by either visual or textual guidance. For simplicity, we perform linear interpolation between two embeddings. As shown in Fig. 7, the interpolation can happen from a frog sound to a sound indicated by an image of a man playing flute, or to a target specified by a description. It is noticeable that vision, text, and audio are semantically gathered to the same space without actual training with three modalities. We hear a relatively smooth transition during the interpolation, which indicates auditorily that the V2A-Mapper does learn the translation from CLIP space to CLAP space. Examples are provided in “Latent Space Interpolation” section of our demo website.

Limitation and Future Work

While our approach has achieved considerable success, it is important to acknowledge several limitations. First, the system can not achieve finer control. The generated sound exhibits semantic relevance in a general sense, but it lacks controllability over specific details. Second, the system fails when the visual cues involve unclear subjects (e.g., multiple objects, blurry/damaged images). Third, the system does not explicitly handle the temporal alignment between audio and visual signals. All of these could be interesting future directions in this area. Enforcing text into the condition could be a starting point for explicit controllability by considering both visual and textual features. To incorporate the language information of high-level semantic meaning into the system, recent multimodal foundation models such as Meta-Transformer (Zhang et al. 2023) and VATLM (Zhu et al. 2023a) could be taken into consideration. They learn the representation across vision, language, and audio which could shape a common space for different modalities.

Conclusion

In this paper, we explore the feasibility and efficiency of adapting foundation models (FMs) in the challenging open-domain vision-to-audio generation task. We propose a simple yet effective mapper mechanism (V2A-Mapper) to connect the representative visual FM CLIP and the generative auditory FM AudioLDM. Learning to translate visual features from CLIP space to the auditory CLAP space, the V2A-Mapper successfully passes visual information to its auditory counterpart from which the AudioLDM can synthesize high-fidelity and visually-aligned sound. Our method is relatively lightweight to train because it only requires optimization of the V2A-Mapper. Despite this simplicity, it achieves superior performance compared to current state-of-the-art approaches with far more complex training regimes as demonstrated by both subjective and objective evaluation.

Ethical Statement

Our method aims to leverage foundation models for efficient vision-to-audio generation. It can be used to enhance the immersion of human experience, such as video editing and foley design. Nevertheless, the application of this technology poses a risk if being maliciously misused on social platforms, potentially resulting in negative outcomes for society. Although significant strides have been made in audio deepfake detection research to mitigate such concerns (Yamagishi et al. 2021), the availability of ample datasets remains pivotal for improving detection accuracy. In light of this, we are committed to presenting our synthesized audio samples, intending to contribute to the advancement and fine-tuning of existing detection algorithms.

Acknowledgments

We extend our sincere appreciation to Xiaoyu Liu, Hannes Muesch, Adam Mater, Chunghsin Yeh, Michael Eckert, and Lie Lu for their valuable comments, corrections and inspiration. In addition, we thank Xiaoyu for providing the code of CLIPSonic-IQ and we are grateful to Hannes and Adam for discussing, designing, and analyzing the subjective testings.

References

Agresti, A. 1992. Analysis of ordinal paired comparison data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 41(2): 287–297.

Alayrac, J.-B.; Donahue, J.; Luc, P.; Miech, A.; Barr, I.; Hasson, Y.; Lenc, K.; Mensch, A.; Millican, K.; Reynolds, M.; et al. 2022. Flamingo: a visual language model for few-shot learning. In *NeurIPS*, volume 35, 23716–23736.

Ao, J.; Wang, R.; Zhou, L.; Wang, C.; Ren, S.; Wu, Y.; Liu, S.; Ko, T.; Li, Q.; Zhang, Y.; et al. 2022. SpeechT5: Unified-modal encoder-decoder pre-training for spoken language processing. In *ACL (Long Papers)*, 5723–5738.

Awais, M.; Naseer, M.; Khan, S.; Anwer, R. M.; Cholakkal, H.; Shah, M.; Yang, M.-H.; and Khan, F. S. 2023. Foundational models defining a new era in vision: A survey and outlook. *arXiv preprint arXiv:2307.13721*.

Bommasani, R.; Hudson, D. A.; Adeli, E.; Altman, R.; Arora, S.; von Arx, S.; Bernstein, M. S.; Bohg, J.; Bosselut, A.; Brunskill, E.; et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Cao, Y.; Li, S.; Liu, Y.; Yan, Z.; Dai, Y.; Yu, P. S.; and Sun, L. 2023. A comprehensive survey of AI-generated content (AIGC): A history of generative AI from GAN to ChatGPT. *arXiv preprint arXiv:2303.04226*.

Chen, H.; Xie, W.; Vedaldi, A.; and Zisserman, A. 2020a. VGGSound: A large-scale audio-visual dataset. In *ICASSP*, 721–725. IEEE.

Chen, K.; Zhang, C.; Fang, C.; Wang, Z.; Bui, T.; and Nevatia, R. 2018. Visually indicated sound generation by perceptually optimized classification. In *ECCV Workshops*.

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

Chen, P.; Zhang, Y.; Tan, M.; Xiao, H.; Huang, D.; and Gan, C. 2020b. Generating visually aligned sound from videos. *IEEE Transactions on Image Processing*, 29: 8292–8302.

Di, S.; Jiang, Z.; Liu, S.; Wang, Z.; Zhu, L.; He, Z.; Liu, H.; and Yan, S. 2021. Video background music generation with controllable music transformer. In *ACM Multimedia*, 2037–2045.

Dong, H.-W.; Liu, X.; Pons, J.; Bhattacharya, G.; Pascual, S.; Serrà, J.; Berg-Kirkpatrick, T.; and McAuley, J. 2023. CLIPSonic: Text-to-audio synthesis with unlabeled videos and pretrained language-vision models. *arXiv preprint arXiv:2306.09635*.

Du, Y.; Liu, Z.; Li, J.; and Zhao, W. X. 2022. A survey of vision-language pre-trained models. *arXiv preprint arXiv:2202.10936*.

Ghosal, D.; Majumder, N.; Mehrish, A.; and Poria, S. 2023. Text-to-Audio Generation using Instruction Guided Latent Diffusion Model. In *ACM Multimedia*, 3590–3598.

Ghose, S.; and Prevost, J. J. 2020. AutoFoley: Artificial synthesis of synchronized sound tracks for silent videos with deep learning. *IEEE Transactions on Multimedia*, 23: 1895–1907.

Hao, W.; Zhang, Z.; and Guan, H. 2018. CMCGAN: A uniform framework for cross-modal visual-audio mutual generation. In *AAAI*, volume 32.

Hershey, S.; Chaudhuri, S.; Ellis, D. P.; Gemmeke, J. F.; Jansen, A.; Moore, R. C.; Plakal, M.; Platt, D.; Saurous, R. A.; Seybold, B.; et al. 2017. CNN architectures for large-scale audio classification. In *ICASSP*, 131–135. IEEE.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *NeurIPS*, 33: 6840–6851.

Houlsby, N.; Giurgiu, A.; Jastrzebski, S.; Morrone, B.; De Laroussilhe, Q.; Gesmundo, A.; Attariyan, M.; and Gelly, S. 2019. Parameter-efficient transfer learning for NLP. In *ICML*, 2790–2799. PMLR.

Hu, E. J.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W.; et al. 2021. LoRA: Low-rank adaptation of large language models. In *ICLR*.

Huang, R.; Huang, J.; Yang, D.; Ren, Y.; Liu, L.; Li, M.; Ye, Z.; Liu, J.; Yin, X.; and Zhao, Z. 2023. Make-An-Audio: Text-to-audio generation with prompt-enhanced diffusion models. In *ICML*.

Iashin, V.; and Rahtu, E. 2021. Taming visually guided sound generation. In *BMVC*.

International Telecommunication Union. 1996. ITU-T Recommendation P.800: Methods for Subjective Determination of Transmission Quality.

Kilgour, K.; Zuluaga, M.; Roblek, D.; and Sharifi, M. 2019. Fréchet audio distance: A metric for evaluating music enhancement algorithms. *Interspeech*, 2350–2354.

Kong, Q.; Cao, Y.; Iqbal, T.; Wang, Y.; Wang, W.; and Plumbley, M. D. 2020. PANNs: Large-scale pretrained audio neural networks for audio pattern recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28: 2880–2894.

Kreuk, F.; Synnaeve, G.; Polyak, A.; Singer, U.; Défossez, A.; Copet, J.; Parikh, D.; Taigman, Y.; and Adi, Y. 2023. AudioGen: Textually guided audio generation. In *ICLR*.

Le Scao, T.; and Rush, A. M. 2021. How many data points is a prompt worth? In *ACL*, 2627–2636.

Lee, S.-g.; Ping, W.; Ginsburg, B.; Catanzaro, B.; and Yoon, S. 2022. BigVGAN: A universal neural vocoder with large-scale training. *arXiv preprint arXiv:2206.04658*.

Li, B.; Hwang, D.; Huo, Z.; Bai, J.; Prakash, G.; Sainath, T. N.; Sim, K. C.; Zhang, Y.; Han, W.; Strohman, T.; et al. 2023. Efficient domain adaptation for speech foundation models. In *ICASSP*, 1–5. IEEE.

Li, J.; Li, D.; Xiong, C.; and Hoi, S. 2022. BlIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, 12888–12900. PMLR.

Liang, V. W.; Zhang, Y.; Kwon, Y.; Yeung, S.; and Zou, J. Y. 2022. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. In *NeurIPS*, volume 35, 17612–17625.

Liu, H.; Chen, Z.; Yuan, Y.; Mei, X.; Liu, X.; Mandic, D.; Wang, W.; and Plumbley, M. D. 2023a. AudioLDM: Text-to-audio generation with latent diffusion models. In *ICML*, 21450–21474. PMLR.

Liu, H.; Tian, Q.; Yuan, Y.; Liu, X.; Mei, X.; Kong, Q.; Wang, Y.; Wang, W.; Wang, Y.; and Plumbley, M. D. 2023b. AudioLDM 2: Learning holistic audio generation with self-supervised pretraining. *arXiv preprint arXiv:2308.05734*.

Lu, K.; Grover, A.; Abbeel, P.; and Mordatch, I. 2021. Pretrained transformers as universal computation engines. In *AAAI*, 7628–7636.

Luo, S.; Yan, C.; Hu, C.; and Zhao, H. 2023. Diff-Foley: Synchronized video-to-audio synthesis with latent diffusion models. *arXiv preprint arXiv:2306.17203*.

McCloskey, M.; and Cohen, N. J. 1989. Catastrophic interference in connectionist networks: The sequential learning problem. *The Psychology of Learning and Motivation*, 24: 109–165.

Mokady, R.; Hertz, A.; and Bermano, A. H. 2021. ClipCap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734*.

Nichol, A. Q.; and Dhariwal, P. 2021. Improved denoising diffusion probabilistic models. In *ICML*, 8162–8171. PMLR.

Owens, A.; and Efros, A. A. 2018. Audio-visual scene analysis with self-supervised multisensory features. In *ECCV*, 631–648.

Owens, A.; Isola, P.; McDermott, J.; Torralba, A.; Adelson, E. H.; and Freeman, W. T. 2016. Visually indicated sounds. In *CVPR*, 2405–2413.

Paaß, G.; and Giesselbach, S. 2023. *Foundation models for natural language processing: Pre-trained language models integrating media*. Springer Nature.

Prajwal, K.; Mukhopadhyay, R.; Namboodiri, V. P.; and Jawahar, C. 2020. Learning individual speaking styles for accurate lip to speech synthesis. In *CVPR*, 13796–13805.

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 *ICML*, 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*.

Sainburg, T.; McInnes, L.; and Gentner, T. Q. 2021. Parametric UMAP embeddings for representation and semisupervised learning. *Neural Computation*, 33(11): 2881–2907.

Sheffer, R.; and Adi, Y. 2023. I hear your true colors: Image guided audio generation. In *ICASSP*, 1–5. IEEE.

Song, Y.; Sohl-Dickstein, J.; Kingma, D. P.; Kumar, A.; Ermon, S.; and Poole, B. 2020. Score-based generative modeling through stochastic differential equations. In *ICLR*.

Van Den Oord, A.; Vinyals, O.; et al. 2017. Neural discrete representation learning. In *NeurIPS*, volume 30.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *NeurIPS*, volume 30.

Wei, Y.; Hu, D.; Tian, Y.; and Li, X. 2022. Learning in audio-visual context: A review, analysis, and new perspective. *arXiv preprint arXiv:2208.09579*.

Wu, H.-H.; Seetharaman, P.; Kumar, K.; and Bello, J. P. 2022. Wav2CLIP: Learning robust audio representations from CLIP. In *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 *ICASSP*, 1–5. IEEE.

Yamagishi, J.; Wang, X.; Todisco, M.; Sahidullah, M.; Patino, J.; Nautsch, A.; Liu, X.; Lee, K. A.; Kinnunen, T.; Evans, N.; and Delgado, H. 2021. ASVspoof 2021: accelerating progress in spoofed and deepfake speech detection. In *Proc. ASVspoof Challenge Workshop*, 47–54.

Yang, D.; Yu, J.; Wang, H.; Wang, W.; Weng, C.; Zou, Y.; and Yu, D. 2023. DiffSound: Discrete diffusion model for text-to-sound generation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Yang, Z.; Gan, Z.; Wang, J.; Hu, X.; Lu, Y.; Liu, Z.; and Wang, L. 2022. An empirical study of GPT-3 for few-shot knowledge-based VQA. In *AAAI*, 3081–3089.

Yin, S.; Fu, C.; Zhao, S.; Li, K.; Sun, X.; Xu, T.; and Chen, E. 2023. A survey on multimodal large language models. *arXiv preprint arXiv:2306.13549*.

Yuan, Y.; Liu, H.; Liang, J.; Liu, X.; Plumbley, M. D.; and Wang, W. 2023. Leveraging pre-trained AudioLDM for sound generation: A benchmark study. In *EUSIPCO*.

Zaken, E. B.; Goldberg, Y.; and Ravfogel, S. 2022. BitFit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. In *ACL (Short Papers)*, volume 2, 1–9.

Zhang, Y.; Gong, K.; Zhang, K.; Li, H.; Qiao, Y.; Ouyang, W.; and Yue, X. 2023. Meta-Transformer: A Unified Framework for Multimodal Learning. *arXiv preprint arXiv:2307.10802*.

Zhou, C.; Li, Q.; Li, C.; Yu, J.; Liu, Y.; Wang, G.; Zhang, K.; Ji, C.; Yan, Q.; He, L.; et al. 2023. A comprehensive survey on pretrained foundation models: A history from BERT to ChatGPT. *arXiv preprint arXiv:2302.09419*.

Zhou, Y.; Wang, Z.; Fang, C.; Bui, T.; and Berg, T. L. 2018. Visual to sound: Generating natural sound for videos in the wild. In *CVPR*, 3550–3558.

Zhu, Q.; Zhou, L.; Zhang, Z.; Liu, S.; Jiao, B.; Zhang, J.; Dai, L.; Jiang, D.; Li, J.; and Wei, F. 2023a. VATLM: Visual-audio-text pre-training with unified masked prediction for speech representation learning. *IEEE Transactions on Multimedia*.

Zhu, Y.; Wu, Y.; Olszewski, K.; Ren, J.; Tulyakov, S.; and Yan, Y. 2023b. Discrete contrastive diffusion for cross-modal and conditional generation. In *ICLR*.