Make-An-Audio: Text-To-Audio Generation with Prompt-Enhanced Diffusion Models

Rongjie Huang [*1] **Jiawei Huang** [*1] **Dongchao Yang** [*2] **Yi Ren** [3] **Luping liu** [1] **Mingze Li** [1] **Zhenhui Ye** [1] **Jinglin Liu** [1] **Xiang Yin** [3] **Zhou Zhao** [1]

Abstract

Large-scale multimodal generative modeling has created milestones in text-to-image and text-to-video generation. Its application to audio still lags behind for two main reasons: the lack of large-scale datasets with high-quality text-audio pairs, and the complexity of modeling long continuous audio data. In this work, we propose Make-An-Audio with a prompt-enhanced diffusion model that addresses these gaps by 1) introducing pseudo prompt enhancement with a distill-then-reprogram approach, it alleviates data scarcity with orders of magnitude concept compositions by using language-free audios; 2) leveraging spectrogram autoencoder to predict the self-supervised audio representation instead of waveforms. Together with robust contrastive language-audio pretraining (CLAP) representations, Make-An-Audio achieves state-of-the-art results in both objective and subjective benchmark evaluation. Moreover, we present its controllability and generalization for X-to-Audio with **"No Modality Left Behind"**, for the first time unlocking the ability to generate high-definition, high-fidelity audios given a user-defined modality input.[1]

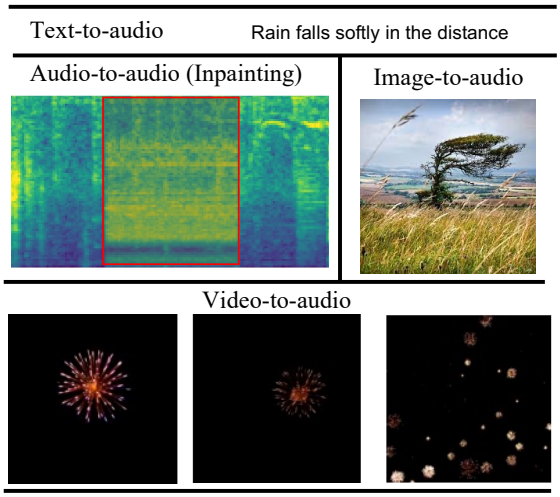


Figure 1. **No Modality Left Behind:** Make-An-Audio generalizes well to X-to-Audio with multiple user-defined inputs (text, audio, image and video), it empowers humans to create rich and diverse audio content, opening up to a various applications with personalized transfer and fine-grained control.

1. Introduction

Deep generative models (Goodfellow et al., 2020; Kingma & Dhariwal, 2018; Ho et al., 2020) have recently exhibited high-quality samples in various data modalities. With large-scale training data and powerful models, kinds of text-to-image (Saharia et al., 2022; Ramesh et al., 2021; Nichol et al., 2021) and text-to-video (Singer et al., 2022; Hong et al., 2022) models are now able to vividly depict the visual scene described by a text prompt, and empower humans to create rich and diverse visual content with unprecedented ease. However, replicating this success for audios is limited for the lack of large-scale datasets with high-quality text-audio pairs, and the extreme complexity of modeling long continuous signal data.

In this work, we propose Make-An-Audio, with a prompt-enhanced diffusion model for text-to-audio (T2A) generation. To alleviate the issue of data scarcity, we introduce a pseudo prompt enhancement approach to construct natural languages that align well with audio, opening up the usage of orders of magnitude unsupervised language-free data. To tackle the challenge of modeling complex audio signals in T2A generation, we introduce a spectrogram autoencoder to predict the self-supervised representations instead of waveforms, which guarantees efficient compression and high-level semantic understanding. Together with the power of contrastive language-audio pretraining (CLAP) (Radford et al., 2021; Elizalde et al., 2022) and high-fidelity diffusion

[*]Equal contribution [1]Zhejiang University [2]Peking University [3]Speech & Audio Team, ByteDance AI Lab. Correspondence to: Zhou Zhao <ZhaoZhou@zju.edu.cn>.

Proceedings of the 40th International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

[1]Audio samples are available at https://Make-An-Audio.github.io

models (Ho et al., 2020; Song et al., 2020; Rombach et al., 2022), it achieves a deep level of language understanding with high-fidelity generation.

While conceptually simple and easy to train, Make-An-Audio yields surprisingly strong results. Both subjective and objective evaluations demonstrate that Make-An-Audio achieves new state-of-the-art in text-to-audio with natural and controllable synthesis. Make-An-Audio exhibits superior audio quality and text-audio alignment faithfulness on the benchmark AudioCaption dataset and even generalizes well to the unsupervised Clotho dataset in a zero-shot fashion.

For the first time, we contextualize the need for audio generation with different input modalities. Besides natural language, Make-An-Audio generalizes well to multiple user-defined input modalities (audio, image, and video), which empowers humans to create rich and diverse audio content and opens up a host of applications for personalized transfer and fine-grained control.

Key contributions of the paper include:

- We present Make-An-Audio – an effective method that leverages latent diffusion with a spectrogram autoencoder to model the long continuous waveforms.
- We introduce a pseudo prompt enhancement with the distill-then-reprogram approach, it includes a large number of concept compositions by opening up the usage of language-free audios to alleviate data scarcity.
- We investigate textual representation and emphasize the advantages of contrastive language-audio pretraining for a deep understanding of natural languages with computational efficiency.
- We evaluate Make-An-Audio and present state-of-the-art quantitative results and thorough evaluation with qualitative findings.
- We generalize the powerful model to X-to-Audio generation, for the first time unlocking the ability to generate high-definition, high-fidelity audios given a user-defined modality input.

2. Related Works

2.1. Text-Guided Image/Video Synthesis

With the rapid development of deep generative models, text-guided synthesis has been widely studied in images and videos. The pioneering work of DALL-E (Ramesh et al., 2021) encodes images into discrete latent tokens using VQ-VAE (Van Den Oord et al., 2017) and considers T2I generation as a sequence-to-sequence translation problem. More recently, impressive visual results have been achieved by leveraging large-scale diffusion models. GLIDE (Nichol et al., 2021) trains a T2I upsampling model for a cascaded generation. Imagen (Saharia et al., 2022) presents T2I with an unprecedented degree of photorealism and a deep level of language understanding. Stable diffusion (Rombach et al., 2022) utilizes latent space diffusion instead of pixel space to improve computational efficiency. A large body of work also explores the usage of T2I models for video generation. CogVideo (Hong et al., 2022) is built on top of a CogView2 (Ding et al., 2022) T2I model with a multi-frame-rate hierarchical training strategy. Make-A-Video (Singer et al., 2022) extends a diffusion-based T2I model to T2V through a spatiotemporally factorized diffusion model.

Moving beyond visual generation, our approach aims to generate high-fidelity audio from arbitrary natural language, which has been relatively overlooked.

2.2. Text-Guided Audio Synthesis

While there is remarkable progress in text-guided visual generation, the progress of text-to-audio (T2A) generation lags behind mainly due to two main reasons: the lack of large-scale datasets with high-quality text-audio pairs, and the complexity of modeling long continuous waveforms data. DiffSound (Yang et al., 2022) is the first to explore text-to-audio generation with a discrete diffusion process that operates on audio codes obtained from a VQ-VAE, leveraging masked text generation with CLIP representations. AudioLM (Borsos et al., 2022) introduces the discretized activations of a masked language model pre-trained on audio and generates syntactically plausible speech or music.

Very recently, the concurrent work AudioGen (Kreuk et al., 2022) propose to generate audio samples autoregressively conditioned on text inputs, while our proposed method differentiates from it in the following: 1) we introduce pseudo prompt enhancement and leverage the power of contrastive language-audio pre-training and diffusion models for high-fidelity generation. 2) We predict the continuous spectrogram representations, significantly improving computational efficiency and reducing training costs.

2.3. Audio Representation Learning

Different from modeling fine-grain details of the signal, the usage of high-level self-supervised learning (SSL) (Baevski et al., 2020; Hsu et al., 2021; Huang et al., 2022d; Xun et al., 2023) has been shown to effectively reduce the sampling space of generative algorithms. Inspired by vector quantization (VQ) techniques, SoundStream (Zeghidour et al., 2021) and HiFi-Codec (Yang et al., 2023) present a hierarchical architecture for high-level representations that carry semantic information. Data2vec (Baevski et al., 2022) uses a fast convolutional decoder and explores the contextualized

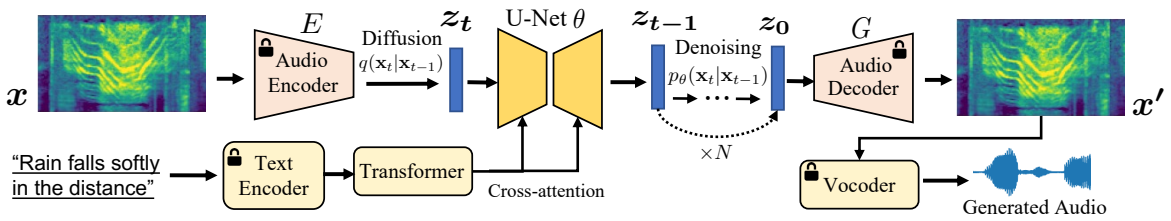


Figure 2. A high-level overview of Make-An-Audio. Note that some modules (printed with a *lock*) are frozen for training the T2A model.

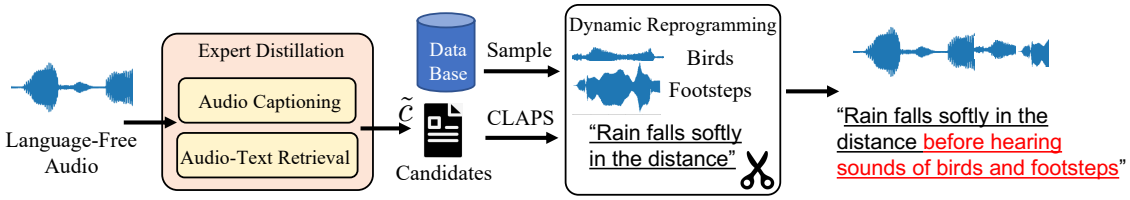


Figure 3. The process of pseudo prompt enhancement. Our semi-parametric diffusion model consists of a fixed expert distillation and a dynamic reprogramming stage. The database D contains audio examples with a sampling strategy ξ to create unseen object compositions. We use CLAPS to denote the CLAP selection.

target representations in a self-supervised manner.

Recently, spectrograms (akin to 1-channel 2D images) autoencoder (Gong et al., 2022; He et al., 2022) with reconstruction objective as self-supervision have demonstrated the effectiveness of heterogeneous image-to-audio transfer, advancing the field of speech and audio processing on a variety of downstream tasks. Among these approaches, Xu et al. (2022) study the Masked Autoencoders (MAE) (He et al., 2022) to self-supervised representation learning from audio spectrograms. Gong et al. (2022) adopt audio spectrogram transformer with joint discriminative and generative masked spectrogram modeling. Inspired by these, we inherit the recent success of spectrogram SSL in the frequency domain, which guarantees efficient compression and high-level semantic understanding.

3. Make-An-Audio

In this section, we first overview the Make-An-Audio framework and illustrate pseudo prompt enhancement to better align text and audio semantics, following which we introduce textual and audio representations for multimodal learning. Together with the power of diffusion models with classifier-free guidance, Make-An-Audio explicits high-fidelity synthesis with superior generalization.

3.1. Overview

Deep generative models have achieved leading performances in text-guided visual synthesis. However, the current development of text-to-audio (T2A) generation is hampered by two major challenges: 1) Model training is faced with data scarcity, as human-labeled audios are expensive to create, and few audio resources provide natural language descriptions. 2) Modeling long continuous waveforms (e.g., typically 16,000 data points for 1s 16 kHz waveforms) poses a challenge for all high-quality neural synthesizers.

As illustrated in Figure 2, Make-An-Audio consists of the following main components: 1) the pseudo prompt enhancement to alleviate the issue of data scarcity, opening up the usage of orders of magnitude language-free audios; 2) a spectrogram autoencoder for predicting self-supervised representation instead of long continuous waveforms; 3) a diffusion model that maps natural language to latent representations with the power of contrastive language-audio pretraining (CLAP) and 4) a separately-trained neural vocoder to convert mel-spectrograms to raw waveforms. In the following sections, we describe these components in detail.

3.2. Pseudo Prompt Enhancement: Distill-then-Reprogram

To mitigate the data scarcity, we propose to construct prompts aligned well with audios, enabling a better understanding of the text-audio dynamics from orders of magnitude unsupervised data. As illustrated in Figure 3, it consists of two stages: an expert distillation approach to produce prompts aligned with audio, and a dynamic reprogramming procedure to construct a variety of concept compositions.

3.2.1. Expert Distillation

We consider the pre-trained automatic audio captioning (Xu et al., 2020) and audio-text retrieval (Deshmukh et al., 2022;

Koepke et al., 2022) systems as our experts for prompt generation. Captioning models aim to generate diverse natural language sentences to describe the content of audio clips. Audio-text retrieval takes a natural language as a query to retrieve relevant audio files in a database. To this end, experts jointly distill knowledge to construct a caption aligned with audio, following which we select from these candidates that endow high CLAP (Elizalde et al., 2022) score as the final caption (we include a threshold to selectly consider faithful results). This simple yet effective procedure largely alleviates data scarcity issues and explicit generalization to different audio domains, and we refer the reader to Section 6.3.2 for a summary of our findings. Details have been attached in Appendix E.2.

3.2.2. Dynamic Reprogramming

To prevent overfitting and enable a better understanding of concept compositions, we introduce a dynamic reprogramming technique that constructs a variety of concept compositions. It proceeds in three steps as illustrated in Figure 3, where we elaborate the process as follows: 1) We first prepare our sound event database D annotated with a single label. 2) Each time N concepts are sampled from the database D, where $N \in \{0, 1, 2\}$. 3) The original text-audio pair data has been randomly concatenated with the sampled events according to the template, constructing a new training example with varied concept compositions. It can be conducted online, significantly reducing the time consumed for data preparation. The reprogramming templates are attached in Appendix F.

3.3. Textual Representation

Text-guided synthesis models need powerful semantic text encoders to capture the meaning of arbitrary natural language inputs, which could be grouped into two major categories: 1) Contrastive pretraining. Similar to CLIP (Radford et al., 2021) pre-trained on image-text data, recent progress on contrastive language-audio pretraining (CLAP) (Elizalde et al., 2022) brings audio and text descriptions into a joint space and demonstrates the outperformed zero-shot generalization to multiple downstream domains. 2) Large-scale language modeling (LLM). Saharia et al. (2022) and Kreuk et al. (2022) utilize language models (e.g., BERT (Devlin et al., 2018), T5 (Raffel et al., 2020)) for text-guided generation. Language models are trained on text-only corpus significantly larger than paired multimodal data, thus being exposed to a rich distribution of text.

Following the common practice (Saharia et al., 2022; Ramesh et al., 2022), we freeze the weights of these text encoders. We find that both CLAP and T5-Large achieve similar results on benchmark evaluation, while CLAP could be more efficient without offline computation of embeddings required by LLM. We refer the reader to Section 6.3.1 for a summary of our findings.

3.4. Audio Representation

Recently, spectrograms (akin to 1-channel 2D images) autoencoder (Gong et al., 2022; He et al., 2022) with reconstruction objective as self-supervision have demonstrated the effectiveness of heterogeneous image-to-audio transfer, advancing the field of speech and audio processing on a variety of downstream tasks. The audio signal is a sequence of mel-spectrogram sample $\boldsymbol{x} \in [0, 1]^{C_{\mathrm{a}} \times T}$, where C_{a}, T respectively denote the mel channels and the number of frames. Our spectrogram autoencoder is composed of 1) an encoder network E which takes samples $\boldsymbol{x}$ as input and outputs latent representations z; 2) a decoder network G reconstructs the mel-spectrogram signals $\boldsymbol{x'}$ from the compressed representation z; and 3) a multi-window discriminator Dis learns to distinguish the generated samples $G(z)$ from real ones in different multi-receptive fields of mel-spectrograms.

The whole system is trained end-to-end to minimize 1) Reconstruction loss $\mathcal{L}_{re}$, which improves the training efficiency and the fidelity of the generated spectrograms; 2) GAN losses $\mathcal{L}_{GAN}$, where the discriminator and generator play an adversarial game; and 3) KL-penalty loss $\mathcal{L}_{KL}$, which restricts spectrogram encoders to learn standard z and avoid arbitrarily high-variance latent spaces.

To this end, Make-An-Audio takes advantage of the spectrogram autoencoder to predict the self-supervised representations instead of waveforms. It largely alleviates the challenges of modeling long continuous data and guarantees high-level semantic understanding.

3.5. Generative Latent Diffusion

We implement our method over Latent Diffusion Models (LDMs) (Rombach et al., 2022), a recently introduced class of Denoising Diffusion Probabilistic Models (DDPMs) (Ho et al., 2020) that operate in the latent space. It is conditioned on textual representation, breaking the generation process into several conditional diffusion steps. The training loss is defined as the mean squared error in the noise $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ space, and efficient training is optimizing a random term of t with stochastic gradient descent:

$$\mathcal{L}_\theta = \left\| \boldsymbol{\epsilon}_\theta(\mathbf{z}_t, t, c) - \boldsymbol{\epsilon} \right\|_2^2, \tag{1}$$

where α denotes the small positive constant, and $\boldsymbol{\epsilon}_\theta$ denotes the denoising network. To conclude, the diffusion model can be efficiently trained by optimizing ELBO without adversarial feedback, ensuring extremely faithful reconstructions that match the ground-truth distribution. Detailed formulation of DDPM has been attached in Appendix D.

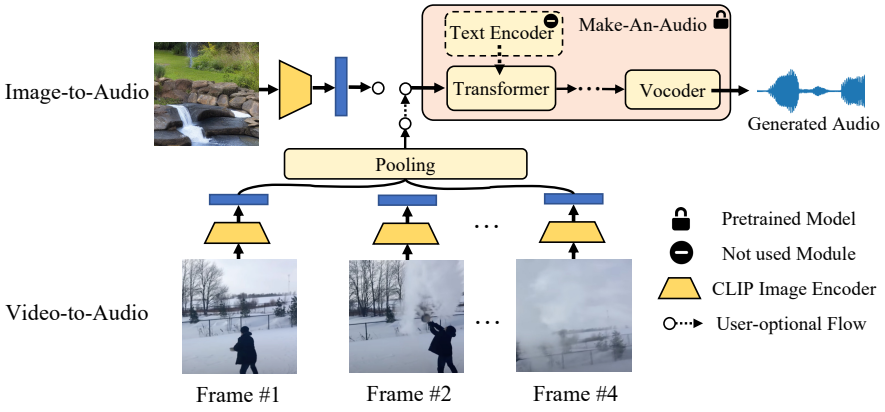


Figure 4. A high-level overview of visual-to-audio generation (I2A/V2A) pipeline using Make-An-Audio.

3.6. Classifier-Free Guidance

For classifier-free guidance shown in (Dhariwal & Nichol, 2021; Ho & Salimans, 2022), by jointly training a conditional and an unconditional diffusion model, it could be possible to combine the conditional and unconditional scores to attain a trade-off between sample quality and diversity. The textual condition in a latent diffusion model $\boldsymbol{\epsilon}_\theta(\mathbf{z}_t, t, c)$ is replaced by an empty prompt $c_\emptyset$ with a fixed probability during training. During sampling, the output of the model is extrapolated further in the direction of $\boldsymbol{\epsilon}_\theta(\mathbf{z}_t, t, c)$ and away from $\boldsymbol{\epsilon}_\theta(\mathbf{z}_t, t, c_\emptyset)$ with the guidance scale $s \geq 1$:

$$\tilde{\boldsymbol{\epsilon}}_\theta(\mathbf{z}_t, t, c) = \boldsymbol{\epsilon}_\theta(\mathbf{z}_t, t, c_\emptyset) + s \cdot (\boldsymbol{\epsilon}_\theta(\mathbf{z}_t, t, c) - \boldsymbol{\epsilon}_\theta(\mathbf{z}_t, t, c_\emptyset)) \quad (2)$$

4. X-To-Audio: No Modality Left Behind

In this section, we generalize our powerful conditional diffusion model for X-To-Audio generation. For the first time, we contextualize the need for audio generation with different conditional modalities, including: 1) text, 2) audio (inpainting), and 3) visual. Make-An-Audio empowers humans to create rich and diverse audio content with unprecedented ease, unlocking the ability to generate high-definition, high-fidelity audio given a user-defined modality input.

4.1. Personalized Text-To-Audio Generation

Adapting models (Chen et al., 2020b; Huang et al., 2022b) to a specific individual or object is a long-standing goal in machine learning research. More recently, personalization (Gal et al., 2022; Benhamdi et al., 2017) efforts can be found in vision and graphics, which allows to inject unique objects into new scenes, transform them across different styles, and even produce new products. For instance, when asked to generate "baby crying" given the initial sound of "thunder", our model produces realistic and faithful audio describing "a baby cries in the thunder day". Distinctly, it has a wide range of uses for audio mixing and tuning, e.g., adding background sound for an existing clip or editing audio by inserting a speaking object.

We investigate the personalized text-to-audio generation by stochastic differential editing (Meng et al., 2021), which has been demonstrated to produce realistic samples with high-fidelity manipulation. Given input audio with a user guide (prompt), we select a particular time t_0 with total denoising steps N, and add noise to the raw data $\mathbf{z}_0$ for $\mathbf{z}_T$ ($T = t_0 \times N$) according to Equation 4. It is then subsequently denoised through a reverse process parameterized by shared θ to increase its realism according to Equation 6.

A trade-off between faithfulness (text-caption alignment) and realism (audio quality) could be witnessed: As T increases, a large amount of noise would be added to the initial audio, and the generated samples become more realistic while less faithful. We refer the reader to Figure 5 for a summary of our findings.

4.2. Audio Inpainting

Inpainting (Liu et al., 2020; Nazeri et al., 2019) is the task of filling masked regions of an audio with new content since parts of the audio are corrupted or undesired. Though diffusion model inpainting can be performed by adding noise to initial audio and sampling with SDEdit, it may result in undesired edge artifacts since there could be an information loss during the sampling process (the model can only see a noised version of the context). To achieve better results, we explicitly fine-tune Make-An-Audio for audio inpainting.

During training, the way masks are generated greatly influences the final performance of the system. As such, we adopt irregular masks (thick, medium, and thin masks) suggested by LaMa (Suvorov et al., 2022), which uniformly uses polygonal chains dilated by a high random width (wide masks) and rectangles of arbitrary aspect ratios (box masks).

In addition, we investigate the frame-based masking strategy commonly adopted in speech liteature (Baevski et al., 2020; Hsu et al., 2021). It is implemented using the algorithm from wav2vec 2.0 (Baevski et al., 2020), where spans of length are masked with a p probability.

4.3. Visual-To-Audio Generation

Recent advances in deep generative models have shown impressive results in the visually-induced audio generation (Su et al., 2020; Gan et al., 2020), towards generating realistic audio that describes the content of images or videos: Hsu et al. (2020) show that spoken language could be learned by a visually-grounded generative model of speech. Iashin & Rahtu (2021) propose a multi-class visual guided sound synthesis that relies on a codebook prior-based transformer.

To pursue this research further, we extend Make-An-Audio for visual-to-audio generation. For the lack of large-scale visual-audio datasets in image-to-audio (I2A) research, our main idea is to utilize contrastive language-image pretraining (CLIP) with CLIP-guided T2A model and leverage textual representations to bridge the modality gap between visual and audio world. As CLIP encoders embed images and text to the joint latent space, our T2A model provides a unique opportunity to visualize what the CLIP image encoder is seeing. Considering the complexity of V2A generation, it is natural to leverage image priors for videos to simplify the learning process. On this account, we uniformly pick up 4 frames from the video and pool these CLIP image features to formulate the "averaged" video representation, which is then deteriorated to I2A generation.

To conclude, the visual-to-audio inference scheme can be formulated in Figure 4. It significantly reduces the requirement for pair visual datasets, and the plug-and-play module with pre-trained Make-An-Audio empowers humans to create rich and diverse audio content from the visual world.

5. Training and Evaluation

5.1. Dataset

We train on a combination of several datasets: AudioSet, BBC sound effects, Audiostock, AudioCaps-train, ESC-50, FSD50K, Free To Use Sounds, Sonniss Game Effects, WeSoundEffects, MACS, Epidemic Sound, UrbanSound8K, WavText5Ks, LibriSpeech, and Medley-solos-DB. For audios without natural language annotation, we apply the pseudo prompt enhancement to construct captions aligned well with the audio. Overall we have $\sim$3k hours with 1M audio-text pairs for training data. For evaluating text-to-audio models (Yang et al., 2022; Kreuk et al., 2022), the AudioCaption validation set is adopted as the standard benchmark, which contains 494 samples with five human-annotated captions in each audio clip. For a more challenging zero-shot scenario, we also provide results in Clotho (Drossos et al., 2020) validation set which contain multiple audio events. A more detailed data setup has been attached in Appendix A.

We conduct preprocessing on the text and audio data: 1) convert the sampling rate of audios to 16kHz and pad short clips to 10-second long; 2) extract the spectrogram with the FFT size of 1024, hop size of 256 and crop it to a mel-spectrogram of size 80×624; 3) non-standard words (e.g., abbreviations, numbers, and currency expressions) and semiotic classes (Taylor, 2009) (text tokens that represent particular entities that are semantically constrained, such as measure phrases, addresses, and dates) are normalized.

5.2. Model Configurations

We train a continuous autoencoder to compress the perceptual space with downsampling to a 4-channel latent representation, which balances efficiency and perceptually faithful results. For our main experiments, we train a U-Net (Ronneberger et al., 2015) based text-conditional diffusion model, which is optimized using 18 NVIDIA V100 GPU until 2M optimization steps. The base learning rate is set to 0.005, and we scale it by the number of GPUs and the batch size following LDM. We utilize HiFi-GAN (Kong et al., 2020) (V1) trained on VGGSound dataset (Chen et al., 2020a) as the vocoder to synthesize waveform from the generated mel-spectrogram in all our experiments. Hyperparameters are included in Appendix B.

5.3. Evaluation Metrics

We evaluate models using objective and subjective metrics over audio quality and text-audio alignment faithfulness. Following common practice (Yang et al., 2022; Iashin & Rahtu, 2021), the key automated performance metrics used are melception-based (Koutini et al., 2021) FID (Heusel et al., 2017) and KL divergence to measure audio fidelity. Additionally, we introduce the CLAP score to measure audio-text alignment for this work. CLAP score is adapted from the CLIP score (Hessel et al., 2021; Radford et al., 2021) to the audio domain and is a reference-free evaluation metric that closely correlates with human perception.

For subjective metrics, we use crowd-sourced human evaluation via Amazon Mechanical Turk, where raters are asked to rate MOS (mean opinion score) on a 20-100 Likert scale. We assess the audio quality and text-audio alignment faithfulness by respectively scoring MOS-Q and MOS-F, which is reported with 95% confidence intervals (CI). More information on evaluation has been attached in Appendix C.

Model	Text-cond	Params	FID	KL	CLAP	MOS-Q	MOS-F	FID-Z	KL-Z
Reference	/	/	/	/	0.526	74.7±0.94	80.5±1.84	/	/
Diffsound	CLIP	520M	7.17	3.57	0.420	67.1±1.03	70.9±1.05	24.97	6.53
Make-An-Audio	**CLAP**	**332M**	**4.61**	**2.79**	0.482	**72.5±0.90**	**78.6±1.01**	17.38	**6.98**
	BERT	809M	5.15	2.89	0.480	70.5±0.87	77.2±0.98	18.75	7.01
	T5-Large	563M	4.83	2.81	**0.486**	71.8±0.91	77.2±0.93	**17.23**	7.02
	CLIP	576M	6.45	2.91	0.444	72.1±0.92	75.4±0.96	17.55	7.09

Table 1. Text-to-audio evaluation. We report the evaluation metrics including MOS(↑), FID(↓), KL(↓), and CLAP(↑). FID-Z and KL-Z denote the zero-shot results in the Clotho dataset. To benchmark the text-to-audio (T2A) generation task, we further compare Make-An-Audio with AudioLDM (Liu et al.) and AudioGen (Kreuk et al., 2022) in Appendix C.

Training Masks	Narrow Masks			Wide Masks		
	FID	KL	MOS-Q	FID	KL	MOS-Q
Irregular (Thin)	1.83	0.46	68.3±1.38	4.01	0.86	66.2±1.20
Irregular (Medium)	1.76	0.31	67.8±1.41	3.93	0.65	66.9±1.22
Irregular (Thick)	1.73	0.32	69.6±1.36	3.83	0.67	69.3±1.05
Frame (p=30%)	1.64	0.29	66.9±1.60	3.68	0.62	66.1±1.29
Frame (p=50%)	1.77	0.32	68.6±1.42	3.66	0.63	67.4±1.27
Frame (p=70%)	1.59	0.32	71.0±1.12	3.49	0.65	70.8±1.50

Table 2. Audio inpainting evaluation with variety masking strategies.

Method	MOS-Q	MOS-F
Image-to-Audio Generation		
Reference	72.0±1.54	76.4±1.83
Make-An-Audio	68.4±1.09	78.0±1.20
Video-to-Audio Generation		
Reference	69.5±1.22	81.0±1.43
Make-An-Audio	60.0±1.31	69.0±1.08

Table 3. Image/Video-to-audio evaluation.

6. Results

6.1. Quantitative Results

Automatic Objective Evaluation The objective evaluation comparison with baseline Diffsound (the only publicly-available T2A generation model) are presented in Table 1, and we have the following observations: 1) In terms of audio qualty, Make-An-Audio achieves the highest perceptual quality in AudioCaption with FID of 4.61 and KL of 2.79. For zero-shot generation, it also demonstrates the outperformed results superior to the baseline model; 2) On text-audio similarity, Make-An-Audio scores the highest CLAP with a gap of 0.037 compared to the ground truth audio, suggesting Make-An-Audio's ability to generate faithful audio that aligns well with descriptions.

Subjective Human Evaluation The evaluation of the T2A models is very challenging due to its subjective nature in perceptual quality, and thus we include a human evaluation in Table 1: Make-An-Audio (CLAP) achieves the highest perceptual quality with MOS-Q of 72.5 and MOS-F of 78.6. It indicates that raters prefer our model synthesis against baselines in terms of audio naturalness and faithfulness.

For audio-inpainting, we compare different masking designs, including the irregular (thick, medium, and thin) strategy from visual world (Suvorov et al., 2022), as well as the frame-based (with varying p) strategy commonly used in speech (Baevski et al., 2020; Hsu et al., 2021). During evaluation, we randomly mask the *wide* or *narrow* regions and utilize FID and KL metrics to measure performance. The results have been presented in Table 2, and we have the following observations: 1) In both frame-based or irregular strategies, larger masked regions in training have witnessed the improved perceptual quality, which force the network to exploit the high receptive field of continuous spectrograms fully. 2) With the similar size of the masked region, the frame-based strategy consistently outperforms the irregular one, suggesting that it could be better to mask the audio spectrograms which align in time series.

We also present our visual-to-audio generation results in Table 3. As can be seen, Make-An-Audio can generalize to a wide variety of images and videos. Leveraging contrastive pre-training, the model provides a high-level understanding of visual input, which generates high-fidelity audio spectrograms well-aligned with their semantic meanings.

6.2. Qualitative Findings

Firstly, we explore the classifier-free guidance in text-to-audio synthesis. We sweep over guidance values and present trade-off curves between CLAP and FID scores in Figure 7. Consistent with the observations in Ho & Salimans (2022), the choice of the classifier guidance weight could scale conditional and unconditional synthesis, offering a trade-off between sample faithfulness and realism with respect to the conditioning text.

For better comparison in audio inpainting, we visualize different masking strategies and synthesis results in Figure 6. As can be seen, given the initial audio with undesired content, our model correctly fills and reconstruct the audio

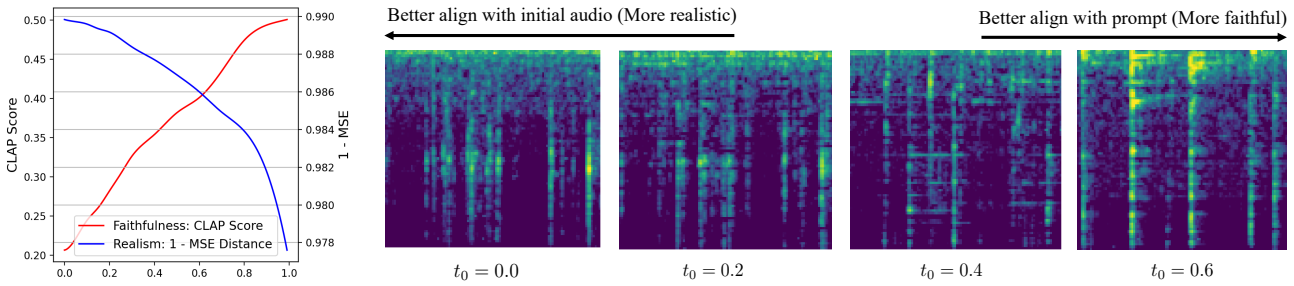


Figure 5. We illustrate personalized text-to-audio results with various t_0 initializations. $t_0 = 0$ indicates the initial audio itself, whereas $t_0 = 1$ indicates a text-to-audio synthesis from scratch. For comparison, realism is measured by the 1-MSE distance between generated and initial audio, and faithfulness is measured by the CLAP score between the generated sample. Prompt: A clock ticktocks.

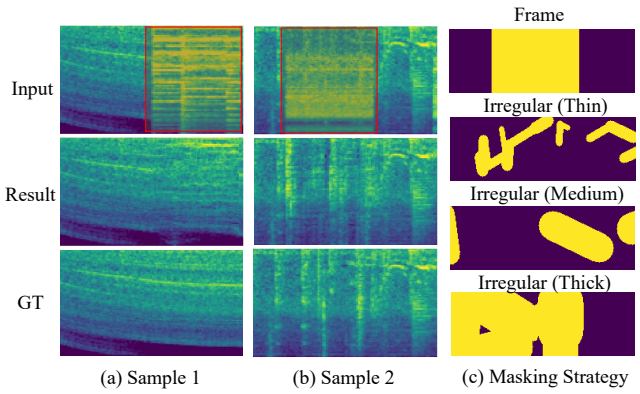


Figure 6. Qualitative results with our inpainting model.

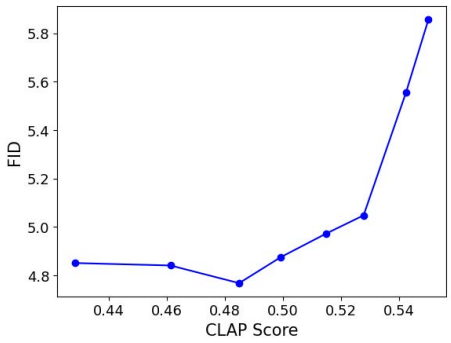


Figure 7. Classifier-free guidance trade-off curves.

robust to different shapes of masked regions, suggesting that it is capable of a high-level understanding of audio content.

On the personalized text-to-audio generation, we explore different $t_0 \in (0, 1)$ to add Gaussian noise and conduct reverse sampling. As shown in Figure 5, a trade-off between faithfulness (measured by CLAP score) and realism (measured by 1-MSE distance) could be witnessed. We find that $t_0 \in [0.2, 0.5]$ works well for faithful guidance with realistic generation, suggesting that audio variants (e.g., speed, timbre, and energy) could be easily destroyed as t_0 increases.

6.3. Analysis and Ablation Studies

To verify the effectiveness of several designs in Make-An-Audio, including pseudo prompt enhancement, textual and audio representation, we conduct ablation studies and discuss the key findings as follows. More analysis on audio representation has been attached in Appendix E.1.

6.3.1. Textual Representation

We explore several pretrained text encoders, including language models BERT (Devlin et al., 2018), T5-Large (Raffel et al., 2020), as well as the multimodal contrastive pre-trained encoder CLIP (Radford et al., 2021) and CLAP (Elizalde et al., 2022). We freeze the weights of text encoders for T2A generation. For easy comparison, we present the results in Table 1 and have the following observations: 1) Since CLIP is introduced as a scalable approach for learning joint representations between text and images, it could be less useful in deriving semantic representation for T2A in contrast to Yang et al. (2022). 2) CLAP and T5-Large achieve similar performances on benchmarks dataset, while CLAP could be more computationally efficient (with only %59 params), without the need for offline computation of embeddings in large-scale language models.

6.3.2. Pseudo Prompt Enhancement

Our prompt enhancement approach alleviates the issue of data scarcity, which consists of two stages with a distill-then-reprogram approach. As shown in Table 5 in Appendix A, we calculate and compare the prompt-audio faithfulness averaged across datasets: The joint expert distillation produces

high-quality captions aligned well with audio, and suggests strong generalization to diverse audio domains.

To highlight the effectiveness of the proposed dynamic reprogramming strategy to create unseen object compositions, we additionally train our Make-An-Audio in the static training dataset, and attach the results in Table 8 in Appendix E: 1) Removing the dynamic reprogramming approach results in a slight drop in evaluation; 2) When migrating to a more challenging scenario to Clotho in a zero-shot fashion, a significant degradation could be witnessed, demonstrating its effectiveness in constructing diverse object compositions for better generalization.

7. Conclusion

In this work, we presented Make-An-Audio with a prompt-enhanced diffusion model for text-to-audio generation. Leveraging the prompt enhancement with the distill-then-reprogram approach, Make-An-Audio was endowed with various concept compositions with orders of magnitude unsupervised data. We investigated textual representation and emphasized the advantages of contrastive pre-training for a deep understanding of natural languages with computational efficiency. Both objective and subjective evaluation demonstrated that Make-An-Audio achieved new state-of-the-art results in text-to-audio with realistic and faithful synthesis. Make-An-Audio was the first attempt to generate high-definition, high-fidelity audio given a user-defined modality input, opening up a host of applications for personalized transfer and fine-grained control. We envisage that our work serve as a basis for future audio synthesis studies.

Acknowledgements

This work was supported in part by the National Key R&D Program of China under Grant No.2022ZD0162000, National Natural Science Foundation of China under Grant No.62222211, Grant No.61836002 and Grant No.62072397.

References

Baevski, A., Zhou, Y., Mohamed, A., and Auli, M. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, 33, 2020.

Baevski, A., Hsu, W.-N., Xu, Q., Babu, A., Gu, J., and Auli, M. Data2vec: A general framework for self-supervised learning in speech, vision and language. *arXiv preprint arXiv:2202.03555*, 2022.

Benhamdi, S., Babouri, A., and Chiky, R. Personalized recommender system for e-learning environment. *Education and Information Technologies*, 22(4):1455–1477, 2017.

Bittner, R. M., Salamon, J., Tierney, M., Mauch, M., Cannam, C., and Bello, J. P. Medleydb: A multitrack dataset for annotation-intensive mir research. In *ISMIR*, volume 14, pp. 155–160, 2014.

Borsos, Z., Marinier, R., Vincent, D., Kharitonov, E., Pietquin, O., Sharifi, M., Teboul, O., Grangier, D., Tagliasacchi, M., and Zeghidour, N. Audiolm: a language modeling approach to audio generation. *arXiv preprint arXiv:2209.03143*, 2022.

Chen, H., Xie, W., Vedaldi, A., and Zisserman, A. Vggsound: A large-scale audio-visual dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 721–725. IEEE, 2020a.

Chen, M., Tan, X., Li, B., Liu, Y., Qin, T., Liu, T.-Y., et al. Adaspeech: Adaptive text to speech for custom voice. In *International Conference on Learning Representations*, 2020b.

Deshmukh, S., Elizalde, B., and Wang, H. Audio retrieval with wavtext5k and clap training. *arXiv preprint arXiv:2209.14275*, 2022.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Dhariwal, P. and Nichol, A. Diffusion models beat gans on image synthesis. In *Proc. of NeurIPS*, volume 34, 2021.

Ding, M., Zheng, W., Hong, W., and Tang, J. Cogview2: Faster and better text-to-image generation via hierarchical transformers. *arXiv preprint arXiv:2204.14217*, 2022.

Drossos, K., Lipping, S., and Virtanen, T. Clotho: An audio captioning dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 736–740. IEEE, 2020.

Elizalde, B., Deshmukh, S., Ismail, M. A., and Wang, H. Clap: Learning audio concepts from natural language supervision. *arXiv preprint arXiv:2206.04769*, 2022.

Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A. H., Chechik, G., and Cohen-Or, D. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022.

Gan, C., Huang, D., Chen, P., Tenenbaum, J. B., and Torralba, A. Foley music: Learning to generate music from videos. In *European Conference on Computer Vision*, pp. 758–775. Springer, 2020.

Gemmeke, J. F., Ellis, D. P., Freedman, D., Jansen, A., Lawrence, W., Moore, R. C., Plakal, M., and Ritter, M. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 776–780. IEEE, 2017.

Ghosal, D., Majumder, N., Mehrish, A., and Poria, S. Text-to-audio generation using instruction-tuned llm and latent diffusion model, 2023.

Gong, Y., Lai, C.-I., Chung, Y.-A., and Glass, J. Ssast: Self-supervised audio spectrogram transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 10699–10709, 2022.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., and Girshick, R. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16000–16009, 2022.

Hessel, J., Holtzman, A., Forbes, M., Bras, R. L., and Choi, Y. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*, 2021.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

Ho, J. and Salimans, T. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. In *Proc. of NeurIPS*, 2020.

Hong, W., Ding, M., Zheng, W., Liu, X., and Tang, J. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. *arXiv preprint arXiv:2205.15868*, 2022.

Hsu, W.-N., Harwath, D., Song, C., and Glass, J. Text-free image-to-speech synthesis using learned segmental units. *arXiv preprint arXiv:2012.15454*, 2020.

Hsu, W.-N., Bolte, B., Tsai, Y.-H. H., Lakhotia, K., Salakhutdinov, R., and Mohamed, A. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460, 2021.

Huang, R., Lam, M. W., Wang, J., Su, D., Yu, D., Ren, Y., and Zhao, Z. Fastdiff: A fast conditional diffusion model for high-quality speech synthesis. *arXiv preprint arXiv:2204.09934*, 2022a.

Huang, R., Ren, Y., Liu, J., Cui, C., and Zhao, Z. Generspeech: Towards style transfer for generalizable out-of-domain text-to-speech synthesis. *arXiv preprint arXiv:2205.07211*, 2022b.

Huang, R., Zhao, Z., Liu, H., Liu, J., Cui, C., and Ren, Y. Prodiff: Progressive fast diffusion model for high-quality text-to-speech. In *Proceedings of the 30th ACM International Conference on Multimedia*, pp. 2595–2605, 2022c.

Huang, R., Zhao, Z., Liu, J., Liu, H., Ren, Y., Zhang, L., and He, J. Transpeech: Speech-to-speech translation with bilateral perturbation. *arXiv preprint arXiv:2205.12523*, 2022d.

Iashin, V. and Rahtu, E. Taming visually guided sound generation. *arXiv preprint arXiv:2110.08791*, 2021.

Kim, C. D., Kim, B., Lee, H., and Kim, G. Audiocaps: Generating captions for audios in the wild. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 119–132, 2019.

Kingma, D. P. and Dhariwal, P. Glow: Generative flow with invertible 1x1 convolutions. *Advances in Neural Information Processing Systems*, 31:10215–10224, 2018.

Koepke, A. S., Oncescu, A.-M., Henriques, J., Akata, Z., and Albanie, S. Audio retrieval with natural language queries: A benchmark study. *IEEE Transactions on Multimedia*, 2022.

Kong, J., Kim, J., and Bae, J. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. *arXiv preprint arXiv:2010.05646*, 2020.

Koutini, K., Schlüter, J., Eghbal-zadeh, H., and Widmer, G. Efficient training of audio transformers with patchout. *arXiv preprint arXiv:2110.05069*, 2021.

Kreuk, F., Synnaeve, G., Polyak, A., Singer, U., Défossez, A., Copet, J., Parikh, D., Taigman, Y., and Adi, Y. Audiogen: Textually guided audio generation. *arXiv preprint arXiv:2209.15352*, 2022.

Liu, H., Chen, Z., Yuan, Y., Mei, X., Liu, X., Mandic, D., Wang, W., and Plumbley, M. D. AudioLDM: Text-to-Audio Generation with Latent Diffusion Models.

Liu, H., Jiang, B., Song, Y., Huang, W., and Yang, C. Rethinking image inpainting via a mutual encoder-decoder with feature equalizations. In *European Conference on Computer Vision*, pp. 725–741. Springer, 2020.

Martín-Morató, I. and Mesaros, A. What is the ground truth? reliability of multi-annotator data for audio tagging. In *2021 29th European Signal Processing Conference (EUSIPCO)*, pp. 76–80. IEEE, 2021.

Meng, C., He, Y., Song, Y., Song, J., Wu, J., Zhu, J.-Y., and Ermon, S. Sdedit: Guided image synthesis and editing with stochastic differential equations. In *International Conference on Learning Representations*, 2021.

Nazeri, K., Ng, E., Joseph, T., Qureshi, F. Z., and Ebrahimi, M. Edgeconnect: Generative image inpainting with adversarial edge learning. *arXiv preprint arXiv:1901.00212*, 2019.

Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., and Chen, M. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021.

Piczak, K. J. Esc: Dataset for environmental sound classification. In *Proceedings of the 23rd ACM international conference on Multimedia*, pp. 1015–1018, 2015.

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

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P. J., et al. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(140):1–67, 2020.

Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., and Sutskever, I. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pp. 8821–8831. PMLR, 2021.

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

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

Ronneberger, O., Fischer, P., and Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241. Springer, 2015.

Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E., Ghasemipour, S. K. S., Ayan, B. K., Mahdavi, S. S., Lopes, R. G., et al. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022.

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

Singer, U., Polyak, A., Hayes, T., Yin, X., An, J., Zhang, S., Hu, Q., Yang, H., Ashual, O., Gafni, O., et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022.

Song, J., Meng, C., and Ermon, S. Denoising diffusion implicit models. In *Proc. of ICLR*, 2020.

Su, K., Liu, X., and Shlizerman, E. Audeo: Audio generation for a silent performance video. *Advances in Neural Information Processing Systems*, 33:3325–3337, 2020.

Suvorov, R., Logacheva, E., Mashikhin, A., Remizova, A., Ashukha, A., Silvestrov, A., Kong, N., Goka, H., Park, K., and Lempitsky, V. Resolution-robust large mask inpainting with fourier convolutions. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 2149–2159, 2022.

Taylor, P. *Text-to-speech synthesis*. Cambridge university press, 2009.

Van Den Oord, A., Vinyals, O., et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.

Xu, H., Li, J., Baevski, A., Auli, M., Galuba, W., Metze, F., Feichtenhofer, C., et al. Masked autoencoders that listen. *arXiv preprint arXiv:2207.06405*, 2022.

Xu, X., Dinkel, H., Wu, M., and Yu, K. A crnn-gru based reinforcement learning approach to audio captioning. In *DCASE*, pp. 225–229, 2020.

Xun, J., Zhang, S., Zhao, Z., Zhu, J., Zhang, Q., Li, J., He, X., He, X., Chua, T.-S., and Wu, F. Why do we click: visual impression-aware news recommendation. In *Proceedings of the 29th ACM International Conference on Multimedia*, pp. 3881–3890, 2021.

Xun, J., Zhang, S., Yang, Y., Zhu, J., Deng, L., Zhao, Z., Dong, Z., Li, R., Zhang, L., and Wu, F. Discover: Disentangled music representation learning for cover song identification. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2023.

Yang, D., Yu, J., Wang, H., Wang, W., Weng, C., Zou, Y., and Yu, D. Diffsound: Discrete diffusion model for text-to-sound generation. *arXiv preprint arXiv:2207.09983*, 2022.

Yang, D., Liu, S., Huang, R., Tian, J., Weng, C., and Zou, Y. Hifi-codec: Group-residual vector quantization for high fidelity audio codec. *arXiv preprint arXiv:2305.02765*, 2023.

Zeghidour, N., Luebs, A., Omran, A., Skoglund, J., and Tagliasacchi, M. Soundstream: An end-to-end neural audio codec. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:495–507, 2021.

Zen, H., Dang, V., Clark, R., Zhang, Y., Weiss, R. J., Jia, Y., Chen, Z., and Wu, Y. Libritts: A corpus derived from librispeech for text-to-speech. *arXiv preprint arXiv:1904.02882*, 2019.

Appendices

Make-An-Audio: Text-To-Audio Generation with Prompt-Enhanced Diffusion Models

A. Detailed Experimental Setup

Dataset	Hours	Type	Source
Clotho	152	Caption	Drossos et al. (2020)
AudioCaps	109	Caption	Kim et al. (2019)
MACS	100	Caption	Martín-Morató & Mesaros (2021)
WavText5Ks	25	Caption	Deshmukh et al. (2022)
BBC sound effects	481	Caption	https://sound-effects.bbcrewind.co.uk/
Audiostock	43	Caption	https://audiostock.net/se
Filter AudioSet	2084	Label	Gemmeke et al. (2017)
ESC-50	3	Label	Piczak (2015)
FSD50K	108	Label	https://annotator.freesound.org/fsd/
Sonniss Game Effects	20	Label	https://sonniss.com/gameaudiogdc/
WeSoundEffects	11	Label	https://wesoundeffects.com/
Epidemic Sound	220	Label	https://www.epidemicsound.com/
UrbanSound8K	8	Label	Salamon et al. (2014)
LibriTTS	300	Language-free	Zen et al. (2019)
Medley-solos-DB	7	Language-free	Bittner et al. (2014)

Table 4. Statistics for the combination of several datasets.

As shown in Table 5, we collect a large-scale audio-text dataset consisting of 1M audio samples with a total duration of $\sim$3k hours. It contains audio of human activities, natural sounds, and audio effects, consisting of several data sources from publicly available websites. For audio with text descriptions, we download the parallel audio-text data. For audios without natural language annotation (or with labels), we discard the corresponding class label (if any) and apply the pseudo prompt enhancement to construct natural language descriptions aligned well with the audio.

As speech and music are the dominant classes in Audioset, we filter these samples to construct a more balanced dataset. Overall we are left with 3k hours with 1M audio-text pairs for training data. For evaluating text-to-audio models (Yang et al., 2022; Kreuk et al., 2022), the AudioCaption validation set is the standard benchmark, which contains 494 samples with five human-annotated captions in each audio clip. In both training and inference, we pad short clips to 10-second long and randomly crop a 624×80 mel-spectrogram from 10-second 16 kHz audio.

Method	FSD50K	ESC-50	Urbansound8k
Original	0.40	0.43	0.33
Captioning	0.35	0.46	0.37
Retrieval	0.31	0.44	0.38
Both + CLAP Select	0.54	0.62	0.55

Table 5. Text-audio alignment CLAP score averaged across the single-label dataset.

B. Model Configurations

We list the model hyper-parameters of Make-An-Audio in Table 6.

Hyperparameter		Make-An-Audio
Spectrogram Autoencoders	Input/Output Channels	1
	Hidden Channels	4
	Residual Blocks	2
	Spectrogram Size	80×624
	Channel Mult	$[1, 2, 2, 4]$
Denoising Unet	Input/Output Channels	4
	Model Channels	320
	Attention Heads	8
	Condition Channels	1024
	Latent Size	10×78
	Channel Mult	$[1, 2]$
CLAP Text Encoder	Transformer Embed Channels	768
	Output Project Channels	1024
	Token Length	77
Total Number of Parameters		332M

Table 6. Hyperparameters of Make-An-Audio models.

C. Evaluation

To probe audio quality, we conduct the MOS (mean opinion score) tests and explicitly instruct the raters to *"focus on examining the audio quality and naturalness."*. The testers present and rate the samples, and each tester is asked to evaluate the subjective naturalness on a 20-100 Likert scale.

To probe text-audio alignment, human raters are shown an audio and a prompt and asked *"Does the natural language description align with audio faithfully?"*. They must respond with "completely", "mostly", or "somewhat" on a 20-100 Likert scale.

Our subjective evaluation tests are crowd-sourced and conducted via Amazon Mechanical Turk. These ratings are obtained independently for model samples and reference audio, and both are reported. The screenshots of instructions for testers have been shown in Figure 8. We paid $8 to participants hourly and totally spent about $750 on participant compensation. A small subset of speech samples used in the test is available at `https://Make-An-Audio.github.io/`.

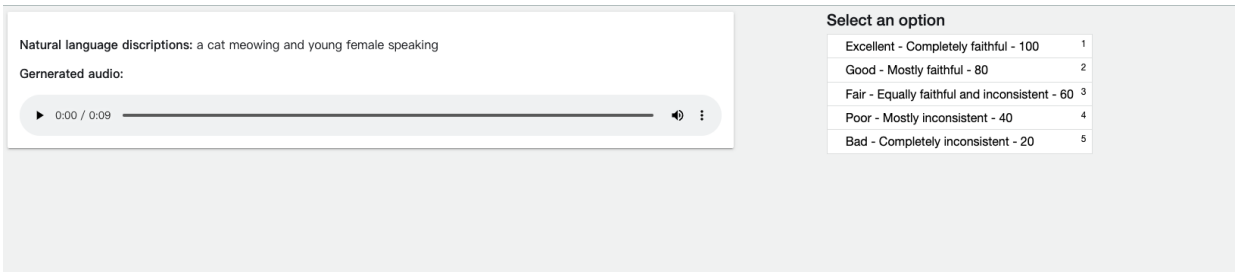
To benchmark the text-to-audio (T2A) generation task, we further evaluate Make-An-Audio and baseline T2A models following AudioLDM (Liu et al.).

Model	FD	IS	KL	FAD
GT	-	-	-	-
AudioGen-S	-	-	2.09	3.13
AudioGen-L	-	-	1.69	1.82
Make-An-Audio	18.32	7.29	1.61	2.66
AudioLDM-S	29.48	6.9	1.97	2.43
AudioLDM-L	23.31	8.13	1.59	1.96

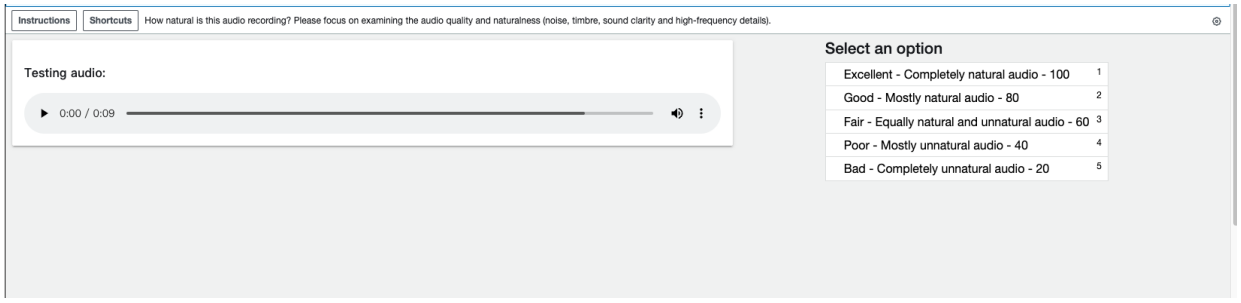
Table 7. The comparison between Make-An-Audio and baseline T2A models on the AudioCaps dataset. We borrowed results from (Liu et al.; Ghosal et al., 2023) and used the model released by the authors on Huggingface to test CLAP Score.

D. Detailed Formulation of DDPM

As a blossoming class of generative models, denoising diffusion probabilistic models (DDPMs) (Ho et al., 2020; Song et al., 2020; **?**) has emerged to prove its capability to achieve leading performances in both image and audio syntheses (Xun et al., 2021; Huang et al., 2022c;a). We define the data distribution as $q(\mathbf{x}_0)$. The diffusion process is defined by a fixed Markov



(a) Screenshot of MOS-F testing.



(b) Screenshot of MOS-Q testing.

Figure 8. Screenshots of subjective evaluations.

chain from data $\mathbf{x}_0$ to the latent variable $\mathbf{x}_T$:

$$q(\mathbf{x}_1, \cdots, \mathbf{x}_T|\mathbf{x}_0) = \prod_{t=1}^{T} q(\mathbf{x}_t|\mathbf{x}_{t-1}), \tag{3}$$

For a small positive constant β_t, a small Gaussian noise is added from $\mathbf{x}_{t-1}$ to the distribution of $\mathbf{x}_t$ under the function of $q(\mathbf{x}_t|\mathbf{x}_{t-1})$.

The whole process gradually converts data $\mathbf{x}_0$ to whitened latents $\mathbf{x}_T$ according to the fixed noise schedule $\beta_1, \cdots, \beta_T$, where $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$:

$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) := \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t \boldsymbol{I}) \tag{4}$$

Efficient training is optimizing a random term of t with stochastic gradient descent:

$$\mathcal{L}_\theta = \left\| \boldsymbol{\epsilon}_\theta \left(\alpha_t \mathbf{x}_0 + \sqrt{1-\alpha_t^2}\boldsymbol{\epsilon} \right) - \boldsymbol{\epsilon} \right\|_2^2 \tag{5}$$

Unlike the diffusion process, the reverse process is to recover samples from Gaussian noises. The reverse process is a Markov chain from x_T to x_0 parameterized by shared θ:

$$p_\theta(\mathbf{x}_0, \cdots, \mathbf{x}_{T-1}|\mathbf{x}_T) = \prod_{t=1}^{T} p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t), \tag{6}$$

where each iteration eliminates the Gaussian noise added in the diffusion process:

$$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) := \mathcal{N}(\mathbf{x}_{t-1}; \mu_\theta(\mathbf{x}_t, t), \sigma_\theta(\mathbf{x}_t, t)^2 \boldsymbol{I}) \tag{7}$$

E. Implementation Details

E.1. Spectrogram Autoencoders

We also investigate the effectiveness of several audio autoencoder variants in Table 8, and find that deeper representation (i.e., 32 or 128) relatively brings more compression, while the information deterioration could burden the Unet model in generative modeling.

Method	Channel	FID	KL
Supervised Evaluation in AudioCaps dataset			
Base	4	5.15	2.89
	32	9.22	3.54
	128	10.92	3.68
w/o PPE	4	5.37	3.05
Zero-Shot Evaluation in Clotho dataset			
Base	4	18.75	7.01
w/o PPE	4	22.31	7.19

Table 8. Audio quality comparisons for ablation study with Make-An-Audio BERT. We use PPE to denote pseudo prompt enhancement.

E.2. Text-to-audio

We first encode the text into a sequence of K tokens, and utilize the cross-attention mechanism to learn a language and mel-spectrograms representation mapping in a powerful model. After the initial training run, we fine-tuned our base model to support unconditional generation, with 20% of text token sequences being replaced with the empty sequence. This way, the model retains its ability to generate text-conditional outputs, but can also generate spectrogram representation unconditionally.

We consider the pre-trained automatic audio captioning (Xu et al., 2020) and audio-text retrieval (Deshmukh et al., 2022; Koepke et al., 2022) systems as our experts for prompt generation. Regarding automatic audio captioning, the model consists of a 10-layer convolution neural network (CNN) encoder and a temporal attentional single-layer gated recurrent unit (GRU) decoder. The CNN encoder is pre-trained on a large-scale Audioset dataset. As for audio-text retrieval, the model leverages BERT with a multi-modal transformer encoder for representation learning. It is trained on AudioCaps and Clotho datasets.

E.3. Visual-to-audio

For visual-to-audio (image/video) synthesis, we utilize the CLIP-guided T2A model and leverage global textual representations to bridge the modality gap between the visual and audio worlds. However, we empirically find that global CLIP conditions have a limited ability to control faithful synthesis with high text-audio similarity. On that account, we use the 110h FSD50K audios annotated with a class label for training, and this simplification avoids multimodal prediction (a conditional vector may refer to different concepts) with complex distribution.

We conduct ablation studies to compare various training settings, including datasets and global conditions. The results have been presented in Table 9, and we have the following observations: 1) Replacing the FSD50K dataset with AudioCaps (Kim et al., 2019) have witnessed a significant decrease in faithfulness. The dynamic concepts compositions confuse the global-condition models, and the multimodal distribution hinders its capacity for controllable synthesis; 2) Removing the normalization in the condition vector has witnessed the realism degradation measured by FID, demonstrating its efficiency in reducing variance in latent space.

Training/Testing Dataset	Condition	FID	KL	CLAP
AudioCaption	Global	/	/	0.12
FSD50k	Global	40.7	8.2	0.40
FSD50k	NormGlobal	31.1	8.0	0.42

Table 9. Ablation studies for training Make-An-Audio with global conditions.

F. Dynamic Reprogramming Templates

Below we provide the list of text templates used when providing dynamic reprogramming:

- before v q a n of &, X

- X before $v\ q\ a\ n$ of &,
- in front of $v\ q\ a\ n$ of &, X
- first is X second is $q\ a\ n$ of &
- after X, $v\ q\ a\ n$ of &
- after $v\ q\ a\ n$ of &, X
- behind $v\ q\ a\ n$ of &, X
- $v\ q\ a\ n$ of &, then X
- $v\ q\ a\ n$ of &, following X
- $v\ q\ a\ n$ of &, later X
- X after $v\ q\ a\ n$ of &
- before X, $v\ q\ a\ n$ of &

Specifically, we replace X and &, respectively, with the natural language of sampled data and the class label of sampled events from the database.

For verb (denoted as v), we have {'hearing', 'noticing', 'listening to', 'appearing'}; for adjective (denoted as a), we have {'clear', 'noisy', 'close-up', 'weird', 'clean'}; for noun (denoted as n), we have {'audio', 'sound', 'voice'}; for numeral/quantifier (denoted as q), we have {'a', 'the', 'some'};

G. Potential Negative Societal Impacts

This paper aims to advance open-domain text-to-audio generation, which will ease the effort of short video and digital art creation. The efficient training method also transfers knowledge from text-to-audio models to X-to-audio generation, which helps avoid training from scratch, and thus reduces the issue of data scarcity. A negative impact is the risk of misinformation. To alleviate it, we can train an additional classifier to discriminate the fakes. We believe the benefits outweigh the downsides.

Make-An-Audio lowers the requirements for high-quality text-to-audio synthesis, which may cause unemployment for people with related occupations, such as sound engineers and radio hosts. In addition, there is the potential for harm from non-consensual voice cloning or the generation of fake media, and the voices in the recordings might be overused than they expect.

H. Limitations

Make-An-Audio adopts generative diffusion models for high-quality synthesis, and thus it inherently requires multiple iterative refinements for better results. Besides, latent diffusion models require typically require more computational resources, and degradation could be witnessed with decreased training data. One of our future directions is to develop lightweight and fast diffusion models for accelerating sampling.