INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

AUDIOTOKEN: Adaptation of Text-Conditioned Diffusion Models for Audio-to-Image Generation

Guy Yariv [♡,♣], *Itai Gat* [♢], *Lior Wolf* [♠], *Yossi Adi* [♡,*], *Idan Schwartz* [♠,♣,*]

[♡]The Hebrew University of Jerusalem , [♢]Technion - Israel Institute of Technology
[♠]Tel-Aviv University, [♣]NetApp

Abstract

In recent years, image generation has shown a great leap in performance, where diffusion models play a central role. Although generating high-quality images, such models are mainly conditioned on textual descriptions. This begs the question: *how can we adopt such models to be conditioned on other modalities?*. In this paper, we propose a novel method utilizing latent diffusion models trained for text-to-image-generation to generate images conditioned on audio recordings. Using a pre-trained audio encoding model, the proposed method encodes audio into a new token, which can be considered as an adaptation layer between the audio and text representations. Such a modeling paradigm requires a small number of trainable parameters, making the proposed approach appealing for lightweight optimization. Results suggest the proposed method is superior to the evaluated baseline methods, considering objective and subjective metrics. Code and samples are available at: `https://pages.cs.huji.ac.il/adiyoss-lab/AudioToken`.

Index Terms: Diffusion models, Audio-to-image.

1. Introduction

Neural generative models have changed the way we consume digital content. From generating high-quality images [1, 2, 3], though coherence of long spans of text [4, 5, 6], up to speech and audio [7, 8, 9, 10]. In recent years, diffusion-based generative models have emerged as the preferred approach, showing promising results on various tasks [11].

During the diffusion process, the model learns to map a predefined noise distribution to the target data distribution. In every step of the diffusion process, the model learns to predict the noise at a given step to finally generate the signal from the target distribution [12, 13, 14]. Diffusion models operate on different forms of data representations, e.g., raw input [15, 12], latent representations [16], etc.

When considering controllable generative models, the common practice these days is to condition the generation on a textual description of the input data; this is especially noticeable in image generation [1, 17, 18]. Recently, several methods proposed using different modalities to condition the generative process such as image-to-audio [19, 20], image-to-speech [21, 22], image-to-text [23, 24], or audio-to-audio [25, 26]. However, such research direction is less explored by the community.

In this work, we focus on the task of audio-to-image generation. Given an audio sample contains an arbitrary sound, we aim to generate a high-quality image representing the acoustic scene. We propose leveraging a pre-trained text-to-image generation model together with a pre-trained audio representation model to learn an adaptation layer mapping between their outputs and inputs. Specifically, inspired by recent work on textual-inversions [27], we propose to learn a dedicated *audio-token* that maps the audio representations into an embedding vector. Such a vector is then forwarded into the network as a continuous representation, reflecting a new word embedding.

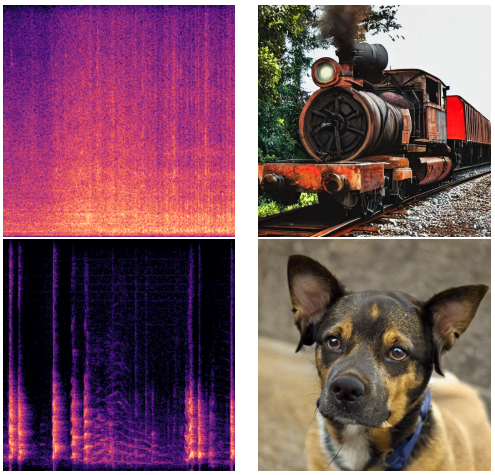

Figure 1: *Generated images (right) and input spectrograms (left) from the proposed method. The model gets as input an audio recording, extracts a representation, and projects into a textual latent space which will be fed into a pre-trained text-conditioned diffusion generative model.*

Several methods for generating audio from image inputs were proposed in prior work. The authors in [28, 29] proposed to generate images based on audio recordings using a Generative Adversarial Network (GAN) based method. Unlike the proposed method, in [28], the authors present results for generating MNIST digits only and did not generalize to general audio sounds. In [29], the authors did generate images from general audio. However, this turned into low-quality images. The most relevant related work to ours is Wav2Clip [30], in which the authors first learn a Contrastive Language-Image Pre-Training (CLIP) [31] like a model for audio-image pairs. Then, later on, such representation can be used to generate images using VQ-GAN [32] under the VQ-GAN CLIP [33] framework.

Why use audio signals as a conditioning to image generation rather than text? Although text-based generative models can generate impressive images, textual descriptions are not naturally paired with the image, i.e., textual descriptions are often added manually. On the other hand, when considering videos, audio, and images capture and represent the same scene,

*Equal Contribution.

10.21437/Interspeech.2023-852

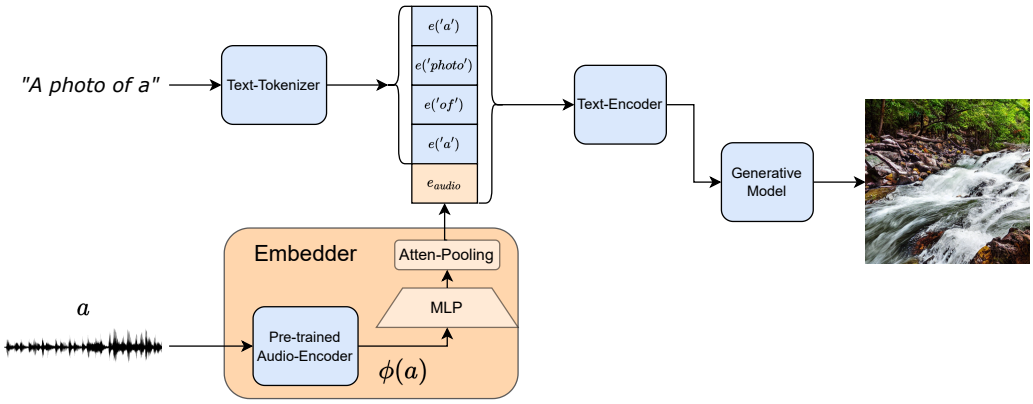

Figure 2: *Architecture overview: We forward an audio recording through a pre-trained audio encoder and then through an Embedder network. A pre-trained text encoder extracts tokens created by a tokenizer and the audio token. Finally, the generative model is fed with the concatenated tensor of representations. It is important to note that only the Embedder parameters are trained during this process.*

hence are naturally paired. Moreover, audio signals can represent complex scenes and objects such as different types of the same instrument (e.g., classic guitar, electric guitar, etc.), or different scenes of the same object (e.g., classic guitar recorded in studio vs. live show). Annotating such fine-grained details of the different objects is labor-intensive, hence hard to scale.

In summary, our contributions are: We propose a novel method AUDIOTOKEN for audio-to-image generation by leveraging a pre-trained text-to-image diffusion model together with a pre-trained audio encoder; We propose a set of evaluation metrics specifically dedicated for the task of an audio-to-image generation. Through extensive experiments, we show that our method is able to generate high-quality and diverse set of images based on audio-scenes.

2. Adaptation of text-conditioned models

Diffusion models are a family of models that are prone to learn the underlying probabilistic model of the data distribution $p(x)$. This is done by learning the reverse Markov process of length T. Given a timestamp $t \in [0, 1]$, the denoising function $\epsilon_\theta : \mathbb{R}^d \to \mathbb{R}^d$ learns to predict a clean version of the perturbed x_t from the training distribution $S = \{x_1, .., x_m\}$:

$$\mathcal{L}_{\text{DM}} \triangleq \mathbb{E}_{x\sim S, t\sim U(0,1), \epsilon\sim\mathcal{N}(0,I)} \left[\|\epsilon - \epsilon_\theta(x_t, t)\|_2^2\right]. \quad (1)$$

Empirical results showed that learning diffusion models on top of latent spaces of autoencoders can produce results in a higher quality than those that are trained on the raw input [16]. Intuitively, this process can be done on a latent representation of encoder-decoder architecture. Latent diffusion operates on top of a representation given by an encoder f:

$$\mathcal{L}_{\text{LDM}} \triangleq \mathbb{E}_{x\sim S, t\sim U(0,1), \epsilon\sim\mathcal{N}(0,I)} \left[\|\epsilon - \epsilon_\theta(f(x_t), t)\|_2^2\right]. \quad (2)$$

The output of the diffusion can later be forwarded through the decoder to obtain the raw result (e.g., audio, image, text).

An important component of modern generative models is conditioning. This allows the generative process to be conditioned on a given input, i.e., modeling $p(x|y)$ where y is a data entry. For example, in a text-based visual generation, the generative process is conditioned on text. There are many types of conditioning, such as text, time, style, etc. [16]. Usually, the conditioning component is done by an injection of a condition representation from an encoder τ to the attention mechanism of ϵ_θ. Conditioning the diffusion process yields the following diffusion process, $\mathcal{L}_{\text{CLDM}} \triangleq$

$$\mathbb{E}_{(x,y)\sim S, t\sim U(0,1), \epsilon\sim\mathcal{N}(0,I)} \left[\|\epsilon - \epsilon_\theta(f(x_t), t, \tau(y))\|_2^2\right]. \quad (3)$$

In the following, we propose a method that leverages a conditional generative model to produce high-quality and diverse images that are based on audio-scenes.

2.1. AUDIOTOKEN

Audio signals contain information that can help us imagine the scene that produced them. This makes it tempting to use a generative model that is conditioned on audio recordings to generate a scene. However, models that generate high-quality images commonly rely on large-scale text-image pairs to generate images using text. We thus propose a method named AUDIOTOKEN that effectively projects audio signals into a textual space, enabling us to leverage existing text-conditioned models to generate images based on audio-based tokens.

Our objective is to investigate the feasibility of directly encoding any audio signals into a dedicated representation that will fit as an additional token for text-conditioning. By doing so, we can leverage existing pre-trained models and not learn a new generative model with audio-visual pairs. Furthermore, we are not required to learn a new token for each individual class of audio or type of scene (as opposed to textual inversion-based methods [27]). Instead, we develop an audio-to-image generator capable of handling a wide range of diverse concepts.

The input to our method is a short video input (i, a), where i represents a frame from the video and a represents its corresponding audio recording. We are aiming to create a generative process that is audio-conditioned, i.e., $p(i|a)$. To achieve this, we utilize a text-conditioned generative model. Thus, we need to associate the audio signal a with the text conditioning.

The process begins with a transformer model that encodes the initial prompt "A photo of a" into a representation $e_{\text{text}} \in \mathbb{R}^{4\times d_a}$, where d_a is the embedding dimension of the text input. Afterward, we concatenate to e_{text}, an extra latent representation of the audio signal, denoted as $e_{\text{audio}} \in \mathbb{R}^{d_a}$. We utilize an Embedder, which is composed of a pre-trained audio encoding network and a small projection network. This results in:

$$e_{\text{audio}} = \text{Embedder}(a). \quad (4)$$

Next, we describe the Embedder network and the optimization process of our method.

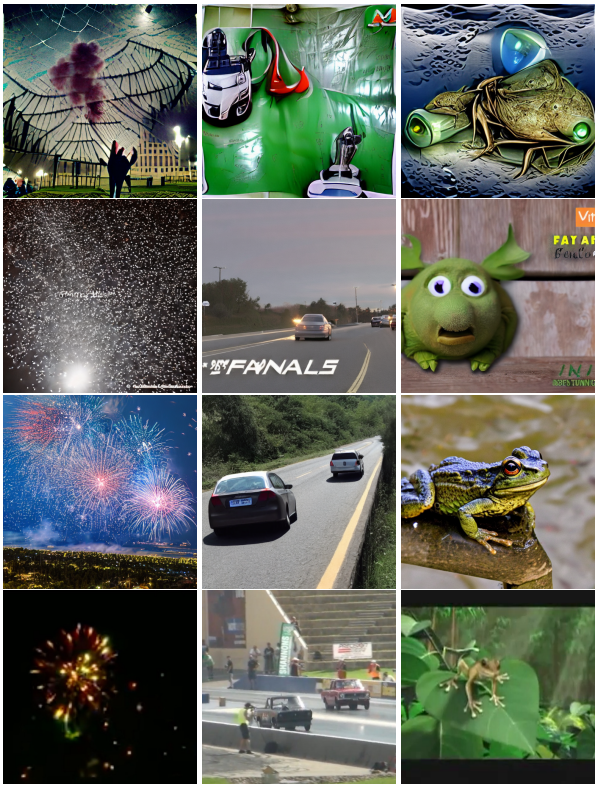

Figure 3: *Qualitative results for Wav2Clip (first row), Image-Bind (second row),* AUDIOTOKEN *(third row), and the original reference images (last row).*

Audio encoding: The Embedder leverages a pre-trained audio classification network ϕ to represent the audio. The discriminative network's last layer is typically used for classification, and thus it tends to diminish important audio information which is irrelevant to the discriminative task. Thus, we take a concatenation of earlier layers and the last hidden layer (specifically selecting the fourth, eighth, and twelfth layers out of a total of twelve). This results in a temporal embedding of the audio $\phi(a) \in \mathbb{R}^{\hat{d} \times n_a}$, where n_a is the temporal audio dimension. Then, to learn a projection into the textual embedding space, we forward $\phi(a)$ in two linear layers with a GELU function between them:

$$\bar{e}_{\text{audio}} = W_2 \sigma(W_1 \phi(a)), \tag{5}$$

where $W_1 \in \mathbb{R}^{\hat{d} \times \hat{d}}, W_1 \in \mathbb{R}^{\hat{d} \times d_{\text{audio}}}$, and σ is a GELU nonlinearity [34]. Finally, we apply an attentive pooling layer [35], reducing the temporal dimension of the audio signal, i.e.,

$$e_{\text{audio}} = \text{Atten-Pooling}(\bar{e}_{\text{audio}}). \tag{6}$$

Optimization: During training, we update only the weights of the linear and attentive pooling layers within the Embedder network during the optimization process. The pre-trained audio network and the generative network remain frozen. We adopt the loss function employed by the original model $\mathcal{L}_{\text{LDM}}$ (Equation 2), maintaining consistency in the training scheme. Furthermore, we introduce an additional loss function that complements the original one, which involves encoding the label of the video, denoted by $l \in \mathbb{R}^{n_l \times d_a}$, where n_l represents the label's length (e.g., the size of the 'acoustic guitar' label is two). The label is encoded using the generative model's textual encoder, and then the spatial dimension is reduced using average pooling, i.e., $\hat{l} = \text{Avg-Pooling}(l)$. The classification loss is defined as follows:

$$\mathcal{L}_{\text{CL}} = \left(1 - \frac{\langle e_{\text{audio}}, \hat{l} \rangle}{\|e_{\text{audio}}\| \|\hat{l}\|}\right)^2. \tag{7}$$

Intuitively, this term ensures that the audio embedding remains close to the video's concept, facilitating faster and more stable convergence. Finally, we also add an ℓ_1 regularization to the encoded audio token, which encourages the audio token to be more evenly distributed. The overall loss that is optimized for AUDIOTOKEN is given by

$$\mathcal{L} = \mathcal{L}_{\text{LDM}} + \lambda_{\ell_1} \|e_{\text{audio}}\|_1. \tag{8}$$

The overall loss that is optimized for AUDIOTOKEN with classification loss is given by

$$\mathcal{L} = \mathcal{L}_{\text{LDM}} + \lambda_{\ell_1} \|e_{\text{audio}}\|_1 + \lambda_{\text{CL}} \mathcal{L}_{\text{CL}}. \tag{9}$$

2.2. Evaluation functions

The evaluation of a visual generation from audio-scene is yet opened. Such evaluation setup is challenging since a well-performed model is excepted to generate images that will (i) capture the most prominent object in the audio recording; (ii) be semantically correlate with the input audio; and (iii) be semantically similar to the "ground truth" / target image. Lastly, we require evaluating the general quality of the generated image. To mitigate that, we propose to use the following evaluation functions.

Audio-Image Similarity (AIS) ideally measures the similarity between the semantic input audio and generated image features. We employ the Wav2CLIP model [30]. The Wav2CLIP model enables to measure of the similarity between representations of an audio and image pair. This allows us to quantify to which extent the generated image describes the audio. Quantifying only the correlation score is not telling the whole story since the score scale may vary. Thus, it is unclear what is a good score. Instead, we compare the similarity between a generated image and its input audio and the similarity between the generated image and arbitrary audio from the data. The AIS score is then averaged over all data entries in the validation set.

Image-Image Similarity (IIS) measures the semantic similarity between the generated image and the "ground truth" one. This information is crucial since it allows quantifying the semantic similarity to a "ground truth" scene.

We employ the same reference-based method as in the AIS metric. Thus, we measure the CLIP [31] score between the (i) generated image and its "ground truth" and (ii) generated image and an arbitrary image from the data. The IIS score is then averaged over all data entries in the validation set.

Audio-Image Content (AIC). To account for the image content, we measure the level of agreement between the predicted class of an image classifier and the ground-truth audio label. However, since there might not be a complete correlation between the image classifier classes and the audio labels, an additional CLIP-based score is employed to determine agreement. If the CLIP-based matching score exceeds a threshold of 0.75, the image and audio class are considered in agreement.

Fréchet Inception Distance (FID). In order to evaluate the quality of the generated images, we adopt the standard FID score [36]. Such reference-free metric compares the distribution of the generated images against the original images using

Table 1: *We report AIC, FID, AIS, and IIS for* AUDIOTOKEN *(with and without Classification Loss (CL)), together with Wav2Clip. For reference, we additionally report results for the original images (reference) and images generated by Stable Diffusion (SD) with text labels.*

Method	Metric			
	AIC ↑	FID ↓	AIS ↑	IIS ↑
Reference	54.66	-	-	-
SD (Text)	71.28	52.85	-	-
Wav2Clip [30]	29.32	99.89	47.76	51.11
ImageBind [37]	39.15	67.42	67.48	75.50
AUDIOTOKEN with CL	**48.01**	66.08	62.28	76.40
AUDIOTOKEN	45.48	**56.65**	**68.23**	**76.66**

an internal representation obtained from a pre-trained model. In this work, we use the Inception model.

Human Evaluation. Lastly, we run a subjective test to evaluate the adherence of the generated images to their labels. For each method, annotators were shown a generated image and asked to rate its relevance to a given label on a scale of 1-5.

3. Results

In the following, we study our method from objective and subjective points of view. We begin by describing details regarding the experimental setup. Then, we report results for our method and baselines using the evaluation framework proposed in Section 2.2. We show that our method outperforms the current baselines. We finally subjectively evaluate and find that annotators agree that our method describes the audio the best.

Baselines. Wav2Clip [30] employs a CLIP-based loss for audio-text pairs. Then, they use this representation to generate an image from a text that is highly correlated with the audio using VQ-GAN [32]. ImageBind [37] combines information from six different modalities (text, image/video, audio, depth, thermal, and inertial measurement units (IMU)) into a single representation space. We used ImageBind's unified latent with stable-diffusion-2-1-unclip [1] to generate images from audio samples.

Data. We use the VGGSound [38] dataset, which is derived from a collection of YouTube videos with corresponding audio-visual data. The dataset contains $200,000$, each in the length of ten seconds. The dataset is also annotated with 309 classes.

Hyperparameters. The Embedder network comprises 3 layers, with attention pooling applied to a sequence of 248. For the generative model, we use Stable Diffusion [16]. This results in an $8,853,507$ parameters model. During training, we randomly crop five-second audio clips and select the frame with the highest CLIP score corresponding to the VGGSound label. We also filter out frames with inconsistent classifications from both the image and audio classifiers. We train the model for $60,000$ steps with a learning rate of 8e-5 and batch size of 8 on Nvidia A6000 GPU.

Objective evaluation. We start by comparing the proposed method, with and without the Classification Loss (CL), against Wav2Clip and ImageBind, considering FID, AIS, AIC, and IIS. For reference, we additionally include a topline of results of generating images directly from textual description (text labels) using Stable Diffusion (SD). Results are reported in Table 1.

[1] https://github.com/Zeqiang-Lai/Anything2Image

Figure 4: *Qualitative results of speaker generation for* AUDIOTOKEN *(first row), and reference images (second row).*

Results suggest that AUDIOTOKEN is superior to Wav2Clip and ImageBind, considering all evaluation metrics. Interestingly, AUDIOTOKEN also performs better when considering the AIS metric, which leverages the Wav2Clip and ImageBind models to obtain the similarity score. This result demonstrates accurate audio detail identification (e.g., distinguishing various guitars) and considers multiple entities (e.g., multiple flying planes or a single plane). As expected, using textual labels reaches a higher accuracy and pushes the model toward learning representation which is more discriminative but less correlated with the target video. Generated images from all methods can be seen in Figure 3.

Subjective evaluation. We compare AUDIOTOKEN against Wav2Clip, and SD using textual descriptions. We randomly sample 15 images from the test set and ask human annotators to rank their relevance to their textual labels on a scale between 1 and 5. We enforce at least 17 annotations for each of the evaluated images and compute the mean score together with its standard deviations. AUDIOTOKEN outperforms Wav2Clip (4.07 ± 0.83 vs. 1.85 ± 0.46). When considering comparison to SD using text labels, AUDIOTOKEN is reaches comparable performance and yields slightly worse subjective scores (4.07 ± 0.83 vs. 4.58 ± 0.60). These findings are especially encouraging, as these suggest users found the images generated by AUDIOTOKENto capture the main objects in the audio scene similarly to using textual labels, which serves as a topline.

Speaker image generation. We investigate its potential to create visuals of various speakers. We gathered samples from two 30-minute videos per person that showcased Barack Obama, Donald Trump, Emma Watson, and David Beckham to achieve this goal and extracted the audio representation from X-Vector [39]. Our results in Fig. 4 indicate that our approach accurately represents Barack Obama and Donald Trump. We postulate that this could be due to their distinct voices. However, with Emma Watson and David Beckham, the method mainly captures their gender.

4. Conclusions

In this paper, we present a method for leveraging text-conditioned generative models for audio-based conditioning. Our method produces high-quality images which describe a scene from the audio recording. In addition, we propose a comprehensive evaluation framework that takes into account the semantics of the images generated. Our method presents a first step toward audio-conditioned image generation. The hidden information in the audio is richer than the observed one in the text. Hence, we think that this problem is interesting and should get more focus from the community.

5. References

[1] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. Denton, S. K. S. Ghasemipour, B. K. Ayan, S. S. Mahdavi, R. G. Lopes *et al.*, "Photorealistic text-to-image diffusion models with deep language understanding," in *NeurIPS*.

[2] O. Gafni, A. Polyak, O. Ashual, S. Sheynin, D. Parikh, and Y. Taigman, "Make-a-scene: Scene-based text-to-image generation with human priors," in *ECCV*, 2022.

[3] H. Chang, H. Zhang, J. Barber, A. Maschinot, J. Lezama, L. Jiang, M.-H. Yang, K. Murphy, W. T. Freeman, M. Rubinstein *et al.*, "Muse: Text-to-image generation via masked generative transformers," *arXiv preprint arXiv:2301.00704*, 2023.

[4] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar *et al.*, "Llama: Open and efficient foundation language models," *arXiv preprint arXiv:2302.13971*, 2023.

[5] T. L. Scao, A. Fan, C. Akiki, E. Pavlick, S. Ilić, D. Hesslow, R. Castagné, A. S. Luccioni, F. Yvon, M. Gallé *et al.*, "Bloom: A 176b-parameter open-access multilingual language model," *arXiv preprint arXiv:2211.05100*, 2022.

[6] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

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

[8] F. Kreuk, Y. Taigman, A. Polyak, J. Copet, G. Synnaeve, A. Défossez, and Y. Adi, "Audio language modeling using perceptually-guided discrete representations," *arXiv preprint arXiv:2211.01223*, 2022.

[9] K. Lakhotia, E. Kharitonov, W.-N. Hsu, Y. Adi, A. Polyak, B. Bolte, T.-A. Nguyen, J. Copet, A. Baevski, A. Mohamed *et al.*, "On generative spoken language modeling from raw audio," *Transactions of the Association for Computational Linguistics*, vol. 9, pp. 1336–1354, 2021.

[10] C. Wang, S. Chen, Y. Wu, Z. Zhang, L. Zhou, S. Liu, Z. Chen, Y. Liu, H. Wang, J. Li *et al.*, "Neural codec language models are zero-shot text to speech synthesizers," *arXiv preprint arXiv:2301.02111*, 2023.

[11] H. Cao, C. Tan, Z. Gao, G. Chen, P.-A. Heng, and S. Z. Li, "A survey on generative diffusion model," *arXiv preprint arXiv:2209.02646*, 2022.

[12] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *NeurIPS*, 2020.

[13] P. Dhariwal and A. Nichol, "Diffusion models beat gans on image synthesis," *Advances in Neural Information Processing Systems*, vol. 34, pp. 8780–8794, 2021.

[14] A. Q. Nichol and P. Dhariwal, "Improved denoising diffusion probabilistic models," in *International Conference on Machine Learning*. PMLR, 2021, pp. 8162–8171.

[15] Z. Kong, W. Ping, J. Huang, K. Zhao, and B. Catanzaro, "Diffwave: A versatile diffusion model for audio synthesis," *arXiv preprint arXiv:2009.09761*, 2020.

[16] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *CVPR*, 2022.

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

[18] A. Q. Nichol, P. Dhariwal, A. Ramesh, P. Shyam, P. Mishkin, B. Mcgrew, I. Sutskever, and M. Chen, "Glide: Towards photorealistic image generation and editing with text-guided diffusion models," in *International Conference on Machine Learning*. PMLR, 2022, pp. 16 784–16 804.

[19] R. Sheffer and Y. Adi, "I hear your true colors: Image guided audio generation," *arXiv preprint arXiv:2211.03089*, 2022.

[20] V. Iashin and E. Rahtu, "Taming visually guided sound generation," in *British Machine Vision Conference (BMVC)*, 2021.

[21] R. Gao and K. Grauman, "Visualvoice: Audio-visual speech separation with cross-modal consistency," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2021, pp. 15 490–15 500.

[22] W.-N. Hsu, T. Remez, B. Shi, J. Donley, and Y. Adi, "Revise: Self-supervised speech resynthesis with visual input for universal and generalized speech enhancement," *arXiv preprint arXiv:2212.11377*, 2022.

[23] Y. Tewel, Y. Shalev, I. Schwartz, and L. Wolf, "Zerocap: Zero-shot image-to-text generation for visual-semantic arithmetic," in *CVPR*, 2022.

[24] J. Li, D. Li, S. Savarese, and S. Hoi, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," *arXiv preprint arXiv:2301.12597*, 2023.

[25] F. Kreuk, A. Polyak, J. Copet, E. Kharitonov, T. A. Nguyen, M. Rivière, W.-N. Hsu, A. Mohamed, E. Dupoux, and Y. Adi, "Textless speech emotion conversion using discrete & decomposed representations," in *EMNLP*, 2022.

[26] G. Maimon and Y. Adi, "Speaking style conversion with discrete self-supervised units," *arXiv preprint arXiv:2212.09730*, 2022.

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

[28] M. Żelaszczyk and J. Mańdziuk, "Audio-to-image cross-modal generation," in *IJCNN*, 2022.

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

[30] H.-H. Wu, P. Seetharaman, K. Kumar, and J. P. Bello, "Wav2clip: Learning robust audio representations from clip," in *ICASSP*, 2022.

[31] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *ICML*, 2021.

[32] P. Esser, R. Rombach, and B. Ommer, "Taming transformers for high-resolution image synthesis," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 12 873–12 883.

[33] K. Crowson, S. Biderman, D. Kornis, D. Stander, E. Hallahan, L. Castricato, and E. Raff, "Vqgan-clip: Open domain image generation and editing with natural language guidance," in *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXVII*. Springer, 2022, pp. 88–105.

[34] D. Hendrycks and K. Gimpel, "Gaussian error linear units (gelus)," *arXiv preprint arXiv:1606.08415*, 2016.

[35] I. Schwartz, S. Yu, T. Hazan, and A. G. Schwing, "Factor graph attention," in *CVPR*, 2019.

[36] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," *NeurIPS*, 2017.

[37] R. Girdhar, A. El-Nouby, Z. Liu, M. Singh, K. V. Alwala, A. Joulin, and I. Misra, "Imagebind: One embedding space to bind them all," in *CVPR*, 2023.

[38] H. Chen, W. Xie, A. Vedaldi, and A. Zisserman, "Vggsound: A large-scale audio-visual dataset," in *ICASSP*, 2020.

[39] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2018, pp. 5329–5333.