MM-Diffusion: Learning Multi-Modal Diffusion Models for Joint Audio and Video Generation

Ludan Ruan[1]*, Yiyang Ma[2], Huan Yang[3]†, Huiguo He[3],
Bei Liu[3], Jianlong Fu[3], Nicholas Jing Yuan[3], Qin Jin[1], Baining Guo[3]
[1]Renmin University of China, [2]Peking University, [3]Microsoft Research
[1]{ruanld,qinj}@ruc.edu.cn, [2]myy12769@pku.edu.cn,
[3]{huayan,v-huiguohe,bei.liu,nicholas.yuan,jianf,bainguo}@microsoft.com

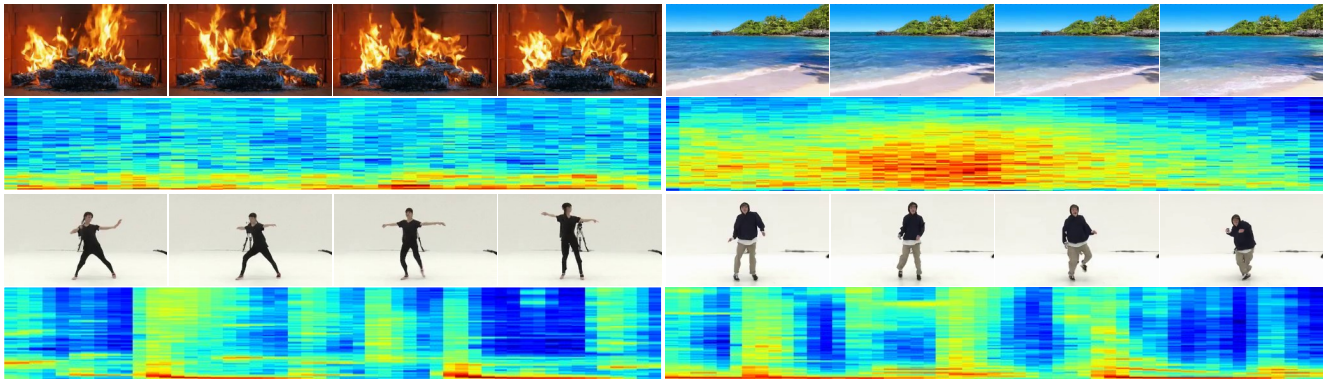

Figure 1. Examples of generated video frames (256×256) and audio spectrograms from Landscape [21] and AIST++ datasets [22]. We can see vivid bonfires burning, beautiful sea wave moving, and elegant dancing. Matched audio is generated with video appearances (e.g., the periodical rhythm for dancers). The complete high-fidelity videos and audio can be found in supplementary materials.

Abstract

*We propose the first joint audio-video generation framework that brings engaging watching and listening experiences simultaneously, towards high-quality realistic videos. To generate joint audio-video pairs, we propose a novel **M**ulti-**M**odal **Diffusion** model (*i.e.*, **MM-Diffusion**), with two-coupled denoising autoencoders. In contrast to existing single-modal diffusion models, MM-Diffusion consists of a sequential multi-modal U-Net for a joint denoising process by design. Two subnets for audio and video learn to gradually generate aligned audio-video pairs from Gaussian noises. To ensure semantic consistency across modalities, we propose a novel random-shift based attention block bridging over the two subnets, which enables efficient cross-modal alignment, and thus reinforces the audio-video fidelity for each other. Extensive experiments show superior results in unconditional audio-video generation, and zero-shot conditional tasks (e.g., video-to-audio). In particular, we achieve the best FVD and FAD on Landscape and AIST++ dancing datasets. Turing tests of 10k votes further demonstrate dominant preferences for our model. The code and pre-trained models can be downloaded at https://github.com/researchmm/MM-Diffusion.*

1. Introduction

AI-powered content generation in image, video, and audio domains has attracted extensive attention in recent years. For example, DALL·E 2 [33] and DiffWave [19] can create vivid art images and produce high-fidelity audio, respectively. However, such generated content can only provide single-modality experiences either in vision or audition. There are still large gaps with plentiful human-created contents on the Web which often involve multi-modal contents, and can provide engaging experiences for humans to perceive from both sight and hearing. In this paper, we take

*This work was performed when Ludan Ruan was visiting Microsoft Research Asia as research interns.

†Corressponding author.

one natural step forward to study a novel multi-modality generation task, in particular focusing on joint audio-video generation in the open domain.

Recent advances in generative models have been achieved by using diffusion models [14, 40]. From task-level perspectives, these models can be divided into two categories: unconditional and conditional diffusion models. In particular, unconditional diffusion models generate images and videos by taking the noises sampled from Gaussian distributions [14] as input. Conditional models usually import the sampled noises combined with embedding features from one modality, and generate the other modality as outputs, such as text-to-image [30, 33, 37], text-to-video [13, 39], audio-to-video [53], etc. However, most of the existing diffusion models can only generate single-modality content. How to utilize diffusion models for multi-modality generation still remains rarely explored.

The challenges of designing multimodal diffusion models mainly lie in the following two aspects. First, video and audio are two distinct modalities with different data patterns. In particular, videos are usually represented by 3D signals indicating RGB values in both spatial (*i.e.*, height $\times$ width) and temporal dimensions, while audio is in 1D waveform digits across the temporal dimension. How to process them in parallel within one joint diffusion model remains a problem. Second, video and audio are synchronous in temporal dimension in real videos, which requires models to be able to capture the relevance between these two modalities and encourage their mutual influence on each other.

To solve the above challenges, we propose the first **M**ulti-**M**odal **Diffusion** model (*i.e.*, **MM-Diffusion**) consisting of two-coupled denoising autoencoders for joint audio-video generation. Less-noisy samples from each modality (*e.g.*, audio) at time step $t-1$, are generated by implicitly denoising the outputs from both modalities (audio and video) at time step t. Such a design enables a joint distribution over both modalities to be learned. To further learn the semantic synchronousness, we propose a novel cross-modal attention block to ensure the generated video frames and audio segments can be correlated at each moment. We design an efficient random-shift mechanism that conducts cross-attention between a given video frame and a randomly-sampled audio segment in a neighboring period, which greatly reduces temporal redundancies in video and audio and facilitates cross-modal interactions efficiently.

To verify the proposed **MM-Diffusion** model, we conduct extensive experiments on Landscape dataset [21], and AIST++ dancing dataset [22]. Evaluation results over SOTA modality-specific (video or audio) unconditional generation models show the superiority of our model, with significant visual and audio gains of 25.0% and 32.9% by FVD and FAD, respectively, on Landscape dataset. Superior performances can be also observed in AIST++ dataset [22], with large gains of 56.7% and 37.7% by FVD and FAD, respectively, over previous SOTA models. We further demonstrate the capability of zero-shot conditional generation for our model, without any task-driven fine-tuning. Moreover, Turing tests of 10k votes further verify the high-fidelity performance of our results for common users.

2. Related Work

Diffusion Probabilistic Models. Diffusion Probabilistic Models (DPMs) [14, 40] are a new type of generative model which have achieved impressive results. They consist of a forward process (mapping signal to noise) and a reverse process (mapping noise to signal). It has further proved that the forward and reverse processes of DPMs can be done by solving differential equations [42]. They usually perform better with a reweighted objective during training [14]. In the aspect of generation quality and diversity, DPMs have outperformed other generative models with appropriate design of the denoising model [5]. It has shown that DPMs can perform well in several image generation tasks, such as image inpainting [28], super-resolution [32, 38, 50], image restoration [17], image-to-image translation [36], *etc*. Because of the character of DPMs which infer the denoising model repeatedly hundreds of times, their sampling speed of them is slow compared to other generative models such as GANs [9] and VAEs [18]. In order to make DPMs more practical, many methods have been proposed. Denoising Diffusion Implicit Models [41] first proposed a method of sampling through a DPM in an implicit way and accelerated the sampling speed. DPM Solver [26, 27] solved the ordinary differential equation of the reverse process of DPMs [42], gave high-order approximated solutions of these equations, and got high-quality results with only about 10-20 evaluations. Stable Diffusion [34] built DPMs on latent spaces so that the number of pixels was decreased. Along with the exploration and perfection of DPMs theories, it is becoming more popular to apply diffusion models in multiple domains.

Cross-Modality Generation. Cross-modal generation such as text-to-visual [7, 11, 29, 30, 39], text-to-audio [20], audio-to-visual [4, 8, 12], visual-to-audio [4, 6, 12, 52, 53], and visual transfer [16, 23, 24, 46–48, 51] has drawn great attention. In the aspect of audio-to-visual generation, Sound2Sight [3] first proposed a method of generating aligned videos from audio. TATS [8] proposed a time-sensitive transformer projecting audio latent embeddings to video embeddings and achieved SOTA results. For visual-to-audio generation, CMT [6] modeled music rhythms and proposed a method of generating background music corresponding to given videos with controllable music transformers. CDCD [53] applied DPMs and proposed a contrastive diffusion loss to improve the alignment of the gen-

erated audio and the given videos. For bidirectional conditional generation, Chen et al. [4] first propose 2 separate frameworks for audio-to-image and image-to-audio generation. CMCGAN [12] further combines audio-image bidirectional transfer with a unified framework and prove it is better than separate frameworks. However, previous works could only generate one modality at a time, while our work can generate two modalities simultaneously.

3. Approach

This section presents our proposed novel **M**ulti-**M**odal **Diffusion** model (i.e., **MM-Diffusion**) for realistic audio-video joint generation. Before diving into specific designs, we first briefly recap the preliminary knowledge of diffusion models in Sec. 3.1. Then, we introduce the proposed **MM-Diffusion** by further developing vanilla diffusion models to enable semantically-consistent multiple-modality generation in Sec. 3.2. After that, we illustrate a coupled U-Net architecture for joint audio-video data modeling by design in Sec. 3.3. In Sec. 3.4, we finally discuss the generation capability of our model for conditional multi-modality generation (*i.e.* audio-to-video and video-to-audio) in a zero-shot manner.

3.1. Preliminaries of Vanilla Diffusion

Diffusion-based models [14,40] refer to a class of generation algorithms that first transfer a given data distribution $\boldsymbol{x}$ into unstructured noise (Gaussian noise in practice), and further learn to recover the data distribution by reversing the above forward process. The original forward process of Denoising Diffusion Probabilistic models (DDPMs) [14] is performed over a discrete T time step. Define x_0 as a sample from X, and x_T as the sample that fits standard Gaussian distribution and is independent from x_0 using the Markovian forward process, which can be expressed as follows:

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t\mathbf{I}), \quad (1)$$

$$q(x_{1:T}|x_0) = \prod_{t=1}^{T} q(x_t|x_{t-1}), \quad (2)$$

where $t \in [1, T]$, and $\beta_0, \beta_1, ..., \beta_T$ are pre-defined variance schedule sequences. We follow previous works [14,42] and use the linear noise schedule to increase β_t.

To recover an original image, learning to reverse the forward process can be simplified as training a model θ to fit $p_\theta(x_{t-1}|x_t)$ that approximates with $q(x_{t-1}|x_t, x_0)$ for all given t and x_t. Thus the reverse process can be formulated as Equation 3, and x_0 can be recovered from a probability density $p(x_T)$ with Equation 4, which are shown as follows:

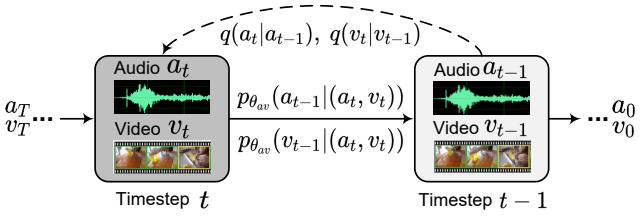

Figure 2. An illustration of multi-modal denoising diffusion process. Forward diffusion (dotted arrow) maps audio & video data to noise independently, while the reverse process (solid arrow) gradually reconstructs multimodal contents by a unified model θ_{av}.

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)), \quad (3)$$

$$p_\theta(x_{0:T}) = p(x_T) \prod_{t=1}^{T} p_\theta(x_{t-1}|x_t), \quad (4)$$

where μ_θ denotes the Gaussian mean value predicted by θ. And finally, x_0 can be obtained. In practice, we remove the variance prediction as it only leads to minor improvement [1,31]. We also omit this term in the following.

3.2. Multi-Modal Diffusion Models

With the forward and reverse process in diffusion defined above, we further present the proposed MM-Diffusion formulations in this section. As shown in Figure 2, different from the vanilla diffusion where a single modality is generated, our target is to recover two consistent modalities (*i.e.*, audio and video) within one diffusion process.

Given a paired data (a, v) from a 1D audio set A ($a \in A$), and a 3D video set V ($v \in V$), we consider that the forward processes of each modality are independent, since they are in different distributions. Taking the audio a as an example, its forward process at time step t is defined as:

$$q(a_t|a_{t-1}) = \mathcal{N}_a(a_t; \sqrt{1-\beta_t}a_{t-1}, \beta_t\mathbf{I}), \quad t \in [1, T]. \quad (5)$$

For simplicity, we omit the forward process for videos v, as they share a similar formulation. We can further calculate any a_t, v_t using Equation 2. It is worth noting that we empirically set a shared schedule for hyper-parameters β across audio and video to simplify the process definition.

Different from the forward process that models audio and video independently, the correlation between the two modalities should be considered during their reverse processes. Therefore, instead of directly fitting $q(a_{t-1}|a_t, a_0)$ and $q(v_{t-1}|v_t, v_0)$, we propose a **unified model** θ_{av} to take both modalities as inputs and reinforce audio and video generation quality for each other. In particular, for a given time step t, the reverse process $p_{\theta_{av}}(a_{t-1}|(a_t, v_t))$ for obtaining a_{t-1} in audio domains is formulated as follows:

$$p_{\theta_{av}}(a_{t-1}|(a_t, v_t)) = \mathcal{N}(a_{t-1}; \mu_{\theta_{av}}(a_t, v_t, t)), \quad (6)$$

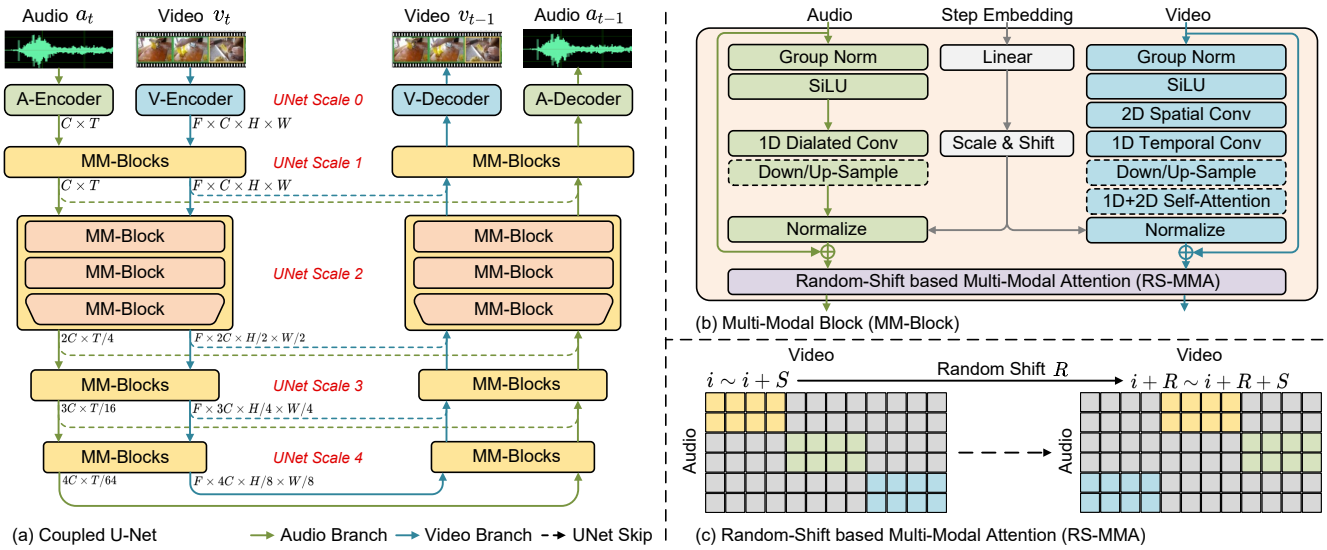

Figure 3. Overview of the proposed **MM-Diffusion** framework. Coupled U-Net contains coupled audio and video streams (indicated by green and blue blocks respectively) at each denoising diffusion step in (a). Each MM-Block encodes audio and video by 1D dilated audio convolutions, and 2D+1D spatial-temporal visual convolutions in (b). An efficient random-shift based multi-modal attention module is further proposed in (c) to facilitate specific inter-modality alignment and avoid redundant computations.

where a_{t-1} is generated from a Gaussian distribution jointly determined by both a_t and v_t. To optimize the whole network, we use ϵ-prediction that is defined as:

$$\mathcal{L}_{\theta_{av}} = \mathbb{E}_{\epsilon\sim\mathcal{N}_a(0,I)}\left[\lambda(t)\,||\tilde{\epsilon}_\theta(a_t, v_t, t) - \epsilon||_2^2\right], \quad (7)$$

where $t \in [0, T]$, and λ_t is an optional weighting function. We omit the video formulations since they share similar representations to audio.

The core advantage of multi-modality generation lies in the unified model θ_{av} that enables jointly reconstructing audio-video pairs from independent Gaussian distributions. Our designed model MM-Diffusion is capable of adapting these two types of input modalities with completely different shapes and patterns.

3.3. Coupled U-Net for Joint Audio-Video Denoising

Previous works [5, 14, 19, 34] have demonstrated the effectiveness of using U-Nets as the model architecture to generate single modality (*e.g.*, 2D U-Net [5, 14] for image generation and 1D U-Net [19] for audio generation). Inspired by these works, we propose a coupled U-Net (shown in Figure 3 (a)), which consists of two single-modal U-Nets for audio and video generation. In particular, we formulate input audio and video as a tensor pair $(a, v) \in (A, V)$. On one hand, $a \in \mathbb{R}^{C\times T}$ refers to audio input, where C and T are the channel and temporal dimensions, respectively. On the other, $v \in \mathbb{R}^{F\times C\times H\times W}$ refers to video input, where F, C, H, and W are frame number, channels, height, and width dimensions, respectively.

Efficient Multi-Modal Blocks. As shown in Figure 3 (b), for **video** sub-network design, to efficiently model the spatial and temporal information, we follow Jonathan et al. [15] to decompose the spatial and temporal dimensions. Specifically, we stack 1D convolutions followed by 2D convolutions as video encoders instead of using the heavy 3D convolutions. Similarly, video attention modules are composed of 2D and 1D attention. Different from videos, the **audio** signal is a 1D long sequence with higher demand for long-term dependency modeling. Therefore, we have two special designs for audio blocks. First, inspired by Kong [19], we stack dilated convolution layers instead of adopting pure 1D convolutions. The dilation is doubled from 1 to 2^N, where N is a hyper-parameter. Second, we delete all temporal attentions in audio blocks, which are computationally heavy and showed a limited effect in our preliminary experiments. Previous studies [13, 19] drew similar conclusions as well.

Random-Shift based Multi-Modal Attention. To bridge the two sub-networks of audio and video, and jointly learn their alignment, the most straightforward way is to perform cross-attention to their features. However, the original attention map for these two modalities is too huge to calculate, with the computational complexity of $O((F \times H \times W) \times T)$. Meanwhile, both video and audio are temporal redundant, which means that not all cross-modal attention computations are necessary.

To solve the above issues, we propose a **M**ulti-**M**odal **A**ttention mechanism with **R**andom **S**hift-based attention masks to align video and audio in an efficient way (denoted

as **RS-MMA**), as shown in Figure 3 (c). Specifically, given the l^{th} layer of the coupled U-Nets, with its output of the shape $\{H^l, W^l, C^l, T^l\}$, a 3D video input tensor v with F frames is represented by $F \times H^l \times W^l$ patches, and a 1D audio input tensor is represented by $C^l \times T^l$.

To better align video frames and audio signals, we propose a random-shift attention scheme with following steps:
Step I: we first split audio stream into segments $\{a_1, a_2, ...a_F\}$ along the time steps of video frames, where each segment a_i is in the shape of $C^l \times \frac{T^l}{F}$.
Step II: we set a window size S that is much smaller than frame number F, and set a random-shift number $R \in [0, F-S]$. The attention weights from audio to video are calculated between each audio segment a_i and video segment v_j starting from frame f_s to frame f_e, where $f_s = (i+R)\%F$ and $f_e = (i+R+S)\%F$.
Step III: the cross attention of audio segment a_i and sampled video segment $v_j = v_{f_s:f_e}$ is formulated as:

$$\mathrm{MMA}(a_i, v_j) = \mathrm{softmax}(\frac{Q_i^a K_j^{vT}}{\sqrt{d_k}})V_j^v, \tag{8}$$

$$K_j^v = \mathrm{linear}(\mathrm{flatten}(v_j)), \tag{9}$$

where d_k is the dimension of K. We omit $\mathrm{MMA}(v_j, a_i)$, since the cross attention from video to audio is symmetrical.

This attention mechanism brings two advantages. First, by using such designs, the computation complexity can be reduced to $O((S \times H \times W) \times (S \times \frac{T}{F}))$. Second, the design maintains global attention capabilities within a neighboring period. Since Multi-Modal Diffusion allows iterating from step T to step 0, video and audio can fully interact with each other during the reverse process. In practice, we set a smaller S at the top of the U-Net to capture fine-grained correspondence, and a larger S at the bottom of the U-Net to capture high-level semantic correspondence in an adaptive way. Detailed settings are described in experiments.

3.4. Zero-Shot Transfer to Conditional Generation

Although the MM-Diffusion model is trained for unconditional audio-video pair generation, it can also be utilized for conditional generation (*i.e.*, audio-to-video or video-to-audio) in a zero-shot transfer manner. Because the model has learned the correlation between these two modalities, a strong zero-shot conditional generation performance can help to verify the superior modeling capability of MM-Diffusion. In practice, inspired by Video Diffusion [15], we take two ways for conditional generation, including a replacement-based method and an improved gradient-guided method.

For **replacement-based** method, to generate an audio a conditioned by a video v, *i.e.*, $a \sim p_{\theta_{av}}(a|v)$, we replace v from the reverse process $p_{\theta_{av}}(a_t|(a_{t+1}, v_{t+1}))$ with samples in forward process $q(\hat{v}_{t+1}|v)$ at each diffusion step t. A similar operation can be conducted for video-to-audio generation. However, the replacement-based method predicts the target audio distribution from $a_t \sim \mathbb{E}_q(a_t|(a_{t+1}, \hat{v}_{t+1}))$, while the original v that can intuitively provides stronger conditional guidance is neglected. Therefore, we add this condition and reformulate it as **gradient-guided** method as follows:

$$\mathbb{E}_q(a_t|(a_{t+1}, \hat{v}_{t+1}, v)) = \mathbb{E}_q(a_t|(a_{t+1}, \hat{v}_{t+1})) + \frac{1}{\sqrt{1-\overline{\alpha}_t}}\nabla_{a_t}\log q(v_t|(a_{t+1}, \hat{v}_{t+1})), \tag{10}$$

where $\alpha = 1-\beta$ and $\overline{\alpha_t} = \alpha_t * \alpha_{t-1} * ... * \alpha_1$. Thus, we get generated audio $\tilde{a}_t$ from the following formulation:

$$a_t, v_t = \theta_{av}(a_{t+1}, \hat{v}_{t+1}), \tag{11}$$

$$\tilde{a}_t = a_t - \lambda\sqrt{1-\overline{\alpha}_t}\nabla_{a_t}||v_t - \hat{v}_t||_2^2. \tag{12}$$

This formulation is also similar to classifier-free conditional generation [25], in which λ plays the role of a gradient weight to control the intensity of the conditioning. The main difference is that traditional conditional generation models often require explicit training to fit the condition data. Thus their updating process of the sampling procedure does not need to change the condition. On contrary, to fit the unconditional training process, the conditional input of our gradient-guided method requires constant replacement as the backward process progresses. As a result, we do not need additional training to adapt to conditional inputs which shows a significant merit.

4. Experiments

In this section, we evaluate the proposed **MM-Diffusion** model, and compare its joint audio and video generation performance with the SOTA generative models. Visual results can be found in Figure 4, and more results in the open domain can be found in supplementary materials.

4.1. Implementation Details

Diffusion model. To make fair comparison, we follow previous works [26, 27] to use a linear noise schedule and the noise prediction objective in Sec. 3.1 for all experiments. The diffusion step T is set as 1,000. To accelerate sampling, we use DPM-Solver [26] as the default sampling method unless otherwise specified.
Model architecture. Our whole pipeline contains a coupled U-Net to generate video of $16 \times 3 \times 64 \times 64$, audio of $1 \times 25,600$, and a **S**uper **R**esolution model to scale image from 64 to 256. For the base coupled U-Net, we set 4 scales of MM-Blocks, and each is stacked by 2 normal MM-Blocks and 1 down/up-sample block. Only on U-Net scale of [2,3,4], video attention and cross-modal attention are applied, and the window size for cross-modal attention

is [1,4,8] corresponding to each scale. The whole model contains 115.13M parameters. For the SR model, we follow the structure and setting of ADM [5] with 311.03M parameters. All details of model architecture and training configuration can refer to supplementary materials.

Evaluation. To keep consistency, we randomly generate 2,048 samples with each model in the objective evaluation. For fair comparison, metrics are calculated on 64×64 resolution for all methods. In the main results of Sec. 4.4, we calculate 6 runs on average for reducing randomness. For the ablation study in Sec. 4.5, we sample 2,048 samples from the base coupled U-Net for efficiency.

4.2. Datasets

Previous works on video or audio generation mainly focus on one modality. Existing video datasets have problems such as low audio quality, missing audio, and visual-audio mismanagement (e.g., half audio is missing in UCF-101 [43]). To facilitate multi-modal generation and extensively compare with different methods, we conduct our experiments on two high-quality video-audio datasets with different types: Landscape [21] and AIST++ [22].

Landscape dataset is a high-fidelity audio-video dataset with nature scenes. We crawl 928 source videos from Youtube with the URLs provided by [21], then divide into 1,000 non-overlapped clips of 10 seconds. The total duration is about 2.7 hours of 300K frames. Landscape dataset contains 9 diverse scenes, including explosion, fire cracking, raining, splashing water, squishing water, thunder, underwater burbling, waterfall burbling, and wind noise.

AIST++ [22] is a subset of AIST dataset [44], which contains street dance videos with 60 copyright-cleared dancing songs. The dataset includes 1,020 video clips with 5.2 hours duration of about 560K frames in total. To generate clear characters, we uniformly crop out a 1024×1024 picture from the center of videos for all methods in training.

4.3. Evaluation Metrics

Objective Evaluation. We measure the quality of generated audio and videos separately for the objective evaluation. For videos, we follow prior settings [8, 49] to use **Fréchet video distance (FVD)** and **kernel video distance (KVD)** with the I3D [2] classifier pre-trained on Kinetics-400 [2]. For audio evaluation, previous works of unconditional audio generation are prone to generate audio in specific domain (e.g., SC09 [45] for spoken digits). Their evaluation metric based on a specifically trained audio classifier is not suitable for our generated audio in the open domain [35]. Inspired by FID for image evaluation and FVD for video evaluation, we propose to compute a similar **Fréchet audio distance (FAD)** between features of generated audio and ground-truth audio (all FAD numbers need to be multiplied by 1e4). We select AudioCLIP [10], a pre-trained audio model that achieved SOTA in the Environmental Sound Classification tasks, as the audio feature extractor.

Table 1. Comparison with single-modal methods on **Landscape** dataset. * denotes complete ddpm sampling.

#	Method	FVD ↓	KVD ↓	FAD ↓
1	Ground-truth	17.83	-0.12	7.51
2	DIGAN [49]	305.36	19.56	-
3	TATS-base [8]	600.30	51.54	-
4	Diffwave [19]	-	-	14.00
5	Ours-v	238.33	15.14	-
6	Ours-a	-	-	13.6
7	Ours	229.08	13.26	**9.39**
8	Ours*	**117.20**	**5.78**	10.72

Table 2. Comparison with single-modal generation methods on **AIST++** dancing dataset.

#	Method	FVD ↓	KVD ↓	FAD ↓
1	Ground-truth	8.73	0.036	8.46
2	DIGAN [49]	119.47	35.84	-
3	TATS-base [8]	267.24	41.64	-
4	Diffwave [19]	-	-	15.76
5	Ours-v	184.45	33.91	-
6	Ours-a	-	-	13.30
7	Ours	176.55	31.92	12.90
8	Ours*	**75.71**	**11.52**	**10.69**

Subjective Evaluation. We also conduct user studies on Amazon Mechanical Turk to measure both the quality and relevance of generated audio-video pairs. Specifically, for each audio-video pair, three tasks are formed to measure the quality of the audio, the video, and the relevance of the pair. For each task, we ask users to assign scores ranging from 1 (bad) to 5 (good). We average the scores as the final score, namely **Mean Opinion Score (MOS)**. Moreover, we perform Turing Test for audio-video pairs generated by our model and the ground-truth data. We mix them up and ask users to judge whether they are generated or not.

4.4. Objective Comparison with SOTA methods

To evaluate the quality of audio and video generated by MM-Diffusion, we compare it with SOTA unconditional video generation methods DIGAN [49], TATS [8], and audio generation method Diffwave [19]. Note that we select these baselines as they are widely-used and have released official codebases for standard replacement on our datasets. To further explore the effectiveness of joint-learning in MM-Diffusion and to make fair comparisons to single-modality generation with the same backbone, we decompose the coupled U-Nets into audio sub-network (Ours-a) and video sub-network (Ours-v) for modality-independent

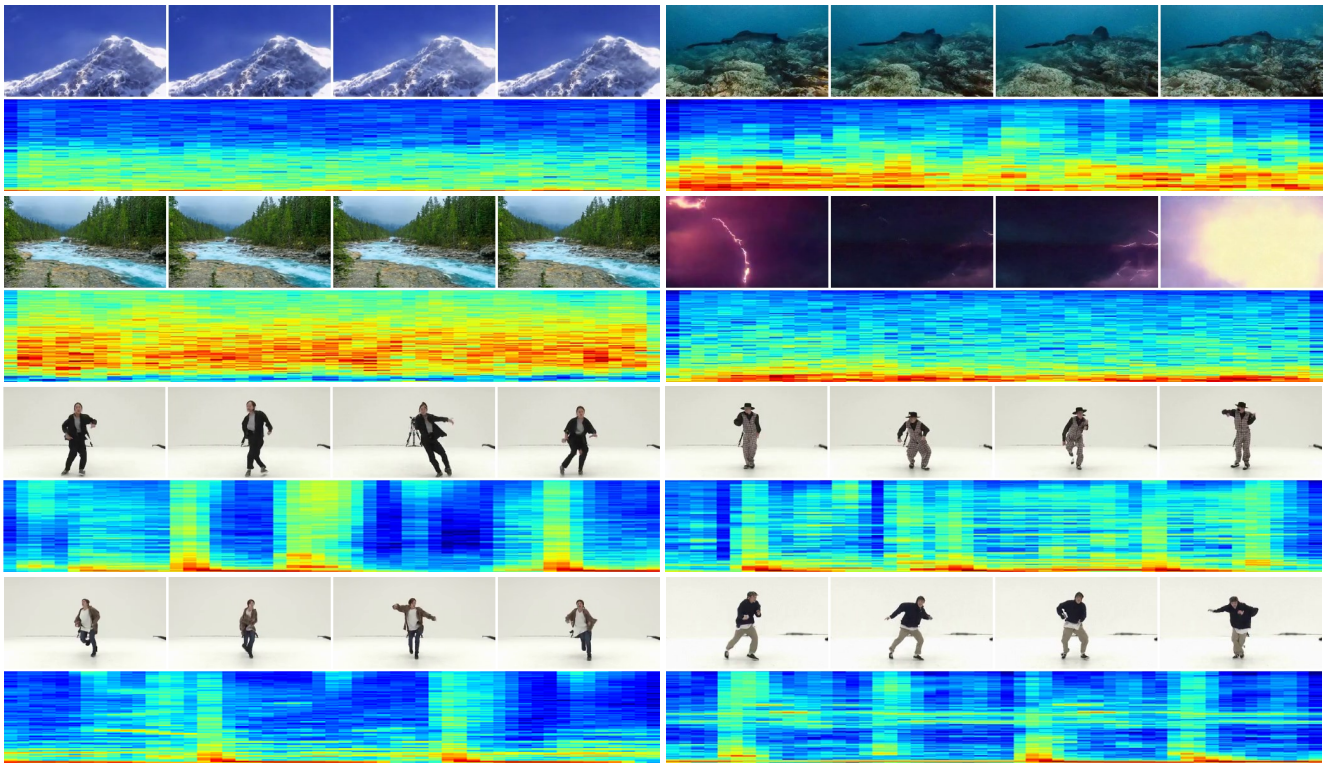

Figure 4. More visual examples of generated video frames (256×256) with semantic-consistent audio (shown in spectrograms). Some cases vividly show the wind blowing in snow mountains, and some show continuous river sound with beautiful scenes.

generation. The results on Landscape and AIST++ are shown in Table 1 and Table 2.

From these two tables, we can draw the following conclusions: (1) Our model significantly outperforms SOTA single-modal generation methods on both video and audio generation. In particular, our model elevates the SOTA FAD towards ground-truth quality. It demonstrates the effectiveness of the proposed MM-Diffusion and coupled U-Net. (2) Our model with only video generation (Ours-v) even outperforms SOTA methods DIGAN and TATS-base on most metrics in both tables (comparing #5 to #2 and #3). This indicates that the diffusion-based method can improve the quality of generated videos compared to traditional methods. (3) By comparing our full setting (#7) to one-stream U-Net (Ours-v and Ours-a), we can see that Coupled U-Nets that jointly learn cross-modality alignment bring further benefit for both video and audio generation. Moreover, complete sampling strategy (#8) will obtain samples of better quality than Dpm-Solver.

4.5. Ablation Studies

Random-Shift based Multi-modal Attention. We have demonstrated the effectiveness of our proposed Random-shift based Multi-Modal Attention mechanism (RS-MMA) in Sec. 4.4. We further conduct two ablation experiments to explore different window sizes and the effectiveness of random shift mechanisms. **(1) Different window sizes.** We first set different window sizes to scale [2,3,4] of the Coupled U-Net. All the experiments are training with 80K steps to save cost and the results are shown in Table 3. From the first three lines, we can see that larger window sizes bring more improvement. The best performance of the adaptive window size according to the channel scale in U-Net shows the effectiveness of this efficient design, especially for improving video generation quality. **(2) Random shift mechanism.** Table 4 shows the results of whether to use random shift (RS) during training. From the comparison, we can find that RS helps generate audio with better quality compared with no shifting, and the convergence of audio is also accelerated. This also demonstrates that our proposed RS-MMA encourages more efficient joint cross-modality learning. Meanwhile, we can see the improvement is more significant in audio quality by using RS. Because the video appearance can provide more information for its paired audio, compared with the effect of audio on the paired video.

Zero-Shot Conditional Generation. We validate the effectiveness of the two methods for zero-shot transferring and find that both can generate high-quality audio using videos as the condition. For audio-based video generation, the gradient-guided method is better than the replacement

Table 3. Ablation study on the window size of multi-modal attention, corresponding to U-Net scale [2,3,4], at 80k training steps.

#	W-Size	FVD ↓	KVD ↓	FAD ↓
1	[1,1,1]	374.18	22.26	9.81
2	[4,4,4]	361.65	21.64	9.65
3	[8,8,8]	350.60	21.47	**9.50**
4	[1,4,8]	**303.19**	**17.26**	10.20

Table 4. Video and audio quality at different training steps, affected by random-shift attention.

Method	60K/ 80K/ 100K Iteration	
	FVD↓	FAD↓
w/o RS	415.50/ 303.19/ 271.64	12.55/ 10.20/ 9.94
w/ RS	440.78/ 306.42/ 267.58	13.29/ 9.67/ 9.10

method to get consistent videos that are semantic and temporal aligned with given audio. The results also show that our model has the ability of modality transfer even without extra training. Figure 5 illustrates that our model can generate videos of similar scenes (sea) from audios with similar patterns, or generate audio that matches the rhythm of the input dancing videos. This further verifies that our joint learning can enhance single-modality generation.

4.6. User Studies

Comparison with other Methods. As we are the first to jointly generate audio-video pairs, there is no direct baselines to compare. Hence we choose a 2-stage pipeline by using an existing single-modality model. In particular, we take a noise-audio-video order as the pipeline. Specifically, we utilize Diffwave [19] for unconditional audio generation and TATS [8] to transfer audio to video. For each dataset, we randomly sampled 1,500 audio-video pairs from our model, baseline, and ground-truth data, each with 500 samples. As we explained in Sec. 4.3, each pair is divided into 3 tasks. Each task was assigned to 5 users. Thus, we have 45k votes from 9,000 tasks in total. From the results in Table 5, we can see that the quality of audio-video pairs generated by our method on both datasets is much better than the 2-stage baseline method, and our results have much smaller gap with the ground-truth data. **Turing Test.** To evaluate the realisticness of our generated videos, we further conduct a Turing test. For each dataset, we randomly sampled 500 audio-video pairs from our generated results and ground-truth data, respectively. Each sample was assigned to 5 users and we have 10k votes in total. From the results shown in Table 6, we can see that over 80% generated sound videos in Landscape can successfully cheat the subjects. Even in AIST++, almost half of the generated sound videos can fool users although the fine-grained parts of persons are difficult to be well generated. This test provides strong validation of the high quality and realism of sound videos we generated for common users.

Table 5. Mean opinion score (5 is the highest). VQ/AQ denotes video and audio quality. A-V denotes cross-modal alignment.

Method	Landscape			AIST++		
	VQ↑	AQ↑	A-V↑	VQ↑	AQ↑	A-V↑
GT	3.84	4.22	4.52	3.79	3.89	4.15
2-Stage	1.61	1.74	1.72	2.31	2.27	1.81
Ours	3.75	3.93	4.33	3.48	3.50	3.87

Table 6. Turing Test on Landscape and AIST++, the number is the percentage of data that is considered to be from real world.

	Landscape	AIST++
Ours	84.9	49.6
Ground-truth	92.5	84.7

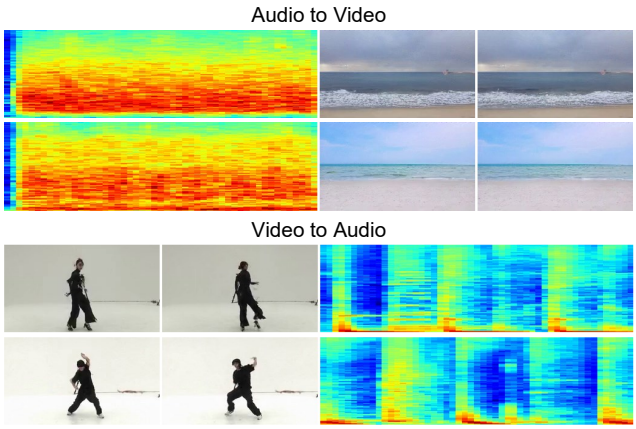

Figure 5. Illustration of several randomly-selected examples generated by zero-shot transferring to conditional generation. We adopt the gradient-guided method for better results.

5. Conclusion

In this paper, we propose **MM-Diffusion**, a novel multimodal diffusion model for joint audio and video generation. Our work pushes the current content generation based on single-modality diffusion models one step forward, and the proposed MM-Diffusion can generate realistic audio and videos in a joint manner. Superior performances are achieved over widely-used audio-video benchmarks by objective evaluations and Turing tests, which can be attributed to the new formulation for multimodal diffusion, and the designed coupled U-Net. In the future, we will add text prompts to guide audio-video generation as a more user-friendly interface, and further develop various video editing techniques (e.g., video inpainting, background music synthesis) by multi-modal diffusion models.

References

[1] Fan Bao, Chongxuan Li, Jun Zhu, and Bo Zhang. Analytic-DPM: an analytic estimate of the optimal reverse variance in diffusion probabilistic models. In *ICLR*, 2022. 3

[2] João Carreira and Andrew Zisserman. Quo vadis, action recognition? A new model and the kinetics dataset. In *CVPR*, pages 4724–4733, 2017. 6

[3] Moitreya Chatterjee and Anoop Cherian. Sound2Sight: Generating visual dynamics from sound and context. In *ECCV*, pages 701–719, 2020. 2

[4] Lele Chen, Sudhanshu Srivastava, Zhiyao Duan, and Chenliang Xu. Deep cross-modal audio-visual generation. In *ACM MM*, 2017. 2, 3

[5] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. In *NeurIPS*, volume 34, pages 8780–8794, 2021. 2, 4, 6

[6] Shangzhe Di, Zeren Jiang, Si Liu, Zhaokai Wang, Leyan Zhu, Zexin He, Hongming Liu, and Shuicheng Yan. Video background music generation with controllable music transformer. In *ACM Multimedia*, pages 2037–2045, 2021. 2

[7] Oran Gafni, Adam Polyak, Oron Ashual, Shelly Sheynin, Devi Parikh, and Yaniv Taigman. Make-A-Scene: Scene-based text-to-image generation with human priors. *arXiv preprint arXiv:2203.13131*, 2022. 2

[8] Songwei Ge, Thomas Hayes, Harry Yang, Xi Yin, Guan Pang, David Jacobs, Jia-Bin Huang, and Devi Parikh. Long video generation with time-agnostic VQGAN and time-sensitive transformer. In *ECCV*, 2022. 2, 6, 8

[9] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NeurIPS*, volume 27, 2014. 2

[10] Andrey Guzhov, Federico Raue, Jörn Hees, and Andreas Dengel. AudioCLIP: Extending clip to image, text and audio. In *ICASSP*, pages 976–980, 2022. 6

[11] Tiankai Hang, Huan Yang, Bei Liu, Jianlong Fu, Xin Geng, and Baining Guo. Language-guided face animation by recurrent StyleGAN-based generator. *arXiv preprint arXiv:2208.05617*, 2022. 2

[12] Wangli Hao, Zhaoxiang Zhang, and He Guan. CMCGAN: A uniform framework for cross-modal visual-audio mutual generation. In *AAAI*, 2018. 2, 3

[13] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022. 2, 4

[14] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, volume 33, pages 6840–6851, 2020. 2, 3, 4

[15] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. In *NeurIPS*, 2022. 4, 5

[16] Kibeom Hong, Seogkyu Jeon, Huan Yang, Jianlong Fu, and Hyeran Byun. Domain-aware universal style transfer. In *ICCV*, pages 14589–14597, 2021. 2

[17] Bahjat Kawar, Michael Elad, Stefano Ermon, and Jiaming Song. Denoising diffusion restoration models. In *NeurIPS*, 2022. 2

[18] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *ICLR*, 2014. 2

[19] Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. Diffwave: A versatile diffusion model for audio synthesis. In *ICLR*, 2021. 1, 4, 6, 8

[20] Felix Kreuk, Gabriel Synnaeve, Adam Polyak, Uriel Singer, Alexandre Défossez, Jade Copet, Devi Parikh, Yaniv Taigman, and Yossi Adi. AudioGen: Textually guided audio generation. In *ICLR*, 2023. 2

[21] Seung Hyun Lee, Gyeongrok Oh, Wonmin Byeon, Chanyoung Kim, Won Jeong Ryoo, Sang Ho Yoon, Hyunjun Cho, Jihyun Bae, Jinkyu Kim, and Sangpil Kim. Sound-guided semantic video generation. In *ECCV*, pages 34–50, 2022. 1, 2, 6

[22] Ruilong Li, Shan Yang, David A. Ross, and Angjoo Kanazawa. AI Choreographer: music conditioned 3d dance generation with AIST++. In *ICCV*, 2021. 1, 2, 6

[23] Chengxu Liu, Huan Yang, Jianlong Fu, and Xueming Qian. 4D LUT: learnable context-aware 4d lookup table for image enhancement. *arXiv preprint arXiv:2209.01749*, 2022. 2

[24] Chengxu Liu, Huan Yang, Jianlong Fu, and Xueming Qian. TTVFI: learning trajectory-aware transformer for video frame interpolation. *arXiv preprint arXiv:2207.09048*, 2022. 2

[25] Xihui Liu, Dong Huk Park, Samaneh Azadi, Gong Zhang, Arman Chopikyan, Yuxiao Hu, Humphrey Shi, Anna Rohrbach, and Trevor Darrell. More control for free! image synthesis with semantic diffusion guidance. *CoRR*, abs/2112.05744, 2021. 5

[26] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. DPM-Solver: a fast ode solver for diffusion probabilistic model sampling in around 10 steps. In *NeurIPS*, 2022. 2, 5

[27] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. DPM-Solver++: fast solver for guided sampling of diffusion probabilistic models. *arXiv preprint arXiv:2211.01095*, 2022. 2, 5

[28] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. RePaint: Inpainting using denoising diffusion probabilistic models. In *CVPR*, pages 11461–11471, 2022. 2

[29] Yiyang Ma, Huan Yang, Bei Liu, Jianlong Fu, and Jiaying Liu. AI illustrator: Translating raw descriptions into images by prompt-based cross-modal generation. In *ACM MM*, pages 4282–4290, 2022. 2

[30] Yiyang Ma, Huan Yang, Wenjing Wang, Jianlong Fu, and Jiaying Liu. Unified multi-modal latent diffusion for joint subject and text conditional image generation. *arXiv preprint arXiv:2303.09319*, 2023. 2

[31] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *ICML*, volume 139, pages 8162–8171, 2021. 3

[32] Zhongwei Qiu, Huan Yang, Jianlong Fu, and Dongmei Fu. Learning spatiotemporal frequency-transformer for com-

pressed video super-resolution. In *ECCV*, pages 257–273, 2022. 2

[33] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with CLIP latents. *arXiv preprint arXiv:2204.06125*, 2022. 1, 2

[34] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695, 2022. 2, 4

[35] Ludan Ruan, Anwen Hu, Yuqing Song, Liang Zhang, Sipeng Zheng, and Qin Jin. Accommodating audio modality in CLIP for multimodal processing. In *AAAI*, 2023. 6

[36] Chitwan Saharia, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. In *ACM SIGGRAPH*, pages 1–10, 2022. 2

[37] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022. 2

[38] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *TPAMI*, 2022. 2

[39] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-A-Video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022. 2

[40] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *ICML*, pages 2256–2265. PMLR, 2015. 2, 3

[41] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*, 2021. 2

[42] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-Based Generative Modeling through Stochastic Differential Equations. In *ICLR*, 2021. 2, 3

[43] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 6

[44] Shuhei Tsuchida, Satoru Fukayama, Masahiro Hamasaki, and Masataka Goto. AIST dance video database: Multi-genre, multi-dancer, and multi-camera database for dance information processing. In *ISMIR*, 2019. 6

[45] Pete Warden. Speech Commands: A dataset for limited-vocabulary speech recognition. *CoRR*, 2018. 6

[46] Hongwei Xue, Bei Liu, Huan Yang, Jianlong Fu, Houqiang Li, and Jiebo Luo. Learning fine-grained motion embedding for landscape animation. In *ACM MM*, pages 291–299, 2021. 2

[47] Fuzhi Yang, Huan Yang, Jianlong Fu, Hongtao Lu, and Baining Guo. Learning texture transformer network for image super-resolution. In *CVPR*, pages 5790–5799, 2020. 2

[48] Fuzhi Yang, Huan Yang, Yanhong Zeng, Jianlong Fu, and Hongtao Lu. Degradation-guided meta-restoration network for blind super-resolution. *arXiv preprint arXiv:2207.00943*, 2022. 2

[49] Sihyun Yu, Jihoon Tack, Sangwoo Mo, Hyunsu Kim, Junho Kim, Jung-Woo Ha, and Jinwoo Shin. Generating videos with dynamics-aware implicit generative adversarial networks. In *ICLR*, 2022. 6

[50] Yanhong Zeng, Huan Yang, Hongyang Chao, Jianbo Wang, and Jianlong Fu. Improving visual quality of image synthesis by A token-based generator with transformers. In *NeurIPS*, pages 21125–21137, 2021. 2

[51] Heliang Zheng, Huan Yang, Jianlong Fu, Zheng-Jun Zha, and Jiebo Luo. Learning conditional knowledge distillation for degraded-reference image quality assessment. In *ICCV*, pages 10222–10231, 2021. 2

[52] Yipin Zhou, Zhaowen Wang, Chen Fang, Trung Bui, and Tamara L. Berg. Visual to sound: Generating natural sound for videos in the wild. In *CVPR*, 2018. 2

[53] Ye Zhu, Yu Wu, Kyle Olszewski, Jian Ren, Sergey Tulyakov, and Yan Yan. Discrete contrastive diffusion for cross-modal and conditional generation. *arXiv preprint arXiv:2206.07771*, 2022. 2