Animate and Sound an Image

Xihua Wang[1] Ruihua Song[1]✉ Chongxuan Li[1] Xin Cheng[1]
Boyuan Li[1] Yihan Wu[1] Yuyue Wang[1] Hongteng Xu[1] Yunfeng Wang[2]
[1]Gaoling School of Artificial Intelligence, Renmin University of China [2]ZHI-TECH GROUP
xihuaw@ruc.edu.cn, songruihua_bloon@outlook.com

Abstract

This paper addresses a promising yet underexplored task, Image-to-Sounding-Video (I2SV) generation, which animates a static image and generates synchronized sound simultaneously. Despite advances in video and audio generation models, challenges remain to develop a unified model for generating naturally sounding videos. In this work, we propose a novel approach that leverages two separate pretrained diffusion models and makes vision and audio influence each other during generation based on the Diffusion Transformer (DiT) architecture. First, the individual video and audio pretrained generation models are decomposed into input, output, and expert sub-modules. We propose using a unified joint DiT block to integrate the expert submodules to effectively model the interaction between the two modalities, resulting in high-quality I2SV generation. Then, we introduce a joint classifier-free guidance technique to boost the performance during joint generation. Finally, we conduct extensive experiments on three popular benchmark datasets, and in both objective and subjective evaluation our method surpass all the baseline methods in almost all metrics. Case studies show our generated sounding videos are high quality and synchronized between video and audio.

1. Introduction

The real world is constantly in motion and sound. Mimicking this dynamic nature is essential for content generation and foundational for building world models. For a long time, the generative community has predominantly focused on single-modality synthesis, exploring architectures [16, 36], generative frameworks [29, 34], and scaling properties [3, 10, 36] to enhance video or audio quality. Though capable of producing high-fidelity video [3, 23, 51] or audio [10, 32] separately, current single-modality generators struggle to create naturally paired sounding videos Bridging two modalities to jointly generate sounding videos remains understudied, persisting as an open challenge.

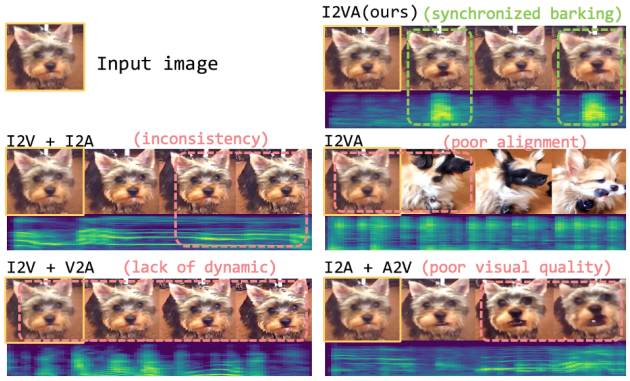

Figure 1. Cases for Image-to-Sounding-Video (I2SV) generation. Comparison with the other four schemes shows that our method significantly outperforms the other schemes regarding synchronization between the generated video and audio when animating and sounding an image in high quality. Zoom in to see details.

Several works initially address sounding video generation, either unconditionally or via text conditioning. Unconditional approaches [39, 43, 47] are limited to specific domains, such as dance [26] or landscapes [25] scenes, hindering an efficient exploration of open-domain capabilities. The unconditional generation also falls short when compared to conditional generation in terms of quality [27]. Text-conditioned approaches [18, 33, 44] offer more flexibility but face bottlenecks in processing the text modality. The performance heavily relies on constructing high-quality video-audio captions and modeling alignment with caption. This diverts the model's focus away from modeling the interaction between video and audio modalities, leading to a lack of sufficient fine-grained cross-modal interactions. Some works explore guidance techniques [13, 49] to enhance performance of aforementioned approaches at inference time, but do not directly address their problems. Both paths achieve generating sounding videos in suboptimal quality, neither shows to be a scalable way to explore sounding video generation with stable generation quality and focusing on audiovisual interaction mechanisms.

We propose formulating sounding video generation as an image-to-sounding-video (I2SV) task, where a static image serves as both the conditional input and the initial frame for generating a synchronized audiovisual sequence. Unlike unconditional or text-conditioned approaches, I2SV emphasizes bridging visual and auditory modalities while enabling open-domain sounding video generation. Based on existing works, there are three potential solutions to this task. (1) *Independent parallel generation* employs separate image→audio [40] and image→video [3] models, but risks temporal or semantic inconsistencies (e.g., a closed-mouth dog paired with barking sounds in Figure 1). (2) *Sequential generation* via cross-modal [32, 48, 55, 56] pipelines (image→audio→video or image→video→audio) reduces inconsistencies but suffers from error propagation—erroneous video outputs could corrupt subsequent audio generation. (3) *Coarse-grained I2SV* frameworks like CoDi [44] utilize contrastive learning for modality interaction, yet their semantic alignment remains insufficient for high-dimensional, temporally precise audiovisual synchronization. Developing a unified model that enables fine-grained cross-modal interaction while preserving temporal and semantic consistency remains a challenge.

To address this challenge, we propose a simple yet effective approach based on the diffusion transformer (DiT) [36] architecture for modeling the interaction between video and audio, using a single model to achieve image-to-sounding video generation. The capabilities of the DiT model have been recently validated across various generative domains, able to generate high-quality single modality data [4, 6, 10] and even fit unconditional video-audio joint [47] or text-image joint [2] distributions. We further investigate the potential of DiT in modeling sounding video. Rather than directly estimating a joint video-audio distribution using a DiT model trained from scratch, we leverage the powerful existing single modality generative models as submodules. Specifically, given two expert diffusion models, a pre-trained video generation model, and an audio generation model, regardless of whether they are based on UNet [3, 31] or DiT architectures, we decompose them into submodules in different types and then use a Joint Block to combine these submodules, forming a single DiT block. We find that such approach can leverage the expertise of pre-trained single-modality models while enabling the fine-grained interaction between the two modalities, thereby achieving high-quality image-to-sounding video generation.

We further improve the inference mechanism for I2SV joint generation. The classifier-free guidance (CFG) [15] for conditional generation is extensively employed in current research, demonstrating strong capabilities in following conditions and producing high-fidelity generation results. Typically, the CFG technique is applied for single-modality generation. We observe that in the joint generation model, the standard vanilla CFG may not be optimal as it only steers alignment with conditions without steering joint interactions. We introduce a joint classifier-free guidance (JointCFG) technique that employs the comparison between marginal distribution and joint distribution to refine the guidance effect. Further, we found that gathering inference results from submodules into JointCFG further improves guidance (JointCFG*).

Our contributions are: (1) We propose an innovative JointDiT architecture that leverages pretrained video and audio diffusion models and uses a Joint Block to model their interaction. (2) We introduce an improved CFG technique in joint generation scenarios that significantly enhances the performance of JointDiT, which is also applicable to other joint frameworks. (3) We benchmark I2SV task. Through detailed experiments, our method achieves state-of-the-art performance, as demonstrated on our demo page [1].

2. Related Work

Video and audio are inherently co-occurring modalities, and modeling their paired relationships is crucial towards estimating the real world. However, research on generation for both modalities has taken a long journey before converging.

Video Generation. In visual generation domain, the focus is on progressively modeling more complex visual distributions. This has evolved from generating low-resolution images [34] to producing high-resolution images [38] and further extending image models towards video generation [6, 12, 41]. As for video, the community concentrates on generating higher-fidelity, higher-resolution and longer-duration videos [3, 23, 51, 54, 58]. Along with this, there is a research line focuses on controllability, such as incorporating image conditions to significantly control the visual content [23, 28, 51], as opposed to relying on text only.

Audio Generation. The audio domain shares a similar trend. Development follows a path of generating higher-quality audio at fixed lengths [10, 30, 50], covering more types (e.g., sound, speech, music) [46], and extending audio length [9]. With the exploration of more complex tasks, along with scaling models and data, both video and audio modalities can now estimate highly complex unimodal distributions, yielding impressive unimodal generation results.

Video-Audio Cross-Generation. While unimodal generation excels, simply combining the outputs of both modalities does not yield a coherent result. To achieve fully consistent (e.g., semantic and synchronization) audiovisual outcomes, recent work has focused on conditioning between modalities to model their relationships. This includes generating videos from sound [25, 52, 55], speech [14] or music [7], or conversely, generating sound for a silent video [8, 32, 48, 56]. Such approaches establish connec-

[1] https://anonymoushub4ai.github.io/JointDiT

tions between the two modalities in a cross conditional manner, aiming for harmonized cross-generation results.

Video-Audio Joint Generation. Direct joint distribution modeling provides inherent modality consistency advantages over sequential generation that risks error propagation (e.g., video generation artifacts affecting subsequent audio synthesis). Current joint generation paradigms fall into two categories: (1) Leveraging two pre-trained video and audio models, where additional classifiers [13, 49] or cross-attentions [18, 33] guide joint generation. This way lacks fine-grained bidirectional interaction between modalities. (2) Training a new model from scratch to directly model the joint distribution [39, 43, 47]. While this approach allows finer-grained interactions, skipping unimodal learning is often inefficient, leading to poor generation quality and being limited to preliminary unconditional generation.

Efficiently modeling the joint distribution and capturing fine-grained interactions between video and audio remains an open challenge. Our work addresses this issue through exploring the I2SV task, focusing on the interaction between visual and audio modalities, especially the efficient interaction way of two pre-trained generative models.

3. Method

3.1. Preliminary: Unimodal Diffusion Models

Basics. Diffusion Models (DMs) are probabilistic generative models that learn to iteratively transform random noise into clean data. Typical DMs estimate single modalities (e.g., video or audio), using various frameworks like DDPMs [34] or EDMs [20]. The forward diffusion process gradually corrupts clean data into Gaussian noise following a predefined schedule. During training, a DM D_θ learns to reverse this corruption in several timesteps. We exemplify this with a video DM D_{θ_v} under the EDM-framework and an audio DM D_{θ_a} under the DDPM-framework. EDM directly adds Gaussian noise $\boldsymbol{n}_\sigma$ with σ^2-variance to clean video latents $\boldsymbol{v}_0$ at continuous timestep $t(\sigma) = 0.25\log\sigma$, where $\log\sigma$ is sampled from $\mathcal{N}(P_{\text{mean}}, P_{\text{std}}^2)$ (P_{mean}, P_{std} are predefined hyperparameters [3]). DDPM combines Gaussian noise ϵ and clean audio latents $\boldsymbol{a}_0$ linearly at discrete timestep t following a predefined schedule $\bar{\alpha}_t$. Both frameworks train denoising networks to predict clean latents from noisy inputs conditioned on c_v or c_a at each continuous or discrete timestep. The training objective minimizes an L2 loss between predicted and ground-truth latents:

$$\mathcal{L}(\theta_v) = \mathbb{E}\left[\|\boldsymbol{v}_0 - D_{\theta_v}\left(\boldsymbol{v}_0 + \boldsymbol{n}_\sigma; c_v, t(\sigma)\right)\|^2\right], \quad (1)$$

$$\mathcal{L}(\theta_a) = \mathbb{E}\left[\|\boldsymbol{a}_0 - D_{\theta_a}\left(\sqrt{\bar{\alpha}_t}\boldsymbol{a}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}; c_a, t\right)\|^2\right]. \quad (2)$$

Guidance Methods. Widely used **Classifier-free guidance (CFG)** [15] is a technique for DM inference that enhances generation quality and conditional alignment. It adjusts the denoising direction through linear extrapolation between conditional and unconditional predictions at each timestep. Take a video DM as illustration:

$$\begin{aligned} D_{\theta_v}^{\text{CFG}}\left(\boldsymbol{v}_t; c_v, t\right) = & D_{\theta_v}\left(\boldsymbol{v}_t; c_v, t\right) \\ & + \omega\left(D_{\theta_v}\left(\boldsymbol{v}_t; c_v, t\right) - D_{\theta_v}\left(\boldsymbol{v}_t; \varnothing, t\right)\right), \end{aligned} \quad (3)$$

where ω governs guidance strength and $\varnothing$ indicates null conditions for unconditional generation. In practice, both conditional and unconditional DMs can be implemented through a shared network D_{θ_v} by randomly replacing c_v with null embeddings during training (e.g., 10% of time).

Recent **Autoguidance (AG)** [21] introduces guidance with a *bad version* of conditional DM rather than with an unconditional version in CFG. It substitutes the unconditional DM term in Equation 3 with an early training-stage checkpoint (less-optimized *bad version*) of the conditional DM. This way demonstrates performance comparable to CFG while eliminating the need for explicit unconditional training.

3.2. JointDiT Architecture

Let i denote conditional image data, $\boldsymbol{v}_0 \in \mathbb{R}^{T_v \times S_v \times C_v}$ and $\boldsymbol{a}_0 \in \mathbb{R}^{T_a \times S_a \times C_a}$ represent paired clean latents of video and audio data respectively, where T, S, and C denote temporal, spatial (or frequential for audio), and channel dimensions. Given input image i and noised inputs $\boldsymbol{v}_t$ and $\boldsymbol{a}_t$ at each timestep (see Section 3.1), the goal of JointDiT is to model cross-modal interactions and predict the clean paired latents $(\hat{\boldsymbol{v}}_0, \hat{\boldsymbol{a}}_0)$, i.e., I2SV, as illustrated in Figure 2(b).

Architecture Overview. As shown in Figure 2(b), JointDiT consists of three main components: an Input Block, multiple Joint Blocks, and an Output Block. To effectively model the two modalities within one model while addressing their heterogeneity, JointDiT incorporates both shared interaction layers and modality-specific layers.

The Input Block contains two independent input layers, the V-Input and A-Input layers, which preprocess video and audio data respectively. The Joint Block comprises three types of layers: (1) modality-specific layers (V-Expert and A-Expert) that focus on intra-modal interactions, such as self-attention and conditional image processing (e.g., concatenation or cross-attention, see Supplementary A); (2) a global full attention layer that models fine-grained cross-modal interactions across temporal, spatial, and frequential dimensions using global attention without inductive biases; and (3) modality-specific feedforward layers (V-Feedforward and A-Feedforward), following the standard DiT design [36]. Each layer is followed by modality-specific adaptive layer normalization (AdaLN) [36]. Finally, the Output Block includes two independent layers, V-Out and A-Out, which decode the processed representations into the clean latents $\hat{\boldsymbol{v}}_0$ and $\hat{\boldsymbol{a}}_0$, respectively.

Input, Expert, and Output Layers. To efficiently model modality-specific pattern, JointDiT directly leverages pre-trained unimodal DMs D_{θ_v} and D_{θ_a} (from any architecture

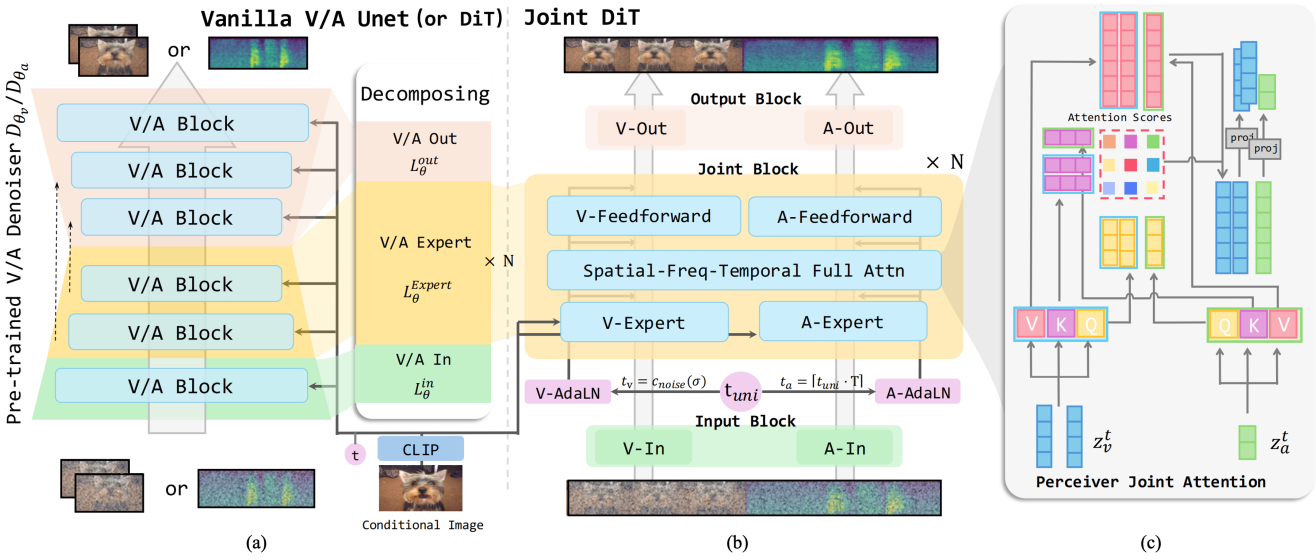

Figure 2. JointDiT. (a) Pretrained Unimodal Denoisers. JointDiT integrates pretrained video and audio denoisers by first decoupling them into specific submodules and then treats them as three types of layers for joint modeling: input layer, expert layer and output layer. (b) JointDiT Structure. The architecture consists of three types of blocks: *Input Block* contains two independent modality-specific input layers derived from the pretrained denoisers. *Joint Block* features a full attention layer to facilitate cross-modal interactions between video and audio. *Output Block* includes two modality-specific output layers to generate the final denoising results for each modality. (c) Implementation of Full Attention Layer in the Joint Block. The Perceiver Joint Attention mechanism is designed to handle the heterogeneous nature of video and audio data, effectively managing the significant channel dimension disparity between the two modalities while interaction.

or training framework), as the modality-specific Input, Expert, and Output layers. Based on previous works [53] that prove intermediate representations of DMs contain high-level semantics, we hypothesize that earlier layers of DMs focus on representation and understanding, while later layers are dedicated to decoding and generation. Based on this intuition, earlier (understanding) layers from D_{θ_v} and D_{θ_a} are used as Input and Expert layers, where cross-modal interactions (needs understanding) are introduced among these layers, while later (generation) layers are employed as Output layers for modality-specific decoding.

Figure 2(a) illustrates this with a six-layer UNet-based DM (each UNet block is treated as a layer in this paper). The first layer serves as the Input layer, the second and third layers act as Expert layers within the first and second Joint Blocks, and the fourth to sixth layers collectively form the Output layer (detailed layer setting in Supplementary A). JointDiT thus provides a framework for decoupling single-modality DMs and effectively integrating them into a unified joint modeling architecture.

Full Attention Layer. At full attention layer in Joint Block, JointDiT firstly flattens video $\boldsymbol{v}_t$ and audio $\boldsymbol{a}_t$ data into $\overline{\boldsymbol{v}}_t \in \mathbb{R}^{(T_v \times S_v) \times C_v}$ and $\overline{\boldsymbol{a}}_t \in \mathbb{R}^{(T_a \times S_a) \times C_a}$ and concatenates them into a sequence of $[\overline{\boldsymbol{v}}_t, \overline{\boldsymbol{a}}_t]$. For their interaction, due to the inherent heterogeneity and different redundancy of video and audio data (e.g., $\boldsymbol{v}_t$ has $\sim$200 times higher spatial dimension but $\sim$4 times lower temporal dimension compared to $\boldsymbol{a}_t$, and $\sim$2 times channels in our setting), a naive full attention mechanism—where both modalities are projected to the same channel dimension and vanilla self-attention is applied to the concatenated sequence—fails to effectively model cross-modal interactions, as evidenced by our experiments (Section 4.5).

To address this limitation, we introduce a balanced interaction mechanism, termed **Perceiver Joint Attention**, which is designed to accommodate modality-specific characteristics as shown in Figure 2(c). Inspired by prior work [19], we modify the shared Q (query) K (key) V (value) projections to be modality-specific. Separate QK projections for video and audio project $\boldsymbol{v}_t$ and $\boldsymbol{a}_t$ into the same QK dimension for Query-Key calculation, while separate V projections map them into a Value dimension distinct from QK. Finally, separate out-projections return them to their original channels after attention operation.

3.3. Training of JointDiT

JointDiT integrates the optimization objectives of two pretrained DMs: DDPM formulation in D_{θ_a} (AudioLDM [31]) and preconditioning EDM formulation in D_{θ_v} (SVD [3]), as detailed in Section 3.1. Our unified training strategy preserves both objectives while enabling joint optimization.

A key distinction between DDPM and EDM lies in their timestep parameterization: DDPM uses discrete timesteps $t \sim \mathcal{U}\{1, ..., 1000\}$, while EDM uses continuous timesteps

$t(\sigma) = 0.25 \log \sigma$ with $\log \sigma \sim \mathcal{N}(P_{\text{mean}}, P_{\text{std}}^2)$. To schedule noise across modalities, we introduce a unified timestep $t_{\text{uni}} \sim \mathcal{U}(0, 1)$ that controls both processes through:

$$t = \lceil 1000 t_{\text{uni}} \rceil, \qquad \text{for DDPM}, \quad (4)$$

$$\log \sigma = P_{\text{mean}} + P_{\text{std}} \cdot \Phi^{-1}(t_{\text{uni}}), \qquad \text{for EDM}, \quad (5)$$

where Φ^{-1} is the inverse cumulative distribution function of the standard normal distribution. This parameterization enables JointDiT to inherit the original training objectives of both base DMs (under any diffusion frameworks) with minimal adaptation, maintaining their pretrained knowledge. The joint optimization objective combines both losses through linear interpolation:

$$(\hat{\boldsymbol{v}}_0, \hat{\boldsymbol{a}}_0) = D_{\theta_{\text{joint}}}\left(\boldsymbol{v}_{t_{\text{uni}}}, \boldsymbol{a}_{t_{\text{uni}}}; i, t_{\text{uni}}\right), \quad (6)$$

$$\mathcal{L}(\theta_{\text{joint}}) = \mathbb{E}\left[\lambda_1 \|\boldsymbol{v}_0 - \hat{\boldsymbol{v}}_0\|^2 + \lambda_2 \|\boldsymbol{a}_0 - \hat{\boldsymbol{a}}_0\|^2\right], \quad (7)$$

where $\boldsymbol{v}_{t_{\text{uni}}}$ and $\boldsymbol{a}_{t_{\text{uni}}}$ are noisy data introduced in Equation 1 and 2 with t_{uni}, λ_1 and λ_2 control modality weighting. This formulation preserves the mathematical foundations of both diffusion frameworks while enabling efficient joint training through a shared timestep coordinate system.

3.4. Inference of JointDiT

For unimodal generation (e.g., videos), vanilla classifier-free guidance (CFG) operates as the following formulation:

$$\underbrace{D_{\theta_v}\left(\boldsymbol{v}_t; i, t\right)}_{\text{Target distribution}} + \omega \Big(\underbrace{D_{\theta_v}\left(\boldsymbol{v}_t; i, t\right) - D_{\theta_v}\left(\boldsymbol{v}_t; \varnothing, t\right)}_{\text{Guidance correction}} \Big), \quad (8)$$

where the first term estimates the conditional distribution, while the second term enhances sample quality by steering predictions toward conditional outputs and away from unconditional counterparts, serving as a *correcting term*. Extending this to joint generation, we substitute D_{θ_v} and $\boldsymbol{v}_t$ with $D_{\theta_{\text{joint}}}$ and $(\boldsymbol{v}_t, \boldsymbol{a}_t)$. Vanilla CFG correction becomes:

$$D_{\theta_{\text{joint}}}\left(\boldsymbol{v}_t, \boldsymbol{a}_t; i, t_{\text{uni}}\right) - D_{\theta_{\text{joint}}}\left(\boldsymbol{v}_t, \boldsymbol{a}_t; \varnothing, t_{\text{uni}}\right). \quad (9)$$

JointCFG and JointCFG*. While effective for condition alignment, the vanilla correction term above neglects modality-interaction enhancement. To address this, we redesign the negative term by simultaneously removing conditional inputs *and* modality interactions (through block-diagonal attention masking on full attention layer):

$$D_{\theta_{\text{joint}}}\left(\boldsymbol{v}_t, \boldsymbol{a}_t; i, t_{\text{uni}}\right) - \left[D_{\theta_{\text{joint}}}\left(\boldsymbol{v}_t; \varnothing, t_{\text{uni}}\right), D_{\theta_{\text{joint}}}\left(\boldsymbol{a}_t; \varnothing, t_{\text{uni}}\right)\right], \quad (10)$$

where the full-attn-masked model generates modality-isolated predictions. This redesigned correction term extends CFG to joint generation, introducing a joint guidance, namely **joint-classifier-free guidance (JointCFG)**. This mechanism simultaneously enforces condition adherence and cross-modal consistency, visualized in Figure 3.

Inspired by Autoguidance [21] (Section 3.1) that *bad version* DMs serve as effective guides, we further develop

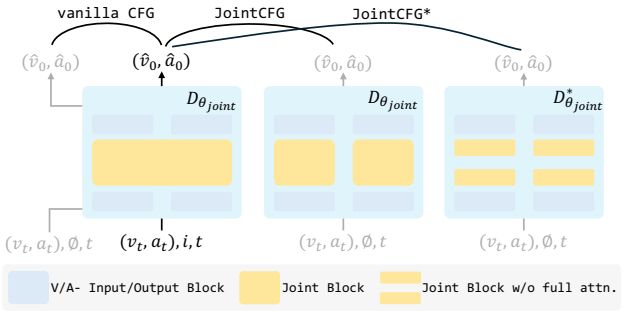

Figure 3. Comparison between refined Joint Classifier-Free Guidance (Joint-CFG*) and conventional CFG in the context of joint generation. We utilize *bad* versions of two unimodal generation results to boost the CFG performance in the joint generation.

modality-specific *bad* model $D^*_{\theta_{\text{joint}}}$ upon JointCFG by removing full attention layers in Joint Blocks. This partial model produces suboptimal single-modality predictions due to the lack of parameters for modality interaction. The enhanced correction term becomes:

$$D_{\theta_{\text{joint}}}\left(\boldsymbol{z}_v, \boldsymbol{z}_a; i, t_{\text{uni}}\right) - \left[D^*_{\theta_{\text{joint}}}\left(\boldsymbol{z}_v; \varnothing, t_{\text{uni}}\right), D^*_{\theta_{\text{joint}}}\left(\boldsymbol{z}_a; \varnothing, t_{\text{uni}}\right)\right]. \quad (11)$$

JointCFG with this correction term is named **JointCFG***. It further improves joint generation quality.

4. Experiments

4.1. Experimental Setup

Implementation Details. For video and audio representations $\boldsymbol{v}_0$ and $\boldsymbol{a}_0$, we utilize latents sampled from the Variational Autoencoder (VAE) from SVD [3] and AudioLDM2 [31], respectively. For processing conditional images: (1) V-Expert layer: Following SVD, the image is encoded using its VAE, duplicated to latent frame length, and concatenated with noisy video latents along the channel dimension. Additionally, CLIP [37] embeddings are extracted for cross-attention. (2) A-Expert layer: We modify AudioLDM2 by encoding the image using CLIP and add the image embeddings to the timestep embeddings. In the full-attention layer, without any conditional image input, this layer focuses on the interaction between the time sequences of the two modalities. We apply 3D and 2D rotary position embeddings (RoPE) [42] to $\boldsymbol{v}_t$ and $\boldsymbol{a}_t$, respectively, for QK computation in this layer. For integrating pretrained unimodal DMs, we preserve the original Input, Expert, and Output layer designs [3, 31] for simplified implementation.
Dataset. We establish the I2SV task using three benchmark datasets: AVSync15 [55], Landscape [25], and GreatestHits [35]. (1) AVSync15: Derived from VGGSound [5], this dataset features synchronized audio and video across 15 categories, with video durations ranging from 2 to 10 seconds. It includes 1350 training videos and 150 test videos.

GenProcess	Method	Video Quality		Audio Quality		VA Consistency		Condition Alignment	
		FVD↓	KVD↓	FAD↓	KL↓	IB-VA↑	AV-Align↓	IB-IV↑	IB-IA↑
$I \to V \parallel I \to A$	SVD[3] + IM2Wav[40]	444	37.7	34.9	3.08	12.8	1.866	87.0	13.6
	SVD[3] + AudioLDM[31]-v	444	37.7	30.3	2.58	24.9	1.352	87.0	24.7
$I \to V + V \to A$	SVD[3] + Diff-Foley[32]	444	37.7	32.2	3.31	19.8	1.693	87.0	18.8
	SVD[3] + FoleyCrafter[56]	444	37.7	31.6	2.19	29.8	1.792	87.0	28.7
	SVD[3] + TiVA[48]	444	37.7	35.5	3.70	17.4	1.834	87.0	17.1
	SVD[3] + SeeingHearing[49]	444	37.7	35.2	3.40	29.6	1.569	87.0	27.4
$I \to A + A^{\dagger} \to V$	IM2Wav[40] + AVSyncD[55]	645	64.4	34.9	3.08	12.9	1.352	83.2	12.7
	AudioLDM[31]-v + AVSyncD[55]	610	64.4	30.3	2.58	25.5	**1.285**	83.7	24.4
$I \to T \to V+$ $I \to T \to A$	Qwen.[1] + CogV.-2B[51] + SDA[10]	1172	157	32.3	3.29	24.8	1.970	68.6	22.1
	Qwen.[1] + CogV.-5B[51] + SDA[10]	1148	138	32.3	3.29	24.9	1.928	70.5	22.1
	Qwen.[1] + Hunyuan.[23] + SDA[10]	1094	140	32.3	3.29	26.0	2.155	70.2	22.1
$I \to VA$	CoDi[44]	1873	369.3	44.8	5.61	5.3	1.772	51.9	4.8
	JointDiT (ours)	**326**	**15.2**	**23.9**	**1.36**	**37.5**	1.296	**87.7**	**36.3**

Table 1. Automatic evaluation on the AVSync15 dataset, assessing video-audio quality, semantic and temporal consistency, and conditional alignment (with input image). **Best** and second-best results are highlighted. $A^{\dagger}$ indicates models requiring text input assistance despite audio. Abbreviations: SVD and SDA denote Stable Video Diffusion [3] and Stable Audio Open 1.0 [10], while Qwen., CogV., Hunyuan. denote Qwen-VL-7B [1], HunyuanVideo [23], CogVideoX [51] (1.0) . For fairness, only T2V mode (not TI2V) is used for HunyuanVideo and CogVideoX respectively. JointDiT achieves state-of-the-art results across all metrics, with competitive synchronization scores.

(2) Landscape: Comprising 928 videos across nine natural landscape categories. (3) Greatest Hits: Contains 977 videos of humans striking objects with drumsticks. Both AVSync15 and Landscape provide category labels in text, but no text information is used in any setting. Training and testing are conducted using the video's first frame as the conditional image and the full video as the ground truth. The task emphasizes the model's ability to understand static visual information and generate dynamic and synchronized visual and audio outputs, without relying on textual cues.

Baseline. We evaluate five I2SV methods: (1) $I \to V \parallel I \to A$: Video and audio are independently generated using Stable Video Diffusion (SVD) [3] for I2V and IM2Wav [40] or AudioLDM-v for I2A. AudioLDM-v is adapted from AudioLDM2 [31] by replacing text features with image features and fine-tuning on VGGSound. (2) $I \to V + V \to A$: Video is generated via SVD, followed by audio generation using V2A models (e.g., Diff-Foley [32], FoleyCrafter [56], TiVA [48], Seeing-and-Hearing [49]). Videos are repeated to match V2A requirements and then truncated to target length. (3) $I \to A + A \to V$: Audio is generated via IM2Wav or AudioLDM-v, then used with AVSyncD [55] for video generation, requiring audio, image, and text prompts as inputs. Dataset category labels serve as text prompts. (4) $I \to T \to V + I \to T \to A$: Captions are generated using vision large language models (e.g., Qwen-VL [1]), then used for T2V generation via HunyuanVideo [23] or CogVideoX [51], and T2A generation via Stable Audio Open [10]. This pipeline is evaluated on diverse AVSync15 scenarios. (5) $I \to VA$, i.e., I2SV: CoDi [44], a general any-to-any generation model, is tested for direct I2SV generation.

Evaluation Metrics. We assess the generated sounding videos across four perspectives: (1) *Video Quality*: We employ FVD [45] and KVD [45]metrics, commonly used in the video generation domain. (2) *Audio Quality*: Melception [17] computes FAD [22], while PasST [24] calculates the KL score [48]. (3) *Video-Audio Quality*: Semantic alignment is measured using ImageBind [11] (IB-VA), and temporal synchronization is evaluated with a refined AV-Align score [52] (details in Supplementary B). (4) *Condition Alignment*: ImageBind assesses the alignment of generated video or audio with the input image (IB-IV, IB-IA). Additionally, we employ Motion Score (MS) [3] in the ablation study to indicate the dynamics of the generated video.

Method	GreatestHits			LandScape		
	FVD↓	FAD↓	AV↓	FVD↓	FAD↓	AV↓
$I \to V \parallel I \to A$						
SVD[3] + A.LDM[31]-v	441	26.65	1.392	402	2.28	1.414
SVD[3] + IM2Wav[40]	441	29.14	1.886	402	1.61	2.005
$I \to V + V \to A$						
SVD[3] + Diff-Foley[32]	441	27.99	1.706	402	2.81	1.830
SVD[3] + FoleyCrafter[56]	441	29.26	1.936	402	**1.60**	1.996
SVD[3] + TiVA[48]	441	29.51	1.925	402	4.18	1.919
SVD[3] + SeeingHearing[49]	441	24.86	1.666	402	3.59	1.634
$I \to A + A^{\dagger} \to V$						
A.LDM[31]-v + AVSyncD[55]	287	26.65	1.219	697	2.28	1.196
IM2Wav[40] + AVSyncD[55]	320	29.14	1.777	707	1.61	1.256
$I \to VA$						
CoDi[44]	1314	24.97	**1.143**	1233	5.20	1.250
JointDiT	**173**	**1.08**	1.462	**262**	**1.60**	**1.178**

Table 2. Performance on GreatestHits and Landscape. AV denotes AV-Align metric here. **Best** and second-best are highlighted.

4.2. Comparison with Baselines

Table 1 presents a performance comparison between our model and the baseline models on the AVSync15 dataset. In objective evaluations, our model surpasses all baselines in terms of video quality, audio quality, and semantic alignment between video and audio, indicated by FVD, KVD, FAD, KL, and IB-VA metrics.

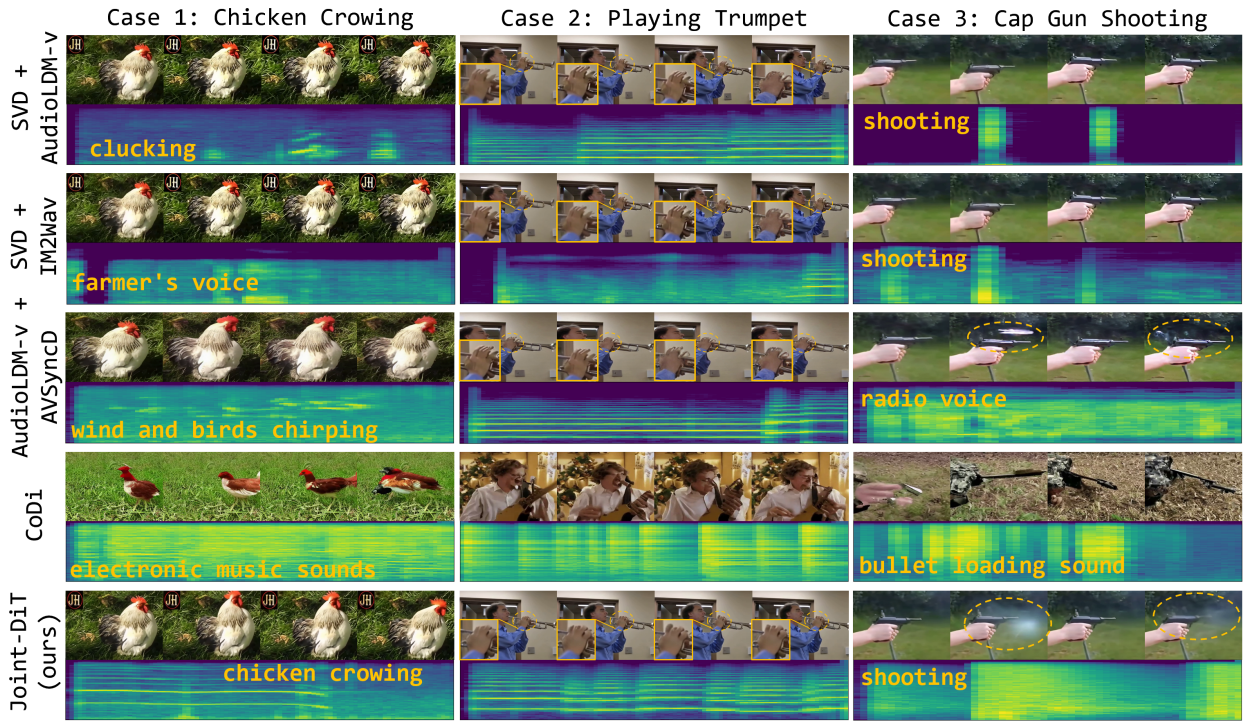

Figure 4. Qualitative Results. We present three instances from AVSync15, each generated by a different model. Instances produced by JointDiT exhibit high visual quality and dynamics, preserving details (e.g., finger movements in Case 2), and ensuring high correlation between audio and video in terms of semantics and temporal synchronization (e.g., rooster raising its head when crowing in Case 1 and cap gun firing when shooting in Case 3). More cases can be viewed at our demopage https://anonymoushub4ai.github.io/JointDiT.

This demonstrates that the generation of video and audio can mutually enhance each other, which also validates the effectiveness of the JointDiT framework. Only AV-Align metric indicates that JointDiT slightly underperforms the AudioLDM-v+AVSyncD model in the image-to-audio-to-video pipeline, achieving second place with a score of 1.296, slightly worse than 1.285. Notably, although HunyuanVideo and CogVideoX demonstrate stronger T2V capabilities, the process of converting images to text in the I2SV task results in significant information loss. This, in turn, makes it challenging for the second-stage T2V results to align with the original image information, as indicated by low FVD, KVD, and IB-IV scores. Table 2 shows different models' performance on the GreatestHits and Landscape dataset. JointDiT achieves the best results on every metric of LandScape dataset, and leads in FVD and FAD metrics of GreatestHits dataset. Comprehensive experiment results on multiple datasets validate that JointDiT is capable of generating high-quelity, semantics-aligned and synchronized video and audio.

4.3. User studies

We also conducted subjective evaluations on AVSync15 in Table 3. We evaluated from five dimensions: Video Quality (VQ), Audio Quality (AQ), Semantic Matching between Video and Audio (Sem), Synchronization between Video and Audio (Sync), and Overall Sounding Video Effect (Overall). The evaluation scores ranged from 0 to 5, with 0.5 as the scoring unit and higher scores indicate better results. We sample 100 results from each model with same condition and have them rated by 10 annotators, with the final score being the average of their ratings. The results of the subjective evaluations demonstrate that JointDiT outperforms other methods in all five dimensions. Interestingly, the the automatic metrics are not entirely consistent with human evaluation results. For example, for video-audio synchronization, our method achieved the best human evaluation result, with AudioLDM-v+AVSyncD coming second, but the objective metrics ranked them in reverse order. This further suggests that despite our improvements on objective metrics (Supplementary B), they

Method	User Study ↑				
	VQ	AQ	Con	Sync	Overall
SVD[3] + AudioLDM[31]-v	1.28	<u>1.33</u>	1.31	1.25	1.26
SVD[3] + SeeingHearing[49]	<u>1.30</u>	1.25	1.24	1.21	1.23
AudioLDM[31]-v + AVSyncD[55]	1.19	<u>1.33</u>	<u>1.36</u>	<u>1.30</u>	<u>1.28</u>
CoDi[44]	0.91	0.95	0.87	0.87	0.92
JointDiT	**1.48**	**1.55**	**1.60**	**1.51**	**1.52**

Table 3. User study results on AVSync15 dataset. We compare with baseline models from different I2SV generation approaches across: VQ (Video Quality), AQ (Audio Quality), Con (Semantic Consistency), Sync (Synchronization), and Overall (Composite Performance). Human ratings are averaged per metric, with **best** and <u>second-best</u> highlighted. JointDiT achieves superior performance across all evaluation dimensions.

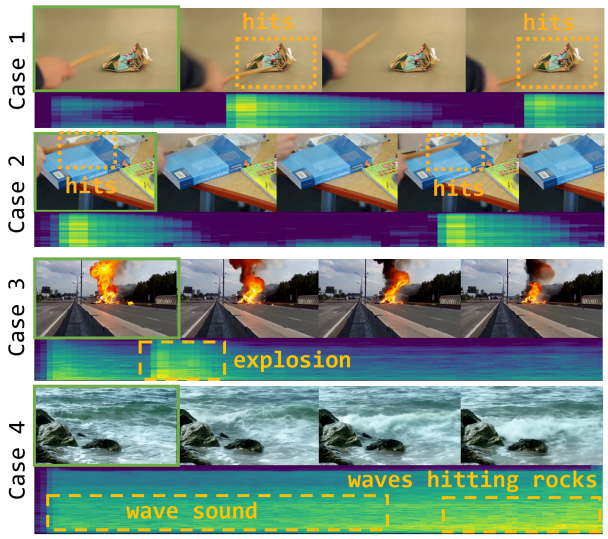

Figure 5. Case study on GreatestHits and Landscape datasets. JointDiT is capable of generating temporally and semantically consistent sounding videos across diverse I2SV scenarios.

still cannot perfectly align with human evaluations. We leave potential improvements to the automatic evaluation for future work.

4.4. Case study on Joint-DiT

Figure 4 illustrates a case comparison between JointDiT and the baseline models. The two types of methods, image-to-video + image-to-audio and image-to-video-to-audio, first generate videos from images. However, the video dynamics produced by SVD, a unimodal video generation model, often lack entity dynamic other than lens and perspective movings. For instance, for an input image of a chicken (case1) and a gun (case3), the SVD-generated results merely move the lens slightly and do not include entity actions such as the chicken raising its head or the gun shaking when fired. The pipeline method of I2A2V often leads to the accumulation of errors and confuses the model in the later stages of the pipeline. For example, in the third line of case 3, the I2A model generated a radio voice but the condition image was of a gun, causing the subsequent IA2V model to produce some unstable images, as shown in the dashed box. The early I2VA method, CoDi, due to its coarse-grained interaction, leads to poor overall generation quality. In contrast, the cases generated by our JointDiT have good visual quality and dynamics, and also maintain details (finger movements in case 2), and can generate sound that matches the video semantics and is synchronized (case 1 crowing when the chicken raises its head, case 3 gun firing when smoke appears). In addition to the general generation performance of AVSync15, Figure 5 further presents the generation results of JointDiT on the GreatestHits and Landscape datasets. The results demonstrate that JointDiT is capable of effectively handling data from specific domains, such as natural scenery near ocean waves and human actions while striking objects. Moreover, it can produce coherent and natural paired video and audio contents that meets with physical laws within the contexts provided by the first frame of a scene.

JointDiT	FVD↓	FAD↓	IB-VA↑	AV-Align↓	MS→
w/ JointCFG*	**653.7**	**27.6**	**37.4**	1.288	10.73
w/ JointCFG	661.6	28.1	36.9	1.273	10.15
w/ vanilla CFG	662.8	28.2	36.1	1.266	10.19
w/o Joint Block	819.5	30.1	36.2	1.363	5.32
w/o Perceiver Attn	921.2	29.3	34.4	**1.240**	5.54

Table 4. Ablation studies on guidance methods and architecture, evaluated on a randomly sampled subset of AVSync15 test for efficiency. MS represents motion scores. The *w/o Joint Block* setting refers to removing the Joint Block from JointDiT, resulting in the fine-tuned setting of two pretrained DMs. The *w/o perceiver attn* setting replaces perceiver joint attention with standard self-attention. These two architecture-ablation settings are inferenced with vanilla CFG as *w/o Joint Block* not applicable with JointCFG.

4.5. Ablation Studies

Guidance Techniques. Table 4 compares three CFG techniques: Enhanced JointCFG (w/ joint-CFG*), vanilla CFG and JointCFG. The table reveals that the enhanced JointCFG* significantly improves image quality, sound quality, and IB-VA, albeit at a slight cost to AV-Align metrics. However, it also consistently enhances the motion score, introducing more dynamic changes to the video.
JiontDiT Architecture. In the bottom part of Table 4, we conducted an ablation study on the JointDiT architecture, comparing three settings: our final version (w/ joint-CFG*), a version without the Joint Block trained on AVSync15 (essentially fine-tuning independent video and audio denoisers on AVSync15), and a version substituting perceiver full attention with vanilla full attention. Tests are performed on the same randomly selected 75 images. The results demonstrate that our final version achieves the highest image quality, audio quality, and IB-VA audio-video semantic matching. Interestingly, our architecture introduces more dynamic changes, as indicated by the motion score, addressing the challenge of lacking of dynamic variation in the image-to-video task [57]. This suggests that a well-designed interaction facilitates more visual dynamics in the joint generation of video and audio.

5. Conclusion and Future Work

To solve the task of Image-to-Sounding Video (I2SV) generation, we propose a novel approach using the DiT architecture to model the interaction between video and audio modalities. We combine the expert pretrained diffusion generation submodules using a unified Joint Block. We further introduce introduce a joint classifier-free guidance technique to refine the guidance effect.

In future work, we plan to extend JointDiT to joint or cross-generation tasks involving the text modality. We also intend to further investigate the relationships among the three modalities and explore the scalability of JointDiT using larger-scale datasets across additional modalities, aiming to develop a more foundational text-video-audio model towards the world model.

Acknowledgement

This work is supported by the National Natural Science Foundation of China (No. 62276268), Public Computing Cloud, Renmin University of China and ZHI-TECH GROUP.

References

[1] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023. 6

[2] Fan Bao, Shen Nie, Kaiwen Xue, Chongxuan Li, Shi Pu, Yaole Wang, Gang Yue, Yue Cao, Hang Su, and Jun Zhu. One transformer fits all distributions in multi-modal diffusion at scale. In *ICML*, 2023. 2

[3] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023. 1, 2, 3, 4, 5, 6, 7

[4] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators, 2024. 2

[5] Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. Vggsound: A large-scale audio-visual dataset. In *ICASSP*, 2020. 5

[6] Shoufa Chen, Mengmeng Xu, Jiawei Ren, Yuren Cong, Sen He, Yanping Xie, Animesh Sinha, Ping Luo, Tao Xiang, and Juan-Manuel Perez-Rua. Gentron: Diffusion transformers for image and video generation. In *CVPR*, 2024. 2

[7] Zeyuan Chen, Hongyi Xu, Guoxian Song, You Xie, Chenxu Zhang, Xin Chen, Chao Wang, Di Chang, and Linjie Luo. X-dancer: Expressive music to human dance video generation. *arXiv preprint arXiv:2502.17414*, 2025. 2

[8] Xin Cheng, Xihua Wang, Yihan Wu, Yuyue Wang, and Ruihua Song. Lova: Long-form video-to-audio generation. In *ICASSP*, 2025. 2

[9] Zach Evans, CJ Carr, Josiah Taylor, Scott H. Hawley, and Jordi Pons. Fast timing-conditioned latent audio diffusion. In *ICML*, 2024. 2

[10] Zach Evans, Julian D Parker, CJ Carr, Zack Zukowski, Josiah Taylor, and Jordi Pons. Stable audio open. In *ICASSP*, 2025. 1, 2, 6

[11] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *CVPR*, 2023. 6

[12] Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. In *ICLR*, 2024. 2

[13] Akio Hayakawa, Masato Ishii, Takashi Shibuya, and Yuki Mitsufuji. Discriminator-guided cooperative diffusion for joint audio and video generation. *arXiv preprint arXiv:2405.17842*, 2024. 1, 3

[14] Xu He, Qiaochu Huang, Zhensong Zhang, Zhiwei Lin, Zhiyong Wu, Sicheng Yang, Minglei Li, Zhiyi Chen, Songcen Xu, and Xiaofei Wu. Co-speech gesture video generation via motion-decoupled diffusion model. In *CVPR*, 2024. 2

[15] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 2, 3

[16] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020. 1

[17] Vladimir Iashin and Esa Rahtu. Taming visually guided sound generation. In *BMVC*, 2021. 6

[18] Masato Ishii, Akio Hayakawa, Takashi Shibuya, and Yuki Mitsufuji. A simple but strong baseline for sounding video generation: Effective adaptation of audio and video diffusion models for joint generation. *arXiv preprint arXiv:2409.17550*, 2024. 1, 3

[19] Andrew Jaegle, Sebastian Borgeaud, Jean-Baptiste Alayrac, Carl Doersch, Catalin Ionescu, David Ding, Skanda Koppula, Daniel Zoran, Andrew Brock, Evan Shelhamer, Olivier J. Hénaff, Matthew M. Botvinick, Andrew Zisserman, Oriol Vinyals, and João Carreira. Perceiver IO: A general architecture for structured inputs & outputs. In *ICLR*, 2022. 4

[20] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In *NeurIPS*, 2022. 3

[21] Tero Karras, Miika Aittala, Tuomas Kynkäänniemi, Jaakko Lehtinen, Timo Aila, and Samuli Laine. Guiding a diffusion model with a bad version of itself. In *NeurIPS*, 2024. 3, 5

[22] Kevin Kilgour, Mauricio Zuluaga, Dominik Roblek, and Matthew Sharifi. Fréchet audio distance: A reference-free metric for evaluating music enhancement algorithms. In *Interspeech*, 2019. 6

[23] Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024. 1, 2, 6

[24] Khaled Koutini, Jan Schlüter, Hamid Eghbal-zadeh, and Gerhard Widmer. Efficient training of audio transformers with patchout. In *Interspeech*, 2022. 6

[25] 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*, 2022. 1, 2, 5

[26] Ruilong Li, Shan Yang, David A. Ross, and Angjoo Kanazawa. AI choreographer: Music conditioned 3d dance generation with AIST++. In *ICCV*, 2021. 1

[27] Tianhong Li, Dina Katabi, and Kaiming He. Return of unconditional generation: A self-supervised representation generation method. In *NeurIPS*, 2024. 1

[28] Zongyu Lin, Wei Liu, Chen Chen, Jiasen Lu, Wenze Hu, Tsu-Jui Fu, Jesse Allardice, Zhengfeng Lai, Liangchen Song, Bowen Zhang, et al. Stiv: Scalable text and image conditioned video generation. *arXiv preprint arXiv:2412.07730*, 2024. 2

[29] Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *ICLR*, 2023. 1

[30] Haohe Liu, Zehua Chen, Yi Yuan, Xinhao Mei, Xubo Liu, Danilo P. Mandic, Wenwu Wang, and Mark D. Plumbley. Audioldm: Text-to-audio generation with latent diffusion models. In *ICML*, 2023. 2

[31] Haohe Liu, Yi Yuan, Xubo Liu, Xinhao Mei, Qiuqiang Kong, Qiao Tian, Yuping Wang, Wenwu Wang, Yuxuan Wang, and Mark D. Plumbley. Audioldm 2: Learning holistic audio generation with self-supervised pretraining. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:2871–2883, 2024. 2, 4, 5, 6, 7

[32] Simian Luo, Chuanhao Yan, Chenxu Hu, and Hang Zhao. Diff-foley: Synchronized video-to-audio synthesis with latent diffusion models. In *NeurIPS*, 2023. 1, 2, 6

[33] Yuxin Mao, Xuyang Shen, Jing Zhang, Zhen Qin, Jinxing Zhou, Mochu Xiang, Yiran Zhong, and Yuchao Dai. Tavgbench: Benchmarking text to audible-video generation. In *ACM MM*, 2024. 1, 3

[34] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *ICML*, pages 8162–8171, 2021. 1, 2, 3

[35] Andrew Owens, Phillip Isola, Josh McDermott, Antonio Torralba, Edward H Adelson, and William T Freeman. Visually indicated sounds. In *CVPR*, pages 2405–2413, 2016. 5

[36] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *ICCV*, 2023. 1, 2, 3

[37] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 5

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

[39] Ludan Ruan, Yiyang Ma, Huan Yang, Huiguo He, Bei Liu, Jianlong Fu, Nicholas Jing Yuan, Qin Jin, and Baining Guo. Mm-diffusion: Learning multi-modal diffusion models for joint audio and video generation. In *CVPR*, 2023. 1, 3

[40] Roy Sheffer and Yossi Adi. I hear your true colors: Image guided audio generation. In *ICASSP*, 2023. 2, 6

[41] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, Devi Parikh, Sonal Gupta, and Yaniv Taigman. Make-a-video: Text-to-video generation without text-video data. In *ICLR*, 2023. 2

[42] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024. 5

[43] Mingzhen Sun, Weining Wang, Yanyuan Qiao, Jiahui Sun, Zihan Qin, Longteng Guo, Xinxin Zhu, and Jing Liu. MM-LDM: multi-modal latent diffusion model for sounding video generation. In *ACM MM*, 2024. 1, 3

[44] Zineng Tang, Ziyi Yang, Chenguang Zhu, Michael Zeng, and Mohit Bansal. Any-to-any generation via composable diffusion. In *NeurIPS*, 2023. 1, 2, 6, 7

[45] Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018. 6

[46] Apoorv Vyas, Bowen Shi, Matthew Le, Andros Tjandra, Yi-Chiao Wu, Baishan Guo, Jiemin Zhang, Xinyue Zhang, Robert Adkins, William Ngan, et al. Audiobox: Unified audio generation with natural language prompts. *arXiv preprint arXiv:2312.15821*, 2023. 2

[47] Kai Wang, Shijian Deng, Jing Shi, Dimitrios Hatzinakos, and Yapeng Tian. Av-dit: Efficient audio-visual diffusion transformer for joint audio and video generation. *arXiv preprint arXiv:2406.07686*, 2024. 1, 2, 3

[48] Xihua Wang, Yuyue Wang, Yihan Wu, Ruihua Song, Xu Tan, Zehua Chen, Hongteng Xu, and Guodong Sui. Tiva: Time-aligned video-to-audio generation. In *ACM MM*, 2024. 2, 6

[49] Yazhou Xing, Yingqing He, Zeyue Tian, Xintao Wang, and Qifeng Chen. Seeing and hearing: Open-domain visual-audio generation with diffusion latent aligners. In *CVPR*, 2024. 1, 3, 6, 7

[50] Dongchao Yang, Jianwei Yu, Helin Wang, Wen Wang, Chao Weng, Yuexian Zou, and Dong Yu. Diffsound: Discrete diffusion model for text-to-sound generation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:1720–1733, 2023. 2

[51] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024. 1, 2, 6

[52] Guy Yariv, Itai Gat, Sagie Benaim, Lior Wolf, Idan Schwartz, and Yossi Adi. Diverse and aligned audio-to-video generation via text-to-video model adaptation. In *AAAI*, 2024. 2, 6

[53] Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin, and Saining Xie. Representation alignment for generation: Training diffusion transformers is easier than you think. *arXiv preprint arXiv:2410.06940*, 2024. 4

[54] Yan Zeng, Guoqiang Wei, Jiani Zheng, Jiaxin Zou, Yang Wei, Yuchen Zhang, and Hang Li. Make pixels dance: High-dynamic video generation. In *CVPR*, 2024. 2

[55] Lin Zhang, Shentong Mo, Yijing Zhang, and Pedro Morgado. Audio-synchronized visual animation. In *ECCV*, 2024. 2, 5, 6, 7

[56] Yiming Zhang, Yicheng Gu, Yanhong Zeng, Zhening Xing, Yuancheng Wang, Zhizheng Wu, and Kai Chen. Foleycrafter: Bring silent videos to life with lifelike and synchronized sounds. *arXiv preprint arXiv:2407.01494*, 2024. 2, 6

[57] Min Zhao, Hongzhou Zhu, Chendong Xiang, Kaiwen Zheng, Chongxuan Li, and Jun Zhu. Identifying and solving conditional image leakage in image-to-video diffusion model. In *NeurIPS*, 2024. 8

[58] Zangwei Zheng, Xiangyu Peng, Tianji Yang, Chenhui Shen, Shenggui Li, Hongxin Liu, Yukun Zhou, Tianyi Li, and Yang You. Open-sora: Democratizing efficient video production for all, 2024. 2