Audio-Synchronized Visual Animation

Lin Zhang[1], Shentong Mo[2], Yijing Zhang[1], and Pedro Morgado[1]

[1] University of Wisconsin-Madison
[2] Carnegie Mellon University

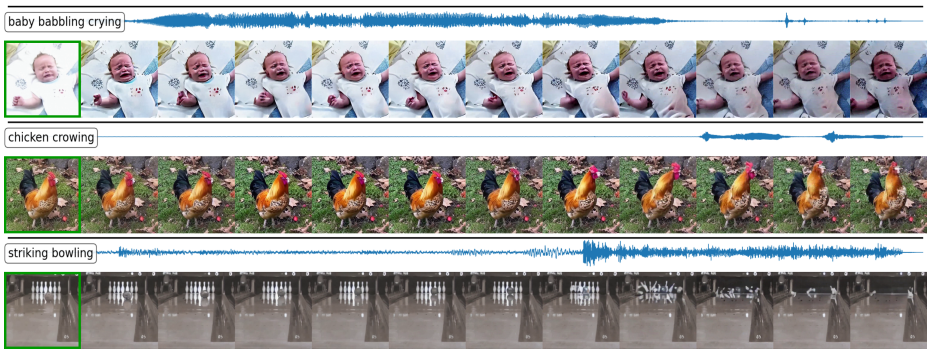


Fig. 1: Given an audio and an image (green box), we produce animations beyond image stylization with complex but natural dynamics, synchronized with audio at each frame. Results were produced by our AVSyncD model trained on the proposed AVSync15. Project webpage: `https://lzhangbj.github.io/projects/asva/asva.html`.

Abstract. Current visual generation methods can produce high-quality videos guided by text prompts. However, effectively controlling object dynamics remains a challenge. This work explores audio as a cue to generate temporally synchronized image animations. We introduce Audio-Synchronized Visual Animation (ASVA), a task that aims to animate a static image of an object with motions temporally guided by audio clips. To this end, we present AVSync15, a dataset curated from VGGSound with videos featuring synchronized audio-visual events across 15 categories. We also present a diffusion model, AVSyncD, capable of generating audio-guided animations. Extensive evaluations validate AVSync15 as a reliable benchmark for synchronized generation and demonstrate our model's superior performance. We further explore AVSyncD's potential in a variety of audio-synchronized generation tasks, from generating full videos without a base image to controlling object motions with various sounds. We hope our established benchmark can open new avenues for controllable visual generation.

Keywords: Audio · Video · Synchronization · Generation

1 Introduction

Generative modeling has witnessed remarkable progress recently, largely due to the emergence of diffusion models [17,34,38]. Conditional generation and, in par-

ticular, text-to-image generation [33,34], has been the focal point given its application potential and availability of high-quality datasets [1,36]. This success has also led to revived interest in video generation, such as text-to-video [5,21,37,45]. While text guidance has been thoroughly investigated, the unique advantages of audio-visual synchrony for video generation remain underexplored. Unlike text, which provides control over global semantics, audio offers both semantic control on videos and precise control at each moment in time.

Most existing works on audio-to-visual generation are however either limited to semantic control [20, 22, 39, 40], or constrained on singular scenarios such as human talking face [9–11, 28, 30, 49, 51, 52]. The former focuses on ambient audio datasets lacking synchronization cues with environmental sounds (ranining, fire crakcling, wind) [22] where shifting the audio temporally does not lead to major changes in visual content. The input audio in such cases thus substitutes text to provide only global semantics. The latter, talking faces, while synchronized, are extremely limited in generation diversity and control.

To bridge this gap, we introduce Audio-Synchronized Visual Animation, ASVA, a task which aims to animate objects depicted in natural static images into a video, with clear motion dynamics that are semantically aligned and temporally synchronized with the input audio. ASVA requires a sophisticated understanding of the audio's temporal structure, as well as of how objects move in synchrony with sound. Prior attempts on visual generation guided by *diverse sounds* fall short in generated visual quality [23] and accurate synchronization control [20, 48] due to two challenges: (1) the lack of high-quality training data and benchmarks for learning audio-synchronized visual dynamics; (2) the development of effective methods capable of generating highly synchronized video motions. Successfully addressing these challenges will expand the scope of current video generation methods to enable more fine-grained semantic and temporal control over the generation process via synchronized audio conditioning.

We address the first challenge by constructing a high-quality diverse dataset with strong correlations between audio and object motions at each moment in time. In an ideal dataset, sound sources should be easily identifiable in the scene. Every visual motion in the video should highly correlate with the audio semantically and temporally, and vice versa. Moreover, the visual content should be of high quality for generation. However, existing audio-visual datasets either are too noisy [7, 13], containing a large number of unassociated audio-visual pairs [31], or predominantly featuring ambient sound categories that lack meaningful synchronized object dynamics [22, 41]. We thus curate a high-quality dataset from VGGSound [7] by deploying an efficient two-step curation pipeline. In the first step, we use a variety of signal processing techniques and foundation models to automatically filter out videos with poor semantic alignment or temporal synchronization, as well as those depicting static scenes or with too fast camera motions. Then, to ensure the highest possible quality as a benchmark, we narrow down the dataset to sound categories with strong audio-visual synchronization cues, and manually verify the quality of each video. We end up obtaining AVSync15 with 15 dynamic sound classes, each with 100 examples rich

and accurate in semantics, object dynamics, and audio-visual synchronization. An overview of AVSync15 classes is in Figs. 2a and 2b and its comparison with prior audio-video generation datasets is in Suppl. Sec. 5.1.

The second challenge pertains to the generation of audio-synchronized motions, which requires a detailed understanding of audio-visual correlations and object dynamics. Take, for instance, a video featuring a dog. To generate realistic video, the model is expected to not only synchronize the dog's mouth with the barking sound, but also accurately depict subtleties in the dog's head pose before and after barking. Furthermore, in a more challenging scenario, the model should discern which object to animate to preserve semantic consistency depending on the input sound. However, existing audio-conditioned visual generation frameworks [14, 39] primarily focus on semantic control, often encoding the audio into a single global semantic feature and thereby neglecting the audio's temporal domain. Even recent attempts at audio-synchronized video generation [20, 23, 48] have not fully realized the potential of audio for fine-grained temporal control, as they either rely on crude audio representations such as audio amplitude [23], learn from weakly synchronized [22], noisy datasets [7], or ignore object dynamics in the generation process [20, 22, 23, 48]. To this end, we introduce Audio-Video Synchronized Diffusion (AVSyncD), a framework improving a pre-trained image latent diffusion model [34] for enhanced audio guidance and motion generation. We employ the pre-trained ImageBind [14] encoder to encode audio into time-aware semantic tokens, then fuse them into each frame's latent features. This allows for precise audio guidance on video semantics and synchronization. To capture complex video motions, we add temporal attention layers to the diffusion model. Finally, to ensure faithful animation of the input image, we incorporate temporal convolutions and attention layers that always reference the input image, i.e., first-frame lookups.

With the carefully designed dataset and architecture, we are able to train a model specialized for ASVA and produce animations with more realistic and audio-synchronized contents than prior works (Figs. 1 and 4). We provide thorough experiments to validate the effectiveness of AVSync15 and AVSyncD and demonstrate how to deploy AVSyncD for controllable generation, including amplifying audio guidance and semantic-aware object animation (Secs. 5.3 and 5.4).

2 Related Work

2.1 Controllable Visual Generation

Many conditional visual generation models based on diffusion process [17, 38] have emerged recently. Benefiting from more efficient architecture, large-scale training data [36], and aligned semantic space [32], Latent Diffusion Model (LDM) [34] has achieved great success to generate realistic images conditioned on text. This inspired researchers to explore various diffusion-based visual generation tasks, such as text-to-video [2, 5, 21, 44–46], audio-to-image [14, 39], and audio-to-video [20, 23, 40, 48]. The architectures can be training-free [21, 23, 45], fully-trained [5, 44], or trained partially, which augments a pre-trained LDM by

carefully adding some trainable layers [2, 20, 46]. Extensive works have also attempted to control the semantics of the generated content [15, 25, 27, 50], while how to apply control in the temporal dimension remains under-explored.

In this work, we developed an image animation model AVSyncD to control generation *semantically* and *temporally* guided by audio. AVSyncD augments pre-trained StableDiffusion [34] with trainable temporal layers and audio conditioning mechanism, preserving training efficiency and generalizing well.

2.2 Audio-to-Video Generation

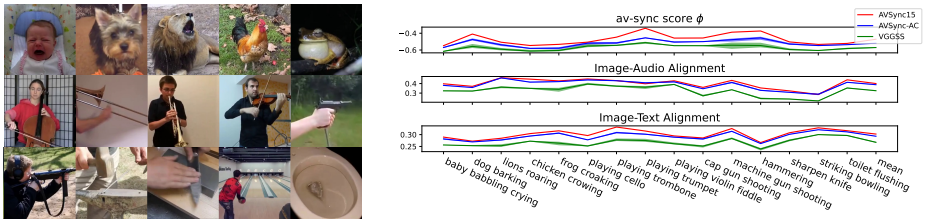
Traditionally, audio has been used as a temporal cue for talking face generation [28, 30, 49, 51, 52], where face and lip actions should be synchronized with audio at each frame. Many works also rely on complex inputs such as 3D meshes and human poses [28,49]. Although accurately synchronized, this line of research is extremely limited in scenarios and cannot generate videos for diverse audios.

A series of works attempted to expand the class of audio by encoding sound into a global semantic condition for video generation [14, 39, 40], however often overlooked the temporal aspect inherent in audio. Some recent works, although divided audio features into time-aware segments as inputs [22,35], failed to achieve promising visual quality or synchronize video motions with audios. AADiff [23] is a training-free method re-weighting the text-image cross-attention map in LDM using audio amplitude at each frame, however can only control styles of each frame. TPoS [20] learns segmented audio features aligned with CLIP [32] using sophisticated modules and training losses, and feeds them into a pre-trained text-to-image model [34] for video generation. TempoToken [48] also learns segmented audio features with a pre-trained audio encoder BEATs [8], and fuses them into a pre-trained text-to-video model [44]. However, primarily focused on monotonous sound classes in Landscapes [22] or noisy audio-visual data in VGGSound [7], these methods are limited to generating video semantics without capturing the natural and synchronized dynamics of video content. Frozen generation architectures also prevent them from generating natural motions.

To address these limitations, we introduce AVSync15, a high-quality dataset specifically designed for ASVA. AVSync15 stands out from previous efforts by focusing on synchronization cues between audio and visual dynamics, allowing for generating motions beyond mere visual effects. Once trained on AVSync15, our AVSyncD can generalize to many applications to control video motions guided by audios, on which previous methods performed poorly.

3 Audio-Synchronized Visual Animation

Formally, the Audio-Synchronized Visual Animation (ASVA) task can be posed as follows. Given an audio clip $\boldsymbol{a}$ of length T seconds and an image $\boldsymbol{x}_1$, the goal is to generate the future video sequence $\boldsymbol{x}_2, \ldots, \boldsymbol{x}_{rT}$ (or $\boldsymbol{x}_{2:rT}$ for short), where r is the desired frame rate. Despite the simple formulation, this is a challenging task as the generated video sequence should (1) be of high visual quality, (2)



(a) (b)

Fig. 2: (a): Overview of 15 categories in AVSync15. Categories are listed below x-axis on the right plot. (b): Categoriy-wise averages of av-sync score ϕ, IA, and IT on AVSync15 and equivalently sized subsets of VGGSS and AVSync-AC. Error bars for VGGSS and AVSync-AC are obtained from 3 random splits.

be semantically aligned with the image $\boldsymbol{x}_1$ and audio $\boldsymbol{a}$, (3) exhibit temporal coherence and (4) natural object motions temporally synchronized with the audio $\boldsymbol{a}$. To facilitate research in ASVA, we introduce a new benchmark that includes a curated high-quality dataset and a suite of evaluation metrics designed to capture the various components of audio-synchronized generation.

3.1 AVSync15: A High-Quality Dataset for Audio Synchronized Video Generation

Existing large-scale audio-visual datasets like VGGSound [6,7,19] and AudioSet [13] often contain amateur videos from platforms like YouTube. These videos, while diverse, can pose challenges for audio-synchronized video generation tasks due to rapid scene changes, camera motion, noisy audio, or out-of-frame sound sources. Prior work [22, 24] has addressed this by focusing on simpler videos, such as those depicting fire crackling or weather patterns. However, such videos often lack strong synchronization cues between audio and visual motion, making them unsuitable for ASVA. To facilitate research in this area, we introduce a high-quality dataset specifically designed for audio-synchronized video generation, ensuring a close synchronization between audio and visuals. More specifically, our selection criteria to create the dataset were: (1) *High Correlation*: Significant visual changes should be closely associated with audio at each timestamp, and vice versa. (2) *Dynamic Content*: We included content rich in temporal changes, excluding ambient or monotonous classes. (3) *Quality and Relevance*: Both video and audio needed to be clean, stable, and semantically aligned.

Preliminary Curation We start from VGGSound [7], a large-scale dataset with 309 diverse audio classes. Similar to VGGSoundSync [6], we first narrow down to 149 classes with potentially clear audio-visual synchronized events, removing ambient classes, which is referred as VGGSS. We then deploy a sequence of automatic cleaning steps and a final manual selection step to identify appropriate videos. The procedures are summarized below and detailed in Suppl. Sec. 5.2.

Automatic Curation We first use PySceneDetect [4] to cut videos with sharp scene changes into different scenes, which are still likely to contain low-quality

short clips. To maximize usage, we split each scene into 3-second clips with 0.5-second strides, and discard unsuitable clips based on the following metrics:

Raw Pixel Difference: We calculate average pixel distances between consecutive frames and remove clips with both small and large values, likely depicting either static or videos with excessive motion.
Image Semantics Difference: Complementing the raw pixel analysis above, we also compute distances on CLIP [32] image features, removing videos with small semantics changes such as zoom in/out or large semantic content transitions.
Audio Waveform Amplitude: We exclude clips whose maximum audio waveform amplitude is low, indicating weak audio cues.
Semantic Alignment: We compute the average image-audio (IA) and image-text (IT) alignment scores [32] in a video using ImageBind [14], removing clips with low scores to ensure cross-modal semantic alignment.
Audio-Video Synchronization: To measure audio-visual synchronization, we follow VGGSoundSync [6] to contrastively train an audio-visual synchronization classifier on VGGSS, ending up with a comparable 40.85% test accuracy. The model outputs an unbounded av-sync score $\phi_{\boldsymbol{a}_i,\boldsymbol{v}_j}$ for an audio-video pair $(\boldsymbol{a}_i, \boldsymbol{v}_j)$. During training, we compute ϕ for a synchronized pair $(\boldsymbol{a}_i, \boldsymbol{v}_i)$ and its temporally-shifted pairs from the same instance. Contrastive loss is then applied to these shifted pairs to maximize the synchronization probability:

$$P_{\text{Sync}}(\boldsymbol{a}_i, \boldsymbol{v}_i) = \frac{\exp(\phi_{\boldsymbol{a}_i,\boldsymbol{v}_i}/\tau)}{\sum_j \exp(\phi_{\boldsymbol{a}_i,\boldsymbol{v}_j}/\tau)} \tag{1}$$

to distinguish the synchronized pair from shifted ones. We use P_{Sync} as a synchronization indicator to remove low-scoring clips. When computing P_{sync}, we discard the temperature parameter τ used to improve training efficacy. We detail the synchronization classifier and P_{sync} in Suppl. Sec. 1 and Sec. 2.1, respectively.

We empirically determine metrics' thresholds by prioritizing quality, acknowledging that some acceptable clips might be discarded. After automatic curation, we merge all continuous 3-second clips from each video and remove categories with less than 100 examples to avoid class imbalance, resulting in AVSync-AC (AVSync w/ Automatic Curation) with 76 categories and 39,902 examples.

Manual Curation We further select 15 diverse categories with clear audio-visual motion cues from AVSync-AC for manual refinement. The categories range from animals and human actions to triggered tools and musical instruments. Manual curation once again seeks to identify appropriate videos for ASVA with the criteria above: high correlation, dynamic content, quality and relevance.

Dataset Summary The final dataset, AVSync15, contains 90 training and 10 testing videos per category, each 2∼10 seconds long. We provide an overview of AVSync15 in Fig. 2a. To validate our curation pipeline, we randomly sample three 1500-video splits on the selected 15 categories from VGGSS and AVSync-AC, and quantitatively compare them with AVSync15 in Fig. 2b and Tab. 2b. We also compare AVSync15 with other audio-visual datasets in Suppl. Sec. 5.1.

3.2 Evaluation Metrics

ASVA is a multi-faceted generation task, necessitating high quality at both image and video level. At the image level, we follow previous conventions [3, 12] to use (1) Fréchet Inception Distance (**FID**) [16] to measure the quality of individual frames; (2) **IA** [14]/**IT** [47] to measure image-audio/image-text semantics alignment on CLIP [32] space. At the video level, we use Fréchet Video Distance (**FVD**) [42] to assess video quality. To measure audio-video synchronization, we compute the following metrics with the trained synchronization classifier:

RelSync During testing, we use the ground truth audio-visual pair $(\boldsymbol{a}, \boldsymbol{v})$ as a reference to measure synchronization of the generated video $\hat{\boldsymbol{v}}$ as follows:

$$\text{RelSync}(\boldsymbol{a}, \hat{\boldsymbol{v}}, \boldsymbol{v}) = \frac{\exp(\phi_{\boldsymbol{a},\hat{\boldsymbol{v}}})}{\exp(\phi_{\boldsymbol{a},\boldsymbol{v}}) + \exp(\phi_{\boldsymbol{a},\hat{\boldsymbol{v}}})} \tag{2}$$

Note that while this reference-based metric normalizes the score by the synchronization of the reference pair $(\boldsymbol{a}, \boldsymbol{v})$, the metric can still be sensitive to the quality of the reference pair. In fact, evaluating synchronization on a dataset where even ground-truth audios and videos are ambiguously synchronized is less informative of the model capabilities, e.g., Landscapes [22].

AlignSync The synchronization classifier is only trained on semantically-aligned and temporally-shifted audio-video pairs sampled from the same instance (See Suppl. Sec. 1.2). RelSync is thus implicitly conditioned on semantics alignment, i.e., P(Sync|Align). To jointly measure semantics alignment and synchronization, we first approximate P_{Align} similarly as RelSync:

$$P_{\text{Align}}(\boldsymbol{a}, \hat{\boldsymbol{v}}, \boldsymbol{v}_1) = \frac{1}{b-1} \sum_{i=2\ldots b} \frac{\exp(\text{IA}_{\boldsymbol{a},\hat{\boldsymbol{v}}_i})}{\exp(\text{IA}_{\boldsymbol{a},\hat{\boldsymbol{v}}_i}) + \exp(\text{IA}_{\boldsymbol{a},\boldsymbol{v}_1})} \tag{3}$$

where $\boldsymbol{v}_1$ is the first frame input for animation, and b is the number of generated frames. The generated first frame $\hat{\boldsymbol{v}}_1$ is eliminated because it is often a replicate of input $\boldsymbol{v}_1$. By multiplying P_{Align} with RelSync, we obtain the joint score:

$$\text{AlignSync}(\boldsymbol{a}, \hat{\boldsymbol{v}}, \boldsymbol{v}) = P_{\text{Align}}(\boldsymbol{a}, \hat{\boldsymbol{v}}, \boldsymbol{v}_1) \cdot \text{RelSync}(\boldsymbol{a}, \hat{\boldsymbol{v}}, \boldsymbol{v}) \tag{4}$$

which jointly measures semantic alignment and temporal synchronization between $\boldsymbol{a}$ and $\hat{\boldsymbol{v}}$.

These automated metrics are not always aligned with human preference. We therefore conduct a user study detailed in Sec. 5.2, asking human raters to compare videos generated by multiple models and select the best one.

4 Audio-Video Synchronized Diffusion

4.1 Preliminary: Text-to-Image Latent Diffusion

Text-to-image latent diffusion models (LDMs [34]) encode images $\boldsymbol{x}$ into a lower-dimensional latent space $\boldsymbol{z} = \mathcal{E}(\boldsymbol{x})$ using a pre-trained perceptual auto-encoder

$\mathcal{E}(\cdot)$, and learn the conditional distribution $p(\boldsymbol{z}|\boldsymbol{\tau})$ given a CLIP-encoded text prompt $\boldsymbol{\tau}$. It gradually denoises latents $\boldsymbol{z}^k$, obtained by corrupting the image latent $\boldsymbol{z}$ with Gaussian noise ϵ, over k time steps. A denoising UNet $\epsilon_\theta(\cdot)$ parameterized by θ is deployed to estimate the added noise ϵ by minimizing

$$\mathcal{L}_{LDM} = \mathbb{E}_{\boldsymbol{z},\epsilon\sim\mathcal{N}(0,1),k}\left[\|\epsilon - \epsilon_\theta(\boldsymbol{z}^k, k, \boldsymbol{\tau})\|_2^2\right] \tag{5}$$

During inference, LDMs start from a random Gaussian noise map $\boldsymbol{z}^K$, and iterate over K reverse diffusion steps [17, 38] to denoise the latents $\boldsymbol{z}^{k-1} = \boldsymbol{z}^k - \epsilon_\theta(\boldsymbol{z}^k, k, \boldsymbol{\tau})$, until $\boldsymbol{z}^0$ is obtained. LDMs then decode the latent into an image $\boldsymbol{x}^0 = \mathcal{D}(\boldsymbol{z}^0)$ using the pre-trained decoder $\mathcal{D}(\cdot)$. For simplicity, we refer to the images by their latents $\boldsymbol{z}$ rather than $\boldsymbol{x}$ throughout the rest of the paper.

4.2 Proposed Architecture

In this work, we extend the capabilities of LDMs for ASVA, by focusing on learning video dynamics and temporal synchronization, unlike existing approaches [14, 39, 40] which primarily use audio to control global semantics. We propose the Audio-Video Synchronized Diffusion model (AVSyncD), which builds upon a pre-trained image LDM and integrates synchronized audio control and temporal layers for improved video consistency. The major component is a UNet ϵ_θ aiming to denoise a video instead of an image. The UNet is trained on synchronized audio-video pairs, with the first frame $\boldsymbol{z}_1$, the corresponding audio $\boldsymbol{a}$, and the CLIP encoded audio category name $\boldsymbol{\tau}$ as input conditions. The LDM denoising objective in Eq. (5) is applied to the remaining frames $\boldsymbol{z}_{2:rT}$ to be predicted. The architecture of AVSyncD is shown in Fig. 3 and discussed below. Further details, for example, highlighting the different attention modules used in the architecture are described in Suppl. Sec. 3.2.

Initial-frame Conditioning To condition LDM on an input image, we feed its latent $\boldsymbol{z}_1$ without noise into the UNet at every diffusion timestep k. For all subsequent frames, we adhere to the original LDM, using independently sampled noisy latents $\boldsymbol{z}^k_{2:rT}$ as inputs and predicting the added noise ϵ^k.

Text Conditioning We retain the text cross-attention layers in the original LDMs [34] (without finetuning) to condition the model on the audio category. However, due to the limited text diversity in training, class conditioning does not bring significant gains (see Suppl. Sec. 6.5).

Audio Conditioning To facilitate audio-synchronized generation, we condition the generation on ImageBind audio embeddings [14]. ImageBind computes an audio classification token, $\boldsymbol{a}^g$, representing global semantics, and local patch tokens, $\boldsymbol{a}_{f,t}$, across T_a timestamps. The original ImageBind only uses $\boldsymbol{a}^g$ for contrastive learning and discards $\boldsymbol{a}_{f,t}$. However, we found these frozen local tokens as efficient synchronization cues. We split the patch tokens temporally into rT segments, corresponding to the same timestamps as the frames $\boldsymbol{z}_{1:rT}$, and append the global token to each. Each frame $\boldsymbol{z}_t$ then learns both semantics and synchronization guidance from its audio segment $\boldsymbol{a}_t$ via cross-attention [43]. In Suppl. Sec. 6.7, we compared ImageBind with CLAP and BEATs as encoders.

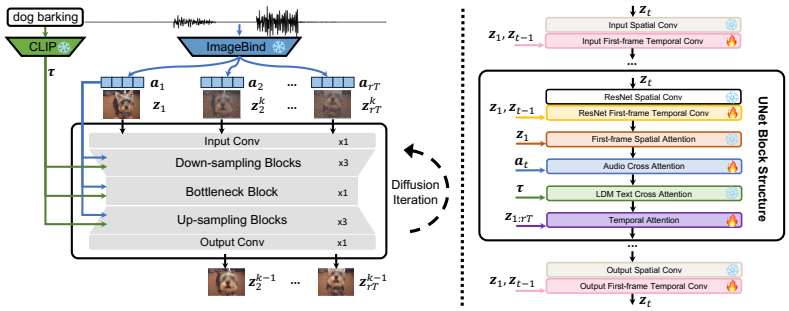


Fig. 3: AVSyncD overview. *Left*: We use ImageBind to encode audio into semantically aware time-dependent tokens $(\boldsymbol{a}_t)_{t=1}^{rT}$ and CLIP to encode the audio category into text embedding $\boldsymbol{\tau}$. In addition, the model receives the latent of the first frame $\boldsymbol{z}_1$, and iteratively denoises noisy latents of the subsequent frames $\boldsymbol{z}_{2:rT}^k$ via reverse diffusion. The denoising UNet, based on LDMs [34], consists of a sequence of downsampling, bottleneck and up-sampling blocks, with structure detailed on the right. *Right*: Anatomy of a UNet block for frame $\boldsymbol{z}_t$. LDM's original spatial conv, spatial attention and text cross attention layers are frozen, while its spatial self-attention layers are adjusted to first-frame spatial attentions, cross-attending to $\boldsymbol{z}_1$ instead. To learn video dynamics, we introduce temporal attention layers, and first-frame lookup temporal convolutions applied to input, output, and ResNet layers. We also train audio cross attentions for audio conditioning and synchronization. Trainable layers are marked with 🔥.

Spatial Convolution The original LDM's pre-trained spatial convolutional layers were frozen and used without modification.

First-frame Temporal Convolution Each spatial convolution block was augmented with a 1D temporal convolution layer (kernel size 3) to capture temporal dependencies. To better adhere to the starting image, $\boldsymbol{z}_1$, we adjusted the receptive field at frame t to include frames $(1, t-1, t)$ as opposed to $(t-1, t, t+1)$. These temporal convolutions with first-frame lookup were applied to three components in the UNet, namely the input/output conv layers and all ResNet convs.

First-frame Spatial Attention We also leveraged the pre-trained LDM's spatial self-attentions. Following [21], we modified the frozen spatial attention layers to cross-attend to the first frame rather than self-attend to the current frame, by computing the key-value pairs from the first frame and the queries from the current one.

Temporal Attention To effectively model long-range visual dependencies, we incorporated temporal attention layers [2]. Each frame index t was converted into a sinusoidal positional embedding, added to the corresponding frame's latents after a learnable projection [43]. Each frame's local representation at spatial position (h, w), $\boldsymbol{z}_{hwt}$, was then updated by attending to all frames at the same position $(\boldsymbol{z}_{hw1}, \boldsymbol{z}_{hw2}, \ldots, \boldsymbol{z}_{hw(rT)})$ through a standard attention mechanism.

Classifier-free Audio Guidance Classifier-free guidance [18] is a technique used to control the impact of input prompts on the generated outputs. We extended it to amplify audio guidance for improved synchronization. To accomplish

this, we trained the model for both audio-conditioned and unconditioned generation, by randomly replacing $\boldsymbol{a}$ with a null audio embedding, $\boldsymbol{a}_\emptyset$, with a 20% likelihood. $\boldsymbol{a}_\emptyset$ was computed by encoding an all-zero audio mel spectrogram via ImageBind. During inference, a factor $\eta \geq 1$ controls the audio guidance:

$$\boldsymbol{z}_{2:rT}^{k-1} = (1-\eta) \cdot \epsilon_\theta(\boldsymbol{z}_{2:rT}^{k}, k; \boldsymbol{z}_1, \boldsymbol{a}_\emptyset, \boldsymbol{\tau}) + \eta \cdot \epsilon_\theta(\boldsymbol{z}_{2:rT}^{k}, k; \boldsymbol{z}_1, \boldsymbol{a}, \boldsymbol{\tau}) \tag{6}$$

where η guides the denoising process towards latents congruent with audio-conditioned generation and away from those of audio-unconditional generation.

5 Experiments

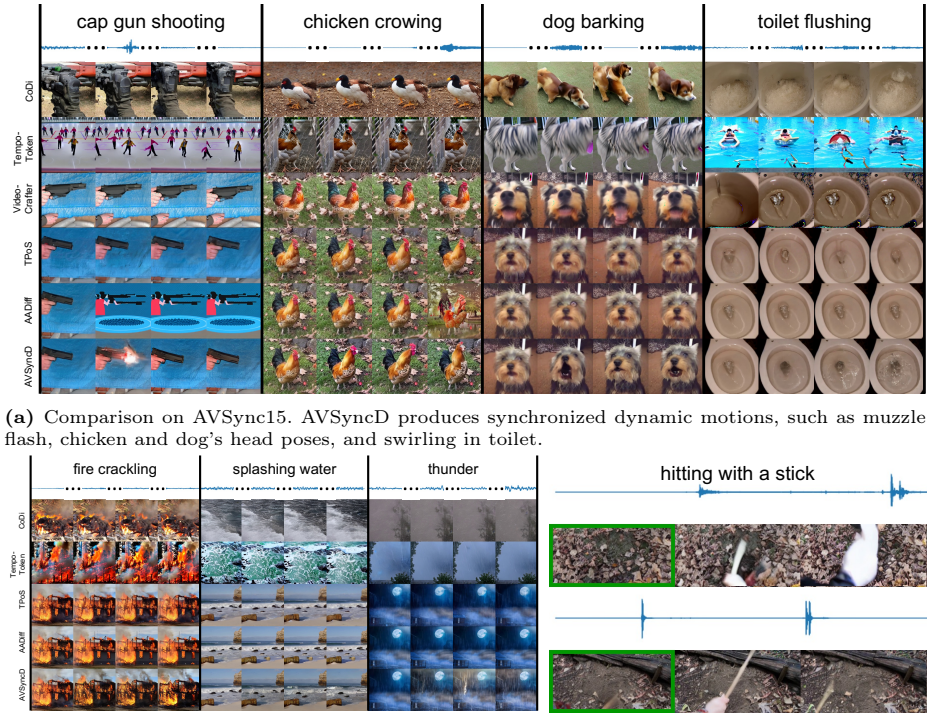
5.1 Implementation

Dataset We conducted experiments on three datasets. *AVSync15*: Our high-quality dataset curated from VGGSound [7], with 15 balanced categories, 1350 training videos, and 150 test videos. We also assessed our data curation pipeline by comparing it to models trained on Landscapes, TheGreatestHits, VGGSS and our AVSync-AC. *Landscapes* [22] is composed of 9 environmental sound classes. We followed the split in [35, 48] with 900 clips for training and 100 for testing. Since Landscapes is full of ambient sounds without synchronized video motion, we mainly use it to evaluate visual quality. *TheGreatestHits* [29] is an audio-video dataset recording humans probing environments with a drumstick, with 733 videos for training and 244 for testing. The videos are synchronized with audio at the moments of impact but contain lots of static moments. It also offers limited diversity, featuring a singular motion of probing. VGGSS and our AVSync-AC were described in Sec. 3.

Data Preprocess We sampled 2-second synchronized audio-video pairs for experiments. Videos were sampled at 6 fps with 12 frames, and resized to 256×256 on AVSync15/Landscapes or 128×256 on TheGreatestHits. Following ImageBind [14], audios were sampled at 16kHz and converted into 128-d spectrograms.

Baselines We first adopted a simple *Static* baseline by repeating the input frame into a video, then compared it to several state-of-the-art works. (1) *Semantic audio-to-video generation* (CoDi [40]): Image, text, and audio are encoded into a shared CLIP [32] space and summed, then fused into a video diffusion model trained on large-scale datasets. (2) *Image-to-video generation* (VideoCrafter [5]): A superior video diffusion model however without audio inputs. It animates images by fusing CLIP-encoded image and text features into the model via modality-dependent cross-attention layers. (3) *Synchronized audio-to-video generation* (TPoS [20], AADiff [23], TempoToken [48]): Audio is encoded into time-dependent segments and fused into frozen text-to-image [34] or text-to-video [44] models. We re-implemented AADiff, and used TPoS and TempoToken's pretrained checkpoints on VGGSound and Landscapes. More details are provided in Suppl. Sec. 4.

Training & Evaluation We adopted the pretrained Stable Diffusion-V1.5 [34] as the diffusion model and ImageBind [14] as the audio encoder. All models



(a) Comparison on AVSync15. AVSyncD produces synchronized dynamic motions, such as muzzle flash, chicken and dog's head poses, and swirling in toilet.

(b) Comparison on Landscapes. Without synchronization cues in audio, AVSyncD still produces evolving visual effects and semantics, such as flames, waves, and lightning.

(c) Generation by AVSyncD on TheGreatestHits. AVSyncD generates *hitting* actions at appropriate moments.

Fig. 4: Qualitative results on three datasets.

were trained using Adam optimizer with a batch size of 64 and a learning rate of 0.0001. Besides metrics in Sec. 3.2, we also provide results using metric in [26] in Suppl. Sec. 3.4. We evaluated on 3 clips uniformly sampled from each video.

5.2 Main Results

Dataset Comparison In Tab. 1, the *Static* baseline is a good indicator of dataset attributes. It has similar IA/IT scores compared to *Groundtruth*, since *Static* is composed of a subset of groundtruth frames, which are obviously semantically aligned. FVD of *Static* on TheGreatestHits is relatively low, since the TheGreatestHits contains frequent static ground truth clips without any moving objects. More importantly, the RelSync and AlignSync of *Static* gradually increase from AVSync15, to TheGreatestHits, to Landscapes, with those on Landscapes even surpassing the ground truth. The fact that static videos perform well in terms of audio synchronization on TheGreatestHits and Landscapes datasets testify to the superiority of AVSync15 for audio-video synchronized generation.

Model Comparison We compared to prior works on AVSync15 and Landscapes in Tabs. 1a and 1b and Figs. 4a and 4b. In Tab. 1a, CoDi achieves inferior

Table 1: Quantitative results. User study shows votes on 3 metrics: image quality, frame consistency, and synchronization. Inputs are combinations of image, text, audio.

Input	Model	FID↓	IA↑	IT↑	FVD↓	AlignSync↑	RelSync↑	User Study		
								IQ↑	FC↑	Sync↑
T+A	TPoS [20]	13.5	23.38	24.83	2671.0	19.52	42.50	-	-	-
	TempoToken [48]	12.2	18.84	17.45	4466.4	19.74	44.05	-	-	-
I+T	VideoCrafter [5]	11.8	-	29.87	840.7	21.28	43.16	38	20	12
	I2VD	12.1	-	30.35	398.2	21.80	43.92	62	90	91
I+T+A	CoDi [40]	14.5	28.15	23.42	1522.6	19.54	41.51	-	-	-
	TPoS [20]	11.9	38.36	**30.73**	1227.8	19.67	39.62	-	-	-
	AADiff [23]	18.8	34.23	28.97	978.0	22.11	45.48	37	4	5
	AVSyncD $\eta = 1$	12.1	38.36	30.34	382.7	22.25	44.81	-	-	-
	AVSyncD $\eta = 4$	**11.7**	**38.53**	30.45	**349.1**	22.62	45.52	**163**	**186**	**192**
	AVSyncD $\eta = 8$	**11.7**	37.99	30.27	420.7	**22.74**	**45.88**	-	-	-
	Static	-	39.76	30.39	1220.4	21.83	43.66			
	Groundtruth	-	40.06	30.31	-	25.04	50.00			

(a) Performance on AVSync15.

Input	Model	FID↓	IA↑	IT↑	FVD↓	AignSync↑	RelSync↑
T+A	TPoS [20]	16.5	15.61	**26.70**	2081.3	23.12	48.15
	TempoToken [48]	16.4	22.58	22.87	2480.0	24.21	48.65
I+T	I2VD	16.7	-	22.56	539.5	24.74	49.89
I+T+A	CoDi [40]	20.5	22.63	24.23	982.9	22.63	45.48
	TPoS [20]	**16.2**	**23.52**	23.20	789.6	23.51	47.05
	AADiff [23]	70.7	22.07	22.92	1186.3	**26.77**	**53.93**
	AVSyncD $\eta = 1$	16.5	22.29	22.81	463.1	24.81	49.96
	AVSyncD $\eta = 4$	**16.2**	22.49	22.79	**415.2**	24.82	49.93
	Static	-	23.60	22.21	1177.5	25.79	51.59
	Groundtruth	-	23.65	22.08	-	25.01	50.00

(b) Performance on Landscapes.

Input	Model	FID↓	IA↑	IT↑	FVD↓	AignSync↑	RelSync↑
I+T	I2VD	9.1	-	**13.42**	425.0	22.05	44.58
I+T+A	AVSyncD $\eta = 1$	9.0	11.85	13.18	313.5	22.59	45.52
	AVSyncD $\eta = 4$	**8.7**	**12.07**	13.31	**249.3**	**22.83**	**45.95**
	Static	-	13.33	16.56	348.9	24.36	48.73
	Groundtruth	-	13.52	16.49	-	25.02	50.00

(c) Performance on TheGreatestHits.

results on almost all metrics. TPoS(I+T+A) shows strong image quality (FID), but is worse in video quality (FVD) and synchronization (RelSync). TempoToken, on the other hand, is better at synchronization rather than visual quality, likely due to the lack of image input. AADiff is competitive on synchronization but extremely bad on image quality (FID). This is expected as AADiff adjusts each frame using audio amplitude, producing visual changes that highly correlate to audio changes temporally but may be overwhelmed by noises, as shown in Fig. 4a. On Landscapes, its similar FVD to *Static* but abnormally higher Align-

Table 2: Effect of (a) first-frame lookups (b) data curation, evaluated on AVSync15.

FF-Conv	FF-Attn	FID↓	FVD↓	AignSync↑
✗	✗	11.8	383.3	22.19
✓	✗	**11.6**	347.4	22.24
✓	✓	11.8	**325.6**	**22.33**

(a)

Dataset	AC	MC	FID↓	FVD↓	AignSync↑
VGGSS	✗	✗	12.9	1307.9	21.50
AVSync-AC	✓	✗	12.0	428.8	22.09
AVSync15	✓	✓	**11.8**	**325.6**	**22.33**

(b)

Sync and RelSync also suggest that it only applied minor modifications to the input image due to lack of sound changes, as shown in Fig. 4b. Without audio input, VideoCrafter performs poorly on synchronization. It also has difficulty faithfully adhering to the input frame, as in Fig. 4a. AVSyncD achieves the best animation results on almost all metrics. On Landscapes, AVSyncD also performs the best on FID and FVD, with other scores being similar to ground truth.

User Study We invited 15 participants to compare 4 animation models with top overall performance (VideoCrafter, AADiff, I2VD, AVSyncD) on AVSync15, based on 3 metrics in Tab. 1. The 4 models generated videos conditioned on the same test examples (audio+image). Each test example was independently evaluated by 2 participants to select their most preferred generation (vote) on each metric. In total, we evaluated all 150 test examples on AVSync15, collecting 150×2=300 votes on each metric. Tab. 1a shows votes each model received.

5.3 Ablation Studies

Audio Conditioning In Tab. 1, AVSyncD outperforms I2VD, especially on AlignSync and RelSync. AVSyncD did not improve RelSync on Landscapes probably due to the lack of synchronization cues on the dataset itself. These results show that audio condition enhances generation quality and synchronization.

Audio Guidance Tab. 1 shows increasing the audio guidance factor η from 1 to 4 improves FID, IA, and FVD on all three datasets. As expected, audio guidance also improved AlignSync and RelSync significantly on AVSync15 and TheGreatestHits, but not on the less synchronized Landscapes dataset. Prior works [20, 23, 39] claimed that increasing audio amplitude can also lead to stronger visual effects. We compare this approach with our audio guidance in Fig. 5a. Audio guidance offers a better control mechanism than audio amplitude.

First-frame (FF) Lookups We validated FF Lookups by replacing them with standard temporal convolutions or spatial self-attention in Tab. 2a.

Data Curation We compared to AVSyncD trained on random subsets from VGGSS and AVSync-AC with equal data scale and balanced categories in Tab. 2b.

5.4 Applications and Extensions

Animate Generated Images When lacking an image as input, we can use existing image generators to generate the image, which AVSyncD can also animate. Fig. 5b shows animations on images generated by StableDiffusion-V1.5 [34].

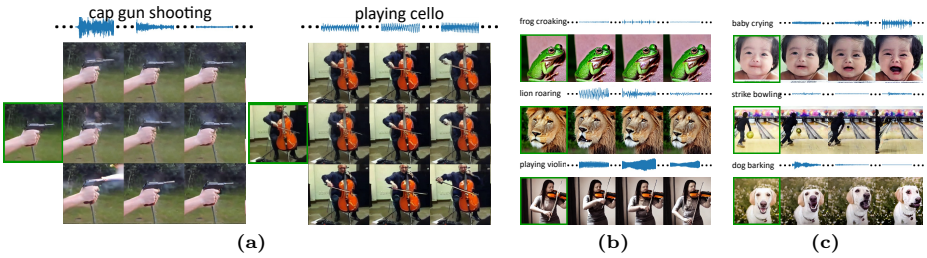
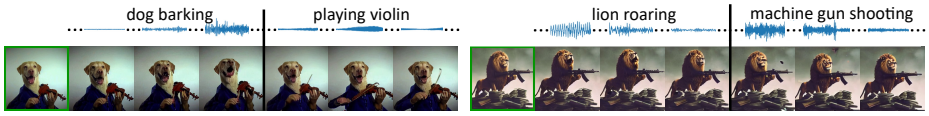


Fig. 5: (a): Effects of audio amplitude vs. classifier-free audio guidance. *top*: original audio with $\eta = 1$; *mid*: $100\times$ amplified audio with $\eta = 1$; *bottom*: original audio with $\eta = 8$. (b): Animate generated images. (c): Animation with internet images and audios.

(a) Controllable animation with (un)related audios.



(b) Animate target objects on internet images with audios.

Fig. 6: Controllable image animation with audios. Key frames are visualized.

Animate Contents from Internet AVSyncD can also generalize well to unseen images and audio, as shown in Fig. 5c.

Control Animations with (Un)Related Audios We can control the motion of an image to follow desired audio, e.g., animate a baby to not only cry but also bark like a dog or roar like a lion, as seen in Fig. 6a. When there is no object related to the audio, the animations do not demonstrate corresponding motion.

Animate Target Objects with Audios When multiple objects exist in the image, a scenario not existing in training data, we can still use audios to only animate the related target object, as shown in Fig. 6b.

6 Conclusion

We tackled the under-explored Audio-Synchronized Visual Animation task, with an emphasis on generating videos with audio-synchronized dynamics. We contributed the high-quality AVSync15 benchmark via careful data curation and proposed the AVSyncD model to animate images with realistic motions. Due to the scale of AVSync15, our work cannot generalize to all audio classes in the world, which requires several orders of magnitude larger datasets. However, we hope our research inspires further work in this direction.

References

1. Bain, M., Nagrani, A., Varol, G., Zisserman, A.: Frozen in time: A joint video and image encoder for end-to-end retrieval. In: IEEE International Conference on Computer Vision (2021)
2. Blattmann, A., Rombach, R., Ling, H., Dockhorn, T., Kim, S.W., Fidler, S., Kreis, K.: Align your latents: High-resolution video synthesis with latent diffusion models. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2023)
3. Brock, A., Donahue, J., Simonyan, K.: Large scale gan training for high fidelity natural image synthesis. In: ICLR (2019)
4. Castellano, B.: Pyscenedetect. `https://www.scenedetect.com/`
5. Chen, H., Xia, M., He, Y., Zhang, Y., Cun, X., Yang, S., Xing, J., Liu, Y., Chen, Q., Wang, X., Weng, C., Shan, Y.: Videocrafter1: Open diffusion models for high-quality video generation (2023)
6. Chen, H., Xie, W., Afouras, T., Nagrani, A., Vedaldi, A., Zisserman, A.: Audio-visual synchronization in the wild. In: Proceedings of the British Machine Vision Conference (BMVC) (2021)
7. Chen, H., Xie, W., Vedaldi, A., Zisserman, A.: Vggsound: A large-scale audio-visual dataset. In: ICASSP (2020)
8. Chen, S., Wu, Y., Wang, C., Liu, S., Tompkins, D., Chen, Z., Wei, F.: Beats: Audio pre-training with acoustic tokenizers. In: ICML (2023)
9. Chung, J.S., Zisserman, A.: Lip reading in the wild. In: ACCV (2016)
10. Chung, J.S., Zisserman, A.: Out of time: automated lip sync in the wild. In: ACCV Workshop (2016)
11. Chung, S.W., Chung, J.S., Kang, H.G.: Perfect match: Improved cross-modal embeddings for audio-visual synchronisation. In: ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (2019)
12. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. In: NeurIPS (2021)
13. Gemmeke, J.F., Ellis, D.P.W., Freedman, D., Jansen, A., Lawrence, W., Moore, R.C., Plakal, M., Ritter, M.: Audio set: An ontology and human-labeled dataset for audio events. In: Proc. IEEE ICASSP 2017 (2017)
14. Girdhar, R., El-Nouby, A., Liu, Z., Singh, M., Alwala, K.V., Joulin, A., Misra, I.: Imagebind: One embedding space to bind them all. In: CVPR (2023)
15. Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., Cohen-Or, D.: Prompt-to-prompt image editing with cross attention control. In: ICLR (2023)
16. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. In: Advances in Neural Information Processing Systems (2017)
17. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: NeurIPS (2020)
18. Ho, J., Salimans, T.: Classifier-free diffusion guidance. In: NeurIPS Workshop on Deep Generative Models and Downstream Applications (2022)
19. Iashin, V., Xie, W., Rahtu, E., Zisserman, A.: Sparse in space and time: Audio-visual synchronisation with trainable selectors. In: British Machine Vision Conference (BMVC) (2022)
20. Jeong, Y., Ryoo, W., Lee, S., Seo, D., Byeon, W., Kim, S., Kim, J.: The power of sound (tpos): Audio reactive video generation with stable diffusion. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7822–7832 (2023)

21. Khachatryan, L., Movsisyan, A., Tadevosyan, V., Henschel, R., Wang, Z., Navasardyan, S., Shi, H.: Text2video-zero: Text-to-image diffusion models are zero-shot video generators. In: ICCV (2023)
22. Lee, S.H., Oh, G., Byeon, W., Bae, J., Kim, C., Ryoo, W.J., Yoon, S.H., Kim, J., Kim, S.: Sound-guided semantic video generation. In: ECCV (2022)
23. Lee, S., Kong, C., Jeon, D., Kwak, N.: Aadiff: Audio-aligned video synthesis with text-to-image diffusion. In: CVPR Workshop on Content Generation (2023)
24. Li, R., Yang, S., Ross, D.A., Kanazawa, A.: Learn to dance with aist++: Music conditioned 3d dance generation. In: ICCV (2021)
25. Li, Y., Liu, H., Wu, Q., Mu, F., Yang, J., Gao, J., Li, C., Lee, Y.J.: Gligen: Open-set grounded text-to-image generation. In: CVPR (2023)
26. Luo, S., Yan, C., Hu, C., Zhao, H.: Diff-foley: Synchronized video-to-audio synthesis with latent diffusion models. In: NeurIPS (2023)
27. Mokady, R., Hertz, A., Aberman, K., Pritch, Y., Cohen-Or, D.: Null-text inversion for editing real images using guided diffusion models. In: CVPR (2023)
28. Ng, E., Romero, J., Bagautdinov, T., Bai, S., Darrell, T., Kanazawa, A., Richard, A.: From audio to photoreal embodiment: Synthesizing humans in conversations. In: ArXiv (2024)
29. Owens, A., Isola, P., McDermott, J., Torralba, A., Adelson, E.H., Freeman, W.T.: Visually indicated sounds. In: CVPR (2016)
30. Park, S.J., Kim, M., Hong, J., Choi, J., Ro, Y.M.: Synctalkface: Talking face generation with precise lip-syncing via audio-lip memory. In: AAAI Conference on Artificial Intelligence (AAAI) (2022)
31. Pedro Morgado, Ishan Misra, N.V.: Robust audio-visual instance discrimination. In: Computer Vision and Pattern Recognition (CVPR), IEEE/CVF Conf. on (2021)
32. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: ICML (2021)
33. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents. In: arXiv (2022)
34. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR (2022)
35. Ruan, L., Ma, Y., Yang, H., He, H., Liu, B., Fu, J., Yuan, N.J., Jin, Q., Guo, B.: Mm-diffusion: Learning multi-modal diffusion models for joint audio and video generation. In: CVPR (2023)
36. Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., Schramowski, P., Kundurthy, S., Crowson, K., Schmidt, L., Kaczmarczyk, R., Jitsev, J.: Laion-5b: An open large-scale dataset for training next generation image-text models. In: NeurIPS (2022)
37. Singer, U., Polyak, A., Hayes, T., Yin, X., An, J., Zhang, S., Hu, Q., Yang, H., Ashual, O., Gafni, O., Parikh, D., Gupta, S., Taigman, Y.: Make-a-video: Text-to-video generation without text-video data (2022)
38. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. In: ICLR (2021)
39. Sung-Bin, K., Senocak, A., Ha, H., Owens, A., Oh, T.H.: Sound to visual scene generation by audio-to-visual latent alignment. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2023)
40. Tang, Z., Yang, Z., Zhu, C., Zeng, M., Bansal, M.: Any-to-any generation via composable diffusion. In: Thirty-seventh Conference on Neural Information Processing Systems (2023), https://openreview.net/forum?id=2EDqbSCnmF

41. Tsuchida, S., Fukayama, S., Hamasaki, M., Goto, M.: Aist dance video database: Multi-genre, multi-dancer, and multi-camera database for dance information processing. In: Proceedings of the 20th International Society for Music Information Retrieval Conference, ISMIR 2019. Delft, Netherlands (Nov 2019)
42. Unterthiner, T., van Steenkiste, S., Kurach, K., Marinier, R., Michalski, M., Gelly, S.: Towards accurate generative models of video: A new metric & challenges. In: arXiv (2019)
43. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: NeurIPS (2017)
44. Wang, J., Yuan, H., Chen, D., Zhang, Y., Wang, X., Zhang, S.: Modelscope text-to-video technical report (2023)
45. Wang, W., Xie, k., Liu, Z., Chen, H., Cao, Y., Wang, X., Shen, C.: Zero-shot video editing using off-the-shelf image diffusion models. arXiv preprint arXiv:2303.17599 (2023)
46. Wu, R., Chen, L., Yang, T., Guo, C., Li, C., Zhang, X.: Lamp: Learn a motion pattern by few-shot tuning a text-to-image diffusion model. arXiv preprint arXiv:2310.10769 (2023)
47. Xu, H., Ghosh, G., Huang, P.Y., Okhonko, D., Aghajanyan, A., Metze, F., Zettlemoyer, L., Feichtenhofer, C.: Videoclip: Contrastive pre-training for zero-shot video-text understanding. In: EMNLP (2021)
48. Yariv, G., Gat, I., Benaim, S., Wolf, L., Schwartz, I., Adi, Y.: Diverse and aligned audio-to-video generation via text-to-video model adaptation (2023)
49. Ye, Z., Jiang, Z., Ren, Y., Liu, J., He, J., Zhao, Z.: Geneface: Generalized and high-fidelity audio-driven 3d talking face synthesis. In: ICLR (2023)
50. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: IEEE International Conference on Computer Vision (ICCV) (2023)
51. Zhou, H., Liu, Y., Liu, Z., Luo, P., Wang, X.: Talking face generation by adversarially disentangled audio-visual representation. In: AAAI Conference on Artificial Intelligence (AAAI) (2019)
52. Zhou, H., Sun, Y., Wu, W., Loy, C.C., Wang, X., Liu, Z.: Pose-controllable talking face generation by implicitly modularized audio-visual representation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2021)