AV-Link: Temporally-Aligned Diffusion Features for Cross-Modal Audio-Video Generation

Moayed Haji-Ali[1,2,*] Willi Menapace[2] Aliaksandr Siarohin[2] Ivan Skorokhodov[2]
Alper Canberk[2] Kwot Sin Lee[2] Vicente Ordonez[1] Sergey Tulyakov[2]

[1]Rice University [2]Snap Inc

Project Webpage: https://snap-research.github.io/AVLink

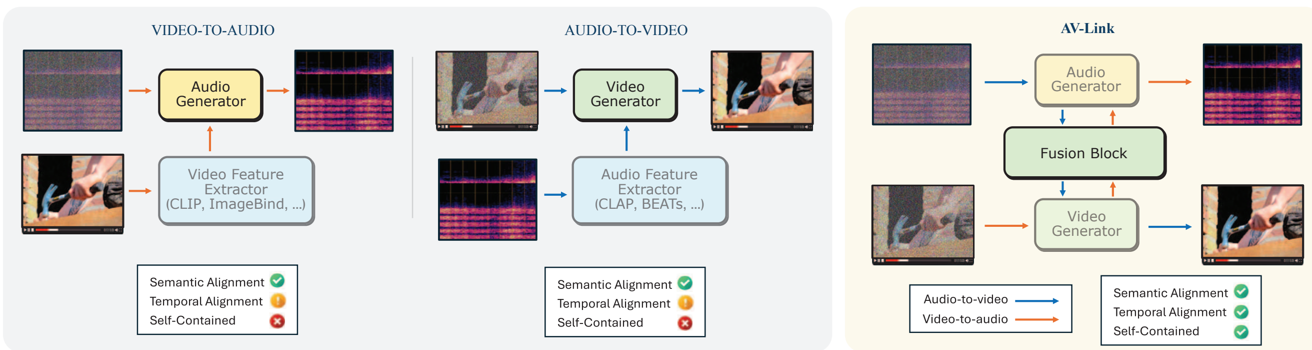

Figure 1. Compared to current Video-to-Audio and Audio-to-Video methods, AV-Link provides a unified framework for these two tasks. Rather than relying on feature extractors pretrained for other tasks (e.g. CLIP [68], CLAP [21]), we directly leverage the activations from pretrained frozen Flow Matching models using a Fusion Block to achieve precise time alignment between modalities. Our approach offers competitive semantic alignment and improved temporal alignment in a self-contained framework for both modalities.

Abstract

We propose AV-Link, a unified framework for Video-to-Audio (A2V) and Audio-to-Video (A2V) generation that leverages the activations of frozen *video and audio diffusion models for temporally-aligned cross-modal conditioning. The key to our framework is a Fusion Block that facilitates bidirectional information exchange between video and audio diffusion models through temporally-aligned self attention operations. Unlike prior work that uses dedicated models for A2V and V2A tasks and relies on pretrained feature extractors, AV-Link achieves both tasks in a single self-contained framework, directly leveraging features obtained by the complementary modality (i.e. video features to generate audio, or audio features to generate video). Extensive evaluations demonstrate that AV-Link achieves substantial improvements in audio-video synchronization, outperforming more expensive baselines such as MovieGen V2A model.*

[*]Work mainly done during an internship at Snap Inc.

1. Introduction

Generative models are becoming proficient at synthesizing high-quality video [1, 7, 38, 58, 62, 75, 84] and audio [23, 27, 53, 54, 85]. Given the intrinsic correlation between audio and video, there has been a growing interest in cross-modal generation including Video-to-Audio (V2A) and Audio-to-Video (A2V). The goal of V2A models is to add sound that is both semantically and temporally aligned to video content [57, 67, 92, 103]. Conversely, A2V models [45, 99, 102] seek to generate video that plausibly represent the sounds of a given audio. Since text has been established as a robust conditioning modality to guide the semantics of generated audio and video [32, 98], achieving accurate audio–video temporal alignment remains as the *most critical* and challenging aspect in cross-modal generation. Current V2A models [57, 67, 92, 103] typically depend on pre-trained visual feature extractors such as CLIP [68], MetaCLIP [93], ImageBind [25], and CAVP [57] to condition audio generation and create soundscapes that are semantically aligned with visual frames. Similarly, A2V methods condition video generators on features from pretrained audio models such as ImageBind [25], and BEATs [11]. However, the reliance on modality-specific feature encoders necessitates distinct models for V2A and A2V, limiting the potential for a unified framework and complicating the deployment for applications requiring comprehensive capabilities: text-to-audio (T2A), text-to-video (T2V), V2A, and A2V.

Recognizing the advantages of a single framework capable of generating both audio and video, recent work explores training a single model for joint audio-video generation [36, 40, 72, 87]. Although promising, these models have a performance gap with respect to conditional models (*i.e*. dedicated V2A and A2V models), and often operate on simpler datasets such as landscape videos [46, 72, 87] or exhibit weaker audio-video synchronization [36].

In this work, we introduce AV-Link, a unified framework that is capable of *semantically* and *temporally* aligned generation for both V2A and A2V as shown in Fig. 1. Starting from two well-trained and *frozen* flow models for audio and video generation, we propose to use their activations as conditioning signals for V2A or A2V synthesis, bypassing the need for pretrained feature extractors. Such activations contain rich semantics [56], but they are also *temporally-aligned* since these activations are trained to generate modalities with a temporal component (*i.e*. audio and video). This is motivated by prior findings showing that activations in image diffusion models encode substantial semantic and spatial information, making them useful for pixel-aligned tasks such as semantic keypoint matching [29, 56] segmentation [97] and depth estimation [39].

To preserve the quality of the frozen unimodal generators while enabling cross-modal generation, we introduce a Fusion Block that connects the audio and video modalities by conditioning the generation of one modality on the activations from the other. Furthermore, we propose time-aligned Rotary Position Embedding (RoPE) as a flexible mechanism to align audio and video tokens across the temporal axis, demonstrating its benefits over prior approaches for temporal alignment. Moreover, although static diffusion model activations serve as a robust conditioning signal, cross-modal generation may benefit from a conditioning signal that dynamically evolves throughout the generation process. To this end, we introduce a *symmetric feature reinjection* strategy that leverages the diffusion prior of the pretrained model for the conditioning modality to refine the conditioning signal, enabling bidirectional information exchange between the conditioning and generated modality. By introducing a relatively small set of parameters (186M parameters), AV-Link offers a compact design that simplifies training and deployment for applications requiring comprehensive T2V, T2A, V2A, and A2V generation. Rather than focusing on improving single-modality generation—a topic extensively explored in concurrent studies [27, 98]—this work builds on those efforts by unifying pretrained audio and video generators in a single framework, emphasizing temporal alignment as the core challenge in cross audio-visual generation.

In summary, we propose AV-Link, a unified framework to address A2V and V2A tasks. We show for the first time that video and audio diffusion activations contain rich temporal information and can replace specialized pretrained feature extractors while improving *temporal* alignment. The key to our method is a novel design of a Fusion Block that enables bidirectional information exchange between the conditioning and generated modalities through a symmetric feature reinjection mechanism. When evaluated against the state-of-the-art V2A method on VGGSounds [10], our method shows the best overall performance and drastically improves temporal alignment as measured by Onset ACC [20] by up to 76.4% over the best baseline. User studies show a strong preference for our method which is preferred over Movie Gen Audio [67] (13B parameters) 63.6% of the time with regards to temporal alignment. In the A2V task, AV-Link surpasses TempoTokens [83] in video quality and audio-video alignment, with user studies showing an overwhelming preference for our method.

2. Related Work

Video-to-Audio Generation The task of generating audio from silent videos recently emerged as a difficult generative task [3, 8, 9, 15, 17, 26, 42, 44, 49, 51, 60, 61, 64, 70, 78, 79, 88, 95, 100] and gained relevance not only for its potential in simplifying movie production but also because of the recent development of high-quality video generation methods which produce no audio output [4, 5, 7, 38, 47, 62, 67, 98, 105]. Video feature representations play a crucial role in achieving highly aligned generation and have been the focus of most recent work [33, 35, 57, 67, 71, 74, 86, 89, 91, 103]. SpecVQGAN [35] employs a cross-modal transformer conditioned on ImageNet-pretrained visual backbones. V2A-Mapper [86] aligns CLIP [68] and CLAP [21] to leverage a pretrained text-to-audio model for V2A. Diff-Foley [57] and Frieren [89] address alignment using CAVP [13], an ad-hoc contrastive video-audio representation. FoleyCrafter [103] augments a frozen audio generator with a learnable semantic adapter and temporal controller. Movie Gen Audio [67] obtains strong video-audio alignment by conditioning on a descriptive text caption and per-frame MetaCLIP [93] features. Concurrent work [14, 16] demonstrated remarkable capabilities by leveraging training on audio-only datasets. Despite such advancements, the unification of A2V and V2A in a single framework remains a challenge. We show that temporally-aligned V2A and A2V generation can be achieved within the same framework by leveraging existing diffusion activations without the need for specialized feature extractors.

Audio-to-Video Generation Similar to the Video-to-Audio task, recent work has focused on building suitable audio representations, paying special attention to time alignment. *Lee et al*. [46] employ a Sound Inversion Encoder to map audio features to the StyleGAN latent space. Seeing-and-Hearing [92] propose a video-audio aligner based on ImageBind [25]. TPoS [37] manipulates generated images conditioned on audio representations extracted through an ad-hoc temporally-aligned audio encoder. TempoTokens [99] aligns a frozen video generator to the audio modality by employing features from a finetuned BEATs encoder. AADif [45] leverages the CLAP [21] audio encoder and audio magnitude to condition generation. Recently, AVSyncD [102] encodes audio into temporally-dependent tokens using ImageBind [25]. We show that using activations from a pretrained audio generator provides a stronger conditioning signal than external feature extractors, yielding improved temporal alignment.

Joint Video-Audio Generation Several works emerged that generate the audio and video modalities jointly. [28, 40, 50, 77, 80, 96] MM-Diffusion [72] proposes a diffusion framework based on a Coupled U-Net for simultaneously denoising the audio and video streams. Seeing-and-Hearing [92] employs ImageBind [25] to establish alignment between modalities. AVDiT [87] introduces a multimodal DiT design and learns multiple conditional distributions over audio and video through a mixture of noise levels, enabling V2A, A2V, and joint audio-video generation. Such approaches, however, present lower performance than conditional methods [67, 103]. Additionally, they are trained from scratch on audio-video datasets. Since most large-scale audio and video datasets are curated around a single modality, the other modality may be missing, have insufficient quality, or unlabeled [12, 27], this limits the data scalability of such methods. Thus, in this work, we target cross-modal generation (*i.e*. V2A and A2V) rather than joint audio-video generation.

3. Method

Starting with pretrained and *frozen* Text-to-Audio and Text-to-Video generators, we aim to combine the two under a unified architecture to produce Video-to-Audio $\mathcal{G}_{V2A}$ and Audio-to-Video $\mathcal{G}_{A2V}$ generators. To achieve this, we propose AV-Link, a symmetric architecture that benefits from the activations of one generator to condition the generator of the other modality. In Sec. 3.1, we describe the background on Flow Matching models. We then describe in Sec. 3.2, the architecture of our pretrained audio and video generators. Finally, in Sec. 3.3, we describe the architecture of the multimodal Fusion Blocks that connect the two generators.

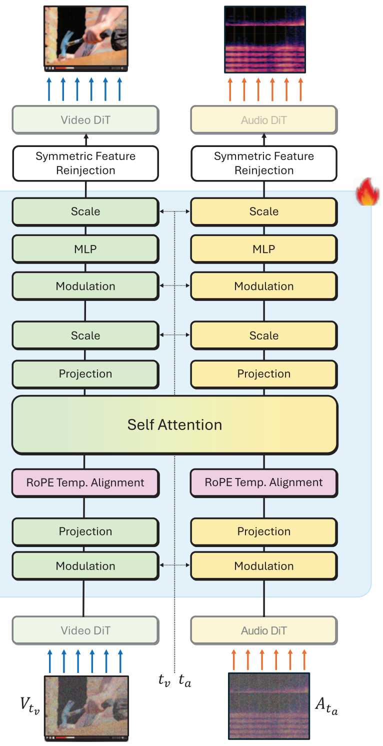

Figure 2. Design of the proposed Fusion Block connecting the frozen video and audio backbones. A RoPE-based temporal alignment mechanism aligns the representation of the two modalities which are processed by self attention. Video and audio features are symmetrically reinjected into the frozen generators. The block is regularly applied multiple times throughout the backbones.

3.1. Background

We base our generative models on the Flow Matching framework [52, 55]. Flow Matching expresses generation of data $\mathbf{X}_1 \sim p_d$ as the progressive transformation of $\mathbf{X}_0$ following a path connecting samples from the two distributions. In its simplest formulation [55], the path is instantiated as a linear interpolation between the samples:

$$\mathbf{X}_t = t\mathbf{X}_1 + (1-t)\mathbf{X}_0, \tag{1}$$

and $\mathbf{X}_0 \sim p_n = \mathcal{N}(0, \mathbb{I})$ originate from a noise distribution.

We can move along the path following the velocity $v_t = \frac{d\mathbf{X}_t}{dt} = \mathbf{X}_1 - \mathbf{X}_0$ approximated by learnable $\mathcal{G}$ minimizing:

$$\mathcal{L} = \mathbb{E}_{t\sim p_t, \mathbf{X}_1\sim p_d, \mathbf{X}_0\sim p_n} \left\|\mathcal{G}(\mathbf{X}_t, t) - v_t\right\|_2^2, \tag{2}$$

where p_t indicates a training distribution over time t, which we instantiate as a logit normal distribution [22]. At inference time, an ODE solver such as a first-order Euler solver can be employed to produce samples $\mathbf{X}_1$ starting from Gaussian noise $\mathbf{X}_0$ using the model velocity estimates.

3.2. Base Models

Audio Model. Given an audio signal from audio dataset $\mathcal{D}_A$, we produce a corresponding Mel-spectrogram and adopt a 1D-VAE [27] to encode it into a 1D sequence of latent representations. This latent audio representation $\mathbf{A}_1$ thus has shape $\mathbb{R}^{T_a \times D_a}$, where T_a and D_a respectively indicate the number of tokens and their dimensionality produced by the autoencoder. The sequence is modeled by a DiT [65] (see Fig. 2) consisting of a stack of identical blocks containing a self attention operation followed by an MLP. Information on the current timestep t is injected through an adaptive LayerNorm [66] and the attention operator is augmented with 1D-RoPE [76] to encode positional information. To condition on text prompts, we encode them using the T5 [69] encoder and insert a cross attention layer attending to the text embeddings after each self attention operation [67]. The model contains 576M parameters and is trained to generate variable-length audio clips at 16kHz following previous work [32, 54]. We train it using the Flow Matching objective (Eq. (2)) over 100k iterations, with a learning rate of $3e-4$ and a batch size of 1024, distributed across 8 A100 GPUs. During inference, a pre-trained Vocoder [41] converts the generated Mel-spectrograms into waveforms.

Video Model. Given an RGB video $\mathbf{V}_1$, we flatten it to shape $\mathbb{R}^{T_v \times 3}$, where T_v indicates the total number of video pixels. We adopt a DiT architecture symmetric to the audio model, and use 3D-RoPE [18]. Due to the computational costs associated with large scale video generators [67, 98] and the orthogonality between improvement of single-modality generation and AV-Link, we adopt a relatively small model (576M parameters) for our experiments, enabling efficient training and inference on the video modality. We found that when using a latent diffusion formulation, its motion and structural quality do not match the ones from an RGB model of the same size. Consequently, we opt to train a base video model to generate 5.16-second videos in the RGB space at a resolution of 64 x 36 pixels and 6 frames-per-second (FPS). The resolution is later increased to 512 x 288 at 24 FPS through a latent video upsampler operating in the latent space of a MAGVIT-v2 [101] autoencoder that is directly finetuned from the video model. The base video model is trained for 250k iterations with a batch size of 512, while the upsampler uses a batch size of 384 for 25k iterations. Both models are trained with a learning rate of $3e-4$ on 16 A100 GPUs.

3.3. Multimodal Fusion Block

Cross-modal generation requires a high-quality representation for the conditioning modality [57, 103]. In the context of V2A, obtaining high-quality video features has been the focus of recent work, which explored the usage of contrastive representations such as ImageBind [92], CLIP [103], MetaCLIP [67], or built ad hoc video representations [57, 103]. While such representations capture the overall video semantic, we observe that they lack precise audio-video temporal alignment, producing characteristic temporal misalignment (see Sec. 4.4). On the other hand, the video generator is capable of generating videos from scratch, implying that its activations contain semantically and temporally aligned information that can be leveraged for V2A. We thus propose a fully symmetric framework for V2A and A2V that leverages activations from the frozen generator of the conditioning modality to achieve aligned cross-modal generation using no ad-hoc feature extractors.

Fusion Blocks We address V2A and A2V by linking frozen audio and video generators $\mathcal{G}_A$, $\mathcal{G}_V$ through the Fusion Block depicted in Fig. 2. Consider the output activations of audio and video DiT blocks $\boldsymbol{x}_a \in \mathbb{R}^{T_a \times D_a}$, $\boldsymbol{x}_v \in \mathbb{R}^{T_v \times D_v}$ at some diffusion timestep t_a, t_v. We derive the activation of the conditional modality from a ground truth input (*e.g.* video for the V2A task). Initially, we project activations into a common dimension D to produce $\bar{\boldsymbol{x}}_a \in \mathbb{R}^{T_a \times D}$, $\bar{\boldsymbol{x}}_v \in \mathbb{R}^{T_v \times D}$ respectively. A multi-head self attention operation is then applied to the concatenation of audio and video features $\{\bar{\boldsymbol{x}}_a, \bar{\boldsymbol{x}}_v\}$. Finally, an MLP projects the output back to the original dimensionality, producing video-aware audio activations $\hat{\boldsymbol{x}}_a \in \mathbb{R}^{T_a \times D_a}$ and audio-aware video activations $\hat{\boldsymbol{x}}_v \in \mathbb{R}^{T_v \times D_v}$ as:

$$\hat{\boldsymbol{x}}_a, \hat{\boldsymbol{x}}_v = \text{FusionBlock}(\boldsymbol{x}_a, \boldsymbol{x}_v, t_a, t_v). \tag{3}$$

The Fusion Block presents a simple and symmetric design that equally treats the different modalities to condition the other modality.

Symmetric Feature Reinjection While the Fusion blocks effectively transfer the conditioning signal to the generated modality, we hypothesize that this signal can benefit from a continuous refinement driven by the generated modality. The symmetric architecture of the Fusion Blocks enables bidirectional information flow between the conditioning and generated modalities, rather than a one-way flow from the conditioning to the generated modality. Therefore, we propose to pass the conditional-aware (*i.e.* $\hat{\boldsymbol{x}}_a$ in the V2A task) and generation-aware activations (*i.e.* $\hat{\boldsymbol{x}}_v$ in V2A task) to subsequent DiT blocks within the respective frozen backbones, thereby injecting cross-modal information. In contrast to using static feature extractors [57, 67, 103], our design, which we call *symmetric feature reinjection*, dynamically improves the quality of the activations extracted from successive blocks of the conditioning modality.

Multimodal Temporal Alignment To facilitate temporal alignment between modalities, we use temporally-aligned 1D RoPE [76] embeddings in the self attention operation. We express 1D temporal RoPE as:

$$\text{RoPE}(\bar{\boldsymbol{x}}_n, \tau(n)) = \bar{\boldsymbol{x}}_n e^{i\tau(n)\theta_{\text{base}}}, \tag{4}$$

where $\boldsymbol{x}$ represents an audio or video token for a simplified two-channel case, n represents its index on the temporal axis, θ_{base} is the base frequency, and τ is a temporal alignment function defined as:

$$\tau(n) = \begin{cases} n, & \text{if } \boldsymbol{x} \text{ is a video token} \\ n\frac{\eta_v}{\eta_a}, & \text{if } \boldsymbol{x} \text{ is an audio token} \end{cases}$$

Where η_a, η_v are the number of tokens to represent one second of audio and video respectively. Intuitively, this approach rotates corresponding audio and video tokens by the same angle proportional to their temporal position, thus establishing temporal correspondence that facilitates temporally-synced multimodal information exchange. Crucially, this design aligns the modalities within the inserted Fusion Blocks while keeping the pretrained generators frozen. We show in Sec. 4.4 that this design is favored over previous temporal alignment approaches.

Time-Aware Feature Fusion We inject the information on the flow timesteps for audio t_a and video t_v through adaptive adaLN-Zero [65], with audio and video parameter sets receiving information on both timesteps. Such design creates conditioning features that are most suited for the current flow time for the generated modality, increasing the effectiveness of *symmetric feature reinjection*.

Training We train the Fusion Block on a paired audio-video dataset $\mathcal{D}_{AV}$ keeping both generators *frozen* and using the rectified flow objective. For the case of V2A, we use the training objective:

$$\mathcal{L} = \mathbb{E}_{t_a,t_v,\mathbf{A}_1,\mathbf{V}_1,\mathbf{A}_0} \left\| \mathcal{G}_{V2A}(\mathbf{A}_{t_a}, \mathbf{V}_{t_v}, t_a, t_v) - (\mathbf{A}_1 - \mathbf{A}_0) \right\|_2^2, \\ t_a \sim p_{t_a}, t_v \sim p_{t_v}, \mathbf{A}_1, \mathbf{V}_1 \sim \mathcal{D}_{AV}, \mathbf{A}_0 \sim p_n, \quad (5)$$

and use a fully symmetric one for A2V. Keeping the audio and video generators frozen, we train the Fusion Blocks for 50k iterations with a learning rate of $3e$–4 and a batch size of 256 for the V2A and A2V tasks. Additionally, we consider using separate parameter sets for the V2A and A2V tasks or training them jointly (see Sec. 4.4). We discuss architectural details in Appx. B, and training details in Appx. C.

Inference In the V2A setting, we consider the input conditioning video $\mathbf{V}_1$ and compute partially noised video $\mathbf{V}_{t_v}$ at the chosen flow time t_v, which we set to a fixed level during the entire inference phase to serve as input for the video backbone. As we later show in Sec. 4.4, it is crucial to inject some noise into the conditioning modality to extract useful features from the conditioning model. We sample $\mathbf{A}_0$ from Gaussian noise at $t_a = 0$ and use velocity predictions from $\mathcal{G}_{V2A}$ to progressively denoise it into $\mathbf{A}_1$. We perform sampling using a first-order Euler solver with 64 sampling steps. Text prompts can optionally be used for both modalities to condition generation. We perform A2V inference completely symmetrically.

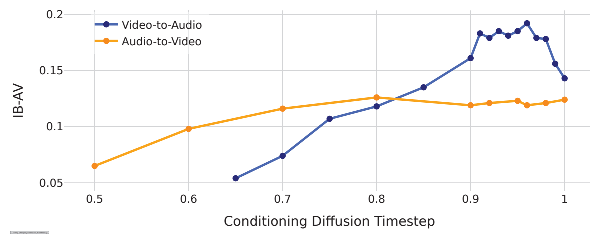

Figure 3. Visualization of Audio-to-Video and Video-to-Audio generation performance for various value of flow timesteps t for conditioning features. Best performance is achieved when conditioning features are close to be fully denoised, *i.e.* $t \in [0.8, 0.98]$.

	Prompt	FAD ↓	FD ↓	CLAP ↑	IS ↑	IB-AI ↑	IB-AV ↑	Ons. ACC ↑
Diff-Foley* [57]		11.00	28.71	-	7.88	0.115	0.121	0.140
S&H* [92]		29.22	66.51	-	2.09	0.179	0.189	0.128
FoleyCrafter* [103]		4.62	18.66	-	9.10	0.204	0.215	0.285
Frieren† [89]		3.52	16.50	-	7.78	0.178	0.184	0.301
V2A-Mapper† [86]		2.77	16.60	-	7.65	0.177	0.183	0.281
Ours (VGGSounds)		2.02	13.68	-	10.06	0.203	0.214	0.405
Ours-Joint		2.19	13.15	-	9.24	0.205	0.215	0.409
Ours		1.58	14.17	-	9.93	0.207	0.223	0.531
S&H* [92]	✓	10.72	25.44	0.186	6.06	0.30	0.318	0.08
FoleyCrafter* [103]	✓	2.99	18.05	0.212	10.88	0.208	0.219	0.307
Ours (VGGSounds)	✓	1.91	13.19	0.236	13.57	0.225	0.238	0.429
Ours-Joint	✓	2.23	13.74	0.224	13.44	0.212	0.224	0.471
Ours	✓	1.33	11.99	0.228	12.40	0.214	0.226	0.540

Table 1. Evaluation results on the V2A task on the VGGSounds [10] benchmark. * evaluated using the released code and checkpoints, † evaluated using results obtained directly from the authors. We report a variant of our method when training on VGGSounds only and when jointly train on A2V and V2A.

4. Experiments

This section introduces datasets (see Sec. 4.1) and evaluation protocol (see Sec. 4.2). We then compare our method to the state-of-the-art in V2A and A2V in Sec. 4.3 and present ablations in Sec. 4.4. We present extensive qualitative results supporting our evaluation in Appx. E and the *Website*.

4.1. Datasets

We train the base video model on an internal automatically captioned video dataset and the base audio model on ambient and music sound clips sourced from AutoReCap and AutoReCap-XL [27], FMA [19], Magna [43], MTG [6], and Song Describer [59]. The Fusion blocks are trained on 200,000 videos from VGGSounds [10] and the temporally-strong AudioSet [24] dataset. VGGSounds is an audio-visual dataset containing 126,000 samples. The temporally-aligned AudioSet includes 103,000 audio-video clips with strong audio-visual alignment. Additionally, we finetune the model on a high-quality internal captioned audio-video dataset comprising 250,000 clips. Audio captions are generated using the AutoCap [27] audio captioning model.

	IS ↑	IB-AI ↑	IB-AV ↑	FAD ↓	FD ↓	IS ↑	IB-AI ↑	IB-AV ↑	Ons. ACC ↑
	Movie Gen Benchmark			*VGGSounds*					
Conditioning timestep t:									
- Uniform samp.	4.19	0.103	0.111	2.25	16.41	8.47	0.179	0.192	0.419
- Fixed (ours)	6.53	0.150	0.155	2.05	14.25	9.82	0.208	0.220	0.409
Conditioning features type:									
- CAVP	4.62	0.116	0.120	1.41	15.52	9.13	0.173	0.180	0.370
- CLIP	3.26	0.143	0.150	1.82	14.44	9.56	0.229	0.240	0.388
- Diffusion features (ours)	6.53	0.150	0.155	2.05	14.25	9.82	0.208	0.220	0.409
Fusion block arrangement:									
- After Block-1	5.67	0.129	0.135	2.31	17.17	8.94	0.173	0.179	0.402
- After Block-11	5.92	0.138	0.140	2.30	17.36	9.10	0.187	0.192	0.393
- After Block-22	4.29	0.120	0.123	1.82	17.22	8.22	0.165	0.177	0.378
- Interleaved (ours)	6.53	0.150	0.155	2.05	14.25	9.82	0.208	0.220	0.409
Feature injection:									
- Concat. to text w/FT	4.12	0.124	0.128	1.72	15.74	10.64	0.187	0.196	0.365
- Direct alignment	2.38	0.029	0.030	3.95	23.22	5.94	0.091	0.098	0.294
- Direct alignment w/FT	3.15	0.028	0.030	2.40	17.71	8.75	0.134	0.138	0.262
- w/o symm. reinj.	4.57	0.059	0.062	1.83	15.47	8.55	0.142	0.151	0.392
- Symm. cross attention	5.70	0.118	0.123	2.17	16.04	8.65	0.192	0.197	0.401
- Fusion blocks (ours)	6.53	0.150	0.155	2.05	14.25	9.82	0.208	0.220	0.409

Table 2. V2A ablation results of our method. Variants marked with *FT* indicate backbone finetuning.

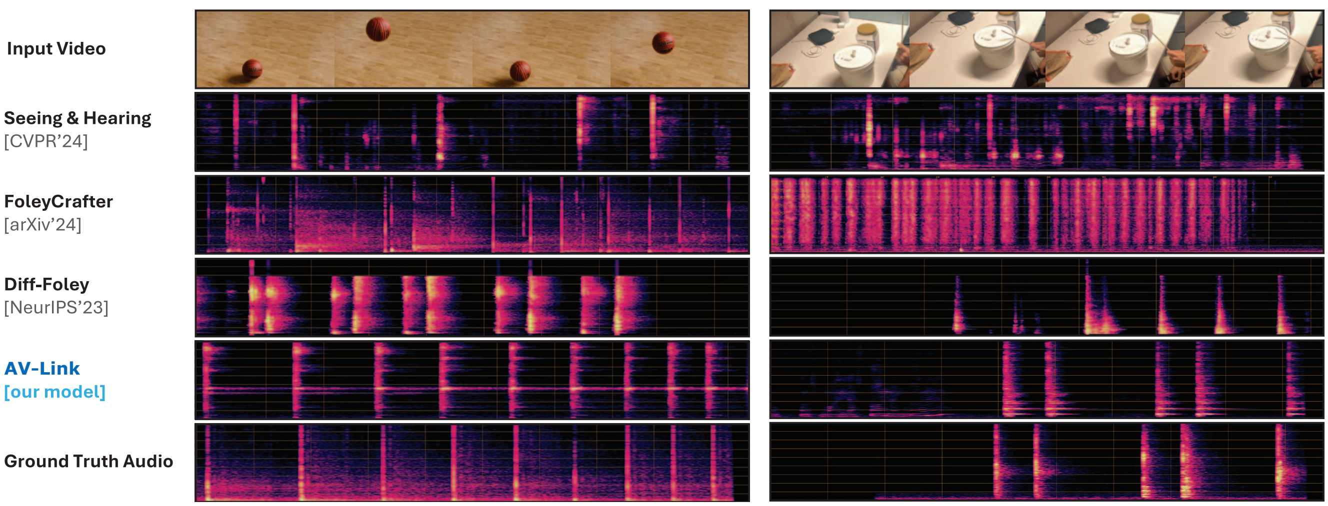

Figure 4. Qualitative V2A results. Our model achieved the best temporal alignment, matching closely the "bouncing" and "drumming" sounds entailed by the video modality. See the *Appendix* and *Website* for additional results.

4.2. Evaluation Protocol

We evaluate our model on the V2A and A2V tasks. We consider a conditioning video or audio input and sample a corresponding audio or video using our V2A or A2V model respectively. As prior work does not consistently adopt conditioning text prompts at inference [57, 67, 103], we report results for both settings. In all experiments, we use 8 fusion blocks (186M parameters), a learning rate of $1e-4$, and a batch size of 128 on 8 A100 GPUs for the ablation, increasing to 256 on 16 A100 GPUs for final experiments.

Baselines. For the task of V2A, we compare with Diff-foley [57], FoleyCrafter [103], Seeing and Hearing [92] (S&H), Frieren [89], V2A Mapper [86] and Movie Gen [67]. Diff-foley and V2A-Mapper conditions an audio generator on CAVP and CLIP, respectively by concatenating it with text embeddings. FoleyCrafter adapts a pre-trained audio generator for A2V using a semantic adapter with frame-based CLIP features [68] and trains a temporal adapter that aligns audio based on an onset mask predicted from the video, inspired by ControlNet. S&H uses a training-free approach that optimizes cosine similarity in ImageBind [25] between video and generated audio. Movie

	Prompt	FID ↓	FVD_{12} ↓	IB-AI ↑	IB-AV ↑
TempoTokens [99]		81.43	1489.35	0.112	0.114
Ours-Joint		51.63	297.24	0.112	0.126
Ours		39.10	266.93	0.160	0.168
TempoTokens [99]	✓	77.87	1275.50	0.167	0.173
Ours-Joint	✓	40.72	206.30	0.185	0.192
Ours	✓	37.62	158.71	0.191	0.201

Table 3. A2V Comparison on VGGSounds [10].

Configuration	Prompt	*A-Qual.*	*V-Qual.*	*AV-Qual.*	*Sem. Align.*	*Temp. Align*
Video-to-Audio:						
-Diff-Foley [57]		78.0	-	86.1	84.9	83.7
-V2A-Mapper [86]		64.8	-	66.8	65.2	71.6
-Frieren [89]		66.4	-	64.4	63.2	70.4
-Seeing and Hearing [92]		86.5	-	97.1	95.1	95.5
-Seeing and Hearing [92]	✓	76.2	-	87.7	86.8	88.1
-FoleyCrafter [103]		66.5	-	76.3	75.5	80.0
-FoleyCrafter [103]	✓	57.4	-	64.3	67.2	65.5
-Movie Gen [67]	✓	34.4	-	52.8	56.8	63.6
Audio-to-Video:						
-TempoTokens [99]		-	75.5	78.0	73.9	74.5
-TempoTokens [99]	✓	-	96.2	83.8	75.2	76.2

Table 4. User study comparing AV-Link against baselines. Results are in % of votes in favor of our method.

Gen and Frieren proposed to concatenate frame-wise features with audio tokens. For A2V, we compare with TempoToken [99], which optimizes latent tokens on a pre-trained video generator using BEATs [11] audio features.

Metrics We follow the protocol of previous work [57, 92, 99] and adopt 2048 videos from the VGGSounds [10] test set. We employ Fréchet Audio Distance (FAD) [53], Inception Score (IS) [73] and CLAP [90] score as measures for audio quality and prompt following. We also report FID [30] and FVD [81] as measures of video generation quality for the A2V task. Additionally, we leverage ImageBind [25] similarity scores between audio-image (IB-AI) and audio-video (IB-AV) modalities to measure semantic alignment. Onset detection accuracy (Onset ACC) [20] is used as the metric for measuring temporal alignment for the V2A task. For user studies, we ask evaluators to express preferences for one of two results based on *Audio Quality*, *Video Quality*, paired *Audio-Video Quality*, *Semantic Alignment* and *Temporal Alignment* between the two modalities. We report full details on user studies in Appx. E. Furthermore, We adopt the newly introduced Movie Gen V2A sound effects Benchmark [67], consisting of 527 *generated* video samples, for ablation studies and evaluating against Movie Gen. Note that since ground truth audio is not available for this benchmark, we omit audio-aligned metrics.

4.3. Results

Video-to-Audio We compare AV-Link with baseline approaches on the V2A task in Tab. 1. Our method achieves the highest scores across all metrics, except for IB-AI and IB-AV, where S&H reports superior values. This is expected, as S&H explicitly optimizes for these metrics as part of its approach, making these scores less indicative of overall quality. Notably, our method demonstrates significantly better temporal alignment, with an Onset Accuracy of 0.53 compared to the top baseline score of 0.3. Such improvements are evident with and without providing audio text descriptions. For a more balanced comparison, we also evaluate our model trained only on VGGSounds, showing that it still substantially outperforms baselines. This advantage is further confirmed by our user studies in Tab. 4, where AV-Link is consistently preferred, particularly for temporal alignment. Additionally, while Movie Gen achieves better sound quality due to its larger model (13B parameters), our method's audio-video temporal alignment is preferred 63.6% of the time. We also illustrate our model's strong temporal alignment in Fig. 4 and on the *Website*, where we perform inference on videos requiring precise temporal audio correspondence such as 'bouncing' and 'tapping' sounds. While baselines capture semantic content, they fall short in synchronizing the precise onset of these events.

Audio-to-Video On the A2V task, our method surpasses TempoTokens [83] in both video quality and audio-video alignment, as shown in Tab. 3. User studies in Tab. 4 also indicate that our approach is preferred for generation quality, semantic, and temporal alignment, with a preference greater than 73.9% across all settings. Additionally, qualitative results in Fig. 5 and on the *Website* demonstrate that our method produces videos that align closely with the audio signal, both semantically and temporally. Our generated video events (e.g., explosions) occur precisely in sync with corresponding audio events, showcasing the strength of our AV-Link in maintaining temporal coherence.

4.4. Ablation studies

This section evaluates the contribution of each method's design choices to output quality. Unless otherwise specified, we employ frozen audio and video generators, and focus on the V2A task, whose findings we expect to generalize to A2V due to the symmetric nature of our approach. We report the main ablation results in Tab. 2 for models trained on VGGSounds dataset. To obtain more accurate qualitative and quantitative comparisons between baselines, we perform evaluation using a single fixed seed for all test videos.

Which is the optimal conditioning flow timestep? Diffusion models learn distinct features at different layers and for different diffusion timesteps [2, 56, 94]. Thus, the choice of flow time distribution for the conditioning modality $t \sim p_t$ determines the quality of the conditioning features. Taking the V2A example, using a high-noise flow timestep t_v would destroy the video signal, leading to loss of conditioning information, while a flow time $t_v = 1$ (*i.e.* with no noise) would produce uninformative video activations, as for $t_v = 1$, Eq. (2) is minimized when the generator function is the identity $\mathcal{G}_V(\mathbf{V}_1, 1) = \mathbf{V}_1 - \mathbb{E}_{\mathbf{V}_0 \sim p_n = \mathcal{N}(0,\mathbb{I})}[\mathbf{V}_0] = \mathbf{V}_1$. We study the optimal choice of t_v empirically by training a generator $\mathcal{G}_{V2A}$ where the video time is uniformly sampled $t_v \sim \mathcal{U}(0, 1)$ during training. At inference, we conduct evaluation for different values of t_v and notice that the optimal value is close to a fully denoised video (see Fig. 3). Intuitively, small changes in the video can determine large variations in the sound, such as smoke from a gunshot, thus the model benefits from a noise level where high-frequency details are still not erased by noise. A symmetric analysis in Fig. 3 shows a similar behavior for t_a in the A2V task. In addition, training with a uniform distribution of t for the conditioning modality rather than fixing it at its optimal value led to a slower convergence and reduced performance (see Tab. 2). Therefore, after identifying the optimal conditioning timestep to be 0.96 for V2A and 0.8 for A2V, we train subsequent models using these fixed timesteps.

Can V2A and A2V models share the same parameters? We jointly train V2A and A2V models by sharing Fusion Block parameters of the two tasks. As reported in Tab. 1 and Tab. 3, using separate parameters contributes to a marginal but consistent improvement over parameter sharing.

Can both modalities be generated at the same time? Fig. 3 suggests a key issue with joint video-audio diffusion: Audio often depends on element that becomes only visible at the end of the sampling trajectory. For instance, accurately generating the sound of a gunshot requires high-frequency visual details, such as the smoke of an explosion, Similarly, sounds like thunder and explosions only become distant at the later sampler steps, making jointly denoising the two modalities challenging. This may explain the lower performance of current methods in this task [72, 87, 92], and suggest that audio-video generation benefits from factorization into video generation followed by V2A and vice-versa. Indeed, when we train AV-Link for joint audio-video generation, we observed a substantial drop in performance.

Are Fusion Blocks necessary? We evaluate several methodologies for injecting conditioning features into the model for which we show results in the *Feature injection* section of Tab. 2. *Concat. to text* follows Diff-foley [57] in concatenating conditioning features to the text embeddings. *Direct alignment* follows Movie Gen [67] in repeating the visual features to match the number of audio frames and sum them together. *Symm. cross attention* replaces the symmetric design of Fusion Blocks based on self attention with two separate blocks based on cross attention: the first for injecting features into the audio backbone, the second for the video. The Fusion Block design consistently outperforms the baselines. Additionally, we explore using video backbone as a static feature extractor without reinjecting the output back to the subsequent video DiT blocks. While this approach allows for caching the video features and thus improves sampling efficiency, it significantly underperforms our method that progressively refines the conditioning features by reinjecting them back to subsequent DiT blocks.

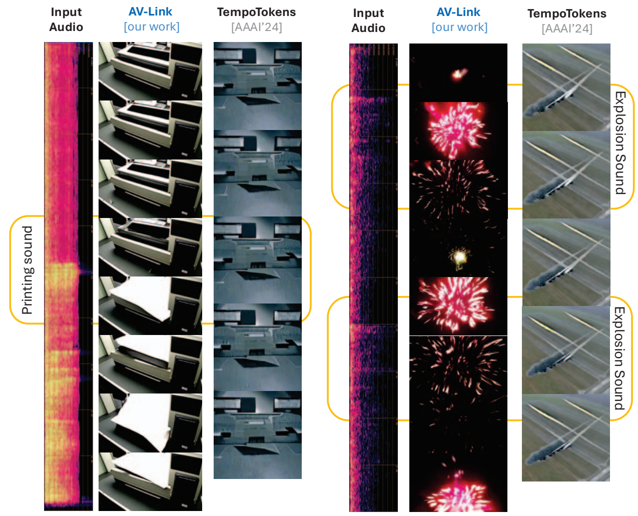

Figure 5. Qualitative results of A2V generation. Our model generates semantically and temporally aligned content capturing temporal events implied by the audio modality such as explosions and printing. See the *Website* for additional results.

How do diffusion features compare to pretrained representations? We experimented with various pretrained feature extractors in place of frozen generator activations as inputs to the Fusion Blocks. As shown in Tab. 2, diffusion features outperform all representations on the Movie Gen benchmark. On VGGSounds it shows the strongest temporal alignment as expressed by Onset ACC. User studies against baselines employing such representations (see Tab. 4) confirm improved temporal and semantic alignment.

Which layers yield the best conditioning features? In Tab. 2, we ablate over four different designs where the multimodal Fusion Blocks are placed at early video DiT layers, in the middle, in the end, or regularly interleaved with the video DiT blocks. Interleaving the Fusion Blocks outperforms the other variants, indicating that feature diversity and regular modality alignment benefit the model.

5. Conclusion

We present AV-Link, a unified approach for V2A and A2V generation. It leverages time-aligned activations from pretrained flow models, bypassing the use of specialized feature extractors. To enable this, we introduce a Fusion Block that allows for bidirectional information exchange between frozen generators for each modality through the use of temporally-aligned self attention and symmetric feature reinjection. Extensive ablations and comparisons show performance improvements over all metrics, with particular regard to temporal alignment.

Acknowledgements. The authors thank Maryna Diakonova and Ke Ma for their help in preparing the data and user studies. We also thank Chaoyang Wang, Sherwin Bahmani, Tsai-Shien Chen and Ziyi Wu for their guidance.

References

[1] Moayed Haji Ali, Andrew Bond, Tolga Birdal, Duygu Ceylan, Levent Karacan, Erkut Erdem, and Aykut Erdem. Vidstyleode: Disentangled video editing via stylegan and neuralodes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7523–7534, 2023. 1

[2] Dmitry Baranchuk, Ivan Rubachev, Andrey Voynov, Valentin Khrulkov, and Artem Babenko. Label-efficient semantic segmentation with diffusion models. *arXiv preprint arXiv:2112.03126*, 2021. 7

[3] Burak Can Biner, Farrin Marouf Sofian, Umur Berkay Karakaş, Duygu Ceylan, Erkut Erdem, and Aykut Erdem. Sonicdiffusion: Audio-driven image generation and editing with pretrained diffusion models. *arXiv preprint arXiv:2405.00878*, 2024. 2

[4] 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. 2

[5] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *CVPR*, 2023. 2

[6] Dmitry Bogdanov, Minz Won, Philip Tovstogan, Alastair Porter, and Xavier Serra. The mtg-jamendo dataset for automatic music tagging. In *Machine Learning for Music Discovery Workshop, ICML (ICML 2019)*, Long Beach, CA, United States, 2019. 5

[7] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, et al. Video generation models as world simulators. *OpenAI Blog*, 1:8, 2024. 1, 2

[8] Changan Chen, Puyuan Peng, Ami Baid, Zihui Xue, Wei-Ning Hsu, David Harwath, and Kristen Grauman. Action2sound: Ambient-aware generation of action sounds from egocentric videos. In *European Conference on Computer Vision*, pages 277–295. Springer, 2024. 2

[9] Gehui Chen, Guan'an Wang, Xiaowen Huang, and Jitao Sang. Semantically consistent video-to-audio generation using multimodal language large model, 2024. 2

[10] Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. Vggsound: A large-scale audio-visual dataset. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020. 2, 5, 7, 4

[11] Sanyuan Chen, Yu Wu, Chengyi Wang, Shujie Liu, Daniel Tompkins, Zhuo Chen, Wanxiang Che, Xiangzhan Yu, and Furu Wei. Beats: audio pre-training with acoustic tokenizers. In *ICML*, pages 5178–5193, 2023. 2, 7

[12] Tsai-Shien Chen, Aliaksandr Siarohin, Willi Menapace, Ekaterina Deyneka, Hsiang-wei Chao, Byung Eun Jeon, Yuwei Fang, Hsin-Ying Lee, Jian Ren, Ming-Hsuan Yang, and Sergey Tulyakov. Panda-70m: Captioning 70m videos with multiple cross-modality teachers. In *International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 3

[13] Yuanhong Chen, Yuyuan Liu, Hu Wang, Fengbei Liu, Chong Wang, Helen Frazer, and Gustavo Carneiro. Unraveling instance associations: A closer look for audio-visual segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26497–26507, 2024. 2

[14] Ziyang Chen, Prem Seetharaman, Bryan Russell, Oriol Nieto, David Bourgin, Andrew Owens, and Justin Salamon. Video-guided foley sound generation with multimodal controls. *arXiv preprint arXiv:2411.17698*, 2024. 2

[15] Ziyang Chen, Prem Seetharaman, Bryan Russell, Oriol Nieto, David Bourgin, Andrew Owens, and Justin Salamon. Video-guided foley sound generation with multimodal controls. *arXiv preprint arXiv:2411.17698*, 2024. 2

[16] Ho Kei Cheng, Masato Ishii, Akio Hayakawa, Takashi Shibuya, Alexander Schwing, and Yuki Mitsufuji. Taming multimodal joint training for high-quality video-to-audio synthesis. *arXiv preprint arXiv:2412.15322*, 2024. 2

[17] Xin Cheng, Xihua Wang, Yihan Wu, Yuyue Wang, and Ruihua Song. Lova: Long-form video-to-audio generation. *arXiv preprint arXiv:2409.15157*, 2024. 2

[18] Xiangxiang Chu, Jianlin Su, Bo Zhang, and Chunhua Shen. Visionllama: A unified llama backbone for vision tasks. In *European Conference on Computer Vision*, pages 1–18. Springer, 2024. 4

[19] Michaël Defferrard, Kirell Benzi, Pierre Vandergheynst, and Xavier Bresson. FMA: A dataset for music analysis. In *18th International Society for Music Information Retrieval Conference (ISMIR)*, 2017. 5

[20] Yuexi Du, Ziyang Chen, Justin Salamon, Bryan Russell, and Andrew Owens. Conditional generation of audio from video via foley analogies. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2426–2436, 2023. 2, 7

[21] Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. Clap learning audio concepts from natural language supervision. In *ICASSP*. IEEE, 2023. 1, 2, 3

[22] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024. 3, 1

[23] Zach Evans, CJ Carr, Josiah Taylor, Scott H. Hawley, and Jordi Pons. Fast timing-conditioned latent audio diffusion. In *ICML*, 2024. 1

[24] Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and

human-labeled dataset for audio events. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017. 5
[25] 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. 1, 2, 3, 6, 7
[26] Wei Guo, Heng Wang, Weidong Cai, and Jianbo Ma. Gotta hear them all: Sound source aware vision to audio generation. *arXiv preprint arXiv:2411.15447*, 2024. 2
[27] Moayed Haji-Ali, Willi Menapace, Aliaksandr Siarohin, Guha Balakrishnan, Sergey Tulyakov, and Vicente Ordonez. Taming data and transformers for audio generation. *arXiv preprint arXiv:2406.19388*, 2024. 1, 2, 3, 4, 5
[28] 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. 3
[29] Eric Hedlin, Gopal Sharma, Shweta Mahajan, Hossam Isack, Abhishek Kar, Andrea Tagliasacchi, and Kwang Moo Yi. Unsupervised semantic correspondence using stable diffusion, 2023. 2
[30] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *NIPS*, 2017. 7
[31] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2022. 1
[32] Jiawei Huang, Yi Ren, Rongjie Huang, Dongchao Yang, Zhenhui Ye, Chen Zhang, Jinglin Liu, Xiang Yin, Zejun Ma, and Zhou Zhao. Make-an-audio 2: Temporal-enhanced text-to-audio generation. *arXiv preprint arXiv:2305.18474*, 2023. 1, 4
[33] Zhiqi Huang, Dan Luo, Jun Wang, Huan Liao, Zhiheng Li, and Zhiyong Wu. Rhythmic foley: A framework for seamless audio-visual alignment in video-to-audio synthesis. *arXiv preprint arXiv:2409.08628*, 2024. 2
[34] Jaesung Huh, Jacob Chalk, Evangelos Kazakos, Dima Damen, and Andrew Zisserman. Epic-sounds: A large-scale dataset of actions that sound. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023. 3, 4
[35] Vladimir Iashin and Esa Rahtu. Taming visually guided sound generation. In *British Machine Vision Conference (BMVC)*, 2021. 2
[36] 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. 2, 4
[37] Yujin Jeong, Wonjeong Ryoo, Seunghyun Lee, Dabin Seo, Wonmin Byeon, Sangpil Kim, and Jinkyu Kim. The power of sound (tpos): Audio reactive video generation with stable diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7822–7832, 2023. 3
[38] Yang Jin, Zhicheng Sun, Ningyuan Li, Kun Xu, Hao Jiang, Nan Zhuang, Quzhe Huang, Yang Song, Yadong Mu, and Zhouchen Lin. Pyramidal flow matching for efficient video generative modeling. *arXiv preprint arXiv:2410.05954*, 2024. 1, 2
[39] Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation. In *CVPR*, 2024. 2
[40] Gwanghyun Kim, Alonso Martinez, Yu-Chuan Su, Brendan Jou, José Lezama, Agrim Gupta, Lijun Yu, Lu Jiang, Aren Jansen, Jacob Walker, et al. A versatile diffusion transformer with mixture of noise levels for audiovisual generation. *arXiv preprint arXiv:2405.13762*, 2024. 2, 3, 4
[41] Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 4
[42] Saksham Singh Kushwaha and Yapeng Tian. Vintage: Joint video and text conditioning for holistic audio generation. *arXiv preprint arXiv:2412.10768*, 2024. 2
[43] Edith Law, Kris West, Michael I. Mandel, Mert Bay, and J. S. Downie. Evaluation of algorithms using games: The case of music tagging. In *International Society for Music Information Retrieval Conference*, 2009. 5
[44] Junwon Lee, Jaekwon Im, Dabin Kim, and Juhan Nam. Video-foley: Two-stage video-to-sound generation via temporal event condition for foley sound. *arXiv preprint arXiv:2408.11915*, 2024. 2
[45] Seungwoo Lee, Chaerin Kong, Donghyeon Jeon, and Nojun Kwak. Aadiff: Audio-aligned video synthesis with text-to-image diffusion. *arXiv preprint arXiv:2305.04001*, 2023. 1, 3
[46] 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 *European Conference on Computer Vision*, pages 34–50. Springer, 2022. 2, 3, 4
[47] Jiachen Li, Qian Long, Jian Zheng, Xiaofeng Gao, Robinson Piramuthu, Wenhu Chen, and William Yang Wang. T2v-turbo-v2: Enhancing video generation model post-training through data, reward, and conditional guidance design. *arXiv preprint arXiv:2410.05677*, 2024. 2
[48] Ruilong Li, Shan Yang, David A. Ross, and Angjoo Kanazawa. Learn to dance with aist++: Music conditioned 3d dance generation. *arXiv*, 2021. 3, 4
[49] Ruiqi Li, Siqi Zheng, Xize Cheng, Ziang Zhang, Shengpeng Ji, and Zhou Zhao. Muvi: Video-to-music generation with semantic alignment and rhythmic synchronization. *arXiv preprint arXiv:2410.12957*, 2024. 2
[50] Susan Liang, Chao Huang, Yapeng Tian, Anurag Kumar, and Chenliang Xu. Language-guided joint audio-visual editing via one-shot adaptation. In *Proceedings of the Asian Conference on Computer Vision*, pages 1011–1027, 2024. 3
[51] Yan-Bo Lin, Yu Tian, Linjie Yang, Gedas Bertasius, and Heng Wang. Vmas: Video-to-music generation via semantic alignment in web music videos. *arXiv preprint arXiv:2409.07450*, 2024. 2

[52] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022. 3

[53] Haohe Liu, Zehua Chen, Yi Yuan, Xinhao Mei, Xubo Liu, Danilo Mandic, Wenwu Wang, and Mark D. Plumbley. Audioldm: Text-to-audio generation with latent diffusion models. *ICML*, 2023. 1, 7

[54] Haohe Liu, Qiao Tian, Yi Yuan, Xubo Liu, Xinhao Mei, Qiuqiang Kong, Yuping Wang, Wenwu Wang, Yuxuan Wang, and Mark D. Plumbley. AudioLDM 2: Learning holistic audio generation with self-supervised pretraining. *arXiv preprint arXiv:2308.05734*, 2023. 1, 4

[55] Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022. 3

[56] Grace Luo, Lisa Dunlap, Dong Huk Park, Aleksander Holynski, and Trevor Darrell. Diffusion hyperfeatures: Searching through time and space for semantic correspondence. *Advances in Neural Information Processing Systems*, 36:47500–47510, 2023. 2, 7

[57] Simian Luo, Chuanhao Yan, Chenxu Hu, and Hang Zhao. Diff-foley: Synchronized video-to-audio synthesis with latent diffusion models. *NeurIPS*, 36, 2024. 1, 2, 4, 5, 6, 7, 8, 3

[58] Zhengxiong Luo, Dayou Chen, Yingya Zhang, Yan Huang, Liang Wang, Yujun Shen, Deli Zhao, Jingren Zhou, and Tieniu Tan. Videofusion: Decomposed diffusion models for high-quality video generation. *CVPR*, 2023. 1

[59] Ilaria Manco, Benno Weck, Seungheon Doh, Minz Won, Yixiao Zhang, Dmitry Bogdanov, Yusong Wu, Ke Chen, Philip Tovstogan, Emmanouil Benetos, Elio Quinton, György Fazekas, and Juhan Nam. The song describer dataset: a corpus of audio captions for music-and-language evaluation. In *Machine Learning for Audio Workshop at NeurIPS 2023*, 2023. 5

[60] 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 *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 6607–6616, 2024. 2

[61] Xinhao Mei, Varun Nagaraja, Gael Le Lan, Zhaoheng Ni, Ernie Chang, Yangyang Shi, and Vikas Chandra. Foleygen: Visually-guided audio generation. In *2024 IEEE 34th International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6. IEEE, 2024. 2

[62] Willi Menapace, Aliaksandr Siarohin, Ivan Skorokhodov, Ekaterina Deyneka, Tsai-Shien Chen, Anil Kag, Yuwei Fang, Aleksei Stoliar, Elisa Ricci, Jian Ren, et al. Snap video: Scaled spatiotemporal transformers for text-to-video synthesis. In *CVPR*, 2024. 1, 2

[63] Andrew Owens, Phillip Isola, Josh McDermott, Antonio Torralba, Edward H. Adelson, and William T. Freeman. Visually indicated sounds. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 3, 4

[64] Santiago Pascual, Chunghsin Yeh, Ioannis Tsiamas, and Joan Serrà. Masked generative video-to-audio transformers with enhanced synchronicity. In *European Conference on Computer Vision*, pages 247–264. Springer, 2024. 2

[65] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4195–4205, 2023. 4, 5, 1

[66] Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *AAAI*, 2018. 4

[67] Adam Polyak, Amit Zohar, Andrew Brown, Andros Tjandra, Animesh Sinha, Ann Lee, Apoorv Vyas, Bowen Shi, Chih-Yao Ma, Ching-Yao Chuang, David Yan, Dhruv Choudhary, Dingkang Wang, Geet Sethi, Guan Pang, Haoyu Ma, Ishan Misra, Ji Hou, Jialiang Wang, Kiran Jagadeesh, Kunpeng Li, Luxin Zhang, Mannat Singh, Mary Williamson, Matt Le, Matthew Yu, Mitesh Kumar Singh, Peizhao Zhang, Peter Vajda, Quentin Duval, Rohit Girdhar, Roshan Sumbaly, Sai Saketh Rambhatla, Sam Tsai, Samaneh Azadi, Samyak Datta, Sanyuan Chen, Sean Bell, Sharadh Ramaswamy, Shelly Sheynin, Siddharth Bhattacharya, Simran Motwani, Tao Xu, Tianhe Li, Tingbo Hou, Wei-Ning Hsu, Xi Yin, Xiaoliang Dai, Yaniv Taigman, Yaqiao Luo, Yen-Cheng Liu, Yi-Chiao Wu, Yue Zhao, Yuval Kirstain, Zecheng He, Zijian He, Albert Pumarola, Ali Thabet, Artsiom Sanakoyeu, Arun Mallya, Baishan Guo, Boris Araya, Breena Kerr, Carleigh Wood, Ce Liu, Cen Peng, Dimitry Vengertsev, Edgar Schonfeld, Elliot Blanchard, Felix Juefei-Xu, Fraylie Nord, Jeff Liang, John Hoffman, Jonas Kohler, Kaolin Fire, Karthik Sivakumar, Lawrence Chen, Licheng Yu, Luya Gao, Markos Georgopoulos, Rashel Moritz, Sara K. Sampson, Shikai Li, Simone Parmeggiani, Steve Fine, Tara Fowler, Vladan Petrovic, and Yuming Du. Movie Gen: A Cast of Media Foundation Models. *arXiv*, 2024. 1, 2, 3, 4, 6, 7, 8

[68] 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. 1, 2, 6

[69] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research (JMLR)*, 2022. 4

[70] Aashish Rai and Srinath Sridhar. Egosonics: Generating synchronized audio for silent egocentric videos. *arXiv preprint arXiv:2407.20592*, 2024. 2

[71] Yong Ren, Chenxing Li, Manjie Xu, Wei Liang, Yu Gu, Rilin Chen, and Dong Yu. Sta-v2a: Video-to-audio generation with semantic and temporal alignment. *arXiv preprint arXiv:2409.08601*, 2024. 2

[72] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10219–10228, 2023. 2, 3, 8, 4

[73] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *NeurIPS*, 2016. 7

[74] Roy Sheffer and Yossi Adi. I hear your true colors: Image guided audio generation. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023. 2

[75] 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. 1

[76] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 2024. 4, 1

[77] Kun Su, Xiulong Liu, and Eli Shlizerman. From vision to audio and beyond: A unified model for audio-visual representation and generation. *arXiv preprint arXiv:2409.19132*, 2024. 3

[78] 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 *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 10853–10861, 2024. 2

[79] Vanessa Tan, Junghyun Nam, Juhan Nam, and Junyong Noh. Motion to dance music generation using latent diffusion model. In *SIGGRAPH Asia 2023 Technical Communications*, New York, NY, USA, 2023. Association for Computing Machinery. 2

[80] Ioannis Tsiamas, Santiago Pascual, Chunghsin Yeh, and Joan Serrà. Sequential contrastive audio-visual learning. *arXiv preprint arXiv:2407.05782*, 2024. 3

[81] 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. 7

[82] Mahtab Alizadeh Vandchali, Anastasios Kyrillidis, et al. One rank at a time: Cascading error dynamics in sequential learning. *arXiv preprint arXiv:2505.22602*, 2025. 2

[83] Ilpo Viertola, Vladimir Iashin, and Esa Rahtu. Temporally aligned audio for video with autoregression. *arXiv preprint arXiv:2409.13689*, 2024. 2, 7

[84] Ruben Villegas, Mohammad Babaeizadeh, Pieter-Jan Kindermans, Hernan Moraldo, Han Zhang, Mohammad Taghi Saffar, Santiago Castro, Julius Kunze, and Dumitru Erhan. Phenaki: Variable length video generation from open domain textual descriptions. In *International Conference on Learning Representations*, 2022. 1

[85] 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. 1

[86] Heng Wang, Jianbo Ma, Santiago Pascual, Richard Cartwright, and Weidong Cai. V2a-mapper: A lightweight solution for vision-to-audio generation by connecting foundation models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 15492–15501, 2024. 2, 5, 6, 7

[87] 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. 2, 3, 8

[88] 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 *Proceedings of the 32nd ACM International Conference on Multimedia (MM '24)*, page 10 pages, Melbourne, VIC, Australia, 2024. ACM. 2

[89] Yongqi Wang, Wenxiang Guo, Rongjie Huang, Jiawei Huang, Zehan Wang, Fuming You, Ruiqi Li, and Zhou Zhao. Frieren: Efficient video-to-audio generation with rectified flow matching. *arXiv e-prints*, pages arXiv–2406, 2024. 2, 5, 6, 7, 3

[90] Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *ICASSP*, 2023. 7

[91] Zhifeng Xie, Shengye Yu, Qile He, and Mengtian Li. Sonicvisionlm: Playing sound with vision language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26866–26875, 2024. 2

[92] 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*, pages 7151–7161, 2024. 1, 3, 4, 5, 6, 7, 8

[93] Hu Xu, Saining Xie, Xiaoqing Ellen Tan, Po-Yao Huang, Russell Howes, Vasu Sharma, Shang-Wen Li, Gargi Ghosh, Luke Zettlemoyer, and Christoph Feichtenhofer. Demystifying clip data. *arXiv preprint arXiv:2309.16671*, 2023. 1, 2

[94] Jiarui Xu, Sifei Liu, Arash Vahdat, Wonmin Byeon, Xiaolong Wang, and Shalini De Mello. Open-vocabulary panoptic segmentation with text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2955–2966, 2023. 7

[95] Qi Yang, Binjie Mao, Zili Wang, Xing Nie, Pengfei Gao, Ying Guo, Cheng Zhen, Pengfei Yan, and Shiming Xiang. Draw an audio: Leveraging multi-instruction for video-to-audio synthesis. *arXiv preprint arXiv:2409.06135*, 2024. 2

[96] Ruihan Yang, Hannes Gamper, and Sebastian Braun. Cmmd: Contrastive multi-modal diffusion for video-audio conditional modeling. *arXiv preprint arXiv:2312.05412*, 2023. 3, 4

[97] Xingyi Yang and Xinchao Wang. Diffusion model as representation learner. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 18938–18949, 2023. 2

[98] 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, 4

[99] 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*, pages 6639–6647, 2024. 1, 3, 7

[100] Fuming You, Minghui Fang, Li Tang, Rongjie Huang, Yongqi Wang, and Zhou Zhao. Momu-diffusion: On learning long-term motion-music synchronization and correspondence. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. 2

[101] Lijun Yu, José Lezama, Nitesh B Gundavarapu, Luca Versari, Kihyuk Sohn, David Minnen, Yong Cheng, Vighnesh Birodkar, Agrim Gupta, Xiuye Gu, et al. Language model beats diffusion–tokenizer is key to visual generation. *arXiv preprint arXiv:2310.05737*, 2023. 4

[102] Lin Zhang, Shentong Mo, Yijing Zhang, and Pedro Morgado. Audio-synchronized visual animation. In *European Conference on Computer Vision*, pages 1–18. Springer, 2024. 1, 3

[103] 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. 1, 2, 3, 4, 5, 6, 7

[104] Yanli Zhao, Andrew Gu, Rohan Varma, Liangchen Luo, Chien chin Huang, Min Xu, Less Wright, Hamid Shojanazeri, Myle Ott, Sam Shleifer, Alban Desmaison, Can Balioglu, Bernard Nguyen, Geeta Chauhan, Yuchen Hao, and Shen Li. Pytorch fsdp: Experiences on scaling fully sharded data parallel. *Proc. VLDB Endow.*, 16:3848–3860, 2023. 2

[105] Yuan Zhou, Qiuyue Wang, Yuxuan Cai, and Huan Yang. Allegro: Open the black box of commercial-level video generation model. *arXiv preprint arXiv:2410.15458*, 2024. 2