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

Supplementary Material

Appendix Contents

We discuss limitations in Sec. A. We then include details on the architecture (Sec. B), training and inference (Sec. C), and evaluation (Sec. E). We *highly* encourage the reader to visit the attached *Website* for extensive qualitative results and comparisons.

A. Limitations

Our base video backbone is a low-resolution and low-fps RGB model coupled with a high-resolution upsampler model. While the model achieves state-of-the-art V2A and A2V performance, leveraging a large high-resolution latent video model may further improve performance. Exploring model scaling to further improve feature quality is an exciting avenue for future work. Additionally, the feature reinjection alters activations in the conditioning modality generator. While beneficial, it introduces additional compute at each sampling step as caching prior activations becomes infeasible. Step distillation techniques reduce this effect by reducing the number of sampling steps and constitute an orthogonal line of work.

B. Architecture Details

We base our audio and video backbones on a shared DiT [65] architecture. For the video model, We use an pixel-based flow matching model with an initial patchification operation of 2×2 using a patch dimension of 1024. We employ 24 DiT blocks. Each DiT block is composed of a self attention operation, followed by a cross attention operation attending to text conditioning signals, and a final MLP. We use 16 heads for each attention operation and a hidden dimension of 4096 for the final MLP. Adaptive layer normalization is used within each block to condition the model on the flow time t. Each attention operator makes use of QK-Normalization [22] to improve training stability when trained in BF16 precision, and 3D RoPE [76] positional embeddings.

For the audio model, we employ a latent model with 24 DiT blocks with hidden dimension of 1024 for the patches. We use 16 attention heads for each self-attention and cross-attention operation and an MLP hidden size of 4096. We follow [27] for encoding the audio and converting the generated Mel-spectrograms to waveform. Both models have 576M trainable parameters.

The Fusion blocks similarly have a hidden dimension of 1024 and 16 heads for the self-attention operation. Their final MLP layers have a hidden dimension of 4096. Each Fusion block has 23.25M parameters and we use 8 Fusion blocks for all of our experiments.

C. Training and Inference Details

This section presents additional details on training and inference and discusses training and inference time.

C.1. Training Details

For all training phases, we train our models using the AdamW optimizer with a learning rate of 3e-4, beta factors of 0.9 and 0.99, epsilon of 1e-8, a weight decay of 0.01, and a 10,000-step warmup.

The base video model is trained for 250,000 steps on an automatically-captioned internal dataset with a total batch size of 512 on 16 A100 GPUs. The base audio models are trained for 100,000 with a total batch size of 1024 on 8 A100 GPUs. We drop text condition 10% of the time to enable classifier-free guidance (CFG) [31] during inference.

The fusion blocks are trained for 50,000 steps on 16 A100 GPUs with a batch size of 256. Ablation experiments are trained on 8 A100s with a total batch size of 128. We drop the generated modality text prompt (*e.g.* audio text prompt in V2A task) 50% of the time and the conditioning modality text prompt (*e.g.* video text prompt in V2A task)

Setting	FAD↓	FD↓	IS↑	IB-AI↑	IB-AV↑	Ons. Acc↑
576M Audio Backbone	2.48	15.87	8.97	0.188	0.198	41.74
288M Audio Backbone	3.02	17.22	9.07	0.168	0.178	40.60

Table 5. Quantitative comparison of AV-Link V2A trained with various Audio backbone sizes.

20% of the time. For all experiments, both the audio and video backbone are kept frozen unless otherwise specified.

C.2. Inference details

We perform inference starting from pure Gaussian noise for the modality to generate and use the model's velocity estimates together with an Euler sampler to progressively transform the noise to the clean generated sample. We found using classifier-free guidance on the conditioning modality to be instrumental in obtaining good multimodal alignment. When conditioning on more than one modality (*e.g.* video and audio text prompt), we drop both conditions simultaneously to compute the unconditional signal. In all of our experiments, we used 64 sampling steps and a CFG weight of 5.0. All of our generated video results are evaluated at the 512 x 288 pixels resolution.

C.3. Training and Inference Time

Training. The base video model was trained for 25 days, while the audio model and the fusion blocks were trained for 8 and 4 days, respectively. All experiments utilized PyTorch Fully Sharded Data Parallel (FSDP) [104] for efficient distributed training.

Inference. We measure a throughput of 27.85s per sample using a batch size of one on an A100 to perform 64 sampling steps for both the A2V and V2A tasks. This is a limitation of our method in the V2A task since it results in a slower sampling time compared with previous approaches [57, 103]. We make, however, two important considerations. First, a major use case for V2A is the sonification of generated soundless videos. Such generation, when performed by state-of-the-art large-scale text-to-video generators usually takes several minutes. Second, since our method relies on a flow model, distillation methods similar to the ones adopted by previous approaches [89] can significantly improve inference time by performing sampling in a few steps. We regard this direction as an interesting avenue for future work that is orthogonal to ours.

D. Additional Experiments

D.1. Freezing Audio and Video Generators

A core motivation behind AV-Link is to connect frozen audio and video generators to enable cross-modal generation without compromising the quality of single-modal generation. However, previous work showed benfitis from finetuning the audio generator for the video-to-audio task [57, 103] using full finetuning or LORA methods [82]. Here, we explore the effect of full fine-tuning of the audio or video backbone for V2A generation. Tab. 6 reports three training paradigms of AV-Link, trained for 35k iterations. Finetuning the Audio backbone improves V2A but hurts T2A performance on out-of-distribution benchmark (Clotho). Finetuning both backbones quickly degrades the conditioning signal (*i.e.* video activations), significantly reducing performance on all tasks, since the video backbone loses its strong features. Similar conclusions can be drawn from finetuning the video backbone for A2V. While finetunning is feasible with AV-Link, it prevents the creation of a modular and self-contained framework for T2V, T2A, V2A, and A2V.

Setting	T2A (Clotho)		T2V (VGGSounds)		V2A (VGGSounds)		
	IS↑	CLAPSIM↑	FVD↓	CLIPSIM↑	IS↑	IB-AV↑	Ons. Acc↑
Frozen Backbones	**9.01**	**24.4**	**1023.9**	**26.5**	8.97	0.198	41.74
FT Audio Backbone	7.05	18.6	**1023.9**	**26.5**	**9.82**	**0.207**	**41.88**
FT Audio&Video Backbone	5.48	14.9	8930.1	21.8	17.04	0.104	32.04

Table 6. Quantitative comparison of various training paradigms of AV-Link for the V2A task

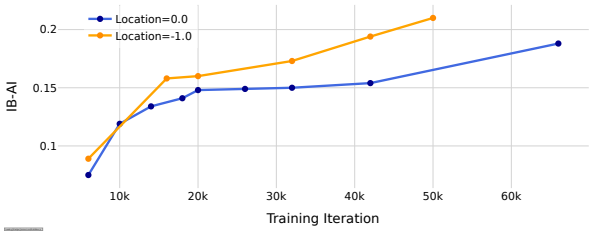

Figure 6. Comparison between different parametrizations for the Logit-Normal training distributions p_t for the flow timestep t. When the location (*i.e.* mean of the normal distribution) is shifted towards higher noise levels, we observe faster model convergence.

D.2. Heterogeneous Audio-Video Backbones

While we base our main experiments on homogeneous audio and video backbone, AV-Link does not require the audio and video backbones to be similar, as the Fusion Blocks can be configured to connect blocks at non-symmetric indices. We demonstrate this by fine-tunubg the base audio backbone with half of its blocks removed (*i.e.* 288M parameters) for 50k steps while keeping the full video backbone (576M parameters) unchanged. Then, the V2A framework is trained with $N = 8$ Fusion Blocks that connect block indices $(3N, \lfloor \frac{3}{2}N \rfloor)$ of the video and audio models, respectively. Tab. 5 shows that performance scales gracefully to the reduced audio capacity, showing that AV-Link is effective in connecting large-scale video generator with a smaller audio generator.

D.3. Noise Sampling Scheduler

For training the video and audio model, we use the Logit-Normal distribution p_t for the flow timestep t with location 0.0 and scale 1.0, following [22]. When training the Fusion block, which builds on pretrained audio and video backbones, we adopt a different approach. We noticed that early diffusion steps are the most critical for correctly following the conditioning modalities. Toward the end of the sampling path, however, the model relies more on the generated modality signal alone. Therefore, since multimodal alignment is the main purpose of the Fusion blocks, we shift the flow time training distribution to sample from more noisy steps. More specifically, Instead of parametrizing the normal distribution component as $\mathcal{N}(0, 1)$, we set it to $\mathcal{N}(-1, 1)$ and observe significantly faster convergence as shown in Fig. 6.

D.4. AV-Link for Joint Audio-Video Generation

A natural next step for AV-Link is joint audio-video diffusion, i.e. enforcing $t_a = t_v$ throughout training and sampling. In practice, however, our joint model lags far behind the cross-modal variants (V2A and A2V): generated clips show little semantic or temporal coherence between sound and videos. We suspect that there is an intrinsic bottleneck. Effective denoising of one modality often requires information that emerges only in later diffusion steps of the other. For example, low-frequency semantic cues in audio (the "boom" of an explosion) depend on high-frequency visual details that are resolved late in the video denoising trajectory—and vice versa. This interdependence makes simultaneous restoration of both streams difficult, a limitation shared by current joint generators [72, 87] while absent in cross-modal pipelines [89, 103]. A more promising avenue may be to diffuse a single, unified audio-video latent that captures their correlations from the outset, which we believe is an interesting direction to explore in future work.

E. Additional Evaluation Results and Details

This section presents additional evaluation results and details. We highly encourage the reader to check out the generated audios and videos on the *Website*.

E.1. Baselines Selection

Below, we include details on baseline selection and inference procedures.

Video-to-Audio. We compare our method against Diff-Foley [57], FoleyCrafter [103], Seeing and Hearing [92], Frieren [89], and Movie Gen A2V [67].

- **Frieren:** We use the Frieren (reflow) model, 64 sampling steps, CFG of 5.0, and the other recommended hyperparameters for inference.
- **Movie Gen A2V:** Since Movie Gen is a closed source model and audio samples are released for the Movie Gen Benchmark only, we include user studies and extensive comparison on the *Website* comparing our method against the released benchmark, showing that AV-Link achieves superior temporal alignment.
- **FoleyCrafter:** We employ FoleyCrafter with default settings for inference.
- **Seeing and Hearing:** We use the official code for V2A and follow default sampling parameters. We exclude Seeing and Hearing qualitative comparison without prompt as it produces barely audible sounds in this setting.
- **Diff-Foley:** We use the official code and set sampling steps to 64 and CFG to 5.0. For the rest of the parameters, we use the default setting.

For all of the baselines, we generate the audio at their recommended length from the full-length videos and crop it to 5.16s for a fair comparison with our method.

Audio-to-Video. To the best of the authors' knowledge, TempoToken [99] is the only in-the-wild A2V baseline with publicly available code. We exclude Seeing and Hearing [92] A2V results as their code is not available for this task. Additionally, we exclude joint audio-video generation methods (see Appx. E.2) such as MMDiffusion [72], as they were trained on the very limited Landscapes Videos dataset [46], which contains only 928 videos and lacks generalization beyond landscape scenarios. We also exclude sound-guided image animation methods [46, 102], as they address a fundamentally different task.

We use the official implementation of TempoToken with default parameters to generate 2-second videos. For the quantitative comparison, we crop our generated videos to 2 seconds for a fair comparison with TempoToken.

E.2. Joint Audio-Video Generation Baselines

In this work, we aim to address the tasks of A2V and V2A generation within a single framework. Some methods for joint video-audio generation [28, 40, 50, 77, 80, 96] are capable of operating under this conditional setting.

However, due to the difficulties in jointly modeling the audio and video modalities, these methods are often trained on domain specific datasets: Landscapes [46] is composed of 928 videos of natural landscapes; AIST++ [48] contains 1020 clips (5.2 hours) of dancing human sequences; GreatestHits [63] is composed of 977 videos featuring a drumstick hitting objects in a scene; EPIC-SOUNDS [34] encompasses 117.6k clips (100 hours) of cooking-related actions; Monologues [40] features 19.1M clips of talking people. The use of narrow distribution datasets coupled with the limited availability of source code and pretrained checkpoints limits the possibility of performing meaningful comparisons of joint audio-video generation on the broad data distribution modeled by AV-Link.

In the following, we discuss the considered joint audio-video baseline methods. MMDiffusion [72] provides checkpoints for models trained on Landscapes [46] and AIST++ [48] datasets only. We find that the Landscapes checkpoint overfits to its training dataset distribution, frequently replicating training samples. By analyzing 500 generated videos using the provided checkpoint, we found that the generated videos achieved an average of 0.825 CLIP similarity between the generated videos and the top-matching video from the training dataset. We include in Fig. 8 examples of such overfitting. This prevents the method from operating on videos outside these domains and does not allow for an informative comparison. Ishii *et al*. [36] provides a checkpoint trained on the GreatestHits [63] dataset only. While trained also on the Landscapes [46] and VGGSound [10] datasets, no A2V or V2A results were reported. Kim *et al*. [40] report results on the Landscapes [46], AIST++ [48] and the Monologues (talking heads) datasets with no code publicly available. CMMD [96] report results on the AIST++ [48] and EPIC-SOUNDS [34] datasets and do not provide code.

E.3. Additional V2A results

Fig. 7 shows additional V2A results comparing our method to baselines. To better showcase the capabilities of AV-Link, we record a series of in-the-wild videos that require a high degree of temporal alignment for the audio modality and run inference for all baselines. Our method produces highly aligned audio results that capture the audio semantic entailed by the visual modality, while baselines produce degraded results. We attribute this phenomenon to the lack of access to visual features that are precisely aligned with the video content. On the contrary, the use of activations from the video generation backbone allows AV-Link to produce precise alignment in this scenario.

E.4. User study details

We hire a team of professional annotators to perform the user studies. A total of 50 samples are generated for each method in the V2A and A2V tasks. We present users with paired videos with accompanying audio generated by different methods, and ask them to express a preference for one of the two based on *Audio Quality*, *Video Quality*, paired *Audio-Video Quality*, *Semantic Alignment* and *Temporal Alignment* between the two modalities. For the case of V2A evaluation, we formulate instructions for annotators for each such aspect as follows:

- **Audio Quality** *"Which audio has the best quality? Only listen to the audio and ignore the video content."*
- **Audio-Video Quality** *"Which audio-video pair has the best quality?"*
- **Semantic Alignment** *"Which audio is semantically closer to the video content?"*
- **Temporal Alignment** *"Which audio is more temporally aligned to the video content?"*

The formulation of questions for the A2V case is completely symmetric. For each generated pair of samples, we ask 5 different users to express a preference to increase the robustness of the evaluation.

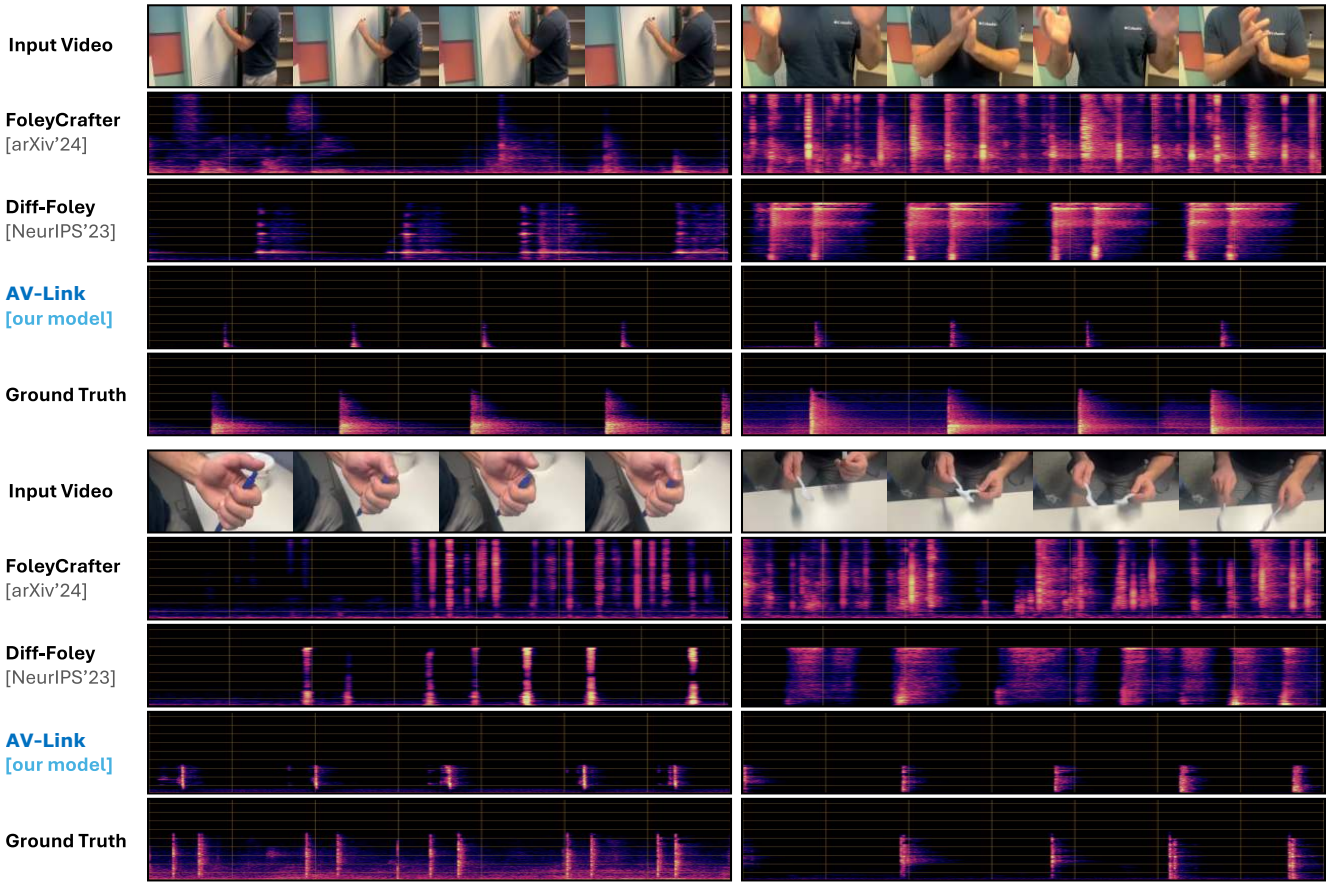

Figure 7. Qualitative V2A results comparing our method to baselines on in-the-wild videos captured by the authors that require precise temporal alignment. AV-Link produces audio signals that closely align to the visual modalities, while baselines often produce audio that is unrelated or not correctly synchronized with the visual content. See the *Website* for more results.

Figure 8. Examples MMdiffusion generated samples using their released checkpoint. We show that their model suffers from severe overfitting due to training on a limited dataset of 900 landscape videos.