Video-Guided Foley Sound Generation with Multimodal Controls

Supplementary Material

A.1. Implementation Details

DAC-VAE. We implemented and trained a modified version of the Descript Audio Codec (DAC) [52] using a variational autoencoder (VAE) [49]. In this approach, we replaced the residual vector quantizer (RVQ) with a VAE objective to encode continuous latents, enabling diffusion models to operate on continuous representations instead of discrete tokens. Our DAC-VAE was trained on audio waveforms at various sampling rates, allowing us to encode a 48kHz waveform into latents at a 40Hz sampling rate, with a feature dimension of 64. We train our DAC-VAE model on a variety of proprietary and licensed data spanning speech, music, and everyday sounds.

DiT architecture. Our DiT model has 12 layers, each with a hidden dimension of 1024, 8 attention heads, and an FFN (Feed-Forward Network) dimension of 3072, totaling 332M parameters. For the audio latents, we use an MLP (Multi-Layer Perceptron) to project them into 512-dimensional features. A separate MLP maps encoded visual features to 512 dimensions, followed by nearest-neighbor interpolation to upsample them fivefold (from 8Hz to 40Hz). Finally, we concatenate the audio and video features along the channel dimension to form 1024-dimensional inputs, which are then fed into the transformer.

Similar to VampNet [29], we use two learnable embeddings to differentiate between conditional input audio latents and noisy latents to be denoised, based on the conditional mask. We then sum the corresponding mask embeddings to the audio latents. During the inference, we create a conditional mask to achieve audio-conditioned generation.

Training details. We use the AdamW optimizer [48, 62] with a learning rate of 10^{-4} and apply a cosine decay schedule. Training begins with a linear warm-up phase for the first 4K iterations, followed by 599.6K iterations. We train our model with Exponential Moving Average (EMA) [68] with EMA decay of 0.99. Throughout the training, we randomly sample from combined datasets where 60% of training examples are from VGGSound and 40% from HQ-SFX. Within VGGSound samples, 60% are dedicated to video-text-tag-to-audio generation, the rest 40% are evenly distributed across different dropout variants (*i.e.*, video+tag, video+text, video-only, text+tag, text-only, tag-only, unconditional). For HQ-SFX samples, 60% are allocated to text-tag-to-audio generation, with the remaining cases divided as follows: 10% for text-only, 15% for tag-only, and 15% for unconditional audio generation.

A.2. Additional Experiments

Guidance scale ablation. We also examine the effect of the classifier-free guidance (CFG) scale, as shown in Tab. 7. The model shows similar performance with guidance weights between 3.0 and 7.0. On the FAD metrics, a higher guidance scale improves FAD@AUD but worsens FAD@VGG, suggesting that the model generates examples that align more with high-quality distributions. We use guidance scales of 3.0 and 5.0 for experiments in the main paper.

A.3. Human Studies

Videos and prompts. We handpicked 10 high-quality videos from the VGGSound test set, choosing examples that span a variety of categories and contain clear, easily perceivable temporal actions. We crafted two text prompts for each video: one matching the original category and another for a different target category, shown in Tab. 6. We then generated four 8-second samples for each video and randomly selected one for the final evaluation in the survey. For our model's generation, we use the "high quality" tag for inference.

Table 6. **Audio prompts for the user studies.** We note that the prompts are paired for the same video.

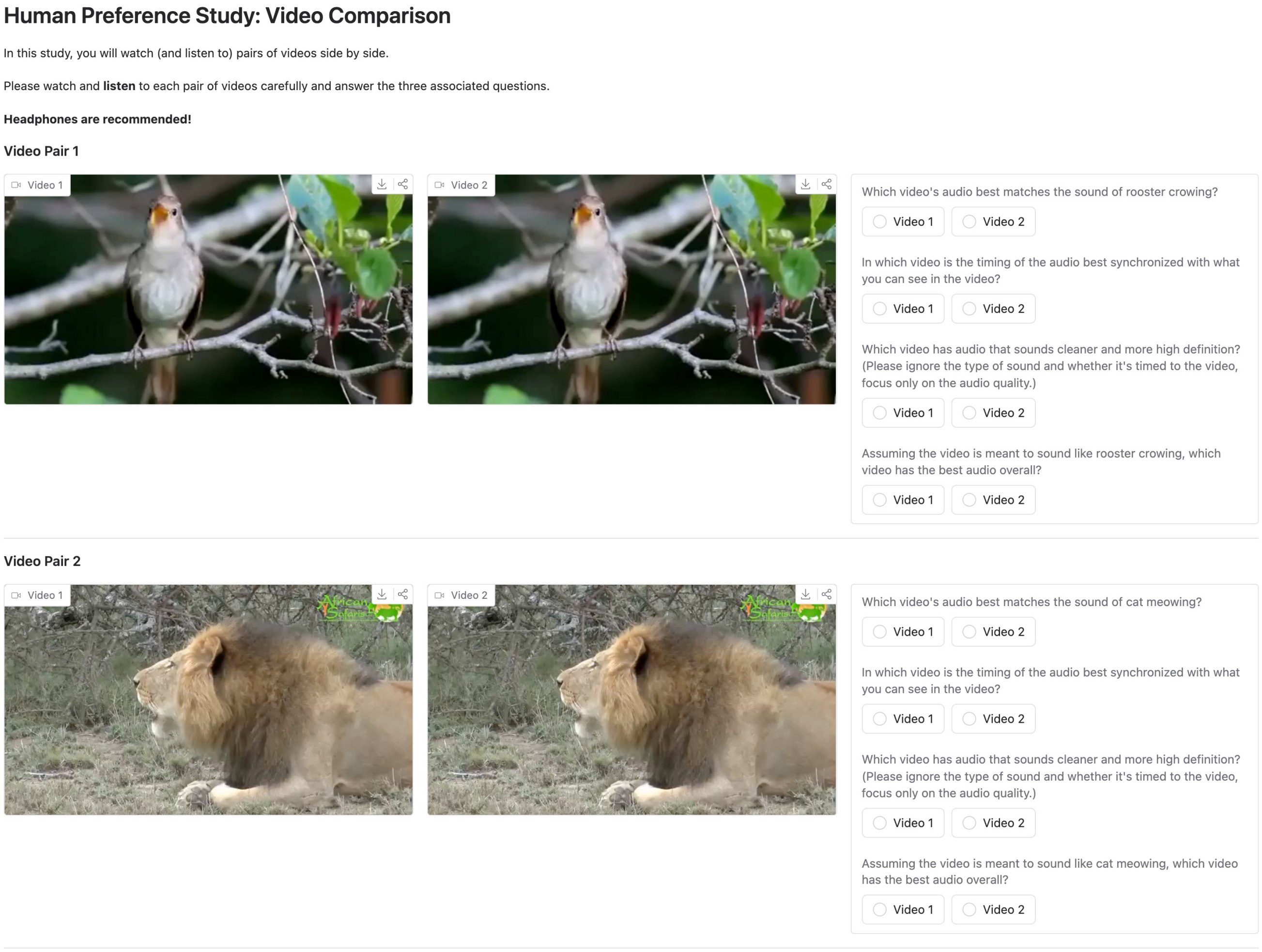
Original prompt	ReFoley prompt
playing cello	playing erhu
bird chirping	rooster crowing
dog barking	playing drum
typewriter	playing piano
gunshot	snare drum playing
chopping wood	kick drum playing
lion roaring	cat meowing
squeezing toys	cracking bones
playing trumpet	playing saxophone
playing golf	explosion

User study survey. In the survey, participants watched and listened to 20 pairs of videos comparing our method with FoleyCrafter [100]. We performed a forced-choice experiment where we randomized the left-right presentation order of the video pairs. For each video pair, participants were asked to respond to four questions:

1. Which video's audio best matches the sound of {`audio prompt`}?
2. In which video is the timing of the audio best synchronized with what you can see in the video?

Table 7. **Ablation study for classifier-free guidance scale on video-to-audio generation.** The best results are in **bold**.

	Variation	ImageBind ↑	CLAP ↑	AV-Sync ↓	FAD@VGG ↓	FAD@AUD ↓	KLD ↓
Ours	$\gamma = 1.0$	26.4	32.4	0.90	3.16	4.27	1.49
	$\gamma = 3.0$	28.0	34.4	0.80	**2.92**	4.62	**1.43**
	$\gamma = 4.0$	**28.1**	34.7	0.77	3.05	4.59	**1.43**
	$\gamma = 5.0$	28.0	**34.8**	0.77	3.27	4.48	**1.43**
	$\gamma = 7.0$	27.5	34.6	**0.75**	3.84	**4.21**	1.44



Human Preference Study: Video Comparison

In this study, you will watch (and listen to) pairs of videos side by side.

Please watch and **listen** to each pair of videos carefully and answer the three associated questions.

Headphones are recommended!

Video Pair 1

Which video's audio best matches the sound of rooster crowing?

○ Video 1 ○ Video 2

In which video is the timing of the audio best synchronized with what you can see in the video?

○ Video 1 ○ Video 2

Which video has audio that sounds cleaner and more high definition? (Please ignore the type of sound and whether it's timed to the video, focus only on the audio quality.)

○ Video 1 ○ Video 2

Assuming the video is meant to sound like rooster crowing, which video has the best audio overall?

○ Video 1 ○ Video 2

Video Pair 2

Which video's audio best matches the sound of cat meowing?

○ Video 1 ○ Video 2

In which video is the timing of the audio best synchronized with what you can see in the video?

○ Video 1 ○ Video 2

Which video has audio that sounds cleaner and more high definition? (Please ignore the type of sound and whether it's timed to the video, focus only on the audio quality.)

○ Video 1 ○ Video 2

Assuming the video is meant to sound like cat meowing, which video has the best audio overall?

○ Video 1 ○ Video 2

Figure 6. **Screenshot of Foley user study.** We show the screenshot from our user study survey. We show the instructions and the first two video pair examples and associated questions.

3. Which video has audio that sounds cleaner and more high definition? (Please ignore the type of sound and whether it's timed to the video, focus only on the audio quality.)
4. Assuming the video is meant to sound like {`audio prompt`}, which video has the best audio overall?

The first question evaluates the semantic alignment between the generated audio and the target audio prompt, ensuring that the sound matches the expected content. The second question evaluates the temporal alignment between the audio and video, focusing on how well the sound synchronizes with visual cues. The third question ignores content and timing to focus specifically on audio quality, examining aspects such as fidelity and production standards. Finally, the last question offers a holistic evaluation, determining which model produces the most effective overall audio. We show a screenshot of our user study survey including the instruction block, the first two video pairs, and associated questions in Fig. 6.