Action2Sound: Ambient-Aware Generation of Action Sounds from Egocentric Videos

Changan Chen[1*], Puyuan Peng[1*], Ami Baid[1], Zihui Xue[1], Wei-Ning Hsu[2], David Harwarth[1], and Kristen Grauman[1]

[1] University of Texas at Austin
[2] FAIR, Meta

Abstract. Generating realistic audio for human interactions is important for many applications, such as creating sound effects for films or virtual reality games. Existing approaches implicitly assume total correspondence between the video and audio during training, yet many sounds happen off-screen and have weak to no correspondence with the visuals—resulting in uncontrolled ambient sounds or hallucinations at test time. We propose a novel *ambient-aware* audio generation model, AV-LDM. We devise a novel audio-conditioning mechanism to learn to disentangle foreground action sounds from the ambient background sounds in in-the-wild training videos. Given a novel silent video, our model uses retrieval-augmented generation to create audio that matches the visual content both semantically and temporally. We train and evaluate our model on two in-the-wild egocentric video datasets Ego4D and EPIC-KITCHENS. Our model outperforms an array of existing methods, allows controllable generation of the ambient sound, and even shows promise for generalizing to computer graphics game clips. Overall, our work is the first to focus video-to-audio generation faithfully on the observed visual content despite training from uncurated clips with natural background sounds.

Keywords: audio-visual learning · egocentric video understanding · action sound generation

1 Introduction

We interact with objects around us in our daily lives and these actions often produce sound as a result of physical interactions, e.g., clicking on a mouse, closing a door, or cutting vegetables. The distinct characteristics of these *action sounds* depend upon the type of action being performed, the shape and material composition of the objects being acted upon, the amount of force being applied, and so forth. Vision not only captures *what* physical interaction happens but also informs us *when* the interaction happens, suggesting the possiblity of synthesizing semantically plausible and temporally synchronous action sounds from silent videos alone. This capability would accelerate many real-world applications, such as text-to-video generation, generating sound effects for films (Foley), or sound effect generation for virtual reality (VR) and video games.

[*] indicates equal contribution.

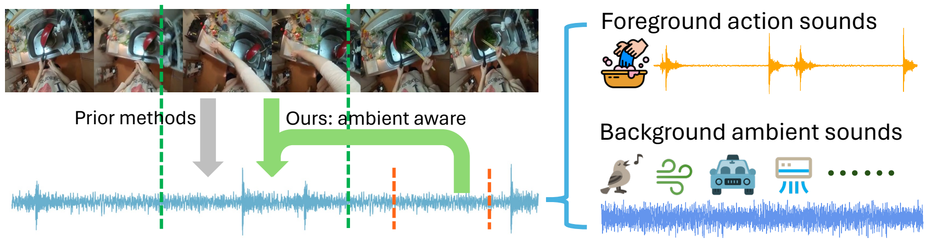


Fig. 1: Real-world audio consists of both foreground action sounds (whose causes are visible in the FoV) and background ambient sounds that are generated by sources offscreen. Whereas prior work is agnostic to this division when performing generation, our method is ambient-aware and disentangles action sound from ambient sound. Our key technical insight is how to train with in-the-wild videos exhibiting natural ambient sounds, while still learning to factor out their effects on generation. The green arrows reference how we condition generation on sound from a related, but time-distinct, video clip to achieve this.

Some prior work studies impact sound synthesis from videos [42, 48] while others target more general video-to-audio generation [25, 38]. All these methods *implicitly assume total correspondence between the video and audio* and aim to generate the whole target audio from the video. However, this strategy falls short for in-the-wild training videos, which are rife with off-screen ambient sounds, e.g., traffic noise, people talking, or A/C running. While some of these ambient sounds are weakly correlated with the visual scene, such as the wind blowing in an outdoor environment, many of them have no visual correspondence, such as off-screen speech or a stationary buzzing noise from the fridge. Existing methods are not able to disentangle action sounds from ambient sounds and treat them as a whole, leading to uncontrolled generation of ambient sounds at test time and sometimes even hallucination, e.g., random action or ambient sounds. This is particularly problematic for generating action sounds because they are often subtle and transient compared to the ambient sounds. For example, trained in the traditional way, a model given a scene that looks like a noisy restaurant risks generating “restaurant-like" ambient sounds, while ignoring the actual movements and activities of the foreground actions, such as a person stirring their coffee with a metal spoon.

How can we disentangle the foreground action sounds from background ambient sounds for in-the-wild video data *without* ground truth separated streams? Simply applying a noise removal algorithm on the target audio does not work well since in-the-wild blind source separation of general sounds from a single microphone is still an open challenge [49]. The key observation we have is that while action sounds are highly localized in time, ambient sounds tend to persist across time. Given this observation, we propose a simple but effective solution to disentangle ambient and action sounds: during training, in addition to the input video clip, we also condition the generation model on an audio clip from the same long video as the input video clip but from different timestamps. See Fig. 1. By doing so, we lift the burden of generating energy-dominating ambient sounds and encourage the model to focus on learning action cues from the visual frames to generate action sounds. At test time, we do not assume access to (even other clips of) the ground truth video/audio. Instead, we propose to retrieve an audio segment from the training set with an audio-visual similarity scoring model, inspired by recent ideas in retrieval-augmented generation (RAG) [19, 29, 34]. This benefits examples where the visual scene has a weak correlation with the ambient sound that is appealing to capture, e.g., outdoor environments.

Existing action sound generation work relies on either clean, manually-collected data that has a limited number of action categories [8,42,48], or videos crawled from YouTube based on predefined taxonomies [5,15,25]. To expand the boundary of action sound generation to in-the-wild human actions, we take advantage of recent large-scale egocentric video datasets [9, 18]. Though our model is not tailored to egocentric video in any way, there are two main benefits of using these datasets: 1) egocentric videos provide a close view of human actions compared to exocentric videos, where hand-object interactions are much smaller from a distance and often occluded, and 2) these datasets have timestamped narrations describing atomic actions. We design an automatic pipeline to extract and process clips from Ego4D, and curate Ego4D-Sounds with 1.2 million audio-visual action clips.

Our idea of disentangling action and ambient sounds implicitly in training is model-agnostic. In this paper, we instantiate it by designing an audio-visual latent diffusion model (AV-LDM) that conditions on both modality streams for audio generation. We evaluate our AV-LDM against recent works on a wide variety of metrics and show that our model outperforms the existing methods significantly on both Ego4D-Sounds and EPIC-KITCHENS. We conduct a human evaluation study that shows our model synthesizes plausible action sounds according to the video. **Please see/listen for yourself in our supplementary video!** We also show promising preliminary results on virtual reality game clips. To the best of our knowledge, this is the first work that demonstrates the disentanglement of foreground action sounds from background sounds for action-to-sound generation on in-the-wild videos.

2 Related Work

2.1 Action Sound Generation

A pioneering work for capturing human-generated action sounds collects videos where people hit, scratch, or prod objects with a drumstick [42]. This is an early inspirational effort, though it is by design limited in the type of actions. The robotics community also studies this problem by using robotic platforms to collect collision sounds and analyze or synthesize them from video [8, 14]. Other works approach this problem by building a simulation for collision events [13]; however, it is hard for computational approaches to simulate the impact or general action sounds due to the complexity of the physical interactions. Existing methods demonstrate good synthesis results when the data are noise-free. However, they are not equipped to learn from in-the-wild action videos, where the action sound is always coupled with ambient sound. We propose an ambient-aware model to deal with this issue head-on and also introduce the Ego4D-Sounds dataset to expand action sound synthesis to in-the-wild actions.

2.2 Egocentric Video Understanding with Audio

Understanding human activities in videos has long been a core challenge of computer vision. Early research studies activity recognition from exocentric videos such as UCF101 [47], Kinetics [27], or ActivityNet [12]. Recent work explores the egocentric setting and introduces large egocentric datasets such as Ego4D [18] or EPIC-KITCHENS [9]. Leveraging both the video and audio streams in egocentric videos, many interesting tasks are enhanced, such as action recognition [28], localization [44], active speaker localization [26], sounding object localization [22], and state-aware visual representations from audible interactions [39]. Most related to our work is SoundingActions [4] that learns visual representations of actions that make sounds, which is valuable for indexing and recognition problem settings, but ill-equipped for generation, as we show later when retrieval models [16] perform poorly for generating synchronized audio. All existing audio-visual learning for egocentric video focuses on perception, i.e., understanding what happens in the video. In contrast, we target the video-to-audio generation problem. Furthermore, relative to any of the above, our idea to implicitly learn to disentangle the action sound from ambient sounds is novel.

2.3 Diffusion Models and Conditional Audio Generation

Diffusion models have attracted significant attention recently because of their high fidelity generation [10,40,45]. Initially proposed for image generation [20,46], dthey have also been successfully applied to speech and audio generation [23, 33, 36, 43, 51]. Benefitting from classifier-free guidance [21] and large-scale representation learning, AudioLDM [36] and Make-An-Audio [23] have demonstrated successful diffusion-based *text*-to-audio generation. More recently, Diff-Foley [38] adapts latent diffusion models for video-to-audio generation by first conducting audio-video contrastive learning and then video-conditioned audio generation. While this approach demonstrates promising results, it does not address the background ambient sound problem. Inspired by recent work on retrieval-augmented generation (RAG) for text [3, 19, 29, 34] and image generation [2, 6], we show how our audio-conditioning insight carries over to inference time via a retrieval component of the model. Conditional video-to-audio generation conditions on either a physics prior to guide diffusion-based impact sound generation [48] or, in CondFoleyGen [11], another video clip to modify the characteristics of the action sound. Our method also considers additional conditioning signals to control the output, but for a very different purpose; our model is the first to address foreground/background sound disentanglement in generation.

3 Ambient-aware Action Sound Generation

We first discuss our high-level idea of how to guide the generation model to disentangle action sounds from ambient sounds. We then extend the latent diffusion models (LDM) to accommodate both audio and video conditions, which we name AV-LDM. We also discuss our pretraining stage.

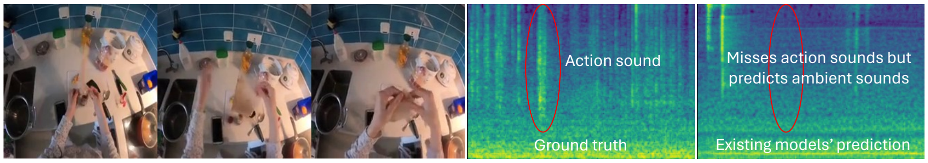


Fig. 2: Illustration of the harm of ambient sound in video-to-audio generation. In this example, this person is closing a packet of ginger powder, which makes some rustling sound (red circled in the middle). There is also some buzzing sound semantically irrelevant to the visual scene in the background, which dominates the energy of the spectrogram. On the right-hand side, we show a prediction made by a vanilla model that misses the action sound but predicts the ambient sound.

3.1 Action-to-Sound Generation

Given a video $V \in \mathbb{R}^{(T*S_V)\times H\times W\times 3}$, where T is the duration of the video and S_V is the video sample rate, and the accompanying audio waveform $A \in \mathbb{R}^{1\times(T*S_A)}$, where S_A is the audio sample rate, our goal is to model the conditional distribution $p(A|V)$ for video-to-audio generation. During training we observe natural video coupled with its audio, whereas at inference time we have only a silent video—e.g., could be an output from text-to-video generation, or a VR/video game clip, or simply a real-world video for which we want to generate new plausible sounds.

3.2 Disentangling Action and Ambient Sounds

Learning a video-to-audio generation model using in-the-wild egocentric videos is challenging because of entangled foreground action and background ambient sounds, as illustrated in Fig. 2. More specifically, the reasons are two-fold: 1) while action sounds are usually of very short duration, ambient sounds can last the entire clip, and therefore dominate the loss, leading to low-quality action sound generation; 2) while some ambient sounds might be semantically related to the visual scene such as bird chirping in the woods, in many cases, ambient sounds are difficult to infer from the visual scene because they are the results of the use of certain microphones, recording conditions, people speaking, off-screen actions, etc. Forcing a generation model to learn those background sounds from video results in hallucinations during inference (see examples in Fig. 6).

Therefore, it's beneficial to proactively disentangle action sounds and ambient sounds during training. However, separating in-the-wild ambient sounds is still an open challenge as recent models rely on supervised training on artificially mixed sounds, for which the ground truth complex masks can be obtained [49]. Simply applying off-the-shelf noise reduction methods to training data leads to poor performance, as we will show in Sec. 5.

While it is difficult to *explicitly* separate the ambient and action sound in the target audio, our key observation is that ambient sounds are usually fairly stationary across time. Given this observation, we propose a simple but effective method to achieve the disentanglement. During training, in addition to video clip V, we also provide the model an audio clip A_n that comes from the same training video but a different timestamp as the input video clip (see Fig. 3). Therefore,

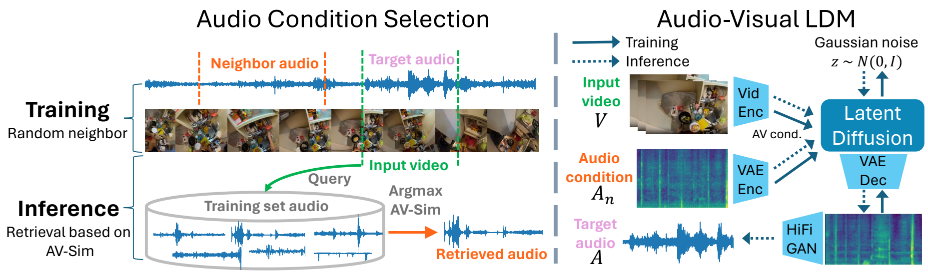


Fig. 3: Audio condition selection and the model architecture. **Left**: During training, we randomly sample a neighbor audio clip as the audio condition. For inference, we query the training set audio with the (silent) input video and retrieve an audio clip that has the highest audio-visual similarity with the input video using our trained AV-Sim model (Sec. 3.5). **Right**: We represent audio waveforms as spectrograms and use a latent diffusion model to generate the spectrogram conditioned on both the input video and the audio condition. At test time, we use a trained vocoder network to transform the spectrogram to a waveform.

instead of modeling $p(A|V)$, we model $p(A|V, A_n)$. Given the hypothesis that A_n is likely to share ambient sound characteristics with A, it can take away the burden of learning weakly correlated or even uncorrelated ambient sounds from visual input alone, and encourages the model to focus on learning action features from the visual input. For the selection of A_n, we randomly sample one audio clip from the nearest X clips in time. While there is no guarantee that the sampled audio shares exactly the same ambient sound with the target audio, their ambient sounds should largely overlap since they are close in time, which provides a consistent learning signal to help the model learn the disentanglement.

3.3 Retrieval Augmented Generation and Controllable Generation

While during training we have access to the clips in the same long video as the input clip, we of course cannot access that information at test time. How we select A_n at test time depends on the purpose of the generation. We consider two use cases: *action-ambient joint* generation and *action-focused* generation. In the first scenario, we would like the model to generate both the action sound and the ambient sound that is plausible for the visual environment. This is, for example, useful for generating sound effects for videos. In the latter scenario, we would like the model to focus the generation on action sounds and *minimize* ambient sounds, which is useful, for example, for generating sounds for games. Fig. 4 depicts the two scenarios.

For action-ambient joint generation, we want A_n to be semantically relevant to the visual scene. Inspired by recent work in retrieval augmented regeneration, we propose to retrieve audio such that:

$$A_n = \arg\max_{A_i \in \mathcal{D}} \text{AV-Sim}(A_i, V), \tag{1}$$

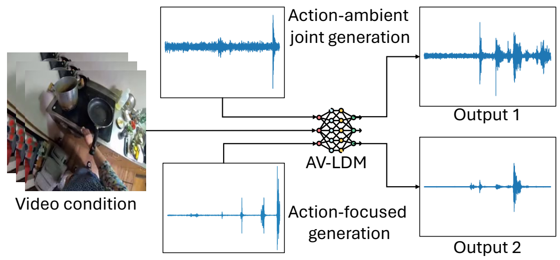


Fig. 4: Two inference settings: "action-ambient joint generation" and "action-focused generation". In the first setting, we condition on audio retrieved from the training set and aim to generate both plausible action and ambient sounds. In the second setting, we specify an audio file with low ambient sound and the model focuses on generating plausible action sounds while minimizing the ambient sounds.

where $\mathcal{D}$ is the dataset of all training audio clips and V is the (silent) input video. AV-Sim(A, V) is a similarity scoring function that measures the similarity between A and V, which we will cover in Sec. 3.5.

For action-focused generation, we want A_n to have minimal ambient level. We find simply filling A_n with all zeros results in poor performance, likely because it is too far out of the training distribution. Instead, we find conditioning the generation on a low-ambient sound will hint the model to focus on action sound generation and generate minimal ambient sound. See Sec. 5.3.

3.4 Audio-Visual Latent Diffusion Model

While the above idea of disentanglement is universal and not specific to any model architecture, here we instantiate this idea on diffusion models due to their success in audio generation [36, 38]. We extend the latent diffusion model to accommodate our audio-visual conditions, thus yielding an audio-visual latent diffusion model (AV-LDM).

Fig. 3 (right) shows the architecture of our model. During training, given audio waveform target A, we first compute the mel-spectrogram $x_0 \in \mathbb{R}^{T \times D_{\text{mel}}}$, where D_{mel} is the number of mel bins. We then use a pretrained Variational Autoencoder (VAE) to compress the mel-spectrogram x_0 to a latent representation $z_0 \in \mathbb{R}^{C' \times H' \times W'}$, where z_0 is the generation target of the LDM. We condition the generation on both the video feature $c_v \in \mathbb{R}^{T_v, D_c}$ and audio feature $c_a \in \mathbb{R}^{T_a, D_c}$. We extract the video feature with a pretrained video encoder (see Sec. 3.5) from V. We extract the audio feature from the audio condition A_n with the same VAE encoder and then transform the feature into 1-d vector with a multilayer perceptron (MLP).

Following [38], we use cross attention where the query is produced by z_t, which is the sample diffusion step t, and key and value are produced by concat$([\text{Pos}_v + c_v; \text{Pos}_a + c_a])$, where Pos denotes learnable positional embeddings. The model is trained with the denoising objective:

$$\mathcal{L} = \mathbb{E}_{t \sim \text{uniform}(1,T), z_0, \epsilon_t} \|\epsilon_t - \epsilon_\theta(\mathbf{x}_t, t, c_v, c_a)\|^2,$$

where ϵ_t is the standard Gaussian noise sampled for diffusion step t, and $\epsilon_\theta(\mathbf{x}_t, t, c_v, c_a)$ is the model estimation of it (θ represents model parameters).

The reverse process can be parameterized as:

$$p(z_T) = \mathcal{N}(0, I),$$
$$p_\theta(z_{t-1}|z_t) = \mathcal{N}(z_{t-1}; \frac{1}{\sqrt{\alpha_t}}\Big(z_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(z_t, t, c_v, c_a)\Big), \sigma_t^2 I),$$

where α_t and σ_t are determined by noise schedule of the diffusion process. To generate audio during inference, we first sample standard Gaussian noise z_T, and then apply classifier free guidance [21] to estimate $\tilde{\epsilon}_\theta$ as

$$\tilde{\epsilon}_t(z_t, t, c_v, c_a) = \omega\epsilon_\theta(z_t, t, c_v, c_a) + (1-\omega)\epsilon_\theta(z_t, t, \emptyset, \emptyset),$$

where $\emptyset$ denotes zero tensor. For the above estimation to be more precise, during training, we randomly replace c_v with $\emptyset$ with probability 0.2. As for c_a, we found dropping it even with even a small probability harms the performance, and therefore we always condition the LDM with c_a.

During inference, we use DPM-Solver [37] on LDM to sample a latent representation, which is then upsampled into a mel-spectrogram by the decoder of VAE. Lastly, we use a vocoder (HiFi-GAN [32]) model to generate waveform from the mel-spectrogram.

3.5 Audio-Visual Representation Learning

Generating semantically and temporally synchronized action sounds from video requires the video encoder to capture these relevant features. In addition, we would like to train a video model and an audio model whose representations align in the embedding space to support retrieval-augmented generation discussed in Sec. 3.3. For this purpose, we train a video encoder and audio encoder contrastively to optimize the following objective:

$$\text{AV-Sim}(A, V) = -\frac{1}{|\mathcal{B}|}\sum_{t\in\mathcal{B}} \log \frac{\exp(e_A^t e_V^t/\tau)}{\sum_{l\in\mathcal{B}} \exp(e_A^t e_V^l/\tau)},$$

where $\mathcal{B}$ is the current batch of data, e_A^t and e_V^t are normalized embeddings of the audio and video features, τ is a temperature parameter. To leverage the full power of narrations on Ego4D, we initialize the video encoder weights from models pre-trained on video and language from [35].

3.6 Implementation Details

We use Ego4D-Sounds (see Sec. 4) to train our AV-LDM. Video is sampled at 5FPS and audio is sampled at 16kHz. Video is passed through the pre-trained video encoder to produce condition features $c_v \in \mathbb{R}^{16\times768}$. The audio waveform is transformed into a mel-spectrogram with a hop size of 256 and 128 mel bins. The mel-spectrogram is then passed to the VAE encoder with padding in the temporal dimension to produce target $z_0 \in \mathbb{R}^{4\times16\times24}$. The audio condition is

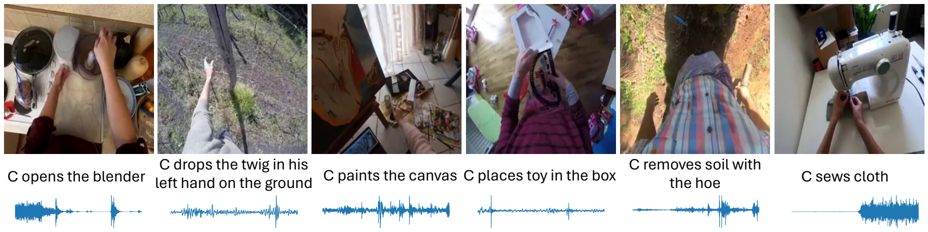


Fig. 5: Example clips in Ego4D-Sounds. We show one video frame, the action description, and the sound for each example. Note how these actions are subtle and long-tail, usually not present in typical video datasets.

Datasets	Clips	Language	Action Types
The Greatest Hits [42]	46.6K	✗	Hit, scratch, prod
VGG-Sound [5]	200K	Video tags	Not action-specific
EPIC-SOUNDS [24]	117.6K	Audio labels	Kitchen actions
Ego4D-Sounds	1.2M	Action narrations	In-the-wild actions

Table 1: Comparison with other audio-visual action datasets. Ego4D-Sounds not only has one order of magnitude more clips, but it is also coupled with language descriptions, supporting evaluation of sound generation based on semantics.

processed the same way except that we use an additional MLP to process VAE's output to produce $c_a \in \mathbb{R}^{24\times768}$. We load the weights of VAE and LDM from the pretrained Stable Diffusion to speed up training, similar to [38], and VAE is kept frozen during training. LDM is trained for 8 epochs with batch size 720 on Ego4D-Sounds with the AdamW optimizer with learning rate $1e-4$. During inference, we use 25 sampling steps with classifier-free guidance scale $\omega = 6.5$. For HiFi-GAN, we train it on a combination of 0.5s segments from Ego4D [18], Epic-Kitchens [24], and AudioSet [15]. We use AdamW to train HiFi-GAN with a learning rate of $2e-4$ and batch size of 64 for 120k steps. We set the number of random nearby audio samples $X = 6$. See more details in Supp.

4 The Ego4D-Sounds Dataset

Next we describe our efforts to curate Ego4D-Sounds, an audio-video dataset for human action sound generation. Our goal is to curate a high-quality dataset for action-audio correspondence for action-to-sound generation, addressing the issue of limited action types in the existing impact sound datasets [7, 42].

Ego4D [18] is an existing large-scale egocentric video dataset that has more than 3,600 hours of video recordings depicting hundreds of daily activities; 2,113 of those hours have audio available. It also has time-stamped narrations that are free-form sentences describing the current activity performed by the camera-wearer. We first utilize the narration timestamps in Ego4D to extract clips. However, not all clips have meaningful action sounds and there are many actions like "talk with someone", "look around", "turn around" that have low audio-visual correspondence. We then use an automatic pipeline to process all extracted clips to create the Ego4D-Sounds dataset, which has 1.2 million audio-visual action clips. Similarly, for the test set, we curate 11k clips for evaluation. See Supp. for more details on the data processing pipeline. We show examples in Fig. 5 and comparison with other datasets in Tab. 1.

For all resulting clips, we extract them as 3s clips with 224×224 image resolution at 30 FPS. For audio, we extract them as a single channel with a 16000 sample rate.

5 Experiments

5.1 Evaluation

To evaluate the performance of our model, we use the following metrics:

1. Fréchet Audio Distance (FAD) [30]: evaluates the quality of generated audio clips against ground truth audio clips by measuring the similarity between their distributions. We use the public pytorch implementation. [3]
2. Audio-visual synchronization (AV-Sync) [38]: a binary classification model that classifies whether the video and generated audio streams are synchronized. Following [38], we create negative examples by either shift audio temporally or sample audio from a different video clip. See more details in Supp.
3. Contrastive language-audio contrastive (CLAP) scores [50]: evaluates the semantic similarity between the generated audio and the action description. We finetune the CLAP model [4] on the Ego4D-Sounds data and compute scores for the generated audio and the narration at test time.

These metrics measure different aspects of generation collectively, including the distribution of generated samples compared to the ground truth clips, synchronization with the video, and the semantic alignment with the action description.

We compare with the following baseline methods:

1. Retrieval: we retrieve the audio from the training set using the AV-Sim model introduced in Sec. 3.5. This method represents retrieval-based generation models such as ImageBind [16].
2. Spec-VQGAN [25]: a video-to-audio model that generates audio based on a codebook of spectrograms. We run their pre-trained model on our test set.
3. Diff-Foley [38]: a recent LDM-based model. We follow their fine-tuning steps on egocentric videos to train on our dataset.

Neither learning-based model has the ability to tackle the ambient sound, whereas our model disentangles it from the action sound.

In addition, we also evaluate the following ablations: "w/o vocoder": we replace the trained HiFi-GAN vocoder with Griffin-Lim; "w/o cond": we remove the audio condition at training time; "w/o cond + denoiser": we use an off-the-shelf model to denoise the target audio [5]; "w/ random test cond": we use random audio from the training set as the condition instead of retrieving audio with the highest AV-Sim score.

[3] https://github.com/gudgud96/frechet-audio-distance
[4] https://github.com/LAION-AI/CLAP
[5] https://github.com/timsainb/noisereduce

	FAD ↓	AV-Sync (%)↑	CLAP↑
Ground Truth (Upper Bound)	0.0000	77.69	0.2698
Retrieval	1.8353	11.84	0.0335
Spec-VQGAN [25]	3.9017	7.12	0.0140
Diff-Foley [38]	3.5608	5.98	0.0346
Ours w/o vocoder	4.9282	29.60	0.1319
Ours w/o cond + denoiser	1.4676	1.09	0.0009
Ours w/o cond	1.4681	39.63	0.1418
Ours w/ random test cond	1.0635	28.74	0.1278
AV-LDM (Ours)	**0.9999**	**45.74**	**0.1435**

Table 2: Results on Ego4D-Sounds test set. We also report the performance of the ground truth audio, which gives the upper bound value for each metric.

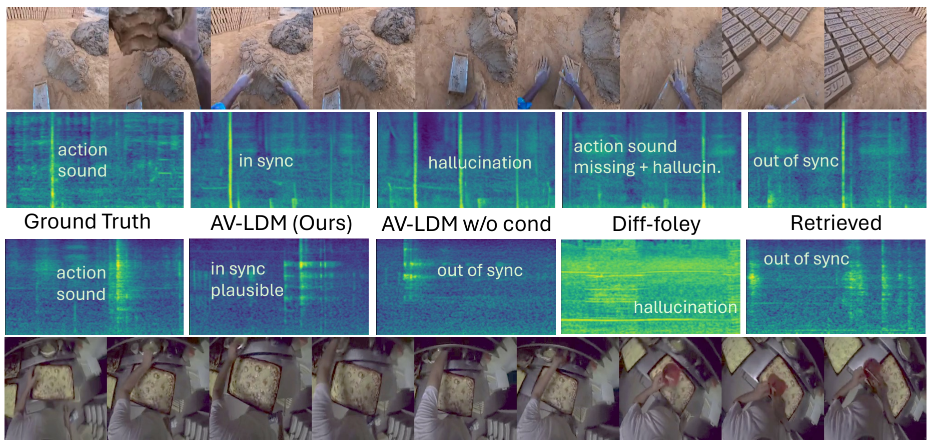
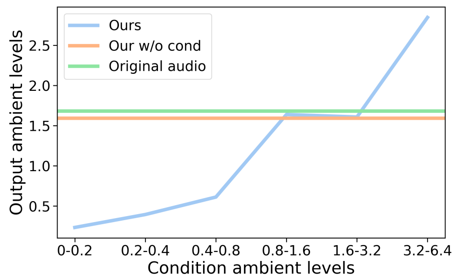


Fig. 6: Qualitative example. We show the frames of each video followed by the waveform/spectrogram of various baseline methods. Our model generates the most synchronized sounds.

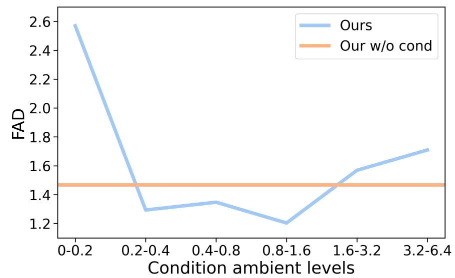
5.2 Results on Ego4D-Sounds

In this section, we evaluate the ambient-sound joint generation setting with retrieval augmented generation. The results are shown in Tab. 2. Compared to all three baselines, we outperform them on all three metrics by a large margin. While the Retrieval baseline retrieves natural sounds from the training set and has a low FAD score compared to Spec-VQGAN and Diff-Foley, both its AV-Sync accuracy and CLAP scores are very low. Diff-Foley has a higher performance than Spec-VQGAN since it has been trained on this task, but it still largely underperforms our model w/o cond, likely because their video features do not generalize to the egocentric setting well.

For ablations, "Ours w/o cond" has a much worse FAD score compared to the full model, showing the importance of our ambient-aware training. As expected, "Ours w/o cond + denoiser" has very low scores on AV-Sync and CLAP since existing noise reduction algorithms are far from perfect. We also test our model by conditioning it on a random audio segment at test time instead of the one



(a) Varying ambient level condition



(b) Audio generation accuracy (FAD)

Fig. 7: The achieved ambient level and accuracy of the prediction as a function of the input ambient levels. (a): we show the ambient level of our model changes according to the ambient level in the audio condition while the ambient level of "Ours w/o cond" and the original audio stay constant, illustrating the controllability of our model. (b) FAD is low for most input ambient levels unless it goes too extreme (too low or too high), showing our model generates high-quality action sounds even when varying output ambient levels.

retrieved with the highest audio-visual similarity and its performance also gets worse, verifying the effectiveness of our retrieval-based solution.

We show two qualitative examples in Fig. 6 comparing our model with several baselines and we show that our model synthesizes both more synchronized and more plausible sounds. To fully evaluate our results, it is important to view the supplementary video.

5.3 Ambient Sound Control

By disentangling action sounds from ambient sounds, our model allows taking any given sound as the condition at test time. To examine whether our model truly relies on the audio condition to learn the ambient sound information, we test the model by providing audio conditions of various ambient levels and then calculate the ambient level in the generated audio. The ambient level is defined as the lowest energy of any 0.5s audio segment in a 3s audio.

The results are shown in Fig. 7, where we also plot the ambient levels of "Ours w/o cond" and the original audio. Our model changes the ambient sound level according to the input ambient (shown in Fig. 7a) while still synthesizing plausible action sounds (shown in Fig. 7b). FAD spikes when the condition ambient is too low or too high, most likely because the generated ambient sound is out of distribution since the original audio always has some ambient sounds.

Fig. 8 shows example outputs from our model and several baselines. The examples show how our model generates plausible action sounds when conditioned on a low-ambient sound for action-focused generation. We can see that the action-focused setting generates similar action sounds as the action-ambient setting while having a minimal ambient level. While by definition we lack a good evaluation of this setting (there is no ground truth audio source separation for the data), our model shows an emerging capability of generating clean action sounds although it has never been explicitly trained to do so.

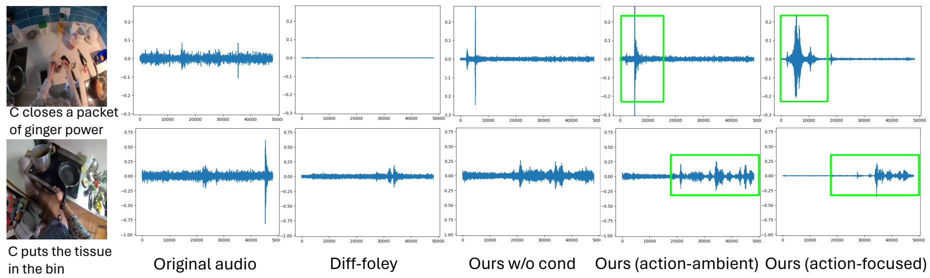


Fig. 8: Visualization of action-focused generation. For both examples, Diff-Foley [38], Ours w/o cond or Ours (action-ambient generation) generate plausible action sounds along with ambient sounds. In contrast, our model conditioned on a low ambient sound generates plausible action sounds (see green boxes) with minimal ambient sound.

	Action sound quality	Least ambient sound
Retrieval	12.5%	12.5%
Diff-Foley [38]	47.5%	12.5%
AV-LDM w/o cond	55.0%	17.5%
AV-LDM (action-focused)	60.0%	**97.5%**
AV-LDM (action-ambient)	**72.5%**	22.5%

Table 3: Survey results showing user preferences. Higher is better. Our model in the action-ambient joint generation setting scores highest for action sound quality, showing its ability to produce action-relevant sounds despite training with in-the-wild data. Ours in the action-focused generation setting scores highest for the least ambient sound, at a slight drop in action sound quality score, showing the ability to eliminate background sounds when requested by the user.

5.4 Human Evaluation

To further validate the performance of various models, we conduct a subjective human evaluation. In each survey, we provide 30 questions and each question has 5 videos with the same visuals but different audio samples. For each video, we ask the participant to select the video(s) whose audio 1) is most semantically plausible and temporally synchronized with the video and 2) has the least ambient sounds. We invite 20 participants to complete the survey and compute the average voting for all 30 examples. See the survey interface and guidelines in Supp.

Tab. 3 shows the results. Overall, all learning-based methods generate reasonable action sounds while our model (action-ambient) has the highest score for action-sound quality compared to other methods. Although ours (action-focused) has a slightly lower action-sound score, it has significantly less ambient sound. This is likely because sometimes the low-ambient condition can lead the model to suppress some minor action sounds.

5.5 Results on EPIC-KITCHENS

To evaluate whether our model generalizes to other datasets, we also test our model on the EPIC-KITCHENS dataset. We first sample 1000 3s clips on EPIC-KITCHENS and then evaluate the retrieval baseline, Diff-Foley, Ours w/o cond,

	GT	Retrieval	Diff-Foley	Ours w/o cond	AV-LDM (Ours)
FAD ↓	0.0000	1.9618	3.4649	1.4731	**1.3200**
AV-Sync (%) ↑	73.94	13.84	14.19	50.42	**59.26**

Table 4: Results on Epic-Kichens. GT stands for Ground Truth.

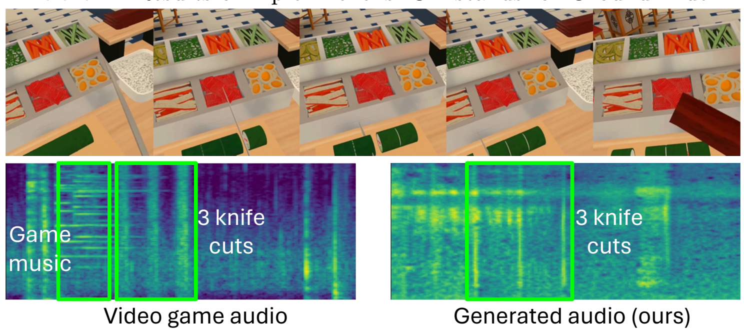


Fig. 9: We apply our model on a VR cooking game clip where the person cuts a sushi roll three times. Our model successfully predicts the 3 cutting sounds.

and our full model on these data and then compute the FAD and AV-Sync scores for them.

Results are shown in Tab. 4. Similar to what we observe on Ego4D-Sounds, our model outperforms other models on FAD and AV-Sync by a large margin, showing ours learns better to generate action sounds from visuals, which also transfer to other datasets.

5.6 Demo on VR Cooking Game

One compelling application of action-to-sound generation is to generate sound effects for games in virtual reality, where simulating complex hand-object interactions is non-trivial. To examine whether our learned model generalizes to VR games, we collect game videos of a cooking VR game "Clash Of Chefs" from YouTube and test our model without fine-tuning. Preliminary results suggest our model can generate synced action sounds (see Fig. 9 and Supp). This suggests promising future in learning action-to-sound models from real-world egocentric videos and applying them to VR games to give a game user an immersive audio-visual experience that dynamically adjusts to their own actions.

6 Conclusion

We investigate the problem of generating sounds for human actions in egocentric videos. We propose an ambient-aware approach that disentangles the action sound from the ambient sound, allowing successful generation after training with diverse in-the-wild data, as well as controllable conditioning on ambient sound levels. We show that our model outperforms existing methods and baselines—both quantitatively and through human subject studies. Overall, it significantly broadens the scope of relevant training sources for achieving action-precise sound generation. In future work we aim to explore the possibilities for sim2real translation of our learned audio generation models to synthetic imagery inputs, e.g., for VR game applications.

Acknowledgments: UT Austin is supported in part by the IFML NSF AI Institute. Wei-Ning Hsu is advising only and all the work and data processing were done outside of Meta.

References

1. Bertasius, G., Wang, H., Torresani, L.: Is space-time attention all you need for video understanding? In: Proceedings of the 38th International Conference on Machine Learning. pp. 813–824 (2021) 19, 20
2. Blattmann, A., Rombach, R., Oktay, K., Ommer, B.: Retrieval-augmented diffusion models. ArXiv **abs/2204.11824** (2022), `https://api.semanticscholar.org/CorpusID:248377386` 4
3. Borgeaud, S., Mensch, A., Hoffmann, J., Cai, T., Rutherford, E., Millican, K., van den Driessche, G., Lespiau, J.B., Damoc, B., Clark, A., de Las Casas, D., Guy, A., Menick, J., Ring, R., Hennigan, T.W., Huang, S., Maggiore, L., Jones, C., Cassirer, A., Brock, A., Paganini, M., Irving, G., Vinyals, O., Osindero, S., Simonyan, K., Rae, J.W., Elsen, E., Sifre, L.: Improving language models by retrieving from trillions of tokens. In: International Conference on Machine Learning (2021), `https://api.semanticscholar.org/CorpusID:244954723` 4
4. Chen, C., Ashutosh, K., Girdhar, R., Harwath, D., Grauman, K.: Soundingactions: Learning how actions sound from narrated egocentric videos. In: CVPR (2024) 4
5. Chen, H., Xie, W., Vedaldi, A., Zisserman, A.: Vggsound: A large-scale audio-visual dataset. In: ICASSP (2020) 3, 9
6. Chen, W., Hu, H., Saharia, C., Cohen, W.W.: Re-imagen: Retrieval-augmented text-to-image generator. ArXiv **abs/2209.14491** (2022), `https://api.semanticscholar.org/CorpusID:252596087` 4
7. Clarke, S., Gao, R., Wang, M., Rau, M., Xu, J., Wang, J.H., James, D.L., Wu, J.: Realimpact: A dataset of impact sound fields for real objects. In: Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (2023) 9
8. Clarke, S., Heravi, N., Rau, M., Gao, R., Wu, J., James, D., Bohg, J.: Diffimpact: Differentiable rendering and identification of impact sounds. In: 5th Annual Conference on Robot Learning (2021) 3
9. Damen, D., Doughty, H., Farinella, G.M., Fidler, S., Furnari, A., Kazakos, E., Moltisanti, D., Munro, J., Perrett, T.: Scaling egocentric vision: The epic-kitchens dataset. In: ECCV (2018) 3, 4
10. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. ArXiv **abs/2105.05233** (2021), `https://api.semanticscholar.org/CorpusID:234357997` 4
11. Du, Y., Chen, Z., Salamon, J., Russell, B., Owens, A.: Conditional generation of audio from video via foley analogies. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2426–2436 (2023) 4
12. Fabian Caba Heilbron, Victor Escorcia, B.G., Niebles, J.C.: Activitynet: A large-scale video benchmark for human activity understanding. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 961–970 (2015) 4
13. Gan, C., Schwartz, J., Alter, S., Mrowca, D., Schrimpf, M., Traer, J., De Freitas, J., Kubilius, J., Bhandwaldar, A., Haber, N., Sano, M., Kim, K., Wang, E., Lingelbach, M., Curtis, A., Feigelis, K., Bear, D.M., Gutfreund, D., Cox, D., Torralba, A.,

DiCarlo, J.J., Tenenbaum, J.B., McDermott, J.H., Yamins, D.L.K.: Threedworld: A platform for interactive multi-modal physical simulation. In: NeurIPS Datasets and Benchmarks Track (2021) 3

14. Gandhi, D., Gupta, A., Pinto, L.: Swoosh! rattle! thump! - actions that sound. In: RSS (2022) 3
15. 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: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 776–780 (2017) 3, 9, 19
16. 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) 4, 10
17. Gong, Y., Chung, Y.A., Glass, J.: Ast: Audio spectrogram transformer. In: InterSpeech (2021) 19, 20
18. Grauman, K., Westbury, A., Byrne, E., Chavis, Z., Furnari, A., Girdhar, R., Hamburger, J., Jiang, H., Liu, M., Liu, X., Martin, M., Nagarajan, T., Radosavovic, I., Ramakrishnan, S.K., Ryan, F., Sharma, J., Wray, M., Xu, M., Xu, E.Z., Zhao, C., Bansal, S., Batra, D., Cartillier, V., Crane, S., Do, T., Doulaty, M., Erapalli, A., Feichtenhofer, C., Fragomeni, A., Fu, Q., Gebreselasie, A., Gonzalez, C., Hillis, J., Huang, X., Huang, Y., Jia, W., Khoo, W., Kolar, J., Kottur, S., Kumar, A., Landini, F., Li, C., Li, Y., Li, Z., Mangalam, K., Modhugu, R., Munro, J., Murrell, T., Nishiyasu, T., Price, W., Puentes, P.R., Ramazanova, M., Sari, L., Somasundaram, K., Southerland, A., Sugano, Y., Tao, R., Vo, M., Wang, Y., Wu, X., Yagi, T., Zhao, Z., Zhu, Y., Arbelaez, P., Crandall, D., Damen, D., Farinella, G.M., Fuegen, C., Ghanem, B., Ithapu, V.K., Jawahar, C.V., Joo, H., Kitani, K., Li, H., Newcombe, R., Oliva, A., Park, H.S., Rehg, J.M., Sato, Y., Shi, J., Shou, M.Z., Torralba, A., Torresani, L., Yan, M., Malik, J.: Ego4d: Around the world in 3,000 hours of egocentric video. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 18973–18990 (2022) 3, 4, 9, 19, 20
19. Guu, K., Lee, K., Tung, Z., Pasupat, P., Chang, M.W.: Realm: Retrieval-augmented language model pre-training. ArXiv **abs/2002.08909** (2020), `https://api.semanticscholar.org/CorpusID:211204736` 3, 4
20. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: NeurIPS (2020) 4
21. Ho, J., Salimans, T.: Classifier-free diffusion guidance (2022) 4, 8
22. Huang, C., Tian, Y., Kumar, A., Xu, C.: Egocentric audio-visual object localization. In: CVPR (2023) 4
23. Huang, R., Huang, J.B., Yang, D., Ren, Y., Liu, L., Li, M., Ye, Z., Liu, J., Yin, X., Zhao, Z.: Make-an-audio: Text-to-audio generation with prompt-enhanced diffusion models. ArXiv **abs/2301.12661** (2023), `https://api.semanticscholar.org/CorpusID:256390046` 4
24. Huh, J., Chalk, J., Kazakos, E., Damen, D., Zisserman, A.: Epic-sounds: A large-scale dataset of actions that sound. In: ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1–5 (2023) 9, 19
25. Iashin, V., Rahtu, E.: Taming visually guided sound generation. In: BMVC (2021) 2, 3, 10, 11
26. Jiang, H., Murdock, C., Ithapu, V.K.: Egocentric deep multi-channel audio-visual active speaker localization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10544–10552 (2022) 4

27. Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., Suleyman, M., Zisserman, A.: The kinetics human action video dataset. CoRR **abs/1705.06950** (2017), http://arxiv.org/abs/1705.06950 4
28. Kazakos, E., Nagrani, A., Zisserman, A., Damen, D.: Epic-fusion: Audio-visual temporal binding for egocentric action recognition. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 5491–5500 (2019) 4
29. Khandelwal, U., Levy, O., Jurafsky, D., Zettlemoyer, L., Lewis, M.: Generalization through memorization: Nearest neighbor language models. ArXiv **abs/1911.00172** (2019), https://api.semanticscholar.org/CorpusID:207870430 3, 4
30. Kilgour, K., Zuluaga, M., Roblek, D., Sharifi, M.: Fréchet audio distance: A metric for evaluating music enhancement algorithms. arxiv (2018) 10
31. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: ICLR (2015) 19
32. Kong, J., Kim, J., Bae, J.: Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. Advances in Neural Information Processing Systems **33**, 17022–17033 (2020) 8, 19
33. Kong, Z., Ping, W., Huang, J., Zhao, K., Catanzaro, B.: Diffwave: A versatile diffusion model for audio synthesis. ArXiv **abs/2009.09761** (2020), https://api.semanticscholar.org/CorpusID:221818900 4
34. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Kuttler, H., Lewis, M., tau Yih, W., Rocktäschel, T., Riedel, S., Kiela, D.: Retrieval-augmented generation for knowledge-intensive nlp tasks. ArXiv **abs/2005.11401** (2020), https://api.semanticscholar.org/CorpusID:218869575 3, 4
35. Lin, K.Q., Wang, J., Soldan, M., Wray, M., Yan, R., Xu, Z., Gao, D., Tu, R.C., Zhao, W., Kong, W., Cai, C., HongFa, W., Damen, D., Ghanem, B., Liu, W., Shou, M.Z.: Egocentric video-language pretraining. In: Advances in Neural Information Processing Systems (2022) 8
36. Liu, H., Chen, Z., Yuan, Y., Mei, X., Liu, X., Mandic, D.P., Wang, W., Plumbley, M..: Audioldm: Text-to-audio generation with latent diffusion models. In: International Conference on Machine Learning (2023), https://api.semanticscholar.org/CorpusID:256390486 4, 7
37. Lu, C., Zhou, Y., Bao, F., Chen, J., Li, C., Zhu, J.: Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. arXiv preprint arXiv:2206.00927 (2022) 8, 19
38. Luo, S., Yan, C., Hu, C., Zhao, H.: Diff-foley: Synchronized video-to-audio synthesis with latent diffusion models. In: NeurIPS (2023) 2, 4, 7, 9, 10, 11, 13, 19, 20
39. Mittal, H., Morgado, P., Jain, U., Gupta, A.: Learning state-aware visual representations from audible interactions. In: Oh, A.H., Agarwal, A., Belgrave, D., Cho, K. (eds.) Advances in Neural Information Processing Systems (2022), https://openreview.net/forum?id=AhbTKBlM7X 4
40. Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., Chen, M.: Glide: Towards photorealistic image generation and editing with text-guided diffusion models. In: International Conference on Machine Learning (2021), https://api.semanticscholar.org/CorpusID:245335086 4
41. Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018) 19
42. Owens, A., Isola, P., McDermott, J., Torralba, A., Adelson, E.H., Freeman, W.T.: Visually indicated sounds. In: CVPR (2016) 2, 3, 9

43. Popov, V., Vovk, I., Gogoryan, V., Sadekova, T., Kudinov, M.A.: Grad-tts: A diffusion probabilistic model for text-to-speech. In: International Conference on Machine Learning (2021), https://api.semanticscholar.org/CorpusID:234483016 4
44. Ramazanova, M., Escorcia, V., Heilbron, F.C., Zhao, C., Ghanem, B.: Owl (observe, watch, listen): Localizing actions in egocentric video via audiovisual temporal context (2022) 4
45. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, S.K.S., Ayan, B.K., Mahdavi, S.S., Lopes, R.G., Salimans, T., Ho, J., Fleet, D.J., Norouzi, M.: Photorealistic text-to-image diffusion models with deep language understanding. ArXiv **abs/2205.11487** (2022), https://api.semanticscholar.org/CorpusID:248986576 4
46. Song, Y., Ermon, S.: Generative modeling by estimating gradients of the data distribution. In: Neural Information Processing Systems (2019), https://api.semanticscholar.org/CorpusID:196470871 4
47. Soomro, K., Zamir, A.R., Shah, M.: Ucf101: A dataset of 101 human actions classes from videos in the wild. CoRR (2012) 4
48. Su, K., Qian, K., Shlizerman, E., Torralba, A., Gan, C.: Physics-driven diffusion models for impact sound synthesis from videos. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 9749–9759 (2023), https://api.semanticscholar.org/CorpusID:257805229 2, 3, 4
49. Wang, D., Chen, J.: Supervised speech separation based on deep learning: An overview. arxiv (201) 2, 5
50. Wu*, Y., Chen*, K., Zhang*, T., Hui*, Y., Berg-Kirkpatrick, T., Dubnov, S.: Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In: IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP (2023) 10
51. Yang, D., Yu, J., Wang, H., Wang, W., Weng, C., Zou, Y., Yu, D.: Diffsound: Discrete diffusion model for text-to-sound generation. IEEE/ACM Transactions on Audio, Speech, and Language Processing **31**, 1720–1733 (2022), https://api.semanticscholar.org/CorpusID:250698823 4

7 Supplementary

In this supplementary material, we provide additional details about:

1. Supplementary video for qualitative examples (referenced in Sec. 1).
2. Additional implementation details (referenced in Sec. 3).
3. Dataset details (referenced in Sec. 4).
4. Evaluation metric details (referenced in Sec. 5).
5. Human evaluation details (referenced in Sec. 5).

7.1 Supplementary Video

In this video, we include examples of Ego4D-Sounds clips, qualitative examples on unseen Ego4D clips, and qualitative examples on VR games. Wear headphones to hear the sound.

7.2 Additional Implementation Details

Audio-Visual LDM. Our Ego4D-Sounds clips are 3 seconds long. For model training and inference, we sample audio waveform at 16000Hz. We use FFT size 1024, mel bins 128, hop size 256 to transform the 3-second audio waveform into a mel-spectrogram of length 188, which we then pad in the temporal dimension to 192. To speed up training, similar to [38], we load VAE and diffusion model weights from the pre-trained Stable Diffusion model. Note that Stable Diffusion expect image as the input/target, and therefore we duplicate the mel-spectrogram in the channel dimension and to achieve size $x_0 \in \mathbb{R}^{3\times128\times192}$, passing x_0 to the VAE encoder, we get compressed latent representation $z_0 \in \mathbb{R}^{4\times16\times24}$. For conditioning, videos are sampled at 5 FPS, passed through the video encoder and a linear projection layer that produces features of size $c_v \in \mathbb{R}^{16\times768}$. Audio condition is also a 3-second clip and is processed the same way as the target audio to get $\tilde{c}_a \in \mathbb{R}^{4\times16\times24}$, it is then projected to a 2 dimensional tensor of shape $c_a \in \mathbb{R}^{24\times768}$. For classifier-free guidance, we set the scale $\omega = 6.5$, and use DPM-Solver [37] for accelerated inference using only 25 sampling steps. For the mel-spectrogram to waveform vocoder HiFi-GAN [32], we train the model from scratch with the mel-spectrogram processing hyperparameters aligned with that of our AV-LDM. During training, we freeze the pre-trained VAE, and train the LDM model on Ego4D-Sounds for 8 epochs with batch size 720. We use the AdamW optimizer with a learning rate of $1e-4$. HiFi-GAN is trained on a combination of 0.5s clips from Ego4D [18], Epic-Kitchens [24], and AudioSet [15]. We use AdamW to train HiFi-GAN with a learning rate of $2e-4$ and a batch size of 64 for 120k steps.

Audio-visual representation learning. We use Timesformer [1] as the video encoder, and AST [17] as the audio encoder. We train video and audio encoders for 5 epochs with batch size 256. We use the InfoNCE [41] loss and Adam optimizer [31] with a learning rate $1e-4$.

7.3 Dataset Details

To evaluate the effectiveness of our proposed ambient-aware action sound generation model, we leverage Ego4D [18], a large-scale egocentric video dataset for daily human activities. While our model is capable of disentangling action sound from ambient sound, there is little value in learning on data that only contain ambient sounds or background speech. Our goal is to curate an in-the-wild action dataset that has meaningful action sounds. We design a four-stage pipeline consisting of both learning-based tagging tools and rule-based filters to curate the Ego4D-Sounds dataset. To be consistent with the public splits of Ego4D benchmarks, we use all 7.5K videos in the training set, where we extract 3.8M clips centered at the narrations' timestamps with the left and right margins being 1.5s, i.e. each clip is 3s long, which we find to be sufficiently long enough to capture the narrated action.

We first remove all clips without sounds, resulting in 2.5M clips. We then filter the above clips based on the scenarios. Each Ego4D video has a scenario label, categorizing the activity depicted in the video. We go through all scenario labels and pick 28 scenarios that are mainly social scenarios, e.g., "playing board games", "attending a party", "talking with friends", where majority of the sounds are speech or off-screen sounds with no on-screen actions. We remove videos with these tags, resulting in 3.1K videos and 1.7M clips.

While the previous stage removed videos for social scenarios as a whole, there are still many clips that have only speech or background music. To detect these clips, we use an off-the-shelf audio tagging tool to tag the remaining clips. The goal is to remove clips that have solely off-screen sounds, i.e., speech and music. So we threshold the tagged probability at 0.5, i.e., removing clips that most likely only contain off-screen sounds and not action sounds. This filtering process further removed 0.5M clips, with 1.23M clips remaining.

Lastly, we also observe that in a long video clip, there are silent periods when no sounding action occurs. Based on this observation, we devise an energy-based filtering process, i.e. we normalize the amplitude of each clip with respect to the maximum amplitude of audio in the video. We then convert the amplitude to dB and remove clips with energy below -60 dB. This results in 1.18M clips.

7.4 Evaluation Metric Details

For the audio-visual synchronization (AV-Sync) [38] metric, we train a synchronization binary classification model on Ego4D-Sounds, and use it to judge whether the generated audio is synchronized with the video. Following [38], to construct the input to the classification model, we input paired and synced video and audio, unpaired video and audio, and paired but unsynced video and audio 50%, 25%, and 25% of the time respectively. The model uses Timesformer [1] as the video encoder, AST [17] as the audio encoder, and a 3 layer MLP as the classification head which takes the CLS tokens from the two encoders, concatenates them in the feature dimension, and produces a probability indicating the synchronization. We train this model for 30k steps with AdamW optimizer, which achieves a classification accuracy of 70.6% on the validation set.

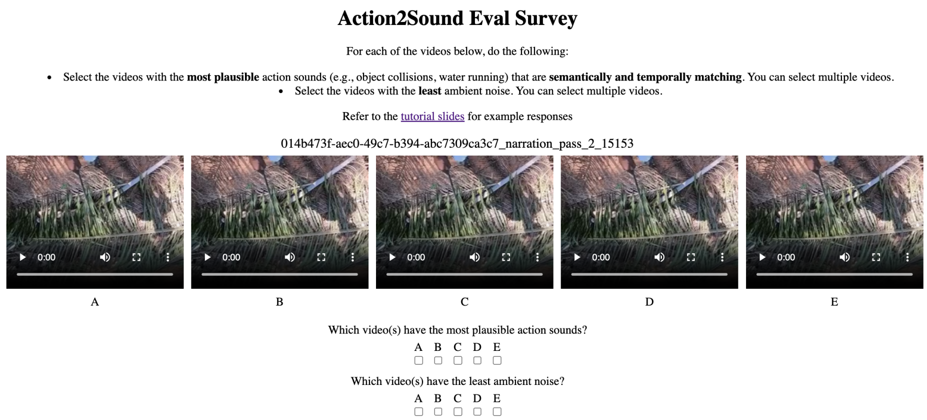


Fig. 10: Survey interface. For each example, we ask participants to select video(s) that are most plausible with the video and the video(s) that have the least ambient sound.

7.5 Human Evaluation Details

For the human evaluation, we first compile a guideline document, clarifying and defining the objectives of the survey and what the participant should be looking for. There are two main objectives: 1) select video(s) with the most plausible action sounds (e.g., object collisions, water running) that are semantically and temporally matching with the visual frames, and 2) select video(s) with the least ambient noise. We also provide multiple positive and negative examples for each criterion in the guideline document. We ask participants to read the guidelines before doing the survey. We show one example of the survey interface in Fig. 10.