

A BASELINES

We compare with these open-sourced VTA baselines:

1. IM2WAV (Sheffer & Adi, 2023) is an open-domain audio generation system which is based on the image or the image sequence. The model uses a language model to generate low-level audio representation. We use the pre-trained checkpoint with the default parameters as the baseline.
2. Diff-Foley (Luo et al., 2024) uses a LDM based on the features extracted by CAVP, an encoder that is contrastively pre-trained to learn temporally and semantically aligned audio-visual features. We use the pre-trained Diff-Foley as our baseline.
3. FoleyCrafter (Zhang et al., 2024) imports semantic and temporal blocks for precise audio-video synchronization, and supports the use of text descriptions to facilitate controllable and diverse TTA generation.
4. Seeing&Hearing (Xing et al., 2024) is built on a multimodality latent aligner with the pre-trained ImageBind model. It also uses a LDM as the generation framework.
5. T2AV (Mo et al., 2024) leverages visual-aligned text embeddings as its conditional foundation in diffusion-based audio generation. Note that, compared with other baselines, T2AV generation is more based on the textual description. The model leverages a pre-trained video-audio CLAP(VA-CLAP) as the vision encoder.

B SUBJECTION EVALUATION

The subjective evaluation aims to gauge the performance of our models from a human perspective, providing insights into the perceived quality and alignment of the generated audio and original video content. Participants undergo a training session where they are introduced to the evaluation objectives, the rating scale, and example demonstrations that illustrate different quality levels. Additionally, they practice rating sample videos to familiarize themselves with the criteria and receive feedback to ensure consistency.

Overall Quality This metric evaluates the general appeal and coherence of the combined video, considering both audio and visual components. It contains how well the audio and video elements fit together without any jarring or inconsistent moments.

Audio Quality Focusing solely on the generated audio, this metric assesses factors such as clarity, fidelity, and naturalness, like how clear and understandable the audio is, free from distortion or muddiness, and how closely the generated audio resembles human speech or natural sounds in terms of intonation, rhythm, and expression.

Video-Audio Semantic Alignment This metric measures how well the audio semantically matches the visual content, ensuring that the sounds correspond appropriately to the actions and scenes depicted in the video, like how relevant the audio is to the visual context, with sounds matching the actions and scenes on screen.

Video-Audio Temporal Alignment The evaluation centers on the synchronization between the audio and video streams, determining how accurately the timing of audio events aligns with visual events, like the precision with which sounds occur in tandem with corresponding visual actions (e.g., a door slamming in sync with the sound of the slam).

C EXPERIMENTS ON VISION ENCODERS

Based on the proposed generation framework, we start by trying ablation results on different vision encoders:

1. Clip4Clip (Luo et al., 2021) focuses on video-text retrieval based on the pretrained Contrastive Language-Image Pre-training (CLIP) (Radford et al., 2021) model. Experiments have shown that CLIP can serve as the backbone to extract knowledge from frame-level video input. We leverage a similar linear projection to map the frame-level video features extracted by CLIP to get the generation condition.
2. Imagebind (Girdhar et al., 2023) is pre-trained on different modalities to learn joint embeddings, including images, text, and audio. All the modalities are bound with the image-paired data. We leverage Imagebind to get the image embeddings of the extracted frames extracted from the input video.
3. LanguageBind (Zhu et al., 2023) is trained similarly with Imagebind but takes the language as the bind across different modalities. As the language modality is well-explored and proved influential as conditions in the previous TTA works, we would like to see whether vision embeddings extracted from LanguageBind can serve as a better condition.
4. V-JEPA (Bardes et al., 2024) is considered as an extension of the I-JEPA (Assran et al., 2023) to video based on self-supervised learning. Improving from I-JEPA which learns semantic image features, V-JEPA deals with self-supervised learning of video representations appropriate for video understanding in a spatio-temporal way.
5. ViViT (Arnab et al., 2021) a pure-transformer based models. Although the model was originally designed for video classification, we would like to check whether the spatio-temporal tokens learned from the input video can help guide the generation of audio. We use the pre-trained ViViT trained on the kinetics400 dataset as the encoder.
6. CAVP (Luo et al., 2024) is trained with two different objects, semantic contrast loss and temporal contrast loss, to improve audio-video features' semantic and temporal alignment. CAVP was first leveraged in the DIFF-FOLEY model to synthesize synchronized video-conditioned content. We use the pre-trained CAVP (also pre-trained on the VGGSound dataset) as our vision encoder.

D EXPERIMENTS ON AUXILIARY EMBEDDINGS

We are also interested in exploring whether auxiliary embeddings, beyond visual features, could enhance the generation process. Numerous studies have demonstrated that additional information can improve generation results in various ways, whether broadly or specifically. We aim to investigate this phenomenon through several auxiliary embeddings:

1. **Text Information:** Additional textual labels can provide valuable context and extra semantic details that may not be immediately discernible from visual features. Moreover, text information can assist in filtering out extraneous information in the video. For instance, the presence of a gun in a video does not necessarily imply that the corresponding audio will be generated. In our experiment, we utilize CLIP to obtain text embeddings and concatenate these with the video embeddings to serve as the condition.
2. **Position Embedding:** Indeed, Position Embedding is critical as it imparts a sense of temporal order or sequence, which is essential in audio generation, particularly when the vision encoders are primarily focused on the semantics of the audio. Intuitively, Position Embedding assists the model in understanding the event sequence, thereby enabling it to generate coherent audio. We utilize sinusoidal positional embedding, akin to the method used in the Transformer model (Vaswani et al., 2023), as it facilitates effortless attention to relative positions.
3. **Optical Flow:** Optical Flow offers valuable insights into the motion and dynamics present in a video sequence. Previous study (Fedorishin et al., 2023) has utilized this information to assist in localizing sound sources within videos. In our approach, we employ optical flow video embeddings as an additional condition for generation. Similar to the previous embeddings, these embeddings are concatenated with the original video embeddings, enriching the input representation and potentially enhancing the audio generation process.

E EXPERIMENTS ON DATA AUGMENTATION

The data quality of the training set is undeniably crucial for a model’s performance, particularly when training large generative models. We explore data augmentation from three different perspectives:

1. **Data Clean:** Data clean ensures that the input data is accurate, consistent, and free from errors or anomalies. We use a CLAP (Wu* et al., 2023) model to help select audio-video pairs with similar semantics based on extra textual labels (with *score* > 0.3). We also use the AV-Align score to filter unmatched video-audio pairs (with *score* > 0.2). For the VGGSound dataset, we filter out about 100k high-quality video-audio pairs.
2. **Concat Augment** The original VGGSound dataset primarily contains audio clips with single audio events. While this makes the dataset a clean, simple test set for audio event generation, it does not evaluate the model’s ability to handle temporal information in complex videos. To simulate complex generation tasks with various audio events during training, we randomly concatenate two videos with different audio events. We also propose a test set to assess our models.
3. **Pretrain** We also try to leverage the power of pretraining. We pretrain our model on two different data corpus separately: a large video-audio corpus consisting of 10k hours of content from YouTube, and a large audio corpus mainly from WavCaps (Mei et al., 2023) and Youtube, amounting to approximately 150k hours of audio with paired captions. The Youtube corpus is processed in the same manner as the VGGSound video data. We filter out talking and music content, using only audio event cases for training. For video-audio pretraining, we leverage video-to-audio supervised training. For audio pretraining, we perform audio self-supervised training. Subsequently, we perform full parameter finetuning of our model on the original training set.

F MORE DEMOS

F.1 SAILENCY MAP

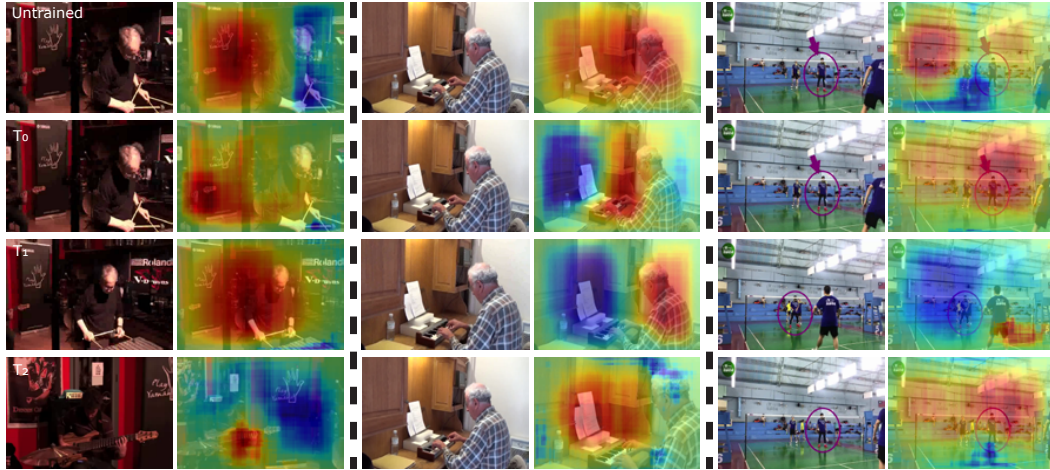


Figure 6: **More demos of the saliency map.** We show that the model focuses on different components in different frames of the given video after training, although the vision encoder parameters are completely frozen during the training time.

F.2 AUDIO GENERATION



Figure 7: **More demos of the VTA generation on open-domain videos.** Refer to the supp materials for more video demos and the inference code.