NEUROBAND: A NEURAL DUBBER OF ANY MUSICAL INSTRUMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Humans are born with the ability to associate synchronized music with musical instrument performance videos. However, existing video-to-audio (V2A) models struggle to generate high-quality and temporally synchronized multi-category music. In this work, we address this issue by introducing an audio generation task based on instrument performance videos. To this end, we present InstruVideo, a novel video-text paired benchmark consisting of over 812 hours of clips across 106 different musical instruments and providing a rich and diverse dataset of musical performances. We propose *NeuroBand*, a diffusion model specifically designed to generate high-fidelity audio from instrument performance videos. NeuroBand leverages Variational Autoencoder (VAE) and vocoders pre-trained on large audio datasets to enhance the quality of the generated audio. Experimental results on three instrumental test sets, *e.g.,* YouTube-Music, Douyin-Music, and MUSIC-solo, demonstrate that NeuroBand pre-trained on large audio datasets can highly enhance the quality of the generated audio. Furthermore, employing flow matching as an optimization technique improves the temporal synchronization between the generated audio and the corresponding video. The dataset, code and models will be made publicly available for further research.

1 INTRODUCTION

Eyes and ears are crucial sensors through which humans perceive the world, enabling individuals to easily associate visual cues with corresponding auditory information. In the context of musical performance, observers can infer the sounds of an instrument being played by recognizing subtle variations in the performer's movements and the type of instrument involved. This ability is extensively applied in fields such as sound restoration for silent films or audio dubbing in video games and animation. However, achieving precise synchronization between video of musical performances and their corresponding audio typically requires frame-by-frame editing and adjustments by skilled musicians or editors, making it both labor-intensive and time-consuming.

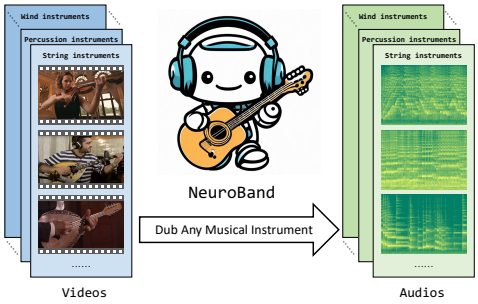


Figure 1: We propose 'dub any musical instrument', a fine-grained audio generation task based on the instrumental performance videos across various instrumental categories.

To unlock the potential of neural networks in the domain of instrument performance dubbing, we are the first to propose the **"dub any musical instrument"** task (fig. 1). In this task, the objective is to generate audios that accurately align with both the motion and the type of musical instruments shown in videos of performances across different instrument categories. While existing video-to-audio (V2A) tasks Luo et al. (2024); Xing et al. (2024); Iashin & Rahtu (2021); Sheffer & Adi (2023); Wang et al. (2024a); Zhang et al. (2024); Jeong et al. (2024) require the model to generate sounds of any given videos, *i.e.,* waves and typhoons, our "dub any musical instrument" task is actually a **fine-grained** audio generation task. For instance, the task expect the model to produce distinct audio outputs for two different string instruments, the Violin and Bouzouki, based on their specific characteristics.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Therefore, different from the general video-to-audio task that is to generate voice by recognize the category of the noisemaker, the new challenge of this task lies in generating music that is consistent with the timbre, pitch, and tonal quality of the instrument depicted in the video, which previous V2A methods overlook.

In this paper, we introduce a new benchmark named InstruVideo with more than 292,000 video-audio pairs spanning 106 types of instruments. InstruVideo includes not only common instruments such as the piano and guitar, but also some special instruments, such as the Matouqin and Balalaika. The raw audio tracks in these videos often contain extraneous audios unrelated to the instrumental performance. For instance, a video might feature a guitarist playing while background audio includes crowd cheering. To improve the relevance and quality of the video-audio pairs in our dataset, we removed human vocal elements from the audio. InstruVideo introduces a comprehensive evaluation with three perspectives: audio quality, semantic alignment and temporal alignment. The audio quality is assessed using metrics such as FAD Kilgour et al. (2018) and KL divergence Iashin & Rahtu (2021) to measure the fidelity of generated samples. Semantic alignment is evaluated with FAVD Shen et al. (2023) to assess the consistency between the video and audio in terms of semantic content. Temporal alignment is measured using AV-Align Yariv et al. (2024) to evaluate the synchronization between video and audio in terms of timing.

Based on Audio latent diffusion model, we propose a simple baseline method named NeuroBand. The input mel spectrogram transformed from the audio send to the VAE Liu et al. (2023) pretrained on audio dataset. Inspired by former work Luo et al. (2024), we employ the video encoder after Contrastive Audio-Visual Pretraining (CAVP) to extract the corresponding video's features, and used it as the condition of the diffusion model. Finally, we introduce flow matching Liu et al. (2022b); Lipman et al. (2022); Esser et al. (2024); Albergo & Vanden-Eijnden (2022) as a novel training strategy for video-to-audio models. By learning a direct path from the model's outputs to noise, flow matching achieves better temporal alignment during inference stage.

We evaluate the performance of our method over three instrumental test sets: YouTube-Music, Douyin-Music and MUSIC-solo Zhao et al. (2018; 2019). Experimental results show that the proposed method overpass previous models on these three test sets with three performance aspects: **(A) Audio Quality:** our model reducing FAD by -2.102, -6.000, -3.984 respectively; **(B) Semantic alignment:** our model reducing FAVD by -1.384, -1.230, -1.230 respectively; **(C) Temporal alignment:** our model improving AV-Align by +0.001, +0.033, +0.059, respectively. Our subjective evaluation also evalidates the effectiveness of our proposed method.

In summary, the contributions of this paper are three folds: (1) We introduce a novel audio generation task named "dub any musical instrument", generate fine-grained instrumental audio based on corresponding video. (2) We propose NeuroBand based on audio latent diffusion model, and using flow matching to optimize it. (3) Our method surpasses former video-to-audio models in three musical test sets under a comprehensive evaluation.

2 Related Work

2.1 Audio Generation from Various Modalities

Recent advances in audio generation have explored a range of input modalities, most notably text-to-audio (T2A) Luo et al. (2024); Xing et al. (2024); Iashin & Rahtu (2021); Sheffer & Adi (2023); Wang et al. (2024a); Zhang et al. (2024); Jeong et al. (2024); Dagli et al. (2024), each posing distinct technical challenges. T2A generation has progressed rapidly with the emergence of diffusion models Ho et al. (2022); Liu et al. (2022a); Kong et al. (2021) and neural audio codecs Kong et al. (2020a); van den Oord et al. (2016); Kumar et al. (2019); Liu et al. (2024), which enable the synthesis of high-fidelity audio from natural language prompts. For example, AudioLDM Liu et al. (2023) utilizes CLAP embeddings to learn continuous latent audio representations. Subsequent approaches have further enhanced generation quality and control by incorporating additional conditioning signals such as pitch, energy contours, and temporal annotations Guo et al. (2024); Ding et al. (2024).

While these models have shown promising results, T2A generation remains inherently constrained by the coarse granularity of textual prompts, which often lack sufficient temporal or semantic resolution. This limitation poses challenges for tasks such as music performance dubbing, where precise

synchronization between visual and auditory modalities is critical Luo et al. (2024). In contrast, V2A generation leverages temporally aligned visual cues from silent performance videos, offering richer and more fine-grained information for conditioned audio synthesis.

2.2 Music Generation

A robust representation of music, such as the MIDI format, has been extensively utilized in computational music modeling. Early approaches convert MIDI data into a piano-roll representation and employ generative models such as Generative Adversarial Networks (GANs) Roberts et al. (2018); Gillick et al. (2019) to generate novel musical sequences. Subsequent research introduces event-based representations to enhance the efficiency and expressiveness of MIDI encoding Oore et al. (2020); Huang & Yang (2020). Additionally, control signals have been incorporated into MIDI-based generative models to facilitate conditioned music generation Engel et al. (2017); Choi et al. (2020); Lattner & Grachten (2019). Beyond symbolic representations, several models have been proposed to generate music directly from raw audio. WaveNet van den Oord et al. (2016) introduces an autoregressive framework capable of synthesizing high-fidelity audio. Jukebox Dhariwal et al. (2020) adopts a hierarchical architecture that generates discrete audio tokens at multiple temporal resolutions, which are then combined to reconstruct coherent musical audio. More recent advances in high-quality audio representation Zeghidour et al. (2021); Défossez et al. (2022); Kumar et al. (2023) utilize residual vector quantization directly on the waveform, enabling efficient audio encoding. These representations have been adopted in transformer-based models for text-to-audio generation, including autoregressive approaches such as AudioLM Borsos et al. (2023), Mubert Zeng et al. (2021), AudioGen Kreuk et al. (2022), and MusicLM Agostinelli et al. (2023), as well as non-autoregressive frameworks Copet et al. (2023); Garcia et al. (2023). Alternatively, diffusion-based methods, such as Riffusion Forsgren & Martiros (2022), have been explored for music generation, along with other recent efforts Huang et al. (2023); Liu et al. (2023); Huang & Yang (2020); Schneider et al. (2023).

3 InstruVideo Benchmark

To achieve audio generation from instrument performance videos, we have decided to train the model using a large dataset of video-audio pairs. We will describe the process of creating the instrument performance-related dataset.

3.1 Data Collection

The examples of our collected dataset can found in fig. 2. We started with those large-scale video datasets Abu-El-Haija et al. (2016); Chen et al. (2020) annotated with video tags, particularly focusing on video understanding datasets Abu-El-Haija et al. (2016), as these types of videos are typically synchronized with corresponding audio tracks. We manually selected all tags related to musical instrument performances, using these tags to filter and retrieve relevant videos. However, a significant limitation arose because these datasets only provided YouTube URLs for videos, many of which were no longer accessible. For the rest of the videos, we proceeded to download them and excluded any that lacked either audio or visual content. To optimize the data for model training, we segmented the downloaded videos into 10-second clips. If slice start and end points were specified in the original dataset, we followed those timestamps; otherwise, we selected a random 10s clip around the midpoint within each video. Although this large video dataset provided an extensive number of clips, the overall audio and video quality was limited by the dataset's original creation date. Additionally, since these datasets were curated for general purposes, the variety of musical instrument types was relatively narrow. To address this, we supplemented the dataset with a collection of high-quality videos sourced from the Bilibili and Douyin, featuring new instrument types, such as the xiao and suona. Unlike the previous slicing method, these high-quality videos were segmented using a sliding window technique to generate 2 to 5 clips of 10 seconds each.

3.2 Protocols

The total datasets contains 292k video-audio pairs, over 812 hours and across 106 instrumental categories. We split 4.6k YouTube clips as one of our test set, named YouTube-Music, and take

Table 1: The comparsion of performance videos in the training sets of different datasets

Datasets	Music videos	Instrument class
YouTube-8M Abu-El-Haija et al. (2016)	280k	73
VGGSound Chen et al. (2020)	48k	63
MUSIC Zhao et al. (2018; 2019)	0.5k	21
InstruVideo	280k	106

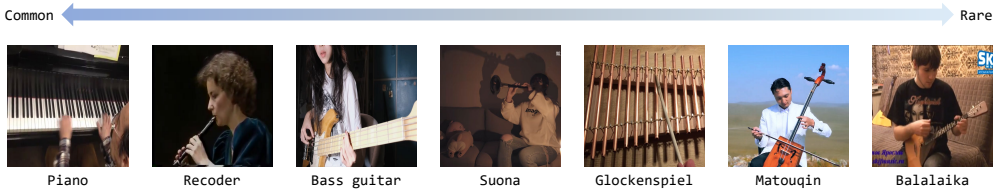


Figure 2: Examples of various instruments in InstruVideo. Instruments on the left are more commonly encountered, such as the piano and flute, while those on the right are less frequently seen, including culturally distinctive instruments like the matouqin and balalaika.

all clips from Douyin to form our second test set, named Douyin-Music. The remaining YouTube clips along with clips from Bilibili, total over 283k video-audio pairs were used for model training. The comparison with the existing video datasets around performance videos is illustrated in table 1. To introduce a comprehensive evaluation, we adopt multiple metrics in three performance aspects: audio quality, semantic consistency, and temporal synchronization.

(1) Audio quality, we use Fréchet Audio Distance (FAD) Kilgour et al. (2018), Kullback–Leibler (KL) divergence Iashin & Rahtu (2021), and Fréchet distance (FD) Kong et al. (2020b) to evaluate the generation quality. FAD measures the distance between generated and real audio distributions in an embedding space, using embeddings extracted with a pre-trained VGGish Hershey et al. (2017) model. KL divergence assesses the similarity between the distributions of generated and real audio embeddings, providing insight into semantic consistency. FD is similar to FAD but uses different embeddings for comparison Kong et al. (2020b). **(2) Semantic alignment**, we use Fréchet audio-visual distance(FAVD) Shen et al. (2023) to assess the semantic alignment, which calculates the Fréchet distance between video and audio embeddings, ensuring alignment of audio and video content. For audio and video embeddings, we employ pre-trained VGGish Hershey et al. (2017) and I3Dmodels Carreira & Zisserman (2017), respectively. **(3) Temporal alignment**, we employ AV-Align Yariv et al. (2024) to evaluate the alignment between video actions and corresponding audio events. This metric is based on detecting energy peaks independently in both audio and visual modalities, under the premise that rapid changes in audio intensity often correlate with visible movements in video. AV-Align first identifies alignment points by detecting peaks in each modality separately: audio peaks are detected using the Onset Detection algorithm, while changes in video are identified through calculating the mean Optical Flow magnitude for each frame, which highlights rapid motion. For each peak in one modality, the algorithm verifies the presence of a corresponding peak in the other modality within a three-frame temporal window, assessing alignment through an Intersection-over-Union (IoU) approach. The final AV-Align score, normalized by the number of peaks, ranges from zero to one, reflecting the model's ability to synchronize audio and video events.

3.3 Human Vocal Removal

The audio in instrument performance videos typically consists of components strongly correlated with the visual motion, such as the sound of the instrument being played, as well as components weakly associated with the video motion, such as audience cheers or off-camera vocals. While it maybe realistic for concert videos to include audience applause, our goal is for the model to output as clean an instrument sound as possible. Furthermore, retaining these motion-unrelated audio components can complicate the model's ability to learn the temporal relationships between the video and the audio. To address this, we applied a publicly available vocal removal model Défossez (2021); Rouard et al. (2023) to clean the audio. This model preserves the instrument sounds while removing as much of the vocal content as possible. The vocal-removed audio is then used as the

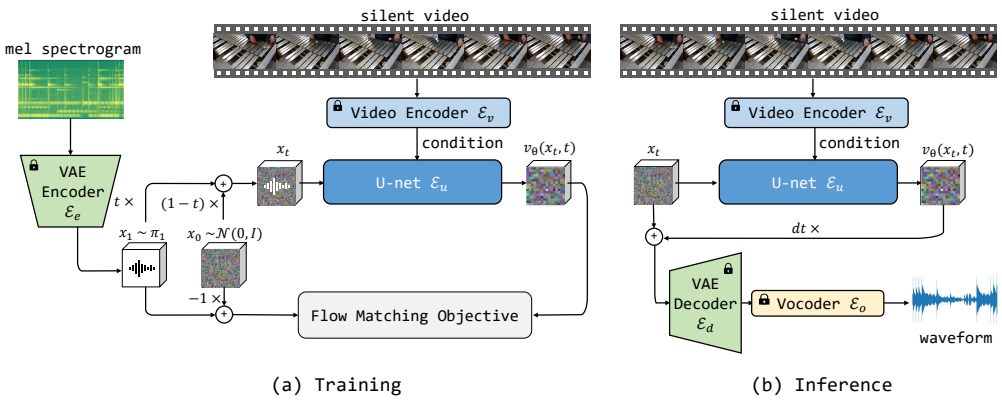


Figure 3: The overall architecture of the proposed method. (a) During training stage, the audio features encoded by the VAE encoder are linearly combined with standard Gaussian noise to obtain the rectified flow. The flow matching objective encourages the U-net to predict the flow's velocity. (b) The rectified flow, initialized with standard Gaussian noise, is updated through several Euler steps. The final flow is decoded by the VAE decoder, and reconstructed into a waveform by a pretrained vocoder.

ground truth for each video, and during testing, the vocal-removed audio serves as the basis for evaluating the model's performance.

4 METHOD

4.1 PRELIMINARY: FLOW MATCHING

Conditional generative model tends to mapping samples x_0 from a noise distribution π_0 to samples x_1 from a data distribution π_1 under the condition c, which can be written in ordinary differential equation (ODE):

$$dx = v_\theta(x, t|c)dt \tag{1}$$

where the velocity v_θ is parameterized by the weights θ of a neural network, and t denotes the time position. Flow matching Liu et al. (2022b); Lipman et al. (2022); Esser et al. (2024); Albergo & Vanden-Eijnden (2022) is an efficient way for regress the velocity v_θ by predefine a specific trajectory for the sample mapping.

$$x_t = a_t x_1 + b_t x_0, \tag{2}$$

where $t \in [0, 1]$ is a time parameter that smoothly transitions the state from x_0 to x_1. The ODE governing this transformation is given by:

$$\frac{dx_t}{dt} = f(x_t, t) \tag{3}$$

where $f(x_t, t)$ is a function that we aim to learn, describing the velocity field of the transformation at each point in time. By solving this ODE, we can generate samples from the target distribution π_1 starting from the source distribution π_0. Now the flow matching objective can be given by using regression:

$$\min_v \int_0^1 \mathbb{E}_{x_0\sim\pi_0, x_1\sim\pi_1} \left[\|f(x_t, t) - v_\theta(x_t, t|c)\|^2 \right] dt. \tag{4}$$

4.2 NEUROBAND

The proposed method aims to solve the problem of recovering audio from silent videos of musical instrument performances. In our task, under the instrument performance videos condition c, we attempt to learn the transformation from noise distribution π_0 to instrumental audio distribution π_1. We introduce a novel end-to-end learning framework capable of generating corresponding audio signals from silent videos.

In the training phase (fig. 3(a)), we first use a Variational Autoencoder (VAE) $\mathcal{E}_e$ derived from audio latent diffusion model(ALDM) to extract audio features. These extracted features x_1 are then combined with standard Gaussian noise x_0 to form the flow. We choose rectified flow Liu et al. (2022b); Esser et al. (2024) as our flow trajectory, which defines a linear interpolation path:

$$x_{\mathrm{t}} = t \cdot x_1 + (1-t) \cdot x_0, \tag{5}$$

where t is a hyper-parameter controlling the noise strength. For video conditions c, we use the video encoder $\mathcal{E}_v$ from Constrastive Audio-Visual Pretraining(CAVP). CAVP was pre-trained by contrastive learning on a large number of video and audio pairs Luo et al. (2024) to extract video features, thus endows video features to synthesis the semantic consistent and temporal aligned audio Luo et al. (2024).

Subsequently, we design a special loss function – the flow matching objective to guide the learning of the U-net Ronneberger et al. (2015) network $\mathcal{E}_u$. The task of the U-net is to predict the velocity field of the input rectified flow based on video embedding conditions. This design enables the U-net to gradually learn how accurately to capture the changing patterns of video and audio features during training. The flow matching objective is defined as:

$$\mathcal{L}_{\mathrm{fm}} = \left\| (x_1 - x_0) - v_\theta(x, t|c) \right\|^2, \tag{6}$$

where x_0, x_1 are from a noise-data pair. Defining the forward process as a straight line path between the data distribution and the standard normal distribution, flow matching speeds up the sampling process.

For faster convergence and stable training, instead of uniformly sampling t, we tends to sample t close to the middle of [0,1] by using a logit-normal sampling based on the former practical experience Esser et al. (2024):

$$u \sim \mathcal{N}(u; m, s) \tag{7}$$

$$t = \frac{1}{1 + e^{-u}} \tag{8}$$

where m, s are hyperparameters.

4.3 Faster sampling

Inference During the inference phase (fig. 3(b)), we no longer require original audio data. Instead, we start directly from standard Gaussian noise $\epsilon \sim \mathcal{N}(0, I)$ to construct the initial state, serving as the starting point of the rectified flow. Following this, we update the state of the rectified flow iteratively through a series of Euler steps until reaching the desired time point:

$$x_0 = \epsilon \tag{9}$$

$$x_{t+\Delta t} = x_t + \Delta t \cdot v_\theta(x_t, t|c), t > 0 \tag{10}$$

where Δt is the step size . Finally, we utilize the VAE decoder $\mathcal{E}_d$ to convert the final rectified flow into an audio feature representation, and subsequently employ a pre-trained vocoder $\mathcal{E}_o$ to transform it into a listenable sound waveform.

Classifier-free guidance In the dub any musical instrument task, we found that the Classifier-Free Guidance (CFG) technology Ho & Salimans (2022) is crucial for understanding instrument timbre and ensuring precise time synchronization. Therefore, when predicting the vector field at each time step, we use CFG technology to guide the model towards generating audio that faithfully matches both the instrument's timbre and the temporal progression of the performance. The specific formula is as follows:

$$v_{\mathrm{cfg}}(x, t|x_c) = \omega v_\theta(x, t|x_c) + (1-\omega) v_\theta(x, t|\varnothing) \tag{11}$$

where ω represents the guidance scale that balances the trade-off between sample diversity and generation quality, with v_{cfg} reducing to the original vector field $v_\theta(x, t|x_c)$ when $\omega = 1$.

The entire process is designed with careful consideration of the requirements of the audio generation task, ensuring both the effectiveness and robustness of the model, as well as significantly improving computational efficiency. Moreover, by introducing the flow matching objective and Euler steps, our method successfully addresses issues such as gradient vanishing commonly encountered in traditional diffusion models, thereby achieving high-quality audio recovery results. Future work will further explore ways to extend this framework to more instruments and complex scenarios.

Table 2: Performance comparison with state-of-the-art methods on YouTube-Music.

Method	Quality			Semantic	Temporal
	FAD ↓	KL ↓	FD ↓	FAVD ↓	AV-Align ↑
Diff-Foley Luo et al. (2024)	17.287	2.549	37.423	5.032	0.190
Im2Wav Sheffer & Adi (2023)	12.944	2.123	23.864	4.453	0.303
Seeing&Hearing Xing et al. (2024)	11.987	2.273	23.861	4.448	0.303
TiVA Wang et al. (2024b)	11.906	2.252	23.734	4.427	0.304
NeuroBand (LDM)	15.936	2.232	37.580	4.825	0.229
NeuroBand (ALDM)	**9.804**	**2.035**	**22.250**	**3.043**	0.304
NeuroBand (FM)	11.780	2.517	24.559	3.923	**0.314**

Table 3: Performance comparison with state-of-the-art methods on Douyin-Music.

Method	Quality			Semantic	Temporal
	FAD ↓	KL ↓	FD ↓	FAVD ↓	AV-Align ↑
Diff-Foley Luo et al. (2024)	9.202	2.370	40.367	3.603	0.186
Im2Wav Sheffer & Adi (2023)	8.531	1.801	24.102	3.722	0.254
Seeing&Hearing Xing et al. (2024)	8.345	1.796	20.365	3.063	0.250
TiVA Wang et al. (2024b)	8.315	1.790	20.476	2.867	0.248
NeuroBand (LDM)	7.710	**1.731**	38.057	3.279	0.196
NeuroBand (ALDM)	**2.315**	1.928	**13.759**	**1.637**	0.245
NeuroBand (FM)	4.295	1.780	19.529	2.271	**0.283**

5 EXPERIMENT

5.1 IMPLEMENTATION DETAILS

Training setting We proposed three training settings for NeuroBand: (1) **LDM:** Using same architecture and same weights from SD-V1.4 Rombach et al. (2021), and using a algorithm-based vocoder Griffin & Lim (1984) as Diff-Foley. (2) **ALDM:** Mainly based on SD-V1.4, but replacing VAE and vocoder from Audio Latent Diffusion Models (ALDM). Liu et al. (2024) (3) **Flow matching (FM):** based on ALDM setting, but replacing DDPM with flow matching. We trained both three models with InstruVideo train-set on 8×A100 for 80 epochs. m and s set to -0.5 and 1.0 in logit-normal samling, respectively. For the audio data in train-set, we cropped each sample to a duration of 9.92 seconds and resampled it to 16,000 Hz. Mel-spectrograms were then generated using a hop size of 160 and 64 Mel filter banks. For the video part, each sample was converted to a frame rate of 4 fps, retaining only the first 39 frames.

Evaluation setting. We also select an out-of-distribution(OOD) dataset MUSIC Zhao et al. (2018; 2019) for evaluation except YouTube-music and Douyin-music in InstruVideo. Specifically, we selected MUSIC21-solo subset and randomly extracted a clip from each videos. We named the 1k result samples as MUSIC-solo for simplicity. All clips in test sets were cut into 8-second duration and then generating audio with 16kHz sample rate for fair comparison. DPM-solver Lu et al. (2022) was used for faster sampling in the NeuroBand (LDM) and NeuroBand (ALDM) settings with 25 sample steps, while NeuroBand (FM) used Euler discrete sampling with 50 steps.

5.2 EXPERIMENTAL RESULTS

To evaluate NeuroBand's performance, we conduct experiments on three distinct instrumental test sets: YouTube-Music, Douyin-Music, and MUSIC-solo. Each dataset introduces distinct difficulties to validate our method's robustness and versatility. (1) **The YouTube-Music dataset** covers diverse solo and ensemble musical performances. As illustrated in table 2, NeuroBand(ALDM) obtains the lowest Fréchet Audio Distance (FAD) of 9.804, outperforming the prior state-of-the-art (SOTA) by 2.102, as well as the minimal Fréchet Distance (FD) of 22.250 with a 1.485 advantage over SOTA. For semantic alignment, NeuroBand(ALDM) reaches the lowest Fréchet Audio-Video Distance (FAVD) at 3.043, reducing the SOTA value by 1.384. Meanwhile, NeuroBand(FM) delivers marginally better temporal alignment with the highest Audio-Visual Alignment (AV-Align) score of 0.314. These outcomes verify NeuroBand's capability to produce high-quality audio with semantic consistency and temporal alignment for complex musical performances. (2) **The Douyin-Music dataset** consists mainly of short-form music videos contaminated by background and crowd noise. As shown in table 3, NeuroBand(ALDM) and NeuroBand(FM) consistently beat baseline approaches. NeuroBand(ALDM) achieves the lowest FAD of 2.315, improving upon SOTA by 6.035, and the minimal FAVD of 1.637, nearly cutting the SOTA value in half for semantic alignment. For temporal alignment, NeuroBand(FM) attains an AV-Align score of 0.283 above other baselines. These observations confirm NeuroBand's robustness under noisy dynamic conditions to keep generated audio well-synchronized with video. (3) **The MUSIC-solo dataset** contains solo instrumental performances, posing challenges from fine-grained, nuanced instrument sounds. According to table 4, NeuroBand(ALDM) substantially surpasses SOTA. Its FAD improvement of 4.392 reflects boosted audio quality and fidelity; the 0.529 reduction in FAVD proves reliable cross-modal semantic alignment between video and audio; and the +0.043 gain in AV-Align demonstrates precise temporal synchronization critical for realistic audiovisual output.

Method	Quality			Semantic	Temporal
	FAD ↓	KL ↓	FD ↓	FAVD ↓	AV-Align ↑
Diff-Foley Luo et al. (2024)	25.061	1.406	41.881	6.793	0.198
Im2Wav Sheffer & Adi (2023)	22.759	1.377	27.673	6.963	0.260
Seeing&Hearing Xing et al. (2024)	21.124	1.371	25.342	6.963	0.260
TiVA Wang et al. (2024b)	21.532	1.373	23.464	5.853	0.261
NeuroBand (LDM)	24.004	1.413	44.045	6.838	0.205
NeuroBand (ALDM)	**17.140**	**1.360**	**24.832**	**5.324**	**0.320**
NeuroBand (FM)	17.526	1.837	43.295	5.525	0.303

Table 4: Performance comparison with state-of-the-art methods on MUSIC-solo.

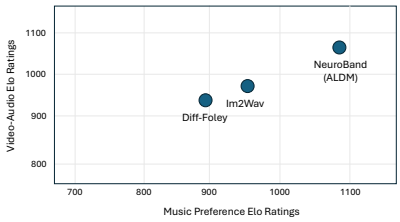


Figure 4: Elo ratings for video-to-audio models in the music preference and video-audio alignment human evaluations.

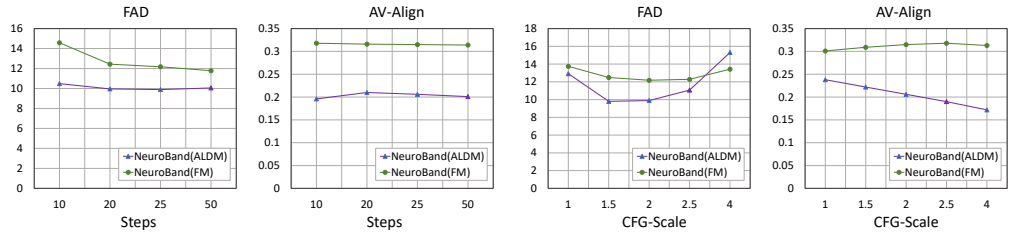


Figure 5: Different Sample steps & CFG Scale: As CFG Scale increases, NeuroBand (FM) initially improves AV-Align before stabilizing, while NeuroBand (ALDM) decreases. For audio quality, both methods show a U-shaped curve, with quality decreasing to an optimal point before rising again. Increasing sampling steps improves quality but also increases computation time.

Collectively, these experimental results confirm NeuroBand's efficacy for generating high-quality, synchronized audio from instrumental performance videos. Flow matching paired with VAE-based mel-spectrogram encoding-decoding brings notable gains in temporal alignment and audio quality. Future work will expand dataset instrument diversity and explore extra modalities such as text descriptions to further improve controllability and generation diversity.

5.3 Subjective Evaluation

To assess the quality of generated music preferences, we evaluate which video-to-audio generation methods produce the most subjectively appealing results. Drawing inspiration from human evaluations of large language model (LLM) chatbots Chiang et al. (2024), we adopt a similar approach and evaluation methodology. Our study employs a survey based on pairwise comparisons, where participants are asked to choose between two 10-second audio clips. For music preference, participants respond to the question: "Which audio clip do you prefer?" For evaluating video-audio alignment, they are asked: "Which audio clip is better described by" followed by a silent video used to generate the clips. Responses are restricted to a binary choice: "Audio Clip 1" or "Audio Clip 2."

We randomly select 100 videos from our proposed InstruVideo dataset, along with the corresponding audio generated by each video-to-audio model. Since each audio track for a given silent video is compared against every other track for that same video, the final survey consists of 300 music preference and 300 video-audio alignment questions. The survey is conducted via the Prolific platform LLC, with participants pre-screened to be between the ages of 15 and 45. Each participant answers three music preference and three video-audio alignment questions, with question order randomized for each individual. We report bootstrapped Elo scores for both music preference and video-audio alignment results (see supplementary materials for formulation), as shown in fig. 4. Our model, NeuroBand, outperforms all others in both music preference and video-audio aligments.

5.4 Ablation Study

Sample Steps. Increasing the number of sample steps leads to a reduction in FAD, which eventually plateaus around 25 steps. This indicates that while more sample steps improve audio quality, they also increase computational cost. For NeuroBand(FM), AV-Align consistently achieves high

scores, benefiting from the efficient sampling process enabled by Flow Matching. This allows for fast and accurate alignment in the generated audio. In contrast, for NeuroBand (ALDM), AV-Align peaks at 20 steps and then stabilizes, due to its dependency on the number of sample steps.

CFG Scale. When increasing the CFG Scale, FAD initially decreases to an optimal value before rising, forming a U-shaped curve. For AV-Align, NeuroBand (FM) shows an initial improvement followed by a decline, while NeuroBand (ALDM) peaks early and continues to decrease. Overall, NeuroBand (FM) consistently outperforms NeuroBand (ALDM) in terms of AV-Align, suggesting that Flow Matching offers superior time alignment, with the cost of slightly reduced audio quality. The experimental results indicate that both NeuroBand (ALDM) and NeuroBand (FM) exhibit relatively high FAD values when CFG ranges from 1.5 to 2.5.

Visualization To provide a detailed visual understanding of the performance of NeuroBand, we examine the generated audio and video results presented in fig. 6. For each instrument, e.g., accordion, piano, congas, and violin, the figure juxtaposes the ground truth (gt) Mel spectrogram against the one produced by our model. The Mel spectrogram provides a visual representation of the spectral power distribution over time, allowing us to compare the frequency content and temporal structure of the generated audio with the original recording. (1) *Accordion*: The generated audio for the accordion closely mirrors the ground truth in terms of the Mel spectrogram. It captures the unique timbral characteristics and dynamic variations typical of accordion music, indicating successful reproduction of the instrument's sound. (2) *Piano*: Our model's output for the piano shows a high degree of similarity to the ground truth. The generated audio not only aligns well with the visual content but also preserves the fine details of the piano's sound, such as the nuances of key strikes and pedal usage. (3) *Violin*: For the violin, the model succeeds in capturing the subtleties of bowing techniques and string vibrations. The generated audio is highly synchronized with the video, showcasing a realistic rendition of the violin's sound. These visual comparisons, supported by the corresponding Mel spectrograms, reinforce the quantitative and qualitative assessments previously discussed. They illustrate the efficacy of NeuroBand in producing high-fidelity and temporally aligned audio for diverse musical performances.

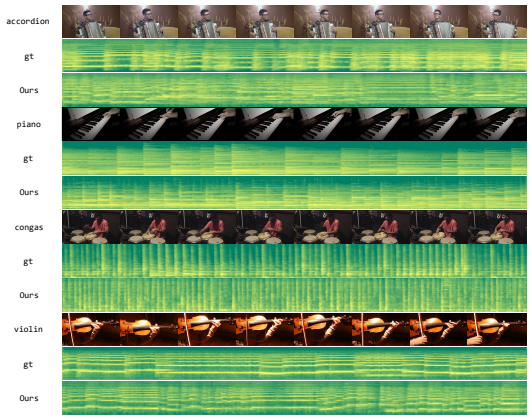


Figure 6: We compared our generated results with the ground truth (gt) in terms of the Mel spectrogram. We select accordion, piano, congas, and violin instrument performance videos as representatives, and the results demonstrate that our model successfully reproduces the timbral characteristics of various instruments while maintaining good synchronization with the video.

6 Conclusion and Limitations

Conclusion. We proposed a novel audio generation task based on instrumental performance videos in this paper. To address this task, we introduced InstruVideo, a large-scale dataset consisting of over 812 hours of instrument performance videos across 106 different instrument types. Using a flow matching-based model, we achieve high-quality audio generation that is temporally aligned with the corresponding video on three test sets. We hope our InstruVideo can facilitate future works in high-fidelity and synchronized music generation.

Limitations. The performance of different components within the model, such as the VAE, needs further validation to ensure optimal integration and efficiency. Expanding the dataset to include different domains, such as anime, could enhance the model's versatility and robustness. These aspects are promising avenues for future exploration.

REPRODUCIBILITY STATEMENT

To help reviewers reproduce our experimental results, we detail the required procedures for dataset construction, model training and evaluation, and will **release full code, dataset split definitions, model checkpoints and evaluation scripts upon paper acceptance.**

For dataset reproduction, reviewers may reconstruct our InstruVideo benchmark covering 106 instrument categories. This dataset sources video-audio clips from filtered YouTube-8M and VGGSound, supplemented by high-quality Bilibili and Douyin footage. All raw videos are uniformly sliced into 10-second segments. A standard vocal-removal pipeline removes human voices and irrelevant background sounds to generate clean instrumental audio as ground-truth. We provide a fixed official train-test split: most YouTube clips and all Bilibili clips constitute the training set, whereas YouTube-Music, Douyin-Music and out-of-distribution MUSIC-solo subsets form the test set.

To replicate the NeuroBand model, reviewers follow our fixed implementation built upon audio latent diffusion. It uses pre-trained audio VAE and vocoder for audio encoding-decoding, and leverages contrastive audiovisual pre-trained encoders for visual conditional input. Rectified flow matching replaces vanilla diffusion sampling for better temporal audiovisual alignment. The full training recipe includes 80-epoch training on 8×A100 GPUs: 16kHz audio sampling rate, standardized mel-spectrogram extraction, 4fps video sampling, together with logit-normal time sampling and classifier-free guidance for both training and inference.

For evaluation, fixed metrics are adopted: FAD, KL divergence and FD measure audio quality; FAVD evaluates semantic alignment; AV-Align assesses temporal synchronization. Baselines run with official public weights under identical settings. Human subjective evaluation uses pairwise Elo-rating comparisons. All ablation and evaluation configurations are documented for consistent reproduction.

AI USE STATEMENT

The large-language models assessed within this study were accessed through their official standard application-programming interface (API). Consistent fixed-form system prompts were applied for all inference runs, alongside deterministic decoding configurations to eliminate randomness in model generation. During the preparation of this manuscript, generative-AI tools were utilized to support textual polishing, sentence-level rephrasing, and document formatting tasks only. Critically, all study-specific experimental workflows, including the design of experimental protocols, construction of structural scoring pipelines, execution of quantitative analyses, and final result evaluations, were fully implemented, manually inspected, and independently verified by the authors. No generative-AI system was involved in hypothesis formulation, raw-data generation, statistical interpretation, or drawing the scientific conclusions reported in this work.

REFERENCES

Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Paul Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan. Youtube-8m: A large-scale video classification benchmark. *arXiv preprint arXiv:1609.08675*, 2016.

Andrea Agostinelli, Timo I Denk, Zalán Borsos, Jesse Engel, Mauro Verzetti, Antoine Caillon, Qingqing Huang, Aren Jansen, Adam Roberts, Marco Tagliasacchi, et al. Musiclm: Generating music from text. *arXiv preprint arXiv:2301.11325*, 2023.

Michael S Albergo and Eric Vanden-Eijnden. Building normalizing flows with stochastic interpolants. *arXiv preprint arXiv:2209.15571*, 2022.

Zalán Borsos, Raphaël Marinier, Damien Vincent, Eugene Kharitonov, Olivier Pietquin, Matt Sharifi, Dominik Roblek, Olivier Teboul, David Grangier, Marco Tagliasacchi, et al. Audiolm: a language modeling approach to audio generation. *IEEE/ACM transactions on audio, speech, and language processing*, 31:2523–2533, 2023.

Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6299–6308, 2017.

Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. Vggsound: A large-scale audio-visual dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 721–725. IEEE, 2020.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael Jordan, Joseph E Gonzalez, et al. Chatbot arena: An open platform for evaluating llms by human preference. In *Forty-first International Conference on Machine Learning*, 2024.

Kristy Choi, Curtis Hawthorne, Ian Simon, Monica Dinculescu, and Jesse Engel. Encoding musical style with transformer autoencoders. In *International conference on machine learning*, pp. 1899–1908. PMLR, 2020.

Jade Copet, Felix Kreuk, Itai Gat, Tal Remez, David Kant, Gabriel Synnaeve, Yossi Adi, and Alexandre Défossez. Simple and controllable music generation. *Advances in Neural Information Processing Systems*, 36:47704–47720, 2023.

Rishit Dagli, Shivesh Prakash, Robert Wu, and Houman Khosravani. See-2-sound: Zero-shot spatial environment-to-spatial sound. *arXiv preprint arXiv:2406.06612*, 2024.

Alexandre Défossez. Hybrid spectrogram and waveform source separation. In *Proceedings of the ISMIR 2021 Workshop on Music Source Separation*, 2021.

Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. High fidelity neural audio compression. *arXiv preprint arXiv:2210.13438*, 2022.

Prafulla Dhariwal, Heewoo Jun, Christine Payne, Jong Wook Kim, Alec Radford, and Ilya Sutskever. Jukebox: A generative model for music. *arXiv preprint arXiv:2005.00341*, 2020.

Shuangrui Ding, Zihan Liu, Xiaoyi Dong, Pan Zhang, Rui Qian, Conghui He, Dahua Lin, and Jiaqi Wang. Songcomposer: A large language model for lyric and melody composition in song generation, 2024. URL https://arxiv.org/abs/2402.17645.

Jesse Engel, Cinjon Resnick, Adam Roberts, Sander Dieleman, Mohammad Norouzi, Douglas Eck, and Karen Simonyan. Neural audio synthesis of musical notes with wavenet autoencoders. In *International conference on machine learning*, pp. 1068–1077. PMLR, 2017.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first International Conference on Machine Learning*, 2024.

Seth Forsgren and Hayk Martiros. Riffusion-stable diffusion for real-time music generation. *URL https://riffusion. com*, 2022.

Hugo Flores Garcia, Prem Seetharaman, Rithesh Kumar, and Bryan Pardo. Vampnet: Music generation via masked acoustic token modeling. *arXiv preprint arXiv:2307.04686*, 2023.

Jon Gillick, Adam Roberts, Jesse Engel, Douglas Eck, and David Bamman. Learning to groove with inverse sequence transformations. In *International conference on machine learning*, pp. 2269–2279. PMLR, 2019.

Daniel Griffin and Jae Lim. Signal estimation from modified short-time fourier transform. *IEEE Transactions on acoustics, speech, and signal processing*, 32(2):236–243, 1984.

Zhifang Guo, Jianguo Mao, Rui Tao, Long Yan, Kazushige Ouchi, Hong Liu, and Xiangdong Wang. Audio generation with multiple conditional diffusion model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 18153–18161, 2024.

Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al. Cnn architectures for large-scale audio classification. In *2017 ieee international conference on acoustics, speech and signal processing (icassp)*, pp. 131–135. IEEE, 2017.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P. Kingma, Ben Poole, Mohammad Norouzi, David J. Fleet, and Tim Salimans. Imagen video: High definition video generation with diffusion models, 2022. URL https://arxiv.org/abs/2210.02303.

Qingqing Huang, Daniel S Park, Tao Wang, Timo I Denk, Andy Ly, Nanxin Chen, Zhengdong Zhang, Zhishuai Zhang, Jiahui Yu, Christian Frank, et al. Noise2music: Text-conditioned music generation with diffusion models. *arXiv preprint arXiv:2302.03917*, 2023.

Yu-Siang Huang and Yi-Hsuan Yang. Pop music transformer: Beat-based modeling and generation of expressive pop piano compositions. In *Proceedings of the 28th ACM international conference on multimedia*, pp. 1180–1188, 2020.

Vladimir Iashin and Esa Rahtu. Taming visually guided sound generation. *arXiv preprint arXiv:2110.08791*, 2021.

Yujin Jeong, Yunji Kim, Sanghyuk Chun, and Jiyoung Lee. Read, watch and scream! sound generation from text and video, 2024. URL https://arxiv.org/abs/2407.05551.

Kevin Kilgour, Mauricio Zuluaga, Dominik Roblek, and Matthew Sharifi. Fr\'echet audio distance: A metric for evaluating music enhancement algorithms. *arXiv preprint arXiv:1812.08466*, 2018.

Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis, 2020a. URL https://arxiv.org/abs/2010.05646.

Qiuqiang Kong, Yin Cao, Turab Iqbal, Yuxuan Wang, Wenwu Wang, and Mark D Plumbley. Panns: Large-scale pretrained audio neural networks for audio pattern recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:2880–2894, 2020b.

Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. Diffwave: A versatile diffusion model for audio synthesis, 2021. URL https://arxiv.org/abs/2009.09761.

Felix Kreuk, Gabriel Synnaeve, Adam Polyak, Uriel Singer, Alexandre Défossez, Jade Copet, Devi Parikh, Yaniv Taigman, and Yossi Adi. Audiogen: Textually guided audio generation. *arXiv preprint arXiv:2209.15352*, 2022.

Kundan Kumar, Rithesh Kumar, Thibault de Boissiere, Lucas Gestin, Wei Zhen Teoh, Jose Sotelo, Alexandre de Brebisson, Yoshua Bengio, and Aaron Courville. Melgan: Generative adversarial networks for conditional waveform synthesis, 2019. URL https://arxiv.org/abs/1910.06711.

Rithesh Kumar, Prem Seetharaman, Alejandro Luebs, Ishaan Kumar, and Kundan Kumar. High-fidelity audio compression with improved rvqgan. *Advances in Neural Information Processing Systems*, 36:27980–27993, 2023.

Stefan Lattner and Maarten Grachten. High-level control of drum track generation using learned patterns of rhythmic interaction. In *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 35–39. IEEE, 2019.

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

Haohe Liu, Zehua Chen, Yi Yuan, Xinhao Mei, Xubo Liu, Danilo Mandic, Wenwu Wang, and Mark D. Plumbley. Audioldm: Text-to-audio generation with latent diffusion models, 2023. URL https://arxiv.org/abs/2301.12503.

Haohe Liu, Yi Yuan, Xubo Liu, Xinhao Mei, Qiuqiang Kong, Qiao Tian, Yuping Wang, Wenwu Wang, Yuxuan Wang, and Mark D Plumbley. Audioldm 2: Learning holistic audio generation with self-supervised pretraining. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024.

Jinglin Liu, Chengxi Li, Yi Ren, Feiyang Chen, and Zhou Zhao. Diffsinger: Singing voice synthesis via shallow diffusion mechanism, 2022a. URL https://arxiv.org/abs/2105.02446.

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

MultiMedia LLC. Prolific · quickly find research participants you can trust. URL https://www.prolific.com/.

Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. *Advances in Neural Information Processing Systems*, 35:5775–5787, 2022.

Simian Luo, Chuanhao Yan, Chenxu Hu, and Hang Zhao. Diff-foley: Synchronized video-to-audio synthesis with latent diffusion models. *Advances in Neural Information Processing Systems*, 36, 2024.

Sageev Oore, Ian Simon, Sander Dieleman, Douglas Eck, and Karen Simonyan. This time with feeling: Learning expressive musical performance. *Neural Computing and Applications*, 32:955–967, 2020.

Adam Roberts, Jesse Engel, Colin Raffel, Curtis Hawthorne, and Douglas Eck. A hierarchical latent vector model for learning long-term structure in music. In *International conference on machine learning*, pp. 4364–4373. PMLR, 2018.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation, 2015. URL https://arxiv.org/abs/1505.04597.

Simon Rouard, Francisco Massa, and Alexandre Défossez. Hybrid transformers for music source separation. In *ICASSP 23*, 2023.

Flavio Schneider, Ojasv Kamal, Zhijing Jin, and Bernhard Schölkopf. Mo\^ usai: Text-to-music generation with long-context latent diffusion. *arXiv preprint arXiv:2301.11757*, 2023.

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

Xuyang Shen, Dong Li, Jinxing Zhou, Zhen Qin, Bowen He, Xiaodong Han, Aixuan Li, Yuchao Dai, Lingpeng Kong, Meng Wang, et al. Fine-grained audible video description. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10585–10596, 2023.

Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio, 2016. URL https://arxiv.org/abs/1609.03499.

Heng Wang, Jianbo Ma, Santiago Pascual, Richard Cartwright, and Weidong Cai. V2a-mapper: A lightweight solution for vision-to-audio generation by connecting foundation models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 15492–15501, 2024a.

Xihua Wang, Yuyue Wang, Yihan Wu, Ruihua Song, Xu Tan, Zehua Chen, Hongteng Xu, and Guodong Sui. Tiva: Time-aligned video-to-audio generation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 573–582, 2024b.

Yazhou Xing, Yingqing He, Zeyue Tian, Xintao Wang, and Qifeng Chen. Seeing and hearing: Open-domain visual-audio generation with diffusion latent aligners, 2024. URL https://arxiv.org/abs/2402.17723.

Guy Yariv, Itai Gat, Sagie Benaim, Lior Wolf, Idan Schwartz, and Yossi Adi. Diverse and aligned audio-to-video generation via text-to-video model adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 6639–6647, 2024.

Neil Zeghidour, Alejandro Luebs, Ahmed Omran, Jan Skoglund, and Marco Tagliasacchi. Soundstream: An end-to-end neural audio codec. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:495–507, 2021.

Mingliang Zeng, Xu Tan, Rui Wang, Zeqian Ju, Tao Qin, and Tie-Yan Liu. Musicbert: Symbolic music understanding with large-scale pre-training. *arXiv preprint arXiv:2106.05630*, 2021.

Yiming Zhang, Yicheng Gu, Yanhong Zeng, Zhening Xing, Yuancheng Wang, Zhizheng Wu, and Kai Chen. Foleycrafter: Bring silent videos to life with lifelike and synchronized sounds, 2024. URL https://arxiv.org/abs/2407.01494.

Hang Zhao, Chuang Gan, Andrew Rouditchenko, Carl Vondrick, Josh McDermott, and Antonio Torralba. The sound of pixels. In *The European Conference on Computer Vision (ECCV)*, September 2018.

Hang Zhao, Chuang Gan, Wei-Chiu Ma, and Antonio Torralba. The sound of motions. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1735–1744, 2019.

A Elo Ratings Calculation

The Elo ratings were initialized with a base rating of 1,000 and a K-factor of 8. With R_A and R_B as the current Elo ratings of model A and B and R'_A and R'_B the updated ratings, the Elo ratings update formula, for each pairwise comparison of two models, is computed as follows:

$$R'_A = R_A + K \cdot \left(S_A - \frac{1}{1 + 10^{(R_B - R_A)/400}} \right) \tag{12}$$

We set $S_A = 1$ if model A wins and $S_A = 0$ if model A loses. The same formula applies to model B, with the variables of A and B switched. We randomly shuffle the pairwise comparisons 10k times and report the mean Elo rating across the bootstrapping procedure for each model.