VinTAGe: Joint Video and Text Conditioning for Holistic Audio Generation

Saksham Singh Kushwaha Yapeng Tian
The University of Texas at Dallas
{sakshamsingh.kushwaha, yapeng.tian}@utdallas.edu
Project page - https://sakshamsingh1.github.io/vintage/

Abstract

Recent advances in audio generation have focused on text-to-audio (T2A) and video-to-audio (V2A) tasks. However, T2A or V2A methods cannot generate holistic sounds (onscreen and off-screen). This is because T2A cannot generate sounds aligning with onscreen objects, while V2A cannot generate semantically complete (offscreen sounds missing). In this work, we address the task of holistic audio generation: given a video and a text prompt, we aim to generate both onscreen and offscreen sounds that are temporally synchronized with the video and semantically aligned with text and video. Previous approaches for joint text and video-to-audio generation often suffer from modality bias, favoring one modality over the other. To overcome this limitation, we introduce VinTAGe, a flow-based transformer model that jointly considers text and video to guide audio generation. Our framework comprises two key components: a Visual-Text Encoder and a Joint VT-SiT model. To reduce modality bias and improve generation quality, we employ pretrained uni-modal text-to-audio and video-to-audio generation models for additional guidance. Due to the lack of appropriate benchmarks, we also introduce VinTAGe-Bench, a dataset of 636 video-text-audio pairs containing both onscreen and offscreen sounds. Our comprehensive experiments on VinTAGe-Bench demonstrate that joint text and visual interaction is necessary for holistic audio generation. Furthermore, VinTAGe achieves state-of-the-art results on the VGGSound benchmark. We will release our pretrained models and the VinTAGe-Bench dataset to facilitate future research in this exciting field.

1. Introduction

Post-production sound design is essential for adding realism and creating an immersive experience in silent videos, including films, music videos, animations, and games [6, 15]. While onscreen sounds provide direct information about key elements of the visual world, offscreen or invisible sounds—ubiquitous in scenes—depict environments beyond the immediate visual frame, enhancing the narrative while adding authenticity and emotional depth. From creating suspense in horror films, through unseen movements in the dark, to suggesting upcoming events, such as a distant roar of an approaching plane hinting at an airport setting, these offscreen sounds are typically added during post-production and are known as non-diegetic sounds.

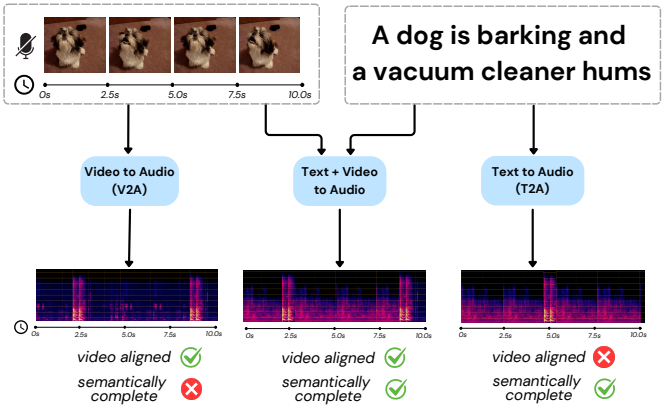

Figure 1. Our VinTAGe model can generate visually aligned and text-corresponding sounds, including both onscreen and offscreen sound sources, providing a more holistic audio experience. However, V2A models can only generate video-aligned but semantically incomplete sounds, while T2A models, though generating semantically complete sounds, are not visually aligned.

In practice, onscreen sounds are re-recorded using traditional Foley techniques [1] or neural Foley methods [5, 30, 53], and then non-diegetic sound effects are created and mixed in. However, creating authentic audio mixes with both onscreen and offscreen sounds using traditional approaches [1, 17] requires specialized skills and is often labor-intensive. Despite recent progress in audio generation, producing a holistic audio experience that seamlessly blends visually aligned onscreen sounds with semantically relevant offscreen sounds remain a significant challenge.

Recent approaches to conditional audio generation broadly focus on text-to-audio (T2A) [11, 25, 27, 32] and video-to-audio (V2A) [18, 30, 45, 53] generation tasks. V2A methods aim to generate temporally and semantically

synchronized audio with on-screen videos, while T2A models learn to generate high-quality audio that is semantically aligned with the provided text input. However, relying only on video or text as input is insufficient to generate a holistic audio experience. For example, as shown in Fig. 1, consider a household scene where a dog barks at approximately 2.5-second and 9-second in a 10-second video, while someone is vacuuming in the background. A V2A model may accurately generate the barking sounds at the appropriate times, but it would omit the ambient vacuum noise. Conversely, T2A model could produce a semantically complete soundscape, including the vacuum cleaner, but will fail to align the dog barking sounds temporally (*e.g.* placing them at 5-second). Therefore, a model that considers both text and video modalities is essential for holistic audio generation.

In this work, we address the task of *holistic audio generation* that is both visually aligned and text-driven. Given a video and a text prompt, our goal is to create high-quality audio that is temporally synchronised with video and semantically consistent with text, *i.e.*, having onscreen and (if any) offscreen sounds. This task presents significant technical challenges, as it requires the model to jointly understand and integrate information from both visual and textual modalities to guide audio generation effectively. Existing approaches [18, 49, 53] that consider both text and video as conditions assume alignment between the input video and text. This assumption leads to modality bias and results in the generation of mostly onscreen sounds. However, in real-world scenarios, this assumption is often invalid, as offscreen sounds are prevalent and may not be directly related to the onscreen content—for example, an ambulance siren (an offscreen sound) might be heard in a video of a dog barking in a home setting. Moreover, models like SonicVLM [49] require explicit knowledge of onscreen and offscreen sounds and generate multiple audio tracks and hence necessitating manual mixing by a sound designer and impractical to scale. These limitations underscore the need of a method tailored to jointly interpret visual and textual inputs for holistic audio generation. Additionally, a new benchmark is needed to effectively evaluate models on both onscreen and offscreen sound generation capabilities.

In this work, we propose VinTAGe, a novel flow-based transformer model for holistic audio generation guided by both video and text. Our architecture has two main components: the VT-Encoder and Joint VT-SiT model. The VT-Encoder efficiently encodes contextual embeddings from cross-interactions between video and text, incorporating motion and frame index information for temporal guidance. We extend the SiT model [31] to condition jointly on the features generated by the VT-Encoder. To address modality bias, improve alignment, and generation quality, we introduce a teacher-student learning framework in which pretrained text-only and video-only audio generators serve as teachers, guiding our VT-SiT model through additional modality alignment losses. Due to the lack of an appropriate benchmark for this task, we introduce VinTAGe-Bench, a new dataset to explore joint video-text to audio generation. VinTAGe-Bench consists of 636 (video, text, audio) pairs, consisting of 14 onscreen and 24 offscreen sound categories. The dataset includes 212 unique videos, each paired with one no-offscreen and two offscreen text captions, along with mixed audio tracks.

Through extensive objective and subjective evaluations, we found that jointly using both video and text features improves audio generation quality and alignment, highlighting the efficacy of the benchmark. Models that do not efficiently integrate text and video features tend to exhibit modality bias. Our carefully designed VT-Encoder and Joint VT-SiT model, enhanced with additional modality guidance losses, achieve superior performance. Furthermore, VinTAGe outperforms previous models on the VGGSound [3] dataset.

Our contributions can be summarized as follows:

- We introduce VinTAGe-Bench, a new benchmark to advance the under-explored task of holistic audio generation conditioned on both text and video inputs.
- We propose VinTAGe, a novel flow-based transformer model specifically designed for this audio generation task, featuring a carefully structured VT-Encoder and Joint VT-SiT model, enhanced with teacher-student guidance.
- Through extensive evaluations, we demonstrate the effectiveness of our approach in generating high-quality audio that is both visually and contextually aligned.

2. Related Works

Text-to-Audio Generation. Initial works on text-to-audio generation, such as AudioGen [21] and DiffSound [51], explored audio representation in discrete space given a text description, which is then decoded into waveforms. The advent of latent diffusion models [40] inspired works like AudioLDM [25], AudioLDM2 [27], Tango [11], Tango2 [32] and Make-An-Audio [14] for text-to-audio generation using latent diffusion techniques. AudioLDM employs an audio-text shared embedding space via CLAP [48], utilizing audio data during training and text during inference. Tango [11] improved audio generation with less data by using an instruction-tuned LLM FLAN-T5 [8] text encoder for both training and inference. However, although these methods can generate high-quality audio, they alone cannot achieve temporal alignment with input visual content.

Video-to-Audio Generation. This field has shown promise in synthesizing sound for silent videos. A pioneering work by Zhou et al. [55] proposed generating raw waveform samples from input video frames using SampleRNN [33]. SpecVQGAN [16] synthesizes sound using RGB and motion features within a transformer-based autoregressive model. Im2Wav [41] uses image frame embeddings from

the CLIP [38] model to condition transformer models for autoregressive audio generation. Diff-Foley [30] improves semantic and temporal synchronization by employing large-scale pretraining on audio-visual pairs and a latent diffusion model. However, these methods have mainly focused on temporal and semantic alignment with onscreen sounds.

Due to the prevalence of offscreen sounds, high-quality audio-visual pairs are scarce in real-world videos, as off-screen sounds are noisy for V2A generation models during training. Therefore, pretrained T2A generation models have been extended for V2A generation [18, 34, 45, 49, 50, 53]. "Seeing and Hearing" [50] uses an off-the-shelf pretrained ImageBind [12] model to convert video frames into text, which is then fed into AudioLDM [25]. Another approach, V2A-Mapper [45], translates CLIP visual embeddings to the CLAP embedding space before using AudioLDM. To incorporate temporal guidance, SonicVLM [49], ReWaS [18], and FoleyCrater [53] train time-conditional estimators. However, these models generally assume that text and video are always aligned—which is not always true—and using a frozen text-to-audio model can lead to one modality overpowering the other. In contrast, our proposed VinTAGe unifies visual and text modalities to simultaneously generate onscreen and offscreen sounds.

Diffusion Transformers (DiTs) and Flow Matching. Transformer-based diffusion models [10, 31, 36] have gained increased interest due to their scalability and impressive results in image generation tasks. Flow matching models [24], which model the vector field of the transport probability path from noise to data, have demonstrated stable training and superior performance compared to score-based models. The Scalable Interpolant Transformers (SiT) [31] employs a flow matching framework to improve the flexibility of connecting one data to another, further enhancing the performance of DiT models [36]. In the realm of audio generation, several models leveraging these advancements have been introduced, such as VoiceBox [23], AudioBox [44], FlashAudio [26], LuminaNext [56], and Diff-SAGe [22]. Building upon the advantages of flow matching in audio generation, we extend this approach to develop our joint visual-text to audio generation framework.

3. Method

3.1. Preliminaries

Flow Matching. Flow matching [24, 28] linearly interpolates between noise and target data in a straight line. Specifically, given data $x \sim p(x)$ and Gaussian noise $\epsilon \sim \mathcal{N}(0, I)$, an interpolation-based forward process is defined as:

$$x_t = \alpha_t x + \beta_t \epsilon, \tag{1}$$

where $\alpha_1 = \beta_0 = 1$, $\alpha_0 = \beta_1 = 0$, to satisfy this interpolation on $t \in [0, 1]$ between $x_0 = \epsilon$ and $x_1 = x$. In our framework, we adopt the linear interpolation schedule between noise and original data, *i.e.,* $x_t = tx + (1 - t)\epsilon$.

This formulation indicates a uniform transformation with constant velocity between noise and data. The corresponding time-dependent velocity field is given by

$$v_t(x_t) = \dot{\alpha}_t x + \dot{\beta}_t \epsilon \tag{2}$$

where $\dot{\alpha}$ and $\dot{\beta}$ denote time derivative of α and β. This velocity can be approximated with a model $v_\theta(x_t, t)$ by minimizing the training objective:

$$\mathcal{L} = \mathbb{E}_{x,\epsilon,t}[\| \, v_\theta(x_t, t) - \dot{\alpha}_t x - \dot{\beta}_t \epsilon \, \|^2], \tag{3}$$

During inference, we integrate the probability flow ODE, starting from a Gaussian noise sample x_0 at $t = 0$ and evolving it to $t = 1$ using an ODE solver such as Euler's method. We iteratively update x_t to obtain a final sample x_1, which aligns with the target data distribution.

Audio VAE and Vocoder. Training generative models directly on waveform or mel-spectrograms (m) is computationally expensive. Hence previous approaches [7, 11, 25] tackled this by training an autoencoder that compresses the spatial representation using Encoder $\mathcal{E}$. Latent generative models are trained on these compressed representations $x = \mathcal{E}(m)$. During inference, the model generates latent $\hat{x}$, which is decoded back into mel-spectrograms $\hat{m} = \mathcal{D}(\hat{x})$ using the decoder $\mathcal{D}$ of the autoencoder. A vocoder is used to reconstruct waveform from synthesized sound spectrogram $\hat{m}$. Similar to previous approaches [11, 32], we use Audio VAE and HiFi-GAN [20] vocoder from Liu et al. [25].

3.2. Proposed Architecture

Our architecture consists of three components: Visual and Text Encoder, Joint Visual-Text (VT) SiT model, and Unimodal teacher models. We will explain them one-by-one.

Visual and Text (VT) Encoder. We employ a VT encoder module to generate embeddings from video frames and text inputs. Within this module as illustrated in Fig. 2A, the cross-modal embeddings interact and are associated to produce conditioning information for audio generation.

Text Encoder. We adopt FLAN-T5 [39] as our text encoder to extract contextual language embeddings. A text prompt T is encoded to $\tau \in \mathbb{R}^{L \times d_{txt}}$, where L is the token count and d_{txt} is the text embedding dimension.

Video Encoder. We utilize the visual encoder from CLIP [38] to extract video representations. Given video frames $V \in \mathbb{R}^{N \times 3 \times H \times W}$, the CLIP visual encoder [38] transforms them into embeddings $\mathcal{V}_c \in \mathbb{R}^{N \times d_{\text{clip}}}$, where each frame is represented by a d_{clip}-dimensional vector capturing its semantic content. To incorporate temporal information, we compute mean of optical flow [13, 52] magnitude values sampled at the same rate as the video frames (N). We also include frame index numbers as explicit temporal position encodings. Both the optical flow values and

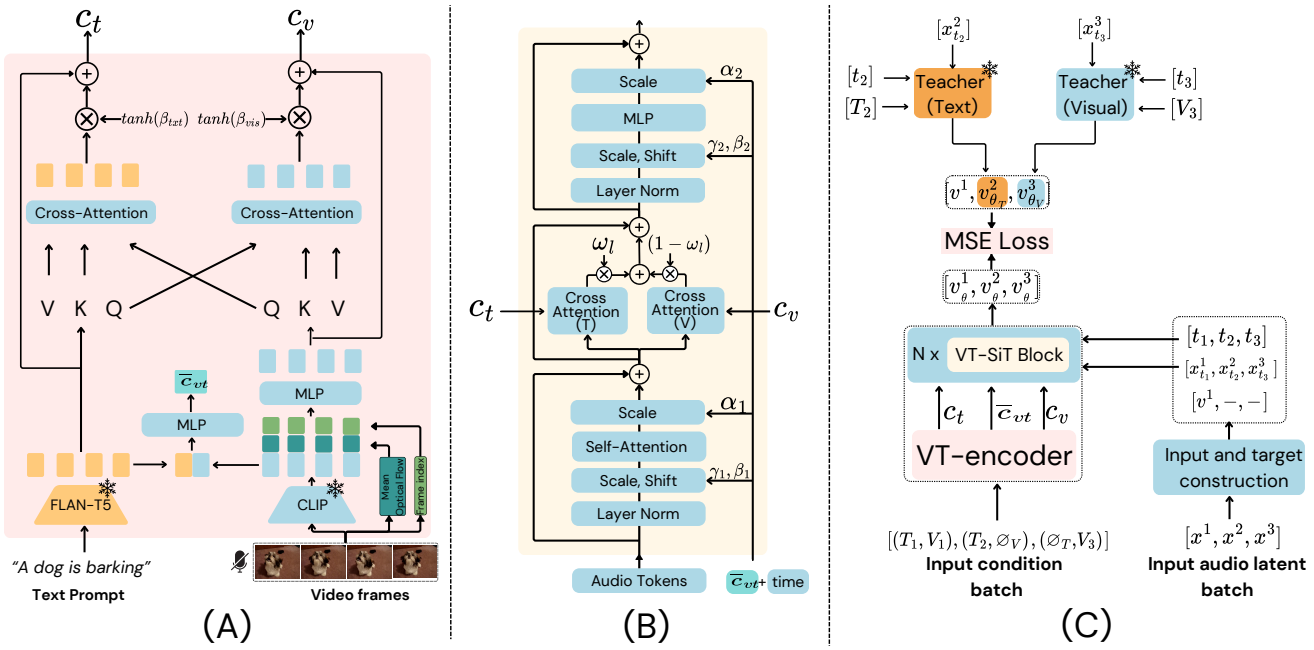

Figure 2. (A) VT-Encoder, (B) Joint VT-SiT block, (C) Overall training pipeline.

frame indices are encoded using sinusoidal embeddings of dimensions d_{opt} and d_{idx} respectively. We concatenate these embeddings with the CLIP embeddings $\mathcal{V}_c$, resulting in $\mathcal{V}_{\text{coi}} \in \mathbb{R}^{N\times(d_{\text{clip}}+d_{\text{opt}}+d_{\text{idx}})}$. Finally, we project the concatenated features using a linear layer for greater flexibility.

Previous approaches to adding temporal control requires extensive pretraining [30] for temporally aligned features or involves training separate modules [18, 49, 53] for onset or energy detection from input video or expects user inputs [49]. In contrast, our use of mean optical flow values provides a temporal energy map that is easier to get and directly guide audio generation.

Visual-Text Cross-Attention. To jointly generate audio from text and video, we utilize cross-attention between the text and visual embeddings to enable cross-modal association. However, there may be scenarios where the two modalities are entirely unrelated or when we want to condition on them individually. To address this, we introduce a gated cross-attention mechanism with tanh gating [9] and skip connections, formulated as:

$$c_v = \mathcal{V}_{\text{coi}} + \tanh(\beta_{vis}) \cdot \text{CA}(Q_{\text{vis}}, K_{\text{txt}}, V_{\text{txt}})$$
$$c_t = \mathcal{T} + \tanh(\beta_{txt}) \cdot \text{CA}(Q_{\text{txt}}, K_{\text{vis}}, V_{\text{vis}})$$

where $CA(Q, K, V)$ denotes the cross-attention operation with *query* Q, *key* K, *and value* V. In this mechanism, β_{txt} and β_{vis} are zero-initialized learnable parameters. They can help to incorporate cross-modal information gradually into the modality sequences and hence helps in stable training.

With the semantic visual and text embeddings, we can integrate them to generate a joint semantic text-visual condition vector: $\bar{c}_{vt} = \text{MLP}\left([\text{AvgPool}(\mathcal{V}_c);\ \text{AvgPool}(\mathcal{T})]\right)$.

This joint pooled semantic embedding $\bar{c}_{vt}$ (along with time embedding) will be leveraged for global modulation using adaptive layer normalization (AdaLN) [37]. Additionally, contextual embeddings (c_v, c_t) will directly interact with the audio latent via cross-attention in VT-SiT.

Joint VT-SiT. Our audio generator is built upon image generation model: SiT [31]. As illustrated in Fig. 2B, input noisy audio latent is passed through N blocks of the VT-SiT model to generate the predicted velocity v_θ.

We enhance the SiT block by adding text and visual cross-attention layers to leverage both modalities in guiding audio generation. In addition to AdaLN-based global modulation [36], this provides direct interaction and contextual guidance via attention [43]. We introduce learnable weight parameters ω_l [7] in each block l to automatically learn the importance of each modality's guidance. The updated intermediate representation $\tilde{x}^{in}$ in block l is computed as:

$$\tilde{x}^{in} = x^{in} + \omega_l \cdot \text{CA}(Q_{x^{in}}, K_{c_t}, V_{c_t}) + (1-\omega_l) \cdot \text{CA}(Q_{x^{in}}, K_{c_v}, V_{c_v}),$$

where x^{in} is the intermediate noisy audio latent before cross-attention, and $\tilde{x}^{in}$ is after cross-attention with text and video embeddings. The learnable parameter ω_l allows the model to balance the contributions from the text and visual modalities dynamically at each layer.

Visual and Text Teachers. Generative models often inherit biases present in the training data [35], which can adversely affect their performance. We observe that previous works [18, 53] using text and video for audio generation suffer from *modality bias*, tending to prioritize visual signals over textual input. To mitigate this issue, we employ

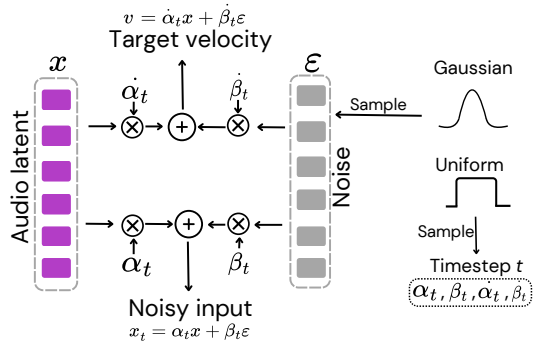

Figure 3. Input and target construction. During training, an audio latent x is interpolated at time t (sampled from uniform distribution), with random noise sampled from gaussion ϵ and generates noisy input and target velocity.

pretrained teacher models that are individually trained for text-to-audio and video-to-audio generation tasks. These teacher models share architectures similar to our VT-SiT model, including cross-attention mechanisms and adaptive layer normalization. By utilizing these teacher models, we guide the joint visual-text audio generation process to effectively balance both modalities. Detailed architectural descriptions are provided in the appendix.

3.3. Training and Inference

In this subsection, we describe the training pipeline of VinTAGe, as illustrated in Fig. 2C, and explain the inference process using classifier-free guidance, shown in Fig. 4.

Our task is to learn velocity field $v_\theta(x_t, t, c_v, c_t)$, given the noisy audio latent input x_t interpolated at time t and given the visual and text contextual conditions: c_v and c_t.

Creating Batches. In order to avoid any modality bias during training, we use teacher models. In order to effectively use them each batch is carefully generated to maximize GPU memory and utilization. Each batch consists of a multiple of 3 inputs, where:

$$(T_i, V_i) = \begin{cases} (T_i, V_i) & \text{if } i \mod 3 = 0, \\ (T_i, \varnothing_V) & \text{if } i \mod 3 = 1, \\ (\varnothing_T, V_i) & \text{if } i \mod 3 = 2. \end{cases} \quad (4)$$

where T_i represents the text input, V_i represents the visual input, $\varnothing_T$ indicates unconditioned text, and $\varnothing_V$ indicates unconditioned visual input.

Input and Target Construction. As shown in Fig. 3, each audio latent x_i is converted to noisy input and target velocity using Eqs. (1) and (2). In our case, we discard the target velocity for cases with $(\varnothing_T, V_i)$ and $(T_i, \varnothing_V)$ and replace them with the predicted velocity from teacher models. We observe that this not only helps in reducing modality bias but also improve generation quality.

Teacher Guided Alignment. To mitigate modality bias and enhance the performance of our audio generation model, we

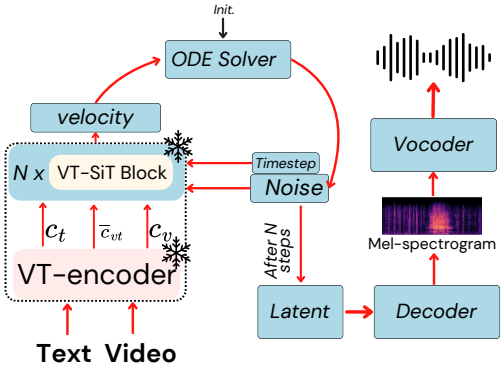

Figure 4. Inference pipeline.

introduce *teacher-guided alignment losses*. This approach leverages pretrained teacher models that are individually trained on T2A and V2A generation tasks.

The alignment losses with the text-only teacher ($\mathcal{L}_t$) and the video-only teacher ($\mathcal{L}_v$) are defined as:

$$\mathcal{L}_t = \mathbb{E}_{x,\epsilon,t}\left[\left\|v_\theta(x_t, t, \varnothing_V, c_t) - v_{\theta_\mathrm{T}}(x_t, t, c_t)\right\|^2\right],$$

$$\mathcal{L}_v = \mathbb{E}_{x,\epsilon,t}\left[\left\|v_\theta(x_t, t, c_v, \varnothing_T) - v_{\theta_\mathrm{V}}(x_t, t, c_v)\right\|^2\right],$$

where, v_θ is the velocity predicted by our joint model. v_{θ_T} and v_{θ_V} are the velocities predicted by the text- and visual-only teacher models, respectively.

The final objective combines the standard loss with these alignment losses:

$$\mathcal{L}_{\text{final}} = \mathbb{E}_{x,\epsilon,t}\left[\left\|v_\theta(x_t, t, c_v, c_t) - \dot{\alpha}_t x - \dot{\beta}_t \epsilon\right\|^2\right] + \lambda_v \mathcal{L}_v + \lambda_t \mathcal{L}_t.$$

We find that incorporating the teacher-predicted velocities helps the model in two ways: regularization and improved quality. By comparing the joint model's outputs with those of the teacher models when only one modality is provided, we regularize the model to prevent dominance by either modality, thus mitigating the modality bias. Additionally, we observe that the model learns more effectively with teacher guidance than with the original velocity targets, resulting in higher-quality generated audio.

Augmentation. To handle cases where text and video may not be fully aligned, we extend T2A augmentation methods [11] by mixing audio, text, and video samples in various combinations, effectively increasing the diversity of our training data and enhancing the model's generation ability. More details are provided in the appendix.

Inference using classifier-free guidance. Our model learns to predict the velocity conditioned on both text and visual inputs. To enhance the quality of the generated audio, we leverage *classifier-free guidance* based on both conditions. During training, in addition to dropping text and visual conditions for teacher alignment, we also independently drop text and visual conditions with a 10% probability to facilitate classifier-free guidance. Following previous

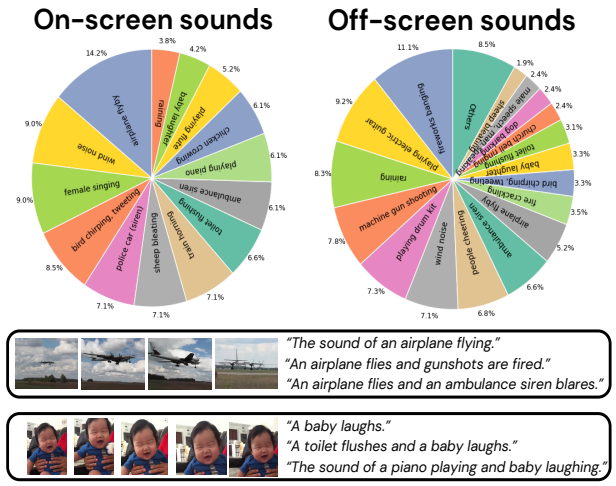

Figure 5. The statistics and examples of our VinTAGe-Bench.

multi-conditional works [2, 4], during inference we introduce two guidance scales, s_{vis} and s_{txt}, to control the influence of each modality. The modified velocity prediction is computed as:

$$\begin{aligned}\tilde{v}_\theta(x_t, c_v, c_t) = v_\theta(x_t, \varnothing_V, \varnothing_T) \\ + s_{\text{vis}} \cdot (v_\theta(x_t, c_v, \varnothing_T) - v_\theta(x_t, \varnothing_V, \varnothing_T)) \\ + s_{\text{txt}} \cdot (v_\theta(x_t, c_v, c_t) - v_\theta(x_t, c_v, \varnothing_T)),\end{aligned}$$

where s_{txt} and s_{vis} represent the weights of visual and text guidance. As illustrated in Fig. 4, we start the generation process from noise and, using the modified velocity function $\tilde{v}_\theta$, we perform N steps. The resulting latent representation is then passed through an Audio VAE Decoder and Vocoder to obtain the synthesized audio waveform. We empirically found s_{vis}=s_{txt}=2.5 to work well in our case and unless mentioned we use these values during inference.

4. Experiments and Results

Datasets. We adopt both VGGSound [3] and our newly established VinTAGe-Bench in our experiments.

VGGSound. This dataset consists of ∼200K 10-second videos spanning 309 classes. Following prior V2A generation work [16, 30], we adopt the original train/test splits. Since the dataset includes only class labels, we enhance it with language descriptions by using LLM-generated captions from Auto-ACD [42]. For any missing captions, we use a text prompt "*The sound of* { *class-label* }".

VinTAGe-Bench. To the best of our knowledge, no suitable dataset exists for benchmarking audio generation conditioned on both text and video. We therefore created a dataset ensuring generated audio is both temporally synchronized with the video content and semantically consistent with accompanying text captions.

Based on the VGGSound test set, we selected 212 videos from 14 diverse classes, excluding those with limited object visibility, static images, and animations. For each class, we added offscreen sounds that could naturally co-occur or be less likely to co-occur with the onscreen sound. After mixing the audio, we filtered out scenarios where (1) Sounds with similar frequencies were mixed (*e.g.*, vacuum cleaner and chainsaw), making them difficult to distinguish; (2) Offscreen sounds were mixed with "busy" onscreen sounds (*e.g.*, drums playing and dog barking), hindering recognition. Offscreen sounds were randomly selected from the VGGSound test set and mixed with the original audio, adjusting gains for balanced pressure levels [11]. Each video was paired with three scenarios: one with only onscreen sounds and two with added offscreen sounds. Text captions were generated for each scenario, with the order of on-screen and off-screen sound descriptions randomly shuffled to prevent bias. This resulted in 636 (video, text, audio) pairs. The distribution of sounds is shown in Fig. 5.

Baselines. We compare our method with recent state-of-the-art V2A generation methods—SpecVQGAN [16], Seeing-and-Hearing [50], and Diff-Foley [30]—and T2A methods like Tango2 [32], AudioLDM2 [27], and Make-an-audio [14]. We also include models that use both text and video inputs, such as FoleyCrafter [53] and ReWaS [18]. To adapt single-condition models, we extend them by combining video and text captions. For Tango2, we generate video captions using LLaVa-Next [54] and concatenate them with the original text captions, creating Tango2+LLaVa. Similarly, for Seeing-and-Hearing, we combine key-frame captions with original text captions to generate audio using AudioLDM, which we call Seeing-and-Hearing-VT.

Evaluation Metrics. We employ both objective and subjective evaluations to measure audio generation performance.

Objective Evaluation. We evaluate perceptual quality using Frechet Audio Distance (FAD) [19] and Melception-based Frechet Distance (FID) [16], two key metrics for T2A and V2A generation tasks. Since FID and FAD capture distribution-level similarity, we also measure paired sample-level similarity with Mean KL Divergence (MKL) [16]. We evaluate audio-video (AV) and audio-text (AT) alignment on CLIP [38] space by calculating cosine similarity times 100, following previous works [41, 45, 53], using Wav2CLIP [47] as audio encoder and CLIP-text and image encoders for visual and text embedding. We compare the results on the mean of AV and AT scores for this joint task. Additionally, we use a classifier to calculate if the generated audio faithfully captures both on-screen and off-screen concepts. Specifically, we separate on-screen and off-screen sounds based on text labels using Separate-Anything Model [29] and then perform classification using ONE-PEACE [46]. Both are pretrained models with open-source and superior performance on VGGSound. We report mean results for this joint task.

Subjective Evaluation. To complement objective metrics, we conduct listening test to measure quality (MOS-Q),

Table 1. Comparison on VinTAGe-Bench. Top-2 results are highlighted.

Model	Txt	Vis	Generation Quality			Alignment			Concept Accuracy(%)			Subjective Metrics		
			FAD$_\downarrow$	FID$_\downarrow$	MKL$_\downarrow$	AT$_\uparrow$	AV$_\uparrow$	Mean$_\uparrow$	On-acc$_\uparrow$	Off-acc$_\uparrow$	Mean$_\uparrow$	MOS-Q$_\uparrow$	MOS-F$_\uparrow$	MOS-T$_\uparrow$
SpecVQGAN [16]	✗	✓	6.10	31.73	6.44	18.73	4.61	11.67	25.0	10.61	17.80	1.36	1.08	1.34
Seeing-and-Hearing [50]	✗	✓	5.08	27.04	6.47	15.78	7.93	11.85	34.9	9.43	22.16	-	-	-
Diff-Foley [30]	✗	✓	6.63	19.78	6.38	18.25	9.39	13.82	40.72	7.05	23.88	1.86	1.26	2.36
Make-an-Audio [14]	✓	✗	**4.05**	**19.01**	5.19	18.54	7.42	12.98	52.04	35.14	47.83	-	-	-
AudioLDM2 [27]	✓	✗	5.40	20.75	5.52	22.02	6.65	14.33	54.4	26.65	40.52	2.42	2.52	2.10
Tango2 [32]	✓	✗	5.85	36.01	4.94	23.84	7.19	15.51	62.57	48.58	**55.57**	2.86	**3.34**	2.56
Seeing-and-Hearing-VT	✓	✓	4.89	22.44	5.00	17.75	8.99	13.37	43.23	17.68	30.45	-	-	-
Tango2 + LLaVA [54]	✓	✓	4.12	29.1	**4.59**	24.14	7.35	15.75	57.86	40.33	49.09	2.88	2.90	2.52
ReWaS [18]	✓	✓	8.01	36.88	7.54	21.03	4.48	12.75	28.14	11.32	19.73	-	-	-
FoleyCrafter [53]	✓	✓	5.81	25.64	4.94	21.36	10.57	**15.96**	64.93	21.69	43.31	**2.92**	2.60	**2.96**
VinTAGe (Ours)	✓	✓	**3.05**	**16.43**	**4.74**	22.29	9.83	**16.06**	57.7	43.63	**50.66**	**3.36**	**3.58**	**3.36**

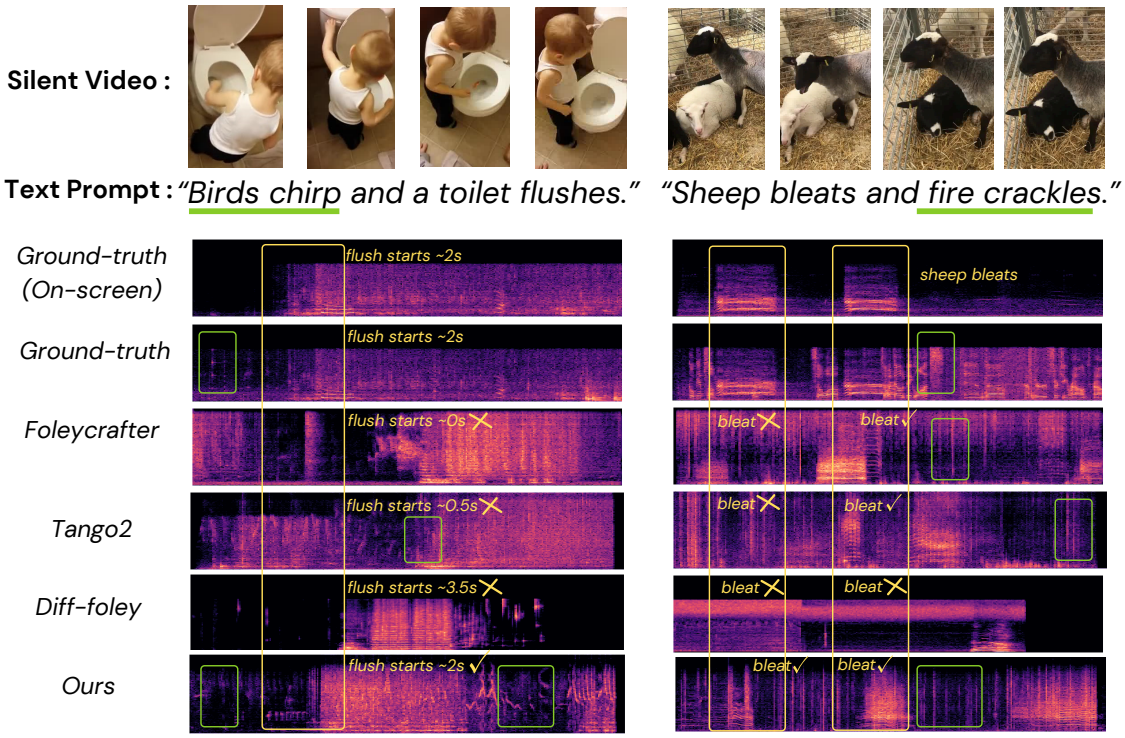

Figure 6. Examples of generated audio from VinTAGe-Bench.

faithfulness (MOS-F) with both onscreen and offscreen sounds, and temporal alignment (MOS-T) of video with on-screen sounds. We ask 10 listeners to rate 35 samples on a discrete 5-point scale and report the mean opinion score. We also note that previous audio-visual temporal metrics [30, 52] are not suitable in our case due to mixed audio containing offscreen sounds. Six baseline methods are compared.

4.1. Comparison Results

Our quantitative results on VinTAGe are shown in Tab. 1.

T2A vs V2A. We observe that T2A models achieve superior audio generation quality compared to V2A models. A key reason is the larger availability of high-quality audio-text pairs, while offscreen sounds remain prevalent in videos.

Joint Generation is Better. Combining visual and text features significantly enhances generation quality. For example, adding video captions to Tango2 (Tango2+LLaVA) improves all quality metrics, boosting FAD by ∼30%. This simple extension even surpasses FoleyCrafter and ReWaS, which also use text and video but less effectively. Our approach achieves the best FAD and FID scores, highlighting the superior quality of the synthesized audio.

Audio-Text and Audio-Visual Alignment. We observe that strong V2A models, such as Diff-Foley, generally achieve better AV alignment than T2A models but lack alignment with the text modality due to missing text information. Similarly, T2A models align well with text but fall short in visual alignment. Notably, Tango2+LLaVA improves upon Tango2 in mean alignment by adding video captions, while our approach achieves the best overall alignment.

Joint Generation Methods Suffer from Modality bias. For

Table 2. Evaluations on VGGSound [3] dataset.

Model	FPS↓	Objective metrics				Subjective metrics	
		FID↓	MKL↓	ISc↑	AV↑	MOS-Q↑	MOS-T↑
SpecVQGAN (RGB+Flow) [16]	21.5	8.93	6.93	30.01	5.07	-	-
SpecVQGAN (ResNet50) [16]	21.5	9.70	7.03	30.80	5.87	-	-
Im2Wav [41]	30	11.44	5.20	39.30	7.82	-	-
Diff-foley [30]	**4**	9.87	6.43	62.37	9.17	3.03	3.46
ReWaS [18]	25	28.28	8.92	12.76	4.69	-	-
FoleyCrafter [53]	15	9.17	4.48	62.49	**10.13**	3.80	3.17
VinTAGe (Ours)	**4**	**6.65**	**4.12**	**63.95**	9.68	**3.86**	**3.88**

Table 3. Ablation on different model components.

	FID↓	Mean-Align↑	Mean-Acc (%)↑	ΔOn-Off(%) ↓
VinTAGe	**16.43**	**16.06**	**50.66**	14.07
w/o guidance loss	19.48	15.95	43.15	18.39
w/o CA VT-Encoder	18.80	16.04	41.66	**13.52**
w/o augmentation	19.22	15.65	38.55	20.99

on- and off-screen accuracy, T2A models generally perform better, benefiting from large-scale, high-quality training data compared to V2A methods. Our mean concept accuracy is the second highest, slightly below Tango2 due to the limited quality of data, amount of data and to consider both text and video modalities. An interesting observation is that the two joint generation methods, ReWaS and FoleyCrafter, exhibit high on-screen accuracy but very low off-screen accuracy, indicating a strong visual bias. In contrast, our proposed approach effectively mitigates this issue.

Subjective Study and Qualitative Results. Our approach outperforms other methods across all three subjective metrics, validating the quality, faithfulness, and temporal sync of the generated sounds. In Fig. 6, we compare our approach with FoleyCrafter (TV2A), Tango2 (T2A), and Diff-foley (V2A). We see that our model can follow both onscreen sounds (*i.e.* toilet flushing starting at ∼2 sec and sheep bleating twice) while other methods, even FoleyCrafter and Diff-foley, fail to follow.

Comparison on VGGSound. We compare our approach on standard V2A generation benchmark: VGGSound [3], with commonly used V2A metrics as shown in Tab. 2. We take the text captions from Auto-ACD [42] and in order to have a fair comparison, we update previous text and video models (*e.g.*, Foleycrafter and ReWaS) with these captions, which originally considered text labels. Our approach achieves the overall best performance. Subjective results also highlight good quality and temporal alignment.

5. Discussion and Analysis

Ablations on Model Components. We do an ablation study to explore the effect of each component for our joint text-video to audio generation. From Tab. 3, we observe that each component positively affects the generation quality, alignment, and concept accuracy. Additionally, we find that augmentation is truly helpful in improving the concept accuracy. As shown in Fig. 7, guidance loss helps in learning better generation quality (FID) and concept accuracy (Mean-Acc), as it is easier to learn from teacher models. We additionally see that the modality bias (On-Offscreen accuracy) is mitigated by guidance loss and augmentation. Furthermore, in Tab. 4, we observe that mean-flow and frame-index features help to improve the generation quality.

Table 4. Ablation on V2A teacher model.

	FID↓	Mean-Align↑	Mean-Acc(%)↑
Teacher - V	**20.32**	**16.04**	**31.56**
w/o mean-flow	26.36	15.46	21.69
w/o frame-idx	25.94	15.30	21.02

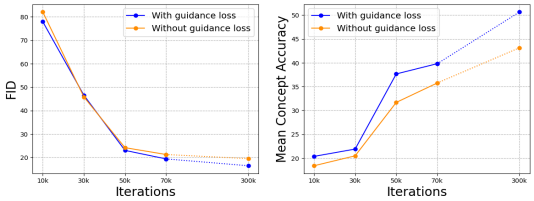

Figure 7. Ablation on the alignment loss.

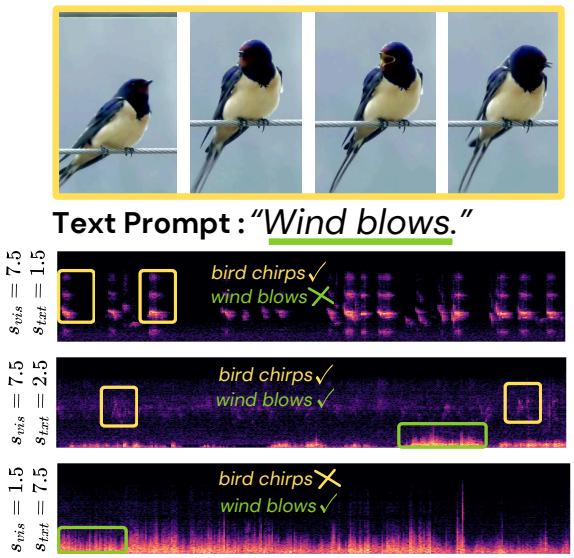

Figure 8. Effect of classifier free guidance.

Effect of s_{vis} and s_{txt}. We further analyse the effect of classifier-free guidance in Fig. 8. We observe that s_{vis} controls the visual content and s_{txt} controls the text content. For a simple scenario in which text and video are unrelated, higher s_{vis} (s_{vis}=7.5,s_{txt}=1.5) would result in only visual sounds and vice-versa. A good balance can result in a mixture of both the sounds (s_{vis}=7.5,s_{txt}=2.5).

6. Conclusion

In this paper, we tackle the task of holistic audio generation, aiming to produce audio that is both visually synchronized and aligned with text content. To address the lack of suitable datasets, we introduce VinTAGe-Bench—a curated dataset from the VGGSound, containing both onscreen and offscreen sounds. Additionally, we propose VinTAGe, a flow-based transformer model that jointly processes text and video inputs, leveraging V2A and T2A teacher models for guided learning. Extensive experiments and ablation studies validate the effectiveness of our approach.

Acknowledgments. This work was supported in part by an Amazon Research Award Fall 2023. Any opinions, findings, and conclusions expressed in this material are those of the authors and do not reflect the views of Amazon.

References

[1] Vanessa Theme Ament. *The Foley grail: The art of performing sound for film, games, and animation*. Routledge, 2014. 1

[2] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18392–18402, 2023. 6

[3] Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. Vggsound: A large-scale audio-visual dataset. In *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2020. 2, 6, 8

[4] Shoufa Chen, Mengmeng Xu, Jiawei Ren, Yuren Cong, Sen He, Yanping Xie, Animesh Sinha, Ping Luo, Tao Xiang, and Juan-Manuel Perez-Rua. Gentron: Diffusion transformers for image and video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6441–6451, 2024. 6

[5] Ziyang Chen, Prem Seetharaman, Bryan Russell, Oriol Nieto, David Bourgin, Andrew Owens, and Justin Salamon. Video-guided foley sound generation with multimodal controls. *arXiv preprint arXiv:2411.17698*, 2024. 1

[6] Michel Chion. *Audio-vision: sound on screen*. Columbia University Press, 2019. 1

[7] Sanjoy Chowdhury, Sayan Nag, K J Joseph, Balaji Vasan Srinivasan, and Dinesh Manocha. Melfusion: Synthesizing music from image and language cues using diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26826–26835, 2024. 3, 4

[8] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53, 2024. 2

[9] Peng Gao, Le Zhuo, Ziyi Lin, Chris Liu, Junsong Chen, Ruoyi Du, Enze Xie, Xu Luo, Longtian Qiu, Yuhang Zhang, et al. Lumina-t2x: Transforming text into any modality, resolution, and duration via flow-based large diffusion transformers. *arXiv preprint arXiv:2405.05945*, 2024. 4

[10] Peng Gao, Le Zhuo, Chris Liu, , Ruoyi Du, Xu Luo, Longtian Qiu, Yuhang Zhang, et al. Lumina-t2x: Transforming text into any modality, resolution, and duration via flow-based large diffusion transformers. *arXiv preprint arXiv:2405.05945*, 2024. 3

[11] Deepanway Ghosal, Navonil Majumder, Ambuj Mehrish, and Soujanya Poria. Text-to-audio generation using instruction tuned llm and latent diffusion model. *arXiv preprint arXiv:2304.13731*, 2023. 1, 2, 3, 5, 6

[12] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *CVPR*, 2023. 3

[13] Berthold KP Horn and Brian G Schunck. Determining optical flow. *Artificial intelligence*, 17(1-3):185–203, 1981. 3

[14] Rongjie Huang, Jiawei Huang, Dongchao Yang, Yi Ren, Luping Liu, Mingze Li, Zhenhui Ye, Jinglin Liu, Xiang Yin, and Zhou Zhao. Make-an-audio: Text-to-audio generation with prompt-enhanced diffusion models. In *International Conference on Machine Learning*, pages 13916–13932. PMLR, 2023. 2, 6, 7

[15] Sander Huiberts. Captivating sound. *Utrecht School of the Arts*, 2010. 1

[16] Vladimir Iashin and Esa Rahtu. Taming visually guided sound generation. In *British Machine Vision Conference (BMVC)*, 2021. 2, 6, 7, 8

[17] Roey Izhaki. *Mixing audio: concepts, practices, and tools*. Routledge, 2017. 1

[18] Yujin Jeong, Yunji Kim, Sanghyuk Chun, and Jiyoung Lee. Read, watch and scream! sound generation from text and video. 2024. 1, 2, 3, 4, 6, 7, 8

[19] 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. 6

[20] Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. *Advances in neural information processing systems*, 33:17022–17033, 2020. 3

[21] 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. 2

[22] Saksham Singh Kushwaha, Jianbo Ma, Mark RP Thomas, Yapeng Tian, and Avery Bruni. Diff-sage: End-to-end spatial audio generation using diffusion models. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2025. 3

[23] Matthew Le, Apoorv Vyas, Bowen Shi, Brian Karrer, Leda Sari, Rashel Moritz, Mary Williamson, Vimal Manohar, Yossi Adi, Jay Mahadeokar, and Wei-Ning Hsu. Voicebox: Text-guided multilingual universal speech generation at scale, 2023. 3

[24] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022. 3

[25] 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. *Proceedings of the International Conference on Machine Learning*, pages 21450–21474, 2023. 1, 2, 3

[26] Huadai Liu, Jialei Wang, Rongjie Huang, Yang Liu, Heng Lu, Wei Xue, and Zhou Zhao. Flashaudio: Rectified flows for fast and high-fidelity text-to-audio generation. *arXiv preprint arXiv:2410.12266*, 2024. 3

[27] 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, 32:2871–2883, 2024. 1, 2, 6, 7

[28] 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*, 2022. 3

[29] Xubo Liu, Qiuqiang Kong, Yan Zhao, Haohe Liu, Yi Yuan, Yuzhuo Liu, Rui Xia, Yuxuan Wang, Mark D Plumbley, and Wenwu Wang. Separate anything you describe. *arXiv preprint arXiv:2308.05037*, 2023. 6

[30] Simian Luo, Chuanhao Yan, Chenxu Hu, and Hang Zhao. Diff-foley: Synchronized video-to-audio synthesis with latent diffusion models, 2023. 1, 3, 4, 6, 7, 8

[31] Nanye Ma, Mark Goldstein, Michael S. Albergo, Nicholas M. Boffi, Eric Vanden-Eijnden, and Saining Xie. Sit: Exploring flow and diffusion-based generative models with scalable interpolant transformers. 2024. 2, 3, 4

[32] Navonil Majumder, Chia-Yu Hung, Deepanway Ghosal, Wei-Ning Hsu, Rada Mihalcea, and Soujanya Poria. Tango 2: Aligning diffusion-based text-to-audio generations through direct preference optimization, 2024. 1, 2, 3, 6, 7

[33] Soroush Mehri, Kundan Kumar, Ishaan Gulrajani, Rithesh Kumar, Shubham Jain, Jose Sotelo, Aaron Courville, and Yoshua Bengio. Samplernn: An unconditional end-to-end neural audio generation model. *arXiv preprint arXiv:1612.07837*, 2016. 2

[34] Shentong Mo, Jing Shi, and Yapeng Tian. Text-to-audio generation synchronized with videos. *arXiv preprint arXiv:2403.07938*, 2024. 3

[35] Rishubh Parihar, Abhijnya Bhat, Abhipsa Basu, Saswat Mallick, Jogendra Nath Kundu, and R Venkatesh Babu. Balancing act: Distribution-guided debiasing in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6668–6678, 2024. 4

[36] William Peebles and Saining Xie. Scalable diffusion models with transformers, 2023. 3, 4

[37] Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI conference on artificial intelligence*, 2018. 4

[38] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 3, 6

[39] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020. 3

[40] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 2

[41] Roy Sheffer and Yossi Adi. I hear your true colors: Image guided audio generation, 2022. 2, 6, 8

[42] Luoyi Sun, Xuenan Xu, Mengyue Wu, and Weidi Xie. Auto-acd: A large-scale dataset for audio-language representation learning. In *ACM Multimedia*, 2024. 6, 8

[43] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017. 4

[44] Apoorv Vyas, Bowen Shi, Matthew Le, Andros Tjandra, Yi-Chiao Wu, Baishan Guo, Jiemin Zhang, Xinyue Zhang, Robert Adkins, William Ngan, Jeff Wang, Ivan Cruz, Bapi Akula, Akinniyi Akinyemi, Brian Ellis, Rashel Moritz, Yael Yungster, Alice Rakotoarison, Liang Tan, Chris Summers, Carleigh Wood, Joshua Lane, Mary Williamson, and Wei-Ning Hsu. Audiobox: Unified audio generation with natural language prompts, 2023. 3

[45] 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*, 2024. 1, 3, 6

[46] Peng Wang, Shijie Wang, Junyang Lin, Shuai Bai, Xiaohuan Zhou, Jingren Zhou, Xinggang Wang, and Chang Zhou. One-peace: Exploring one general representation model toward unlimited modalities. *arXiv preprint arXiv:2305.11172*, 2023. 6

[47] Ho-Hsiang Wu, Prem Seetharaman, Kundan Kumar, and Juan Pablo Bello. Wav2clip: Learning robust audio representations from clip. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022. 6

[48] Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023. 2

[49] Zhifeng Xie, Shengye Yu, Qile He, and Mengtian Li. Sonicvisionlm: Playing sound with vision language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26866–26875, 2024. 2, 3, 4

[50] Yazhou Xing, Yingqing He, Zeyue Tian, Xintao Wang, and Qifeng Chen. Seeing and hearing: Open-domain visual-audio generation with diffusion latent aligners. In *CVPR*, 2024. 3, 6, 7

[51] Dongchao Yang, Jianwei Yu, Helin Wang, Wen Wang, Chao Weng, Yuexian Zou, and Dong Yu. Diffsound: Discrete diffusion model for text-to-sound generation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:1720–1733, 2023. 2

[52] 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*, pages 6639–6647, 2024. 3, 7

[53] Yanhong Zeng Zhening Xing Yuancheng Wang Zhizheng Wu Kai Chen Yiming Zhang, Yicheng Gu. Foleycrafter: Bring silent videos to life with lifelike and synchronized sounds, 2024. 1, 2, 3, 4, 6, 7, 8

[54] Yuanhan Zhang, Bo Li, haotian Liu, Yong jae Lee, Liangke Gui, Di Fu, Jiashi Feng, Ziwei Liu, and Chunyuan Li. Llava-next: A strong zero-shot video understanding model, 2024. 6, 7

[55] Yipin Zhou, Zhaowen Wang, Chen Fang, Trung Bui, and Tamara L. Berg. Visual to sound: Generating natural sound for videos in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2

[56] Le Zhuo, Ruoyi Du, Xiao Han, Yangguang Li, Dongyang Liu, Rongjie Huang, Wenze Liu, et al. Lumina-next: Making lumina-t2x stronger and faster with next-dit. *arXiv preprint arXiv:2406.18583*, 2024. 3