OmniSonic: Towards Universal and Holistic Audio Generation from Video and Text

Supplementary Material

A. Appendix

A.1. Training and Inference Process

Training. During the training of OmniSonic, for each data point, we randomly select one scenario from the three predefined configurations in our UniHAGen task: (1) Scenario 1: on-screen environmental sound + off-screen human speech, (2) Scenario 2: on-screen human speech + off-screen environmental sound, and (3) Scenario 3: on-screen environmental sound + off-screen environmental sound + off-screen human speech. For Scenario 1, we randomly select a video from our environmental audio–visual training set, *i.e.*, VGGSound [4], to obtain the video clip with on-screen environmental sound and its corresponding caption. Then, we randomly sample a human speech audio from the speech training set (LRS3 [1] and CommonVoice [3]) to serve as the off-screen speech component, mixing it with the original video's environmental audio according to a random signal-to-noise ratio (SNR) level. The resulting mixed waveform, along with its corresponding captions, transcriptions, and visual frames, is used as the training input for the model. For Scenario 2, we select a video from the speech audio–visual training set LRS3 [1], where the on-screen visual content corresponds to a speaking person. The associated speech and its transcription serve as the on-screen components. An environmental audio clip is then randomly sampled from the environmental training set and added as the off-screen component. The two sources are mixed at a randomly sampled SNR level, and the corresponding textual and visual conditions are used for conditioning the model. For Scenario 3, we randomly select a video from the environmental audio–visual training set to provide the on-screen environmental sound and caption. An additional environmental audio clip and a human speech clip are then randomly sampled from their respective datasets to serve as the off-screen environmental sound and off-screen speech. All three sources are mixed using randomly sampled SNR levels, forming a complex multi-source auditory scene that mimics real-world conditions. The corresponding captions, transcriptions, and visual frames of each source type are used as conditioning inputs during training.

To enable classifier-free guidance (CFG) [7] during inference, we adopt a condition dropout strategy during training. Specifically, for each condition type, *i.e.*, on-screen environmental caption, off-screen environmental caption, speech transcription, and visual frames, we randomly drop the entire condition with a specified probability, *e.g.* 0.1. This strategy encourages the model to learn both conditional and unconditional generation behaviors, enhancing robustness to partially missing conditions and enabling controllable audio generation via CFG during inference.

To stabilize training and ensure high-quality speech generation, we adopt a two-stage training strategy. In the first stage, the model is trained only on speech data, where environmental sound–related conditions are kept empty (*i.e.*, captions are empty strings). This stage helps the model effectively learn speech representation and synchronization without interference from environmental sound components. In the second stage, we switch to the full UniHAGen training setup described above, jointly learning to generate both speech and environmental sounds under different scenarios. This progressive training scheme prevents unstable optimization and improves the model's ability to synthesize clear and coherent speech in complex multi-source audio scenes.

Inference. During inference, we adopt a multi-condition classifier-free guidance (CFG) [7, 12] strategy to achieve controllable audio generation under different condition types. The modified velocity prediction is computed as:

$$\begin{aligned}\tilde{\mathcal{V}}_\theta(\boldsymbol{x}_t, \boldsymbol{c}_{txt}^{on}, \boldsymbol{c}_{txt}^{off}, \boldsymbol{c}_{txt}^{sp.}, \boldsymbol{c}_v) = \mathcal{V}_\theta(\boldsymbol{x}_t, \boldsymbol{c}_{txt}^{on}, \boldsymbol{c}_{txt}^{off}, \boldsymbol{c}_{txt}^{sp.}, \boldsymbol{c}_v)) \\ + \lambda_{txt}^{on}(\mathcal{V}_\theta(\boldsymbol{x}_t, \boldsymbol{c}_{txt}^{on}, \varnothing, \varnothing, \boldsymbol{c}_v)) - \mathcal{V}'_\theta \\ + \lambda_{txt}^{off}(\mathcal{V}_\theta(\boldsymbol{x}_t, \varnothing, \boldsymbol{c}_{txt}^{off}, \varnothing, \boldsymbol{c}_v)) - \mathcal{V}'_\theta) \\ + \lambda_{txt}^{sp.}(\mathcal{V}_\theta(\boldsymbol{x}_t, \varnothing, \varnothing, \boldsymbol{c}_{txt}^{sp.}, \boldsymbol{c}_v)) - \mathcal{V}'_\theta)\end{aligned} \tag{1}$$

where $\varnothing$ denotes the dropped (unconditioned) inputs, and $\mathcal{V}'_\theta = \mathcal{V}_\theta(\boldsymbol{x}_t, \varnothing, \varnothing, \varnothing, \boldsymbol{c}_v))$ represents the non-text unconditional prediction. The derivation is presented in Sec. A.2.

A.2. Derivation of Our Multi-Condition Classifier-Free Guidance

To extend classifier-free guidance (CFG) [7] to our four-condition setup, we consider the conditional distribution $p_{\boldsymbol{\theta}}(\boldsymbol{x}_t|\boldsymbol{c}_{1:3}, \boldsymbol{c}_v)$, where $\boldsymbol{c}_1$, $\boldsymbol{c}_2$, $\boldsymbol{c}_3$ denote the three text-based conditions (on-screen environmental caption, off-screen environmental caption, and speech transcription), and $\boldsymbol{c}_v$ denotes the video condition. The flow model parameterizes this distribution through its time-dependent vector field. Following the modified CFG in [12] for dual conditions, we further enhance the influence of each conditioning signal by modifying the target conditional distribution.

However, unlike the dual-condition case in [12], the four conditions in our setting are not symmetric nor independent. The visual condition $\boldsymbol{c}_v$ is tightly coupled with one of the text-based conditions (the on-screen environmental caption or the speech transcription): visual frames directly reveal the on-screen sound source, its motion, and its temporal structure. As a result, the likelihood term $p_{\boldsymbol{\theta}}(\boldsymbol{c}_v|\boldsymbol{x}_t)$ is not independent from $p_{\boldsymbol{\theta}}(\boldsymbol{c}_1|\boldsymbol{x}_t)$ or $p_{\boldsymbol{\theta}}(\boldsymbol{c}_3|\boldsymbol{x}_t)$, where $\boldsymbol{c}_1$ and $\boldsymbol{c}_3$ denote conditions of on-screen environmental sound caption and speech transcription, respectively. Applying a CFG-style "condition–unconditional" subtraction to $\boldsymbol{c}_v$ would therefore amplify shared information twice. This double-counting empirically leads to unstable guidance and degraded audio–visual consistency.

In contrast, the three text-based conditions $\boldsymbol{c}_1$, $\boldsymbol{c}_2$, $\boldsymbol{c}_3$ serve as independent semantic instructions: they specify what sound should occur (*e.g.*, "waves crashing", "a dog barking", or speech content), but do not dictate how this sound temporally evolves with the visual scene. Thus, applying guidance to these three conditions is both well-defined and beneficial. Importantly, since the video condition provides the essential scene-level prior, we always retain $\boldsymbol{c}_v$ in both the conditional and the "unconditional" branches, ensuring that the model never loses the scene context during guidance.

Under this formulation, the modified conditional distribution becomes:

$$\tilde{p}_{\boldsymbol{\theta}}(\boldsymbol{x}_t|\boldsymbol{c}_{1:3},\boldsymbol{c}_v) \propto p_{\boldsymbol{\theta}}(\boldsymbol{x}_t|\boldsymbol{c}_{1:3},\boldsymbol{c}_v)\prod_{i=1}^{3} p_{\boldsymbol{\theta}}(\boldsymbol{c}_i|\boldsymbol{x}_t,\boldsymbol{c}_v)^{w_i}. \tag{2}$$

Taking the gradient of the log-density yields

$$\begin{aligned}\nabla_{\boldsymbol{x}_t}\log\tilde{p}_{\boldsymbol{\theta}}(\boldsymbol{x}_t|\boldsymbol{c}_{1:3},\boldsymbol{c}_v) = \nabla_{\boldsymbol{x}_t}\log p_{\boldsymbol{\theta}}(\boldsymbol{x}_t|\boldsymbol{c}_{1:3},\boldsymbol{c}_v)\\ + \sum_{i=1}^{3} w_i\nabla_{\boldsymbol{x}_i}\log p_{\boldsymbol{\theta}}(\boldsymbol{c}_i|\boldsymbol{x}_t,\boldsymbol{c}_v).\end{aligned} \tag{3}$$

Using Bayes' rule, $p_{\boldsymbol{\theta}}(\boldsymbol{c}_i|\boldsymbol{x}_t,\boldsymbol{c}_v) = \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}_t|\boldsymbol{c}_i,\boldsymbol{c}_v)p(\boldsymbol{c}_i|\boldsymbol{c}_v)}{p_{\boldsymbol{\theta}}(\boldsymbol{x}_t|\boldsymbol{c}_v)}$, and noting that $p(\boldsymbol{c}_i|\boldsymbol{c}_v)$ does not depend on $\boldsymbol{x}_t$, we obtain

$$\begin{aligned}\nabla_{\boldsymbol{x}_t}\log p_{\boldsymbol{\theta}}(\boldsymbol{c}_i|\boldsymbol{x}_t,\boldsymbol{c}_v) =& \nabla_{\boldsymbol{x}_t}\log p_{\boldsymbol{\theta}}(\boldsymbol{x}_t|\boldsymbol{c}_i,\boldsymbol{c}_v)\\ &- \nabla_{\boldsymbol{x}_t}\log p_{\boldsymbol{\theta}}(\boldsymbol{x}_t|\boldsymbol{c}_v).\end{aligned} \tag{4}$$

Substituting Eq. 4 into Eq. 3, we have:

$$\begin{aligned}\nabla_{\boldsymbol{x}_t}\log\tilde{p}_{\boldsymbol{\theta}}(\boldsymbol{x}_t|\boldsymbol{c}_{1:3},\boldsymbol{c}_v) = \nabla_{\boldsymbol{x}_t}\log p_{\boldsymbol{\theta}}(\boldsymbol{x}_t|\boldsymbol{c}_{1:3},\boldsymbol{c}_v)\\ + \sum_{i=1}^{3} w_i(\nabla_{\boldsymbol{x}_t}\log p_{\boldsymbol{\theta}}(\boldsymbol{x}_t|\boldsymbol{c}_i,\boldsymbol{c}_v) - \nabla_{\boldsymbol{x}_t}\log p_{\boldsymbol{\theta}}(\boldsymbol{x}_t|\boldsymbol{c}_v)).\end{aligned} \tag{5}$$

Finally, we rewrite Eq. 5 in terms of the model's predicted vector field, yielding the following guided velocity formulation:

$$\begin{aligned}\tilde{\mathcal{V}}_{\boldsymbol{\theta}}(\boldsymbol{x}_t,\boldsymbol{c}_{1:3},\boldsymbol{c}_v) =& \mathcal{V}_{\boldsymbol{\theta}}(\boldsymbol{x}_t,\boldsymbol{c}_{1:3},\boldsymbol{c}_v)\\ &+ \sum_{i=1}^{3} w_i(\mathcal{V}_{\boldsymbol{\theta}}(\boldsymbol{x}_t,\boldsymbol{c}_i,\boldsymbol{c}_v) - \mathcal{V}_{\boldsymbol{\theta}}(\boldsymbol{x}_t,\boldsymbol{c}_v)).\end{aligned} \tag{6}$$

A.3. Evaluation Metrics

We evaluate our OmniSonic and the baseline models using both objective and subjective evaluation metrics. For objective evaluation metrics, we adopt Fréchet Audio Distance (FAD) [10] and Mean Kullback–Leibler Divergence (MKL) [8] assess the perceptual quality and distributional similarity of generated audios. For semantic alignment evaluation, following previous works [11, 21, 23, 25], we use the AV score and AT score to measure the semantic correspondence between audio and video (AV) and between audio and text (AT), respectively. Specifically, we employ Wav2CLIP [24] to encode the generated audio into the CLIP [17] feature space, enabling direct computation of cross-modal similarity with visual and textual embeddings. To evaluate the speech correctness in the generated audio, we utilize Word Error Rate (WER), Character Error Rate (CER), and Phoneme Error Rate (PER) to quantitatively assess the accuracy of synthesized speech content. Following [12], we employ a pretrained Whisper [18] model to transcribe the generated audio and compute the error rates by comparing the transcriptions with the ground-truth speech transcription. To measure audio–visual temporal synchronization, we adopt DeSync [5, 20], which utilizes Synchformer [9] to estimate the temporal misalignment between the generated audio and the corresponding video frames. For subjective evaluation, we conduct human listening tests to assess four aspects: overall quality (MOS-Q), environmental faithfulness (MOS-EF) for on-screen and/or off-screen environmental sounds, speech faithfulness (MOS-SF) for on-screen and/or off-screen speech, and temporal alignment (MOS-T) between the video and the on-screen sound. We randomly select 24 samples from our UniHAGen-Bench and generate the corresponding audio using OmniSonic and the compared baseline models. The all generated audios are randomly distributed among 13 human listeners, who rate them on a discrete 5-point scale. We report the mean opinion scores (MOS) averaged across all ratings for each evaluation aspect. The interface for this subjective evaluation is shown in Fig. 1.

A.4. Baselines

We compare OmniSonic with state-of-the-art audio generation models: AudioLDM 2 [14], VoiceLDM [12], VinTAGe [11], MMAudio [5], and HunyuanVideo-Foley [20].

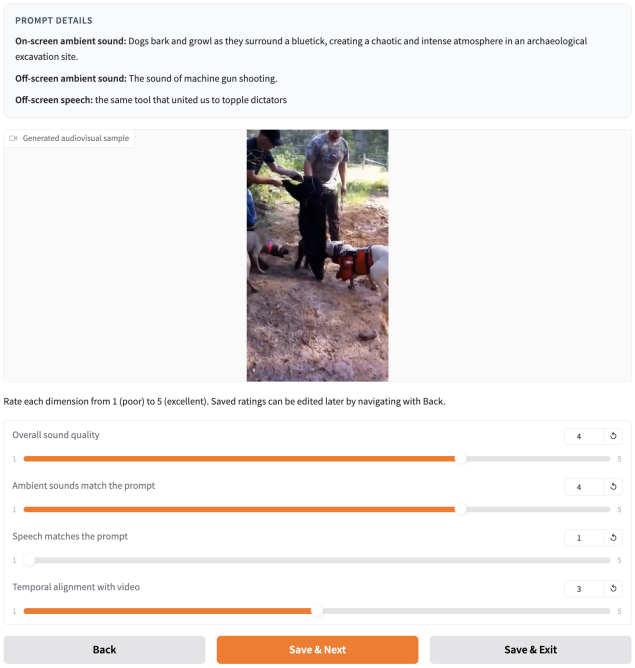

Figure 1. Interface for the subjective evaluation.

Among them, AudioLDM 2 is a general text-to-audio generation model. In addition to its official checkpoint, we also evaluate a speech-adapted version[1] fine-tuned for speech-related tasks. VoiceLDM is a text-to-speech generation model designed to synthesize environmental speech, *i.e.*, speech recorded or simulated within specific acoustic environments. VinTAGe is a recent advancement in joint video–text-to-audio generation for holistic auditory scene synthesis, aiming to produce both on-screen and off-screen sounds simultaneously. MMAudio and HunyuanVideo-Foley are video–text conditioned audio generation models built upon the MM-DiT [6] architecture, which leverages multimodal diffusion transformers to generate temporally aligned and semantically consistent audio from visual and textual cues.

A.5. Implementation Details

We implement OmniSonic in PyTorch [16]. For the model architecture, we use CLIP [17] as the visual frame encoder, FLAN-T5 [19] as the environmental sound caption encoder, and SpeechT5 [2] as the speech transcription encoder. The differentiable Durator follows the design in [12, 22], which consists of a Duration Predictor and a Learnable Upsampling Layer. The dimensions of the visual, speech transcription, and environmental sound caption embeddings are 512, 769, and 1024, respectively. For the audio VAE, we adopt the pre-trained version from AudioLDM [13], which generate audio latent representation with number of channels of 8 and size of 256×16. For the TriAttn-DiT, we set the patch size to 2 and the hidden size to 1152, and stack 28 blocks. We sample the audio at 16kHz, which is then transformed to spectrogram using Short-Time Fourier Transform (STFT) with FFT size of 1024, hop length of 160, and Hann window size of 1024. We pad or crop the waveform to a fixed duration of 1024×160 before applying the STFT, yielding log-Mel spectrograms with $T = 1024$ and $F = 64$. We train our model on 8 NVIDIA A100 GPUs with a global batch size of 64, the learning rate of 5e-5, and the weight decay of 1e-2. The model is optimized using the AdamW [15] optimizer. During training, the visual frame encoder and environmental sound caption encoder are frozen, while the speech transcription encoder and the differentiable Durator are trainable.

For the CFG scales during sampling, we set different groups of values for the three different scenarios. Specifically, for scenario 1 (on-screen environmental sound + off-screen human speech), we set $\lambda_{txt}^{on} = 5.0$, $\lambda_{txt}^{sp.} = 2.5$, and $\lambda_{txt}^{off} = 0.5$. For Scenario 2 (on-screen human speech + off-screen environmental sound), we set $\lambda_{txt}^{sp.} = 7.5$, $\lambda_{txt}^{off} = 2.5$, and $\lambda_{txt}^{on} = 0.5$. For Scenario 3 (on-screen environmental sound + off-screen environmental sound + off-screen human speech), we set $\lambda_{txt}^{on} = 5.0$, $\lambda_{txt}^{off} = 2.5$, and $\lambda_{txt}^{sp.} = 2.5$.

A.6. Parameter Study

We conduct parameter studies on the CFG scales. The results for Scenario 1 are presented in Tab. 1, where we fix $\lambda_{txt}^{off} = 0.5$ and examine how varying λ_{txt}^{on} and $\lambda_{txt}^{sp.}$ affects the performance. We present the results of Scenario 2 in Tab. 2, where we fix $\lambda_{txt}^{on} = 0.5$ and examine how varying $\lambda_{txt}^{sp.}$ and λ_{txt}^{off} affects the performance. The results of Scenario 3 is shown in Tab. 3, in which we investigate the impact of values of λ_{txt}^{on}, λ_{txt}^{off}, and $\lambda_{txt}^{sp.}$ on the final results.

Table 1. Parameter study for Scenario 1 (on-screen environmental sound + off-screen human speech). We fix $\lambda_{txt}^{off} = 0.5$ and examine the effect of varying λ_{txt}^{on} and $\lambda_{txt}^{sp.}$ on the results.

λ_{txt}^{on}	$\lambda_{txt}^{sp.}$	FAD↓	Mean AV AT↑	WER↓
5.0	2.5	3.40	19.62	0.15
5.0	7.5	5.39	19.50	0.14
7.5	2.5	4.14	19.02	0.16
7.5	5.0	4.98	19.32	0.15
7.5	7.5	5.04	19.28	0.15
9.5	2.5	4.23	18.94	0.17
9.5	5.0	5.19	19.32	0.15
9.5	7.5	5.40	19.30	0.15
12.5	2.5	4.06	18.64	0.18
12.5	5.0	4.94	18.82	0.16
12.5	7.5	5.39	18.95	0.16

[1] `audioldm2-speech-gigaspeech` checkpoint

Table 2. Parameter study for Scenario 2 (on-screen human speech + off-screen environmental sound). We fix $\lambda_{txt}^{on} = 0.5$ and examine the effect of varying $\lambda_{txt}^{sp.}$ and λ_{txt}^{off} on the results.

$\lambda_{txt}^{sp.}$	λ_{txt}^{off}	FAD	Mean AV AT	WER
2.5	2.5	2.63	18.13	0.11
2.5	5.5	4.77	17.26	0.12
2.5	7.5	5.32	17.19	0.10
5.5	2.5	2.33	17.95	0.12
5.5	5.5	4.42	17.55	0.11
5.5	7.5	5.36	17.18	0.11
7.5	2.5	2.59	17.95	0.12
7.5	5.5	4.79	17.51	0.11
7.5	7.5	5.09	17.45	0.12

Table 3. Parameter study for Scenario 3 (on-screen environmental sound + off-screen environmental sound + off-screen human speech). We examine the effect of varying λ_{txt}^{on}, λ_{txt}^{off}, and $\lambda_{txt}^{sp.}$ on the results.

λ_{txt}^{on}	λ_{txt}^{off}	$\lambda_{txt}^{sp.}$	FAD↓	Mean AV AT↑	WER↓
5.0	2.5	2.5	3.39	18.26	0.16
5.0	2.5	3.5	3.97	18.55	0.15
5.0	5.0	2.5	3.66	17.00	0.16
5.0	5.0	3.5	3.98	17.50	0.14
7.5	2.5	2.5	3.88	18.30	0.15
7.5	2.5	3.5	4.34	18.54	0.18
7.5	5.0	2.5	3.70	17.48	0.15
7.5	5.0	3.5	4.07	17.64	0.15
7.5	5.0	5.0	4.95	17.70	0.13

A.7. Limitations and Future Work

Although OmniSonic achieves strong performance across diverse mixed-source scenarios, several limitations remain.

First, the training samples used in our UniHAGen task are synthetically constructed by combining audio, text, and video clips from VGGSound, LRS3, and CommonVoice to simulate the three scenario configurations. While this composition strategy enables controlled supervision across on/off-screen speech–environment combinations, it does not fully capture the richness, spontaneity, and acoustic complexity of truly in-the-wild audio–visual scenes. Moreover, synthetic mixing often results in acoustic inconsistencies, such as mismatched loudness, differing recording conditions, or unnatural blending between speech and environmental sounds, which limits the realism of the training distribution.

Future work may explore collecting large-scale natural audio–visual corpora with organically co-occurring speech and environmental events, or developing more advanced simulation pipelines to better approximate real-world multimodal dynamics.

Second, our model relies solely on CLIP visual features for video conditioning. Although CLIP provides strong global semantic understanding, it lacks fine-grained temporal sensitivity. In visually stable or weakly dynamic scenes, where consecutive frames exhibit minimal variation, CLIP features tend to remain nearly invariant, limiting the model's ability to infer subtle temporal cues for precise audio–visual synchronization. This leads to weaker performance on synchronization-focused metrics compared with models that incorporate temporally specialized encoders such as Synchformer.

Future work may integrate more temporally expressive video representations or design dedicated audio-aware video encoders to strengthen fine-grained synchronization without compromising semantic grounding.

References

[1] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. Lrs3-ted: a large-scale dataset for visual speech recognition. *arXiv preprint arXiv:1809.00496*, 2018. 1

[2] Junyi Ao, Rui Wang, Long Zhou, Chengyi Wang, Shuo Ren, Yu Wu, Shujie Liu, Tom Ko, Qing Li, Yu Zhang, et al. Speecht5: Unified-modal encoder-decoder pre-training for spoken language processing. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5723–5738, 2022. 3

[3] Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. Common voice: A massively-multilingual speech corpus. In *Proceedings of the twelfth language resources and evaluation conference*, pages 4218–4222, 2020. 1

[4] 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)*, pages 721–725. IEEE, 2020. 1

[5] Ho Kei Cheng, Masato Ishii, Akio Hayakawa, Takashi Shibuya, Alexander Schwing, and Yuki Mitsufuji. Mmaudio: Taming multimodal joint training for high-quality video-to-audio synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 28901–28911, 2025. 2

[6] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first International Conference on Machine Learning*, 2024. 3

[7] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 1

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

[9] Vladimir Iashin, Weidi Xie, Esa Rahtu, and Andrew Zisserman. Synchformer: Efficient synchronization from sparse cues. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5325–5329. IEEE, 2024. 2

[10] Kevin Kilgour, Mauricio Zuluaga, Dominik Roblek, and Matthew Sharifi. Fréchet audio distance: A metric for evaluating music enhancement algorithms. *arXiv preprint arXiv:1812.08466*, 2018. 2

[11] Saksham Singh Kushwaha and Yapeng Tian. Vintage: Joint video and text conditioning for holistic audio generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13529–13539, 2025. 2

[12] Yeonghyeon Lee, Inmo Yeon, Juhan Nam, and Joon Son Chung. Voiceldm: Text-to-speech with environmental context. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 12566–12571. IEEE, 2024. 1, 2, 3

[13] 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. In *International Conference on Machine Learning*, pages 21450–21474. PMLR, 2023. 3

[14] 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. 2

[15] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. 3

[16] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019. 3

[17] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 2, 3

[18] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR, 2023. 2

[19] 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

[20] Sizhe Shan, Qiulin Li, Yutao Cui, Miles Yang, Yuehai Wang, Qun Yang, Jin Zhou, and Zhao Zhong. Hunyuanvideo-foley: Multimodal diffusion with representation alignment for high-fidelity foley audio generation. *arXiv preprint arXiv:2508.16930*, 2025. 2

[21] 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)*, pages 1–5. IEEE, 2023. 2

[22] Xu Tan, Jiawei Chen, Haohe Liu, Jian Cong, Chen Zhang, Yanqing Liu, Xi Wang, Yichong Leng, Yuanhao Yi, Lei He, et al. Naturalspeech: End-to-end text-to-speech synthesis with human-level quality. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(6):4234–4245, 2024. 3

[23] 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*, pages 15492–15501, 2024. 2

[24] 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)*, pages 4563–4567. IEEE, 2022. 2

[25] 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. *arXiv preprint arXiv:2407.01494*, 2024. 2