Video-to-Audio Generation with Fine-grained Temporal Semantics

1st Yuchen Hu
Tencent AI Lab
Beijing, China
uniychu@tencent.com

2nd Yu Gu
Tencent AI Lab
Beijing, China
colinygu@tencent.com

3rd Chenxing Li
Tencent AI Lab
Beijing, China
chenxingli@tencent.com

4th Rilin Chen
Tencent AI Lab
Beijing, China
rilinchen@tencent.com

5th Dong Yu
Tencent AI Lab
Seattle, USA
dyu@tencent.com

Abstract—With recent advances of AIGC, video generation have gained a surge of research interest in both academia and industry (e.g., Sora). However, it remains a challenge to produce temporally aligned audio to synchronize the generated video, considering the complicated semantic information included in the latter. In this work, inspired by the recent success of text-to-audio (TTA) generation, we first investigate the video-to-audio (VTA) generation framework based on latent diffusion model (LDM). Similar to latest pioneering exploration in VTA, our preliminary results also show great potentials of LDM in VTA task, but it still suffers from sub-optimal temporal alignment. To this end, we propose to enhance the temporal alignment of VTA with frame-level semantic information. With the recently popular grounding segment anything model (Grounding SAM), we can extract the fine-grained semantics in video frames to enable VTA to produce better-aligned audio signal. Extensive experiments demonstrate the effectiveness of our system on both objective and subjective evaluation metrics, which shows both better audio quality and fine-grained temporal alignment.[1]

Index Terms—Video-to-audio generation, latent diffusion model, fine-grained temporal alignment, segment anything model

I. INTRODUCTION

Vision and hearing constitute the main senses of we humans for perceiving the world surrounding us, which are usually complementary to each other and both indispensable [1]–[6]. Recent advances of AIGC enables neural model to produce vivid video from textual descriptions (e.g., Sora [7]), presenting a great potential of AI in benefiting human life. However, they lack the capability to generate synchronized audio signal to make the video close to realistic scenarios. Therefore an effective video-to-audio (VTA) method is expected to have outstanding performance on both audio quality and temporal alignment, in which the generated audio should not only match the contents and semantics of the input video but also align temporally with the given video frames.

Inspired by the success of latent diffusion model (LDM) in text-to-video (TTV) generation [8]–[13], recent works achieve a promising progress in text-to-audio (TTA) generation [10], [14]–[24]. Despite the good performance, TTA can hardly synchronize the generated audio to target video, due to the lack of temporal semantics in the textual input. Therefore, latest efforts pioneer the study in direct VTA generation [25]–[27] with latent diffusion model. To illustrate, [27] investigate various contrastive pre-trained audio/video/text embeddings and data augmentation methods on top of LDM to improve semantic alignment, and [26] propose a sophisticated temporal adapter to model the audio-video alignment. Despite the effectiveness, these pioneering studies fail to explicitly model the frame-level video-audio synchronization, which is hard considering the complicated input video representations.

In this work, we propose a fine-grained video-to-audio generation approach with explicit frame-level synchronization. Specifically, we first investigate the popular latent diffusion model for VTA task, similar to the typical TTA works and latest VTA studies. To enhance the video-audio synchronization, we propose to leverage the grounding segment anything model (SAM) [28], [29] to extract detailed object information from each video frame, thus providing fine-grained semantic information to condition the LDM for audio generation. Extensive experiments demonstrate the effectiveness of our approach in enhancing the temporal semantics, in terms of both objective and subjective evaluation metrics. Analysis shows the superiority of Grounding SAM in providing fine-grained temporal information in guiding LDM generation.

II. METHODOLOGY

A. Overall Framework

Drawing inspirations from the popular TTA works [14], we develop a LDM-based VTA framework as shown in Fig. 1. It consists of several key components: a video encoder, a Grounding SAM, a Transformer UNet based conditional diffusion model, and an audio mel-spectrum variational auto-encoder (VAE). Given the input video, we leverage pre-trained CLIP visual encoder [30] to extract visual embeddings. On the other hand, we introduce Grounding SAM to extract the objects (e.g., "car") in each video frame, as well as their appearing probability predicted by SAM. These information are also transferred into CLIP embedding. The resulted visual and object embeddings are combined together as the conditioner of LDM for audio generation. The output representations are sent into VAE decoder to generate mel-spectrum and finally goes into vocoder to produce waveform.

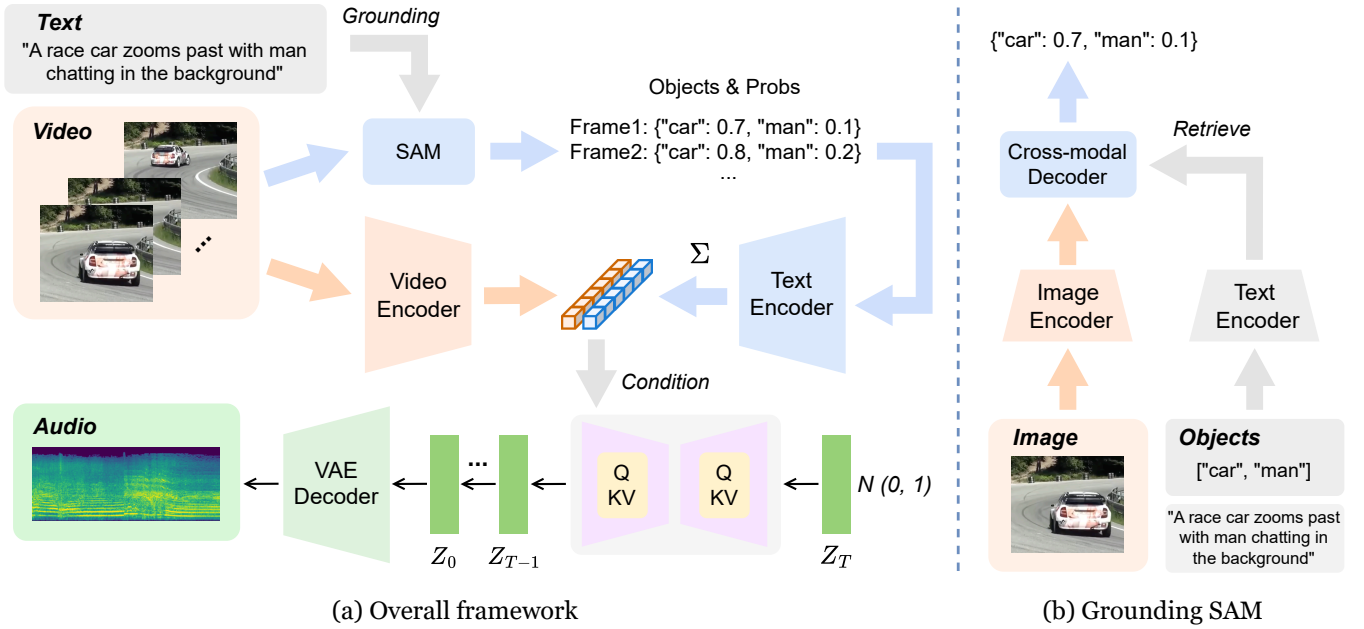[1]Demo page is available at: https://sounddemos.github.io/vta-sam/

Fig. 1: The diagrams of (a) overall framework of our VTA approach, (b) Grounding SAM model [28] from literature. For each video frame, we use Grounding SAM to extract all the objects and their probabilities to represent fine-grained semantics.

B. Visual Encoder

In our VTA framework, visual encoder $\mathcal{E}_v$ is used to encode both the semantic and temporal information in video input for comprehensive audio generation. We leverage the popular CLIP vision encoder in our framework, which is expected to capture the comprehensive visual information including scenes, characters, and events. Necessary linear projectors are added after it to match the dimension of LDM embeddings.

C. Grounding Segment Anything Model (SAM)

The Grounding SAM[2] [28] $\mathcal{S}$ is a open-set object detector that can detect arbitrary objects with human inputs such as category names or referring expressions. In our VTA framework, for better use of the grounding ability of SAM, we send the paired textual description t to ground and detect the appearing objects from i-th video frame v_i, e.g., "car" and "man". Specifically, we first employ nltk toolkit[3] to extract the nouns (i.e., object) $o_i^1, o_i^2, ...$ from input text, thus provides fine-grained prompts for SAM grounding. Thereafter, SAM consists of two modality encoders to extract image and object embeddings respectively, followed by a cross-modal decoder to retrieve the video frame and predict the appearing probability of each object. In this way, SAM helps produce comprehensive semantic information of current video frame.

The j-th detected object o_i^j is then sent into a CLIP text encoder $\mathcal{E}_t$ to produce the semantic embedding, which are

[2]https://github.com/IDEA-Research/GroundingDINO
[3]https://www.nltk.org/index.html

finally weighted summed using the appearing probability p_i^j:

$$(o_i^1, o_i^2, ...), (p_i^1, p_i^2, ...) = \mathcal{S}(v_i, t),$$
$$s_i = \sum_j p_i^j * \mathcal{E}_t(o_i^j), \quad (1)$$

s_i is the semantic embedding of i-th video frame, and we send $c_o = \{s_i\}_{i=1}^T$ as the final sequential object embedding to enhance diffusion VTA. It is worth noting that we use the text instead of detected image region to represent the object, which aims to make more comprehensive conditioners for LDM, since the image region is already covered by previous video embeddings. On the other hand, our incorporation of frame-level object information resolves the limitation of previous TTA methods that relies on utterance-level text.

D. Latent Diffusion Model (LDM)

Diffusion model is currently among the most popular generative approaches, which consists of diffusion process and reverse process. Given the original input x_0, it first follows a Morkov chain of diffusion steps to add noise to x_0 until it finally reaches a pure Guassian distribution, i.e., $x_t \in N(0, 1)$. Thereafter, during the reverse process, the model learns to denoise the x_t step by step until it recovers the original input x_0. To save the computation cost, latent diffusion model propose to conduct these two process on latent space Z instead of original input space X.

In our VTA setting, we have two diffusion conditioners, i.e., $c_v = \mathcal{E}_v(v)$ from video input v, where $\mathcal{E}_v$ is CLIP video encoder for feature extraction, and the $c_o = \{s_i\}_{i=1}^T$ from

Eq. (1). For step-by-step denoising, we train a neural model ϵ_θ to predict the noise with following training objective:

$$\theta = \arg\min \mathbb{E}_{t \sim U(1,T)} \|\epsilon - \epsilon_\theta(Z_t, t, c_v, c_o)\|_2^2, \quad (2)$$

where we use SAM to extract frame-level semantic information in textual format, and then extract the hidden embedding using CLIP text encoder $\mathcal{E}_t$, which shares the same semantic space with CLIP video encoder $\mathcal{E}_v$. Here our consideration in utilizing textual output from SAM instead of the image output is that, the visual information is already covered by video embedding. Therefore, we select the textual output to make comprehensive condition information.

In addition, during reverse process, we adopt the popular classifier-free guidance, which has been proved effective in text-to-audio generation [14]. Given the latent variable z_t and the two conditioners, we perform the generation both conditionally and unconditionally, and then sum them up with a specified weight α.

$$\epsilon'_\theta = \alpha * \epsilon_\theta(Z_t, t, c_v, c_o) + (1 - \alpha) * \epsilon_\theta(Z_t, t), \quad (3)$$

E. LDM Conditioned on Video and Objects

We investigate how to integrate the two conditioners c_v and c_o into diffusion model ϵ_θ. Our LDM follows the UNet structure where each module is a Transformer block (i.e., self-attention, cross-attention, feed-forward network). Denote the latent variable after self-attention module as z_t, now we need to incorporate c_v and c_o during cross-attention computation. Considering that c_o contains more fine-grained temporal information than c_v with the help of SAM, we send the video representations into cross-attention for retrieval and then add the sequential object information to enhance the semantics:

$$z'_t = \text{Attention}(z_t, c_v, c_v) + c_o, \quad (4)$$

We also investigate some other schemes for conditional LDM as discussed in Section. III-B4.

F. Mel-spectrum VAE and Vocoder

As introduced before, our diffusion model is performed in latent space, where the variable z_0 is extracted by VAE encoder from mel-spectrum input. Specifically, we use the pretrained VAE weights from AudioLDM [16]. After generating the target mel-spectrum, we finally send it into Hifi-GAN Vocoder [31] to obtain the waveform output.

III. EXPERIMENTS AND RESULTS

A. Experimental Setup

Dataset. We evaluate the proposed method on public VG-GSound dataset [32]. In particular, this dataset contains 550 hours of videos with paired audio-visual events.

Configurations. Our LDM follows the popular Transformer UNet [33], [34] structure, which is configured with a cross-attention dimension of 1024 and 8 input and output channels. We train our models with learning rate of 3e-5 and no warmup steps, and the batch size is set to 8 per GPU. All models are trained for a total of 40 epochs on 4 NVIDIA-A100-40GB

TABLE I: Performance comparison with baselines from both TTA and VTA literature, in terms of objective metrics.

System	FD ↓	FAD ↓	KL ↓	AV-Align ↑
Text-to-Audio				
Tango [24]	41.7	4.7	5.5	0.173
Video-to-Audio				
Diff-Foley [25]	30.6	4.5	5.2	0.193
VTA-LDM [27]	30.3	4.5	5.0	0.204
FoleyCrafter [26]	28.3	3.8	4.8	0.233
VTA-SAM (ours)	**25.1**	**3.1**	**4.5**	**0.255**

TABLE II: Subjective evaluation results of proposed method and baselines. Larger value means better performance.

System	Audio Quality	Semantic Alignment	Temporal Alignment	Overall Quality
Ground-Truth	81.7	83.8	80.5	79.9
Text-to-Audio				
Tango [24]	72.0	56.2	52.1	57.8
Video-to-Audio				
Diff-Foley [25]	73.0	68.6	57.9	62.7
VTA-LDM [27]	72.5	70.4	60.3	65.2
FoleyCrafter [26]	74.7	70.9	61.4	65.9
VTA-SAM (ours)	**75.7**	**72.4**	**63.5**	**66.7**

GPUs. During inference stage, we select 1,000 samples from official test set for efficient evaluation. Specifically, we set the denoising steps to 300, the number of samples per audio to 1, and the guidance scale to 3 for classifier-free guidance.

Metric. For objective evaluation, we use several quantitative metrics, including Frechet distance (FD), Frechet audio distance (FAD), and Kullback–Leibler (KL) to measure the semantic similarity of generated audio with the ground-truth. Furthermore, we use the AV-Align metric [35] to evaluate the temporal alignment of generated audio with input video. For subjective evaluation, we invite human listeners to evaluate from four different perspectives of perception, i.e., audio quality, semantic alignment, temporal alignment, and the overall quality. Specifically, we select 30 samples and invite five listeners to evaluate them using a 0-100 score scale [27].

B. Results and Analysis

1) Objective Evaluation: Table I illustrates the comparison of our proposed VTA-SAM with previous works. We use objective metrics for evaluation, including FD, FAD, and KL to measure the semantic similarity of generative audio and ground-truth audio, as well as AV-Align to measure the temporal alignment of generated audio and input video. We first reproduce Tango, the SOTA text-to-audio approach, to evaluate the TTA performance under our setting. Compared to that, the first diffusion-based video-to-audio approach Diff-Foley shows significantly better performance in terms of both semantic and temporal alignment, thanks to the much richer information brought by video than text. Thereafter, the work VTA-LDM investigate the impact of vision encoders, auxiliary embeddings, and data augmentation techniques on LDM-based

TABLE III: Ablation study of input modalities. 'V' denotes video, 'T' denotes text, 'O' denotes the detected objects from Grounding SAM, and '+' denotes combination.

Input Modality	FD ↓	FAD ↓	KL ↓	AV-Align ↑
T-Only	41.7	4.7	5.5	0.173
V-Only	30.3	4.5	5.0	0.204
V + T	28.8	3.8	4.9	0.220
V + O	**25.1**	3.1	**4.5**	**0.255**
V + O + T	25.5	**3.0**	4.6	0.251

TABLE IV: Ablation study of conditional LDM. We investigate the optimal method to insert video and object information into LDM as conditioners. 'Z' denotes the latent variable inside LDM, 'V' denotes video, 'O' denotes the detected objects from Grounding SAM, '@' denotes cross-attention operation, '+' denotes concatenation operation.

Method	FD ↓	FAD ↓	KL ↓	AV-Align ↑
Z @ V	30.3	4.5	5.0	0.204
Z @ (V + O)	26.8	3.5	4.7	0.242
Z @ O + V	28.3	3.9	4.6	0.233
Z @ V + O	**25.1**	**3.1**	**4.5**	**0.255**

VTA approach, which moves one step forward on performance. The most recent work FoleyCrafter designs specific semantic and temporal adapters to explicitly improve the objective performance. In this work, our proposed approach leverages the abundant pre-trained knowledge of grounding SAM to extract fine-grained temporal semantics to improve VTA. As a result, it presents better performance in terms of all objective metrics, which demonstrates its effectiveness and provides a new insight for VTA task.

2) Subjective Evaluation: Apart from objective metrics, we also conduct subjective evaluation to verify the effectiveness of our approach. We evaluate from four different perspectives, i.e., 1) audio quality: merely the quality of generated audio, 2) semantic alignment, the consistency of overall semantic information covered by input video and generated audio, 3) temporal alignment, the consistency of frame-level semantic information covered by input video and generated audio, 4) overall quality, the overall quality of generated audio-video pair. We first observe from Table II that the ground-truth audio-video pairs enjoy excellent quality (i.e., around 80) from all perspectives, which sets a high upper-bound for VTA task. In comparison, the TTA baseline performs much worse on these subjective evaluation, except the audio quality metric. In addition, we also observe that temporal alignment is the most difficult metric (i.e., lowest value in all methods), which exactly matches our major research motivation. Previous VTA works post some improvements on these four subjective metrics with sophisticated techniques with diffusion model, while the temporal alignment is still sub-optimal (i.e., around 60). On top of VTA-LDM, our introduction of grounding SAM brings fine-grained semantic information and successfully improves

the quality of temporal alignment (i.e., 60.3→63.5). However, we are also aware that its absolute performance is still limited when compared to the ground-truth. We expect that the next generation of VTA models may depend on large model and abundant data to achieve groundbreaking progress, referring to the recent success of large language models and Sora.

3) Effect of Input Modalities: Table III presents the ablation study of different combination of input modalities. From the first two rows, we observe that VTA outperforms TTA by a large margin, thanks to the richer semantic information brought by videos. Then, we try to combine both modalities and it yields some improvement as indicated by the third row in Table III. However, the improvement of 'V + T' over 'V-Only' is limited since the text description contains little temporal information. In comparison, our approach introduces grounding SAM to extract the object information in each video frame and weight them using predicted appearing probability and CLIP text encoder. As a result, we can obtain fine-grained temporal semantics to align the generated audio to input video, which produces significantly better performance in terms of all objective metrics. In addition, we observe that combine incorporating both fine-grained object information and coarse-grained text information fails to bring further improvement, we speculate that the former is already sufficient to enhance audio generation, where the latter seems to be redundant.

4) Effect of Different Conditioning Methods: Given the optimal combination of 'V + O' for conditioner, Table IV further investigates different methods to incorporate them into LDM. Inspired by previous work [27], we leverage cross-attention operation, with the latent variable Z as query and the video conditioner 'V' as key/value, to fuse V into LDM. On top of that, we try to incorporate our introduced object conditioner 'O' into LDM via three methods, i.e., 1) concatenate V and O as the key/value, 2) leverage O as key/value and add V after cross-attention, 3) leverage V as key/value and add O after cross-attention. Results indicate that the last scheme shows the best performance. We speculate that the extracted object information is already fine-grained enough, which makes the cross-attention retrieval unnecessary and thus may explain why simple addition at last yields the best performance.

IV. CONCLUSION

In this work, we propose a sophisticated video-to-audio generation approach using latent diffusion model with explicit synchronization. Specifically, we first investigate the popular latent diffusion model for VTA task, similar to the typical TTA works and latest VTA studies. To enhance the video-audio synchronization, we propose to leverage the grounding segment anything model to extract detailed object information from video frames, which provides fine-grained semantic information to condition the LDM for audio generation. Experiments demonstrate the effectiveness of our approach in enhancing the temporal semantics compared to baselines, in terms of both objective and subjective evaluation metrics. Further analysis indicates the superiority of grounding SAM in providing fine-grained temporal information in guiding LDM generation.

REFERENCES

[1] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 2, pp. 423–443, 2018.

[2] T. Afouras, J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Deep audio-visual speech recognition," *IEEE transactions on pattern analysis and machine intelligence*, 2018.

[3] D. Hazarika, R. Zimmermann, and S. Poria, "Misa: Modality-invariant and-specific representations for multimodal sentiment analysis," in *Proceedings of the 28th ACM international conference on multimedia*, 2020, pp. 1122–1131.

[4] P. Ma, S. Petridis, and M. Pantic, "End-to-end audio-visual speech recognition with conformers," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 7613–7617.

[5] H. Zhu, M.-D. Luo, R. Wang, A.-H. Zheng, and R. He, "Deep audio-visual learning: A survey," *International Journal of Automation and Computing*, vol. 18, no. 3, pp. 351–376, 2021.

[6] B. Shi, W.-N. Hsu, K. Lakhotia, and A. Mohamed, "Learning audio-visual speech representation by masked multimodal cluster prediction," in *International Conference on Learning Representations*, 2022.

[7] R. Sun, Y. Zhang, T. Shah, J. Sun, S. Zhang, W. Li, H. Duan, B. Wei, and R. Ranjan, "From sora what we can see: A survey of text-to-video generation," *arXiv preprint arXiv:2405.10674*, 2024.

[8] J. Wang, H. Yuan, D. Chen, Y. Zhang, X. Wang, and S. Zhang, "Modelscope text-to-video technical report," *arXiv preprint arXiv:2308.06571*, 2023.

[9] J. Z. Wu, Y. Ge, X. Wang, S. W. Lei, Y. Gu, Y. Shi, W. Hsu, Y. Shan, X. Qie, and M. Z. Shou, "Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 7623–7633.

[10] A. Blattmann, R. Rombach, H. Ling, T. Dockhorn, S. W. Kim, S. Fidler, and K. Kreis, "Align your latents: High-resolution video synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 22 563–22 575.

[11] H. Jeong, G. Y. Park, and J. C. Ye, "Vmc: Video motion customization using temporal attention adaption for text-to-video diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 9212–9221.

[12] W. Weng, R. Feng, Y. Wang, Q. Dai, C. Wang, D. Yin, Z. Zhao, K. Qiu, J. Bao, Y. Yuan *et al.*, "Art-v: Auto-regressive text-to-video generation with diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 7395–7405.

[13] Z. Yang, J. Teng, W. Zheng, M. Ding, S. Huang, J. Xu, Y. Yang, W. Hong, X. Zhang, G. Feng *et al.*, "Cogvideox: Text-to-video diffusion models with an expert transformer," *arXiv preprint arXiv:2408.06072*, 2024.

[14] D. Ghosal, N. Majumder, A. Mehrish, and S. Poria, "Text-to-audio generation using instruction-tuned llm and latent diffusion model," *arXiv preprint arXiv:2304.13731*, 2023.

[15] D. Yang, J. Yu, H. Wang, W. Wang, C. Weng, Y. Zou, and D. Yu, "Diffsound: Discrete diffusion model for text-to-sound generation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 1720–1733, 2023.

[16] H. Liu, Z. Chen, Y. Yuan, X. Mei, X. Liu, D. Mandic, W. Wang, and M. D. Plumbley, "Audioldm: Text-to-audio generation with latent diffusion models," *arXiv preprint arXiv:2301.12503*, 2023.

[17] R. Huang, J. Huang, D. Yang, Y. Ren, L. Liu, M. Li, Z. Ye, J. Liu, X. Yin, and Z. Zhao, "Make-an-audio: Text-to-audio generation with prompt-enhanced diffusion models," in *International Conference on Machine Learning*. PMLR, 2023, pp. 13 916–13 932.

[18] J. Huang, Y. Ren, R. Huang, D. Yang, Z. Ye, C. Zhang, J. Liu, X. Yin, Z. Ma, and Z. Zhao, "Make-an-audio 2: Temporal-enhanced text-to-audio generation," *arXiv preprint arXiv:2305.18474*, 2023.

[19] L. Ruan, Y. Ma, H. Yang, H. He, B. Liu, J. Fu, N. J. Yuan, Q. Jin, and B. Guo, "Mm-diffusion: Learning multi-modal diffusion models for joint audio and video generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 10 219–10 228.

[20] Y. Bai, T. Dang, D. Tran, K. Koishida, and S. Sojoudi, "Accelerating diffusion-based text-to-audio generation with consistency distillation," *arXiv preprint arXiv:2309.10740*, 2023.

[21] J. Xue, Y. Deng, Y. Gao, and Y. Li, "Auffusion: Leveraging the power of diffusion and large language models for text-to-audio generation," *arXiv preprint arXiv:2401.01044*, 2024.

[22] J. Michaels, J. B. Li, L. Yao, L. Yu, Z. Wood-Doughty, and F. Metze, "Audio-journey: Open domain latent diffusion based text-to-audio generation," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 6960–6964.

[23] S. Mo, J. Shi, and Y. Tian, "Text-to-audio generation synchronized with videos," *arXiv preprint arXiv:2403.07938*, 2024.

[24] N. Majumder, C.-Y. Hung, D. Ghosal, W.-N. Hsu, R. Mihalcea, and S. Poria, "Tango 2: Aligning diffusion-based text-to-audio generations through direct preference optimization," *arXiv preprint arXiv:2404.09956*, 2024.

[25] S. Luo, C. Yan, C. Hu, and H. Zhao, "Diff-foley: Synchronized video-to-audio synthesis with latent diffusion models," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[26] Y. Zhang, Y. Gu, Y. Zeng, Z. Xing, Y. Wang, Z. Wu, and K. Chen, "Foleycrafter: Bring silent videos to life with lifelike and synchronized sounds," *arXiv preprint arXiv:2407.01494*, 2024.

[27] M. Xu, C. Li, Y. Ren, R. Chen, Y. Gu, W. Liang, and D. Yu, "Video-to-audio generation with hidden alignment," *arXiv preprint arXiv:2407.07464*, 2024.

[28] S. Liu, Z. Zeng, T. Ren, F. Li, H. Zhang, J. Yang, C. Li, J. Yang, H. Su, J. Zhu *et al.*, "Grounding dino: Marrying dino with grounded pre-training for open-set object detection," *arXiv preprint arXiv:2303.05499*, 2023.

[29] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo *et al.*, "Segment anything," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 4015–4026.

[30] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.

[31] J. Kong, J. Kim, and J. Bae, "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis," *Advances in neural information processing systems*, vol. 33, pp. 17 022–17 033, 2020.

[32] C. Valentini-Botinhao, X. Wang, S. Takaki, and J. Yamagishi, "Investigating rnn-based speech enhancement methods for noise-robust text-to-speech." in *SSW*, 2016, pp. 146–152.

[33] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*. Springer, 2015, pp. 234–241.

[34] O. Petit, N. Thome, C. Rambour, L. Themyr, T. Collins, and L. Soler, "U-net transformer: Self and cross attention for medical image segmentation," in *Machine Learning in Medical Imaging: 12th International Workshop, MLMI 2021, Held in Conjunction with MICCAI 2021, Strasbourg, France, September 27, 2021, Proceedings 12*. Springer, 2021, pp. 267–276.

[35] G. Yariv, I. Gat, S. Benaim, L. Wolf, I. Schwartz, and Y. Adi, "Diverse and aligned audio-to-video generation via text-to-video model adaptation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 7, 2024, pp. 6639–6647.