TAGMO: Temporal Control Audio Generation for Multiple Visual Objects Without Training

Xinyu Zhang[1,2], Keyu Fan[1,2], Yiran Wang[1,2], Yingshan Liang[1,2],
Jiasheng Lu[2], Zhicheng Du[1,2], Qingyang Shi[1,2], Peiwu Qin[1,*]
[1]Shenzhen International Graduate School, Tsinghua University, Shenzhen, China
[2]Huawei Technologies Co., Ltd., Shenzhen, China

***Abstract*—With the great popularity of Sora, video-based audio generation has become indispensable. While numerous video-to-audio generation models have emerged, they frequently face difficulties including semantic incompatibilities and synchronization problems, especially in situations with multiple objects. To address these difficulties, we introduce TAGMO, a novel training-free audio generation method that offers precise time control for multi-object video scenarios. Our approach first employs object detection to obtain the class labels and temporal labels of each object, which are then structured and utilized as control conditions within a latent diffusion model (LDM) to generate multi-object audio. Additionally, we innovatively design a time mask based on the corresponding temporal labels and integrate it into the denoising process of the pre-trained audio generation model to achieve accurate temporal control. Experimental results demonstrate that our method enhances temporal alignment accuracy and semantic consistency. Audio demonstrations are available at https://coco-create.github.io/.**

***Index Terms*—audio generation, multiple objects, temporal control, training-free strategy.**

I. Introduction

The field of Artificial Intelligence Generated Content (AIGC) [1] is rapidly growing with advancements in deep learning and natural language processing technologies. AIGC enables the generation of diverse types of content, including text, images, audio, and video. With video generation technologies like Sora, there is a rising demand for audio generation that complements the video content [2], leading to research and applications in generating audio that matches the video content efficiently and with higher quality [3].

Video-to-audio (V2A) models are used to generate semantically consistent audio with the video, ensuring temporal synchronization between the audio and the video frame. Currently, there are two types of V2A models: one that directly generates audio without text description of the video content [4], [5], and another that utilizes a video comprehension model [2], [6] to generate description for text-to-audio (T2A) generation [7], [8]. However, these models face challenges when it comes to semantic consistency and precise time synchronization in scenarios with multiple objects. As shown in the Fig. 1, (a) shows a video of a dog catching a crowd of cows, (b) shows the generated audio is missing the object *cow* in the video, (c) shows the generated audio and video is not synchronized.

∗Corresponding Author: Peiwu Qin:pwqin@sz.tsinghua.edu.cn
Work done during an internship at Huawei Technologies Co., Ltd.

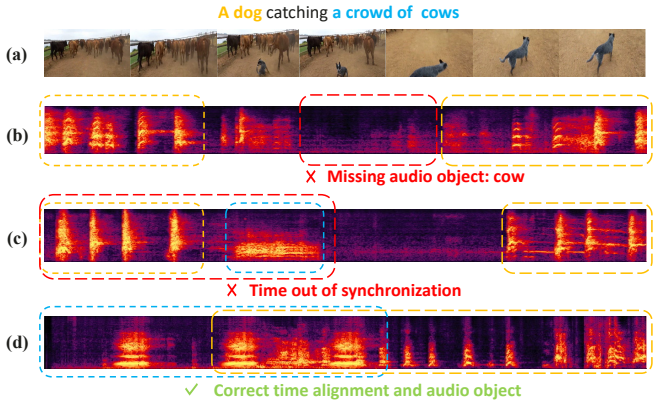

Fig. 1. Two challenges in V2A: semantic consistency and video-audio synchronization.

To address these challenges, a multi-object time-controlled audio generation model called TAGMO is proposed. The model obtains class and time labels of the objects from the input video, which are used to guide the audio generation for multiple objects within a given time duration. Specifically, the model consists of two modules: the first video-to-label module leverages object detection to get semantic and time information labels and structure them into labels [object class, (start time, end time)], and the second label-to-audio module generates audio guided by object class labels and utilizes time masks derived from time labels to achieve precise temporal control. Experiments have demonstrated the effectiveness and superiority of the model in achieving semantic information consistency and temporal alignment, Fig. 1 (d) shows the multi-object time-controlled audio generated by TAGMO. Our main contributions are as follows:

1. To the best of our knowledge, we first propose a novel method TAGMO to achieve fine-grained time-controlled audio generation for multiple visual objects without training.

2. We obtain semantic information from videos through object detection and structure it into class and time labels, generating stable multi-object audio.

3. To achieve precise temporal control, we introduce a time mask based on temporal labels and integrate it into the denoising process of diffusion model, significantly enhancing multi-object video-audio synchronization.

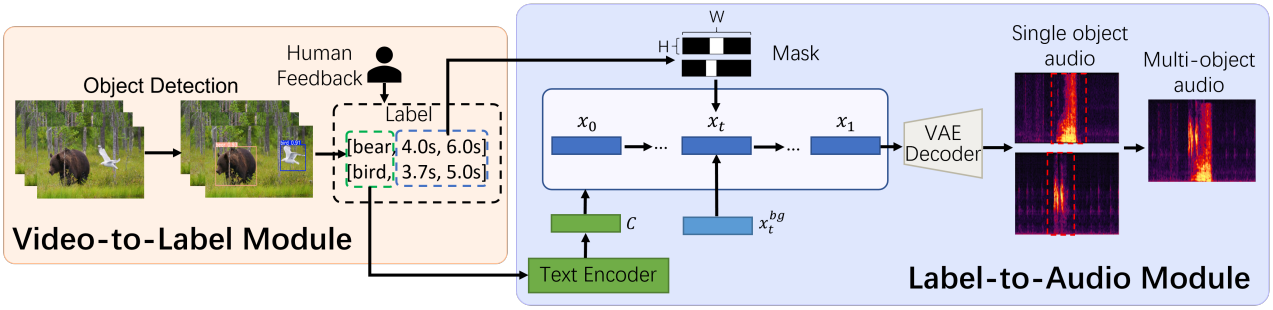

Fig. 2. Overview of the TAGMO system for video-to-audio generation. The pipeline consists of two stages: (i) video-to-label module obtains the content information of the video and structuring it into labels and (ii) label-to-audio utilizes the class and time label to generate multi-object temporal control audio.

II. Related Work

A. Text to audio generation

Recent advancements in AIGC have demonstrated the strong potential of T2A models [9]–[11], significantly improving audio generation capabilities. AudioLDM [12] introduced audio generation using a latent diffusion model (LDM), reducing computational resources. AudioLDM2 [13] proposed a general framework for multimodal audio generation based on LDM, but its generated audio is influenced by non-object information in the textual input. Make-An-Audio [1] leverages the potential diffusion of a spectrogram autoencoder for modeling long continuous waveforms, and Make-An-Audio 2(MAA2) [14] achieves temporal order control of audio events by labeling event and time information in the prompt. However, MAA2 [14] provides only fuzzy timestamps, leading to coarser temporal control. In contrast, TAGMO uses labeled textual information to focus on the object, ensuring the stability of object audio and providing accurate timestamps for precise temporal control.

B. Video to audio generation

Sora sparked a video generation boom [15], highlighting the importance of generating corresponding audio from video. While V2A models have shown promise, stable multi-object audio generation remains challenging. Diff-Foley's [16] CAVP module aligns semantic and temporal features but lacks precise multi-object temporal alignment. Lumina-T2X [17] improves audio quality with flow matching but lacks time control for synchronization. To address this, FoleyCrafter [18] introduced a temporal controller with timestamp detection for better synchronization, though it struggles with accuracy, especially for multiple objects sounding simultaneously. TAGMO, however, uses object labels and time mask to ensure both semantic stability and precise temporal synchronization for multi-object audio generation.

C. Multi-object image generation

Many efforts have been made to improve the controllability of multi-object generation models in the image domain. Zero-Painter [19] method enables precise control over multi-object generation by using object masks and individual descriptions to create objects in specific regions, and then employs an inpainting model to integrate these objects and generate the overall background. Haruka Matsuda et al. introduced a personalized text-to-image model using segmentation and continual learning to maintain visual fidelity across multiple objects [20]. Sen Li et al. developed MuLan, a training-free method that decomposes prompts into sub-tasks for progressive generation and feedback control [21]. Jianxiang Lu et al. addressed one-shot learning by initializing prototypical embeddings and using class-characterizing regularization to enhance generalizability and fidelity [22].

III. Proposed Method

A. Problem setting

Given a silent video V containing N different visual objects, our goal is to generate an audio A that maintains semantic consistency and temporal alignment with the video. Specifically, the generated audio A not only contains the audio of N objects, but also the audio a_i of each object i should exactly match the duration of its appearance in the video, meaning a_i should begin to sound at time t_{start_i} when i first appears in the video and end sounding at time t_{end_i}.

B. Video to structured labels

The input to the video-to-label module of TAGMO is a silent video. Using object detection algorithms, we detect the class and frame-by-frame presence of objects within the video, obtaining their class labels and the start and end time of each object's appearance. This information is structured as [object, start time, end time] and passed to the label-to-audio module for audio generation. For instance, the label [bear, 3.9s, 4.8s] indicates that audio of a bear should be generated from 3.9 to 4.8 seconds. Our framework also supports flexible external editing, allowing manual adjustments to object and time information via an interactive interface.

C. Structured labels to audio

1) Audio generation based on labels: Audio generation involves the reverse process of the diffusion model. After acquiring the structured label, the label-to-audio module uses a text encoder to extract the embedding, which are then input as control condition into the conditional latent diffusion model [12]. Instead of the step-by-step reverse process of diffusion models, we employ flow-matching-based Diffusion

Transformers(DiT) for denoising [17], which utilizes a linear interpolation forward process between noise and data:

$$x_t = tx + (1-t)\epsilon, t \in [0,1] \tag{1}$$

where data $x \sim p(x)$ and Gaussian noise $\epsilon \sim \mathcal{N}(0,1)$, and $t \in [0,1]$ is defined between $x_0 = \epsilon$ and $x_1 = x$ to indicate the interpolation range. The time mask is then obtained from the time labels and added to the denoising process to guide the object audio generation in the specified time period. At the same time, a silent audio clip will be input after VAE encoding to participate in the denoising process as padding. The VAE decoder converts the latent space data into a mel-spectrogram, which is then transformed into a wavform by the vocoder. Finally, audio for multiple objects is mixed from the precisely time-controlled single object audio.

2) Precise temporal control based on mask: Similar to using a layout mask in Zero-Painter [19], [23] to guide image generation at specified locations, we design a time mask set $M_i \in \{0,1\}^{H\times W}, i = 0,1,...N$ for audio generation to guide mel-spectrogram creation within a specific time period, H and W are the shape of the representation in the latent space encoded by a mel-spectrogram-based VAE, and H dimension represents frequency while W dimension represents time, in our model $H = 20$ and $W = 312$. The start and end time labels (t_{start_i}, t_{end_i}) are adjusted to the mel-spectrogram timescale and mapped to the corresponding positions on the latent space's horizontal axis (w_{start_i}, w_{end_i}), $M_i(w_0, h_0) = 1$ when $w_0 \in [w_{start_i}, w_{end_i}]$ and $h_0 \in [0, H]$.

Since the later steps of the denoising process in diffusion model are responsible for generating the detailed information of the objects [19], we only apply the mask on later phase of the denoiseing process, i.e. $t \in [t_0, 1], 0.5 < t_0 < 1$. We first obtain the latent code of a background audio A^{bg} by adding noise on a latent encoding of a constant silence sound:

$$x_{t_0}^{\text{bg}} = \sqrt{\alpha_{t_0}}\mathcal{E}(A^{\text{bg}}) + \sqrt{1-\alpha_{t_0}}\epsilon \tag{2}$$

where $\mathcal{E}()$ is the VAE encoder and α_t are hyperparameters of DDPM [24]. To ensure that the generated audio does not extend beyond the mask as well as make the diffusion process smoother and more natural, we blend the noized latent of the background audio x_t^{bg} and the predicted x_t, and x_t^{bg} is only involved in the denoising process for $t \in [t_0, 1]$:

$$x_t = M_i \odot x_t + (1 - M_i) \odot x_t^{bg}, t \in [t_0, 1] \tag{3}$$

This time-controlled audio generation requires only adding masks and background audio padding in the later denoising process of diffusion. So this method is universal and can be employed to any other DDPM-based audio generation model.

IV. Experiments

A. Experimental setup

1) Dataset: To objectively verify TAGMO's accurate time-control capability for multi-object audio generation, we used a test dataset from Audio-Condition [25]. This dataset includes 1110 10-second audio samples, each with an event description and a corresponding time period. We selected videos containing the categories in object detection dataset COCO [26] to meet the pipeline requirements.

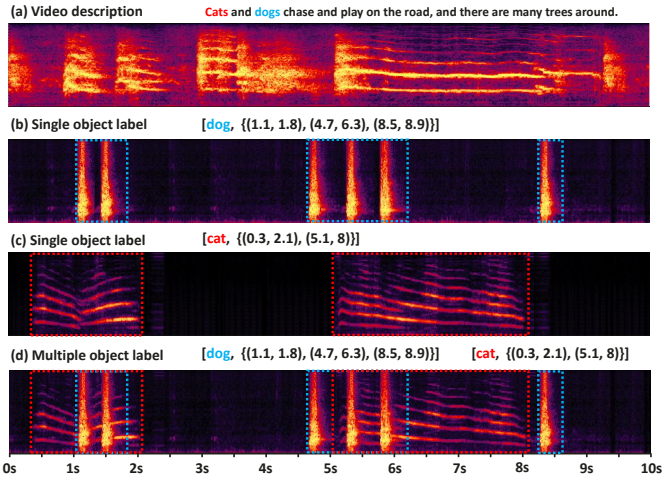

Fig. 3. Effect of inputting video description and label on semantic consistency of generated audio. (a) uses captions from a video understanding model, while (b)(c)(d) use structured labels from object detection. It demonstrates that audio generated from video descriptions is less synchronized with the video compared to using structured labels.

2) Model configurations: We use YOLOv10 [27] for object detection, specifically the YOLOv10-X model, with 29.5 million parameters. Our implementation of training-free, multi-object, time-controlled audio generation is based on the Lumina-T2X [17] framework, so we have not modified the original module and parameter settings of the base model, i.e., we have used the Flag-DiT-B as the generation model, with 8 layers, 12 heads and 768 hidden size; the text encoder is T5-v1.1-XXL, VAE used is proposed by Make-an-Audio 2 [14] finetuned from Make-an-Audio, and BigVGAN [28] is used as a vocoder to transform mel-spectrogram to waveform.

B. Evaluation methods

We evaluate generated audio using both subjective and objective metrics. Objectively, we use Frechet Audio Distance (FAD) [29] to assess audio quality and diversity, Mean KL Divergence (MKL) [30] to measure discrepancies between generated and real data distributions, and Alignment Accuracy (Align Acc) [16] to evaluate synchronization and audio-visual relevance. Subjectively, we employ Mean Opinion Score (MOS) to evaluate audio semantic consistency (MOS-S) and video-audio alignment (MOS-A).

TABLE I
THE AUDIO QUALITY COMPARISONS WITH BASELINE

Model	Objective metrics			Subjective metrics	
	FAD↓	MKL↓	Align Acc(%)↑	MOS-S↑	MOS-A↑
Lumina to audio	1.36	1.92	76.39	80.23	82.97
Diff-Foley	9.31	4.75	87.96	82.39	85.72
FoleyCrafter	5.77	3.58	74.07	81.43	86.19
TAGMO	**1.08**	**1.63**	**89.64**	**83.65**	**90.36**
-w/o mask	1.28	1.79	78.95	84.65	83.53
-w/o padding	1.31	2.03	80.13	81.37	87.46

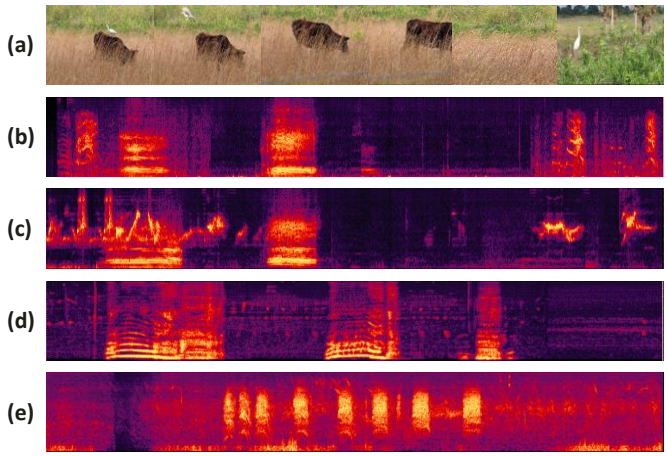

Fig. 4. Video-Audio alignment results compared with other V2A models. (a) a silent video with cows and birds in the frame, (b) ground truth audio, (c) audio generated by TAGMO, (d) audio generated by Diff-Foley, (e) audio generated by Foley-Crafter.

C. Quantitative Results

The results shown in Table I indicate that TAGMO consistently outperforms not only the baseline but also other video-audio generation models Diff-Foley and FoleyCrafter on all metrics. Significant improvements are observed in Align Acc and MOS-A, highlighting TAGMO's superior temporal alignment capabilities. Furthermore, the increase in MOS-S demonstrates TAGMO's effectiveness in maintaining semantic consistency in scenarios with multiple objects.

D. Qualitative Results

1) Semantic information consistency: To illustrate TAGMO's capability in maintaining semantic consistency in multi-object scenarios, we provided video descriptions from Video-LLaMA [31] and class and time labels from YOLOv10 to the label-to-audio module of TAGMO. As shown in Fig. 3, the audio generated from the Video-LLaMA descriptions includes noise and extraneous sounds because Video-LLaMA captures extensive semantic information beyond the vocalizable objects. In contrast, using YOLOv10 for object detection focuses solely on relevant objects, minimizing irrelevant distractions. Additionally, the structured labels improve semantic stability across multiple objects, ensuring no missing or confused elements.

2) Video-Audio Alignment Results: Compared to other V2A models, TAGMO demonstrates superior in temporal alignment for multiple objects. As illustrated in Fig. 4, both TAGMO and FoleyCrafter generate accurate cow moo and bird chirp sounds. However, Diff-Foley fails to produce the bird chirping, and FoleyCrafter introduces additional noise. This highlights TAGMO's better semantic consistency in multi-object scenarios. Additionally, TAGMO achieves fine-grained temporal alignment for each object, while FoleyCrafter's bird chirp sound is out of synchronization. These results underscore the effectiveness of TAGMO's object-level time mask control.

We investigated the effects of applying time masks at different stages of the denoising process. As illustrated in Fig. 5, applying a mask early in the process predominantly results in noise. Introducing the mask during the middle stage produces some recognizable dog barking sounds but with noticeable noise and lower audio quality. In contrast, applying the mask at the final stage yields the clearest results, generating distinct and high-quality dog barking sounds. This outcome is due to earlier stages focusing on reconstructing the general outline and information, while later stages are dedicated to refining the details. Therefore, applying the mask at the final stage proves to be the most effective approach.

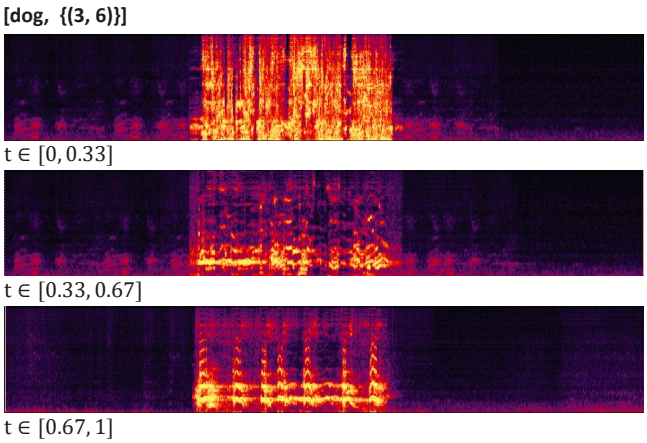

Fig. 5. The results of adding mask at different stages.

E. Ablation study

We compared the results of no mask, adding mask without background padding, and adding mask with background padding in our ablation experiments, as shown in Fig. 6. Without a mask, there is no time control effect. When the mask lacks background padding, barking occurs within the time control range but is accompanied by significant noise outside this period. With background padding, the generated audio quality is the best. This is because the padding ensures that areas outside the mask remain meaningful data, not affecting the masked sections. Setting the non-masked part to 0 in the latent space does not achieve the desired background conditions like silence or background music. Additionally, quantitative results in table I demonstrate significant improvements in Align ACC and MOS-S with the addition of the mask, confirming the effectiveness of our method in enhancing time control.

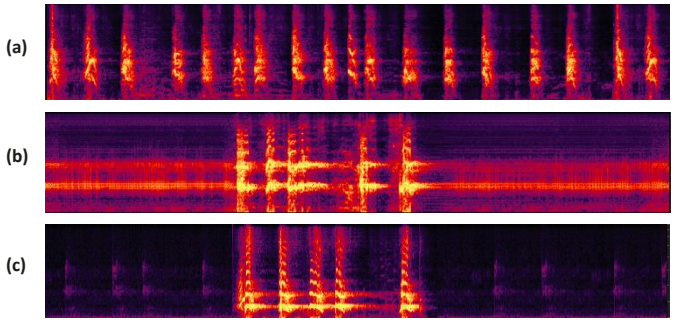

Fig. 6. The results of (a) no mask, (b) adding mask without background audio padding, and (c) mask with background padding.

REFERENCES

[1] 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," 2023. I, II-A

[2] G. Chen, G. Wang, X. Huang, and J. Sang, "Semantically consistent video-to-audio generation using multimodal language large model," *ArXiv*, vol. abs/2404.16305, 2024. I

[3] Y. Du, Z. Chen, J. Salamon, B. C. Russell, and A. Owens, "Conditional generation of audio from video via foley analogies," *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2426–2436, 2023. I

[4] M. Xu, C. Li, Y. Ren, R. Chen, Y. Gu, W. Liang, and D. Yu, "Video-to-audio generation with hidden alignment," *ArXiv*, vol. abs/2407.07464, 2024. I

[5] X. Wang, Y. Wang, Y. Wu, R. Song, X. Tan, Z. Chen, H. Xu, and G. Sui, "Tiva: Time-aligned video-to-audio generation," in *ACM Multimedia 2024*, 2024. I

[6] Z. Xie, S. Yu, Q. He, and M. Li, "Sonicvisionlm: Playing sound with vision language models," *ArXiv*, vol. abs/2401.04394, 2024. I

[7] J. Copet, F. Kreuk, I. Gat, T. Remez, D. Kant, G. Synnaeve, Y. Adi, and A. D'efossez, "Simple and controllable music generation," *ArXiv*, vol. abs/2306.05284, 2023. I

[8] F. Kreuk, G. Synnaeve, A. Polyak, U. Singer, A. D'efossez, J. Copet, D. Parikh, Y. Taigman, and Y. Adi, "Audiogen: Textually guided audio generation," *ArXiv*, vol. abs/2209.15352, 2022. I

[9] D. Ghosal, N. Majumder, A. Mehrish, and S. Poria, "Text-to-audio generation using instruction-tuned llm and latent diffusion model," *ArXiv*, vol. abs/2304.13731, 2023. II-A

[10] 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*, vol. abs/2404.09956, 2024. II-A

[11] H. Liao, H. Han, K. Yang, T. Du, R. Yang, Q. Xu, Z. Xu, J. Liu, J. Lu, and X. Li, "Baton: Aligning text-to-audio model using human preference feedback," *Proceedings of the Thirty-ThirdInternational Joint Conference on Artificial Intelligence*, 2024. II-A

[12] 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," 2023. II-A, III-C1

[13] H. Liu, Y. Yuan, X. Liu, X. Mei, Q. Kong, Q. Tian, Y. Wang, W. Wang, Y. Wang, and M. D. Plumbley, "Audioldm 2: Learning holistic audio generation with self-supervised pretraining," 2024. II-A

[14] 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," 2023. II-A, II-A, IV-A2

[15] Y. Liu, K. Zhang, Y. Li, Z. Yan, C. Gao, R. Chen, Z. Yuan, Y. Huang, H. Sun, J. Gao, L. He, and L. Sun, "Sora: A review on background, technology, limitations, and opportunities of large vision models," *ArXiv*, vol. abs/2402.17177, 2024. II-B

[16] S. Luo, C. Yan, C. Hu, and H. Zhao, "Diff-foley: Synchronized video-to-audio synthesis with latent diffusion models," 2023. II-B, IV-B

[17] P. Gao, L. Zhuo, D. Liu, R. Du, X. Luo, L. Qiu, Y. Zhang, C. Lin, R. Huang, S. Geng, R. Zhang, J. Xi, W. Shao, Z. Jiang, T. Yang, W. Ye, H. Tong, J. He, Y. Qiao, and H. Li, "Lumina-t2x: Transforming text into any modality, resolution, and duration via flow-based large diffusion transformers," 2024. II-B, III-C1, IV-A2

[18] 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," 2024. II-B

[19] M. Ohanyan, H. Manukyan, Z. Wang, S. Navasardyan, and H. Shi, "Zero-painter: Training-free layout control for text-to-image synthesis," 2024. II-C, III-C2, III-C2

[20] H. Matsuda, R. Togo, K. Maeda, T. Ogawa, and M. Haseyama, "Multi-object editing in personalized text-to-image diffusion model via segmentation guidance," in *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 8140–8144, 2024. II-C

[21] S. Li, R. Wang, C.-J. Hsieh, M. Cheng, and T. Zhou, "Mulan: Multimodal-llm agent for progressive and interactive multi-object diffusion," 2024. II-C

[22] J. Lu, C. Xie, and H. Guo, "Object-driven one-shot fine-tuning of text-to-image diffusion with prototypical embedding," 2024. II-C

[23] Y. Balaji, S. Nah, X. Huang, A. Vahdat, J. Song, Q. Zhang, K. Kreis, M. Aittala, T. Aila, S. Laine, B. Catanzaro, T. Karras, and M.-Y. Liu, "ediff-i: Text-to-image diffusion models with an ensemble of expert denoisers," 2023. III-C2

[24] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," 2020. III-C2

[25] Z. Guo, J. Mao, R. Tao, L. Yan, K. Ouchi, H. Liu, and X. Wang, "Audio generation with multiple conditional diffusion model," 2023. IV-A1

[26] T.-Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick, and P. Dollár, "Microsoft coco: Common objects in context," 2015. IV-A1

[27] A. Wang, H. Chen, L. Liu, K. Chen, Z. Lin, J. Han, and G. Ding, "Yolov10: Real-time end-to-end object detection," 2024. IV-A2

[28] S. gil Lee, W. Ping, B. Ginsburg, B. Catanzaro, and S. Yoon, "Bigvgan: A universal neural vocoder with large-scale training," 2023. IV-A2

[29] K. Kilgour, M. Zuluaga, D. Roblek, and M. Sharifi, "Fréchet audio distance: A metric for evaluating music enhancement algorithms," 2019. IV-B

[30] V. Iashin and E. Rahtu, "Taming visually guided sound generation," 2021. IV-B

[31] H. Zhang, X. Li, and L. Bing, "Video-llama: An instruction-tuned audio-visual language model for video understanding," 2023. IV-D1