Interspeech 2025
17-21 August 2025, Rotterdam, The Netherlands

MM-MovieDubber: Towards Multi-Modal Learning for Multi-Modal Movie Dubbing

Junjie Zheng[1], *Zihao Chen*[1], *Chaofan Ding*[1], *Yunming Liang*[1], *Yihan Fan*[2], *Huan Yang*[1], *Lei Xie*[3], *and Xinhan Di*[1]

[1]AI Lab, Giant Network, China
[2]The East China University of Science and Technology, China
[3]ASLP@NPU, Northwestern Polytechnical University, China

*@ztgame.com, Y80220125@mail.ecust.edu.cn, lxie@nwpu.edu.cn

Abstract

Current movie dubbing technology can produce the desired speech using a reference voice and input video, maintaining perfect synchronization with the visuals while effectively conveying the intended emotions. However, crucial aspects of movie dubbing, including adaptation to various dubbing styles, effective handling of dialogue, narration, and monologues, as well as consideration of subtle details such as speaker age and gender, remain insufficiently explored. To tackle these challenges, we introduce a multi-modal generative framework. First, it utilizes a multi-modal large vision-language model (VLM) to analyze visual inputs, enabling the recognition of dubbing types and fine-grained attributes. Second, it produces high-quality dubbing using large speech generation models, guided by multi-modal inputs. Additionally, a movie dubbing dataset with annotations for dubbing types and subtle details is constructed to enhance movie understanding and improve dubbing quality for the proposed multi-modal framework. Experimental results across multiple benchmark datasets show superior performance compared to state-of-the-art (SOTA) methods. In details, the LSE-D, SPK-SIM, EMO-SIM, and MCD exhibit improvements of up to 1.09%, 8.80%, 19.08%, and 18.74%, respectively.

Index Terms: movie dubbing, video understanding, speech generation, multi-modal

1. Introduction

Dubbing involves adding the correct human voice to a video's dialogue, ensuring synchronization with the characters' lip movements, and conveying the emotions of the scene. Existing dubbing methods can be categorized into two groups, each focusing on learning different styles of key prior information to generate high-quality voices. The first group focuses on learning effective speaker style representations [1, 2, 3, 4]. The second group aims to learn appropriate prosody by utilizing visual information from the given video input [5, 6, 7]. However, the accuracy of these priors is often insufficient due to the limitations of various modules, which impacts the overall quality of speech generation. Additionally, with the rapid advancement of large language models (LLMs) and related methods, multi-modal large vision language models (VLMs) have increasingly demonstrated their potential in multi-modal understanding tasks. These models can integrate information from various modalities for effective understanding, such as audio-text [8] and vision-text [9, 10]. These advancements in VLMs understanding capabilities hold promise for accurately providing dubbing types and fine-grained attributes.

Therefore, we introduce a multi-modal generative framework for movie dubbing that incorporates an understanding of dubbing types and fine-grained attributes. First, a multi-modal large vision-language model (VLM) is trained to improve the understanding of dubbing types and fine-grained attributes from video inputs. Second, a large speech generation model is trained with designed control mechanisms guided by multi-modal conditions. Finally, we construct a movie dubbing dataset with diverse annotations, including dubbing types(Dialogue, Narration and Monologue) and fine-grained attributes, to improve the movie understanding and dubbing quality of the proposed model.

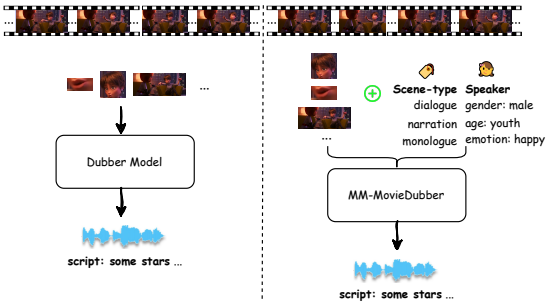


Figure 1: *Curent Dubbing models [11, 12, 13] (Left). Proposed Dubbing Models (Right) For dubbing types and fine-grained attributes.*

2. Method

2.1. Overview

Given a silent video clip V_l, a corresponding subtitle T_v, and the goal of generating a fully dubbed video, the proposed model (MM-MovieDubber) aims to produce speech $\hat{S}$ that matches the video, ensures contextual and prosodic relevance, and maintains speech-video synchronization. The model can be formalized as follows:

$$\hat{S} = F^{n}_{dubber}(V_l, T_v) \tag{1}$$

MM-MovieDubber consists of two modeling stages: i) Multi-modal understanding through in-context learning. In this stage, a multi-modal large language model (mLLM) receives an instruction to generate an understanding conclusion C_v from video inputs, which can be formalized as:

$$C_v = F^{1}_{mm}(V_l, I_v) \tag{2}$$

where V_l represents the input video clip, and I_v represents the specified instruction. ii) Speech Generation Stage: This stage incorporates a conditional DiT based speech generator. The generative model combines the generated understanding conclusion, video clip and video related script to produce speech,

10.21437/Interspeech.2025-1933

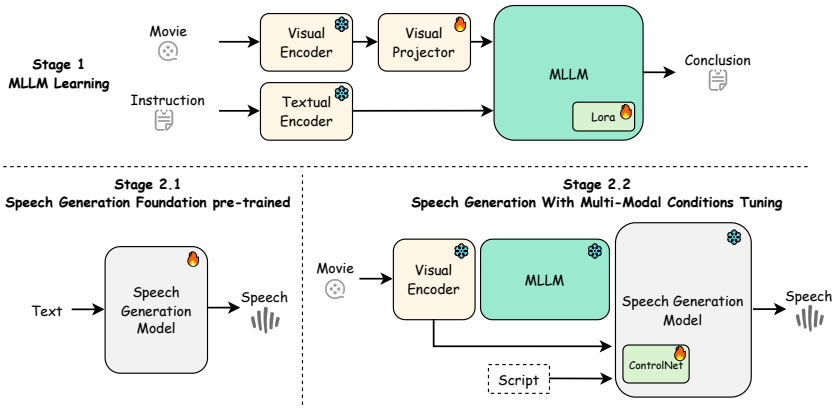


Figure 2: *MM-MovieDubber pipeline with multi-stage, multi-modal training.*

which can be formalized as:

$$\hat{S} = F^{2}_{generator}(V_l, C_v, T_v) \tag{3}$$

where V_l represents the input video clip, C_v represents the understanding conclusion, and T_v represents the expected script. The speech $\hat{S}$ is generated under conditioning.

2.2. Stage 1: Multi-modal Large Language Model (mLLM) Learning

Building upon the success of multi-modal large language models (mLLMs) like Visual-LLM [14, 15], we employ multi-modal instruction tuning to achieve video understanding.

$$C_i = F_{mm}(V_l, I_v) \tag{4}$$

where V_l represents the input video clip, I_v represents the specific instruction, and F_{mm} represents the learning stage of the mLLM. Once the scene-related visual features are aligned with the textual scene features in the mLLM embedding space, the mLLM effectively encodes multi-modal information, which serves as a representation for understanding conclusion generation.

2.3. Stage 2: Mutli-Conditioned Speech Generation

2.3.1. Stage 2.1: Speech Generation Foundation Pre-Training

In the second stage of MM-MovieDubber, the foundational speech generation model will first be trained. We use the same architecture as F5-TTS [16], which employs a diffusion transformer (DiT) as the backbone and is trained to output a vector field v_τ with the conditional flow-matching objective $\mathcal{L}_{CFM}$ [17]. The neural network's training is guided by the conditional flow matching objective:

$$\mathcal{L}^{\text{CFM}}(\theta) = \mathbb{E}_{\tau, q(x_1), p(x|x_1)} \| u_\tau(x|x_1) - v_\tau(x;\theta) \|^2 \tag{5}$$

where p_τ is the probability path at time τ, u_τ is the designated vector field for p_τ, x_1 symbolizes the random variable corresponding to the training data, and q is the distribution of the training data.

2.3.2. Stage 2.2: Speech Generation With Multi-Modal Conditions Tuning

Next, in the ControlNet-transformer tuning stage, the video frames, along with an instruction are fed as inputs to the mLLM model. The sequence of video features and the video understanding conclusion are then combined and passed into the speech generation model. In this context, the provided conclusion helps guide the video-to-speech generation process, as shown in stage 2.2 of Figure 2. The proposed speech generation model takes as input the script, silent video, video understanding conclusion, and an optional reference speech, and generates video-aligned speech context sequences, which can be described as:

$$\hat{S} = F_{generator}(V_l, C_v\{C_1, \ldots, C_n\}_{n=4}, T_v) \tag{6}$$

where C_v is the combination of $\{C_1, \ldots, C_n\}, n = 4$, which represents the scene type condition C_s, speaker gender condition C_g, speaker age condition C_a, and speaker emotion condition C_e. These conditions are combined and encoded by an encoder [18]. The V_l represents the visual features derived from the input video frames, which are encoded by CLIP [19]. We implement a cross-attention mechanism that facilitates the integration of understanding conclusion features C_v and visual features V_l. Additionally, T_v represents the embedded script. Furthermore, we added a duration loss $\mathcal{L}_{dur}$ to constrain the duration consistency, which can be described as:

$$\mathcal{L}_{\text{dur}} = \ell\big(f(V_l, C_l),\, dur\big), \tag{7}$$

The final loss function is constructed as follows:

$$\mathcal{L}_{\text{g}} = \mathbb{E}_{\tau, q(x_1), p(x|x_1)} \| u_\tau(x|x_1, v_l, c_v, t_v) - v_\tau(x;\theta) \|^2 + \mathcal{L}_{\text{dur}} \tag{8}$$

In the training stage, the visual condition V_l, video understanding conclusion condition C_v, and video script condition T_v are each set to ϕ with a 5 % probability. Extending classifier-free guidance from the script condition to visual input and visual understanding enhances both conditional control precision and speech quality. The guidance scales, λ_V, λ_C, and λ_T, correspond to the video clip, video conclusion, and video-related script, respectively, and measure the alignment between the sampling results and conditions. Inspired by [20], during inference, the modified velocity estimate is as follows:

$$\begin{aligned} g'_\theta &= g_\theta(x_\tau, \phi, \phi, \phi) \\ &+ \lambda_V \cdot (v_0(x_\tau, c_v, c_c, c_t) - v_0(x_\tau, \phi, c_c, c_t)) \\ &+ \lambda_C \cdot (v_0(x_\tau, \phi, c_c, c_t) - v_0(x_\tau, \phi, \phi, c_t)) \\ &+ \lambda_T \cdot (v_0(x_\tau, \phi, \phi, c_t) - v_0(x_\tau, \phi, \phi, \phi)) \end{aligned} \tag{9}$$

Table 1: *Objective results on V2C-Animation benchmark. For the Dub 1.0 setting, we use the ground truth audio as reference audio, for the Dub 2.0 setting, we use the non-ground truth audio from the same speaker within the dataset as the reference audio which is more aligned with practical usage in dubbing.*

Setting		Dub 1.0					Dub 2.0				
Methods	Visual	SPK-SIM (%) ↑	WER (%) ↓	EMO-SIM (%) ↑	MCD ↓	MCD-SL ↓	SPK-SIM (%) ↑	WER (%) ↓	EMO-SIM (%) ↑	MCD ↓	MCD-SL ↓
GT	-	100	17.38	100	0.00	0.00	100	17.38	100	0.00	0.00
F5-TTS [16]	✗	89.30	24.41	76.78	8.32	8.32	83.11	24.83	64.91	10.86	10.87
HPMDubbing [11]	✓	73.64	151.02	39.85	8.59	8.32	73.01	150.83	34.69	9.11	12.15
Speaker2Dub [13]	✓	82.15	31.23	65.92	10.68	11.21	79.53	31.28	59.71	11.16	11.70
StyleDubber [12]	✓	82.48	27.36	66.24	10.06	10.52	79.81	26.48	59.08	10.56	11.05
Ours	✓	**89.74**	**22.52**	**78.88**	**6.98**	**6.99**	**83.30**	**24.71**	**64.93**	**8.80**	**8.80**

Table 2: *Results on GRID benchmark with the same dub setting as the V2C-Animation benchmark.*

Setting		Dub 1.0						Dub 2.0					
Methods	Visual	LSE-C ↑	LSE-D ↓	SPK-SIM (%) ↑	WER (%) ↓	MCD ↓	MCD-SL ↓	LSE-C ↑	LSE-D ↓	SPK-SIM (%) ↑	WER (%) ↓	MCD ↓	MCD-SL ↓
GT	-	7.18	13.36	100	13.67	0.00	0.00	7.18	13.36	100	13.67	0.00	0.00
F5-TTS [16]	✗	5.51	14.70	**96.51**	**11.94**	**4.23**	**4.24**	5.10	14.71	94.45	16.75	4.89	4.90
HPMDubbing [11]	✓	**6.35**	14.78	93.64	16.78	4.57	4.85	**6.34**	14.79	92.84	17.40	4.95	5.24
Speaker2Dub [13]	✓	5.64	14.82	96.11	12.11	7.85	8.01	5.56	14.84	94.91	12.89	7.57	7.73
StyleDubber [12]	✓	6.19	14.81	96.40	11.97	7.71	7.81	6.16	14.83	**95.25**	**11.97**	7.34	7.43
Ours	✓	4.87	**14.63**	95.73	14.71	4.48	4.49	4.46	**14.63**	94.71	16.08	**4.73**	**4.74**

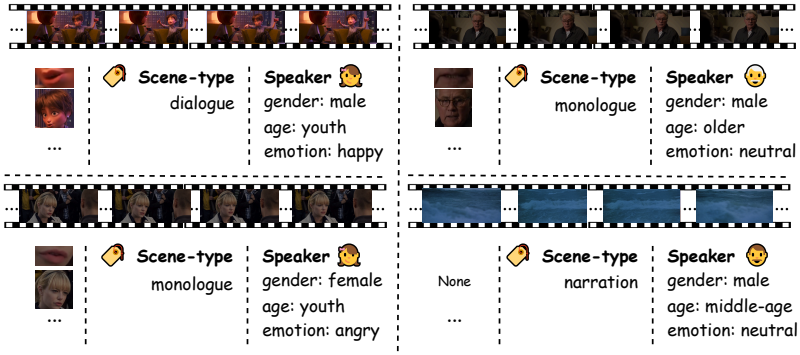


Figure 3: *Proposed dataset with multi-type annotations, including annotation for lips, faces, scene-type, speaker gender, speaker age, voice emotion.*

3. Experiments

3.1. Datasets

Emilia is a comprehensive multilingual speech generation dataset containing a total of 101,654 hours of speech data across six languages [21]. The English portion of this dataset, comprising approximately 46,800 hours, is utilized to train our foundational text-to-speech (TTS) model.

V2C-Animation is a specialized dataset designed for animated movie dubbing, consisting of 10,217 clips from 26 films with synchronized text, audio, and video [22]. The dataset is partitioned into 60% for training, 10% for validation, and 30% for testing.

GRID is a dubbing benchmark for multi-speaker dubbing [23]. The whole dataset has 33 speakers, each with 1000 short English samples. All participants are recorded in studio with unified background. The number of train and test data are 32,670 and 3280, respectively.

Movie Dubbing Dataset. We have built a 7.2-hour multi-modal movie dubbing dataset with multi-type annotations to enhance movie understanding and improve dubbing quality. As shown in Figure 3, the conclusion provides a high-level overview of the entire scene, while other labels describe the scene type, speaker, and speech in the video. The dataset is partitioned as follows: 60% for training, 10% for validation, and 30% for testing.

3.2. Implementation Details

To improve training stability, multi-modal speech generation capability is progressively incorporated throughout the training process using a multi-stage approach. Video frames are sampled at 25 FPS, and all speech is resampled to 22 kHz. The foundational speech generation model is trained on 8 NVIDIA A800 80G GPUs for about one week using Emilia-EN [21]. The mLLM learning stage and the speech generation tuning stage are implemented on a single NVIDIA A800 80G GPU with the proposed movie-dubbing dataset.

4. Evaluation

We evaluate using both objective and subjective metrics. To assess pronunciation accuracy, we use Word Error Rate (WER) with Whisper-V3[24] as the ASR model. Timbre consistency is evaluated with speaker encoder cosine similarity (SPK-SIM) [12]. We also calculate mel cepstral distortion dynamic time warping (MCD) and speech length variance (MCD-SL) [25] for

Table 3: *Subjective evaluation on V2C-Animation and GRID benchmarks.*

Dataset	V2C-Animation		GRID	
Methods	NMOS	SMOS	NMOS	SMOS
GT	4.98±0.01	-	4.99±0.01	-
F5-TTS [16]	4.20±0.68	3.83±0.63	**4.43±0.03**	**3.32±0.05**
HPMDubbing [11]	1.04±0.01	1.02±0.01	3.50±0.10	2.77±0.12
Speaker2Dubber [13]	2.93±0.21	2.58±0.19	4.04±0.07	3.00±0.10
StyleDubber [12]	2.68±0.21	2.39±0.21	4.01±0.03	3.06±0.07
Ours	**4.37±0.35**	**3.91±0.45**	4.33±0.07	3.14±0.08

Table 4: *Results on zero-shot test, which use unseen speaker as reference audio.*

Setting	Dubbing Setting 3.0				
Methods	LSE-C ↑	LSE-D ↓	SPK-SIM ↑ (%)	WER ↓ (%)	MOS ↑
HPMDubbing [11]	1.72	**11.74**	68.14	126.85	1.29±0.60
Speaker2Dub [13]	2.21	12.67	76.10	16.57	3.38±0.14
StyleDubber [12]	2.15	12.76	78.30	19.07	3.30±0.15
Ours	**2.21**	12.59	**83.55**	**15.49**	**4.12±0.16**

spectral and length differences. Emotion similarity is assessed using a speech emotion recognition model [26]. For alignment with video, we use Lip Sync Error Distance (LSE-D) and Lip Sync Error Confidence (LSE-C) metrics on the Grid benchmark, based on the pre-trained SyncNet model [27]. For subjective evaluation, we conduct human evaluations of the Mean Opinion Score (MOS) for naturalness (NMOS) and similarity (SMOS), rated on a 1-to-5 scale with 95% confidence intervals. Following [13], participants evaluate the dubbing quality of 30 randomly selected audio samples from each test set.

4.1. Results

We compare our approach with a recent SOTA TTS model, F5-TTS [16], and three recent SOTA video dubbing models: HPMDubbing [11], StyleDubber [12], and Speaker2Dubber [13]. Tables 1, 2, 4, and 5 present our objective evaluation results. In the subjective evaluation, we randomly select 30 samples from the generated dubbing of each dub setting for human assessment, and the results are shown in Table 3.

4.1.1. Comparison with SOTA

Results on V2C-Animation benchmark. As shown in Table 1, in the comparison with the state-of-the-art models[11, 12, 13], our model achieves improvements across evaluation metrics in the same setting[13]. In details, the SPK-SIM increased from 79.81% to 83.30%, EMO-SIM improved from 59.71% to 64.93%, MCD decreased from 9.11 to 8.80, and WER reduced from 26.48% to 24.71%.

Results on GRID benchmark. As shown in Table 2, our model achieves the best lip-sync performance on the GRID benchmark with the same evaluation of the state-of-the-art models[13], which decreased from 14.79 to 14.63. And the MCD decreased from 4.95 to 4.73. In addition, our method also achieves competitive results in WER, slightly lower than the best fine-tuned F5-TTS model, Speaker2Dub and StyleDubber. However, these models have WER results (11.94%, 12.11% and 11.97%) that exceed the ground truth WER result (13.67%), suggesting that the intelligibility has reached an acceptable range for humans.

Results on Speaker Zero-shot test. This setting uses the audio of unseen speakers as reference audio to measure the generalizability of the dubbing model[13]. Here, we use the audio from GRID as reference audio to measure V2C. We compare LSE-C/D, SPK-SIM, and WER, along with subjective evaluations in the same evaluation setting with the state-of-the-art models[13]. As shown in Table 4, our method outperforms the StyleDubber and Speaker2Dub in both SPK-SIM and WER. In details, the SIP-SIM improved from 78.30% to 83.55%, the WER decreased from 16.57% to 15.49%. Furthermore, the proposed method still maintains competitive performance in audio-visual synchronization (see LSE-C and LSE-D), slightly lower than HPMDubbing.

Table 5: *Results on proposed benchmark, which use silence video and script as inputs.*

Setting	towards fined-grained movie dubbing test				
Methods	LSE-C ↑	LSE-D ↓	SPK-SIM (%) ↑	WER ↓ (%)	MOS
HPMDubbing [11]	1.46	**11.46**	61.06	199.40	1.11±0.06
Speaker2Dub [13]	1.68	12.00	61.73	84.42	1.54±0.22
StyleDubber [12]	**2.35**	13.06	64.03	52.69	2.76±0.16
Ours	1.84	11.88	**70.82**	**27.68**	**4.01±0.04**

Table 6: *Results of ablation study on the proposed dataset.*

	LSE-C ↑	LSE-D ↓	SPK-SIM (%) ↑	EMO-SIM (%) ↑	MCD ↓	MCD-SL ↓
w/o clip	1.99	12.73	82.63	63.24	8.85	8.86
w/o dur	**2.06**	12.81	82.71	64.32	8.82	8.83
w/o conclusion	1.89	12.66	82.59	63.18	8.81	8.82
proposed	2.01	**12.61**	**82.99**	**64.74**	**8.76**	**8.77**

Results on towards fined-grained movie dubbing test. In order to better fit the real-world usage scenarios, we also conducted a test on the proposed dataset, which we call the towards fined-grained movie dubbing test. This setting just uses the silent video and related script as inputs (no reference speech) to measure the performance of the dubbing model for different dubbing types(Dialogue, Narration and Monologue) and fine-grained attributes(gender, age and emotion). The results are shown on Table 5. Our method outperforms the SOTA dubbing methods (StyleDubber and Speaker2Dub) on SPK-SIM, EMO-SIM and WER. In details, the SPK-SIM improved from 64.03% to 70.83%, WER decreaded from 52.69% to 27.68%. Furthermore, the proposed method still maintains the competitive performance in audio-visual synchronization (see LSE-D), 11.88%, which slightly lower than HPMDubbing(11.46%).

Ablation Studies. The ablation results in Table 6 indicate that each condition plays a role in overall performance. Removing the video clip control causes all metrics to drop significantly, showing its importance for speech-video alignment. Adding the video understanding conclusion control improves SPK-SIM and EMO-SIM. Finally, removing the duration predictor leads to the largest drop in LSE-D performance, emphasizing that duration-level consistency learning is essential for synchronizing speech and video.

5. Conclusion

In this paper, we propose a multi-stage, multi-modal generative framework for movie dubbing. Additionally, we have developed a movie dubbing dataset with multi-type annotations to enhance movie understanding and improve dubbing quality. In the future, we plan to incorporate reasoning ability into our movie dubbing model to improve its understanding and reasoning skills, ultimately producing higher-quality dubbing.

6. References

[1] Q. Chen, M. Tan, Y. Qi, J. Zhou, Y. Li, and Q. Wu, "V2c: Visual voice cloning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 21 242–21 251.

[2] M. Hassid, M. T. Ramanovich, B. Shillingford, M. Wang, Y. Jia, and T. Remez, "More than words: In-the-wild visually-driven prosody for text-to-speech," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 10 587–10 597.

[3] L. Wan, Q. Wang, A. Papir, and I. L. Moreno, "Generalized end-to-end loss for speaker verification," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4879–4883.

[4] G. Cong, L. Li, Y. Qi, Z.-J. Zha, Q. Wu, W. Wang, B. Jiang, M.-H. Yang, and Q. Huang, "Learning to dub movies via hierarchical prosody models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 14 687–14 697.

[5] C. Hu, Q. Tian, T. Li, W. Yuping, Y. Wang, and H. Zhao, "Neural dubber: Dubbing for videos according to scripts," *Advances in neural information processing systems*, vol. 34, pp. 16 582–16 595, 2021.

[6] J. Lee, J. S. Chung, and S.-W. Chung, "Imaginary voice: Face-styled diffusion model for text-to-speech," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[7] Y. Zhao, Z. Jia, R. Liu, D. Hu, F. Bao, and G. Gao, "Mcdubber: Multimodal context-aware expressive video dubbing," in *National Conference on Man-Machine Speech Communication*. Springer, 2024, pp. 168–182.

[8] A. S. Penamakuri, K. Chhatre, and A. Jain, "Audiopedia: Audio qa with knowledge," 2024.

[9] Y. Dong, Z. Liu, H.-L. Sun, J. Yang, W. Hu, Y. Rao, and Z. Liu, "Insight-v: Exploring long-chain visual reasoning with multimodal large language models," 2024. [Online]. Available: https://arxiv.org/abs/2411.14432

[10] Z. Cheng, S. Leng, H. Zhang, Y. Xin, X. Li, G. Chen, Y. Zhu, W. Zhang, Z. Luo, D. Zhao, and L. Bing, "Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms," 2024. [Online]. Available: https://arxiv.org/abs/2406.07476

[11] G. Cong, L. Li, Y. Qi, Z. Zha, Q. Wu, W. Wang, B. Jiang, M.-H. Yang, and Q. Huang, "Learning to dub movies via hierarchical prosody models," 2023.

[12] G. Cong, Y. Qi, L. Li, A. Beheshti, Z. Zhang, A. van den Hengel, M.-H. Yang, C. Yan, and Q. Huang, "Styledubber: Towards multi-scale style learning for movie dubbing," 2024. [Online]. Available: https://arxiv.org/abs/2402.12636

[13] Z. Zhang, L. Li, G. Cong, H. YIN, Y. Gao, C. Yan, A. van den Hengel, and Y. Qi, "From speaker to dubber: Movie dubbing with prosody and duration consistency learning," in *ACM Multimedia 2024*, 2024. [Online]. Available: https://openreview.net/forum?id=QHRNR64J1m

[14] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," in *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. [Online]. Available: https://openreview.net/forum?id=w0H2xGHlkw

[15] G. Xu, P. Jin, H. Li, Y. Song, L. Sun, and L. Yuan, "Llava-cot: Let vision language models reason step-by-step," 2025. [Online]. Available: https://arxiv.org/abs/2411.10440

[16] Y. Chen, Z. Niu, Z. Ma, K. Deng, C. Wang, J. Zhao, K. Yu, and X. Chen, "F5-tts: A fairytaler that fakes fluent and faithful speech with flow matching," 2024. [Online]. Available: https://arxiv.org/abs/2410.06885

[17] Y. Lipman, R. T. Q. Chen, H. Ben-Hamu, M. Nickel, and M. Le, "Flow matching for generative modeling," in *The Eleventh International Conference on Learning Representations*, 2023. [Online]. Available: https://openreview.net/forum?id=PqvMRDCJT9t

[18] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *J. Mach. Learn. Res.*, vol. 21, pp. 140:1–140:67, 2020. [Online]. Available: https://jmlr.org/papers/v21/20-074.html

[19] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, ser. Proceedings of Machine Learning Research, M. Meila and T. Zhang, Eds., vol. 139. PMLR, 2021, pp. 8748–8763. [Online]. Available: http://proceedings.mlr.press/v139/radford21a.html

[20] J. Jiang, G. Hong, L. Zhou, E. Ma, H. Hu, xia zhou, J. Xiang, F. Liu, K. Yu, H. Sun, K. Zhan, P. Jia, and M. Zhang, "DiVE: Dit-based video generation with enhanced control," in *ECCV 2024 Workshop on Multimodal Perception and Comprehension of Corner Cases in Autonomous Driving*, 2024. [Online]. Available: https://openreview.net/forum?id=QE5IUwmBuL

[21] H. He, Z. Shang, C. Wang, X. Li, Y. Gu, H. Hua, L. Liu, C. Yang, J. Li, P. Shi, Y. Wang, K. Chen, P. Zhang, and Z. Wu, "Emilia: An extensive, multilingual, and diverse speech dataset for large-scale speech generation," 2024. [Online]. Available: https://arxiv.org/abs/2407.05361

[22] Q. Chen, Y. Li, Y. Qi, J. Zhou, M. Tan, and Q. Wu, "V2c: Visual voice cloning," 2021. [Online]. Available: https://arxiv.org/abs/2111.12890

[23] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *The Journal of the Acoustical Society of America*, vol. 120, no. 5, pp. 2421–2424, 11 2006. [Online]. Available: https://doi.org/10.1121/1.2229005

[24] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," 2022. [Online]. Available: https://arxiv.org/abs/2212.04356

[25] E. Battenberg, R. Skerry-Ryan, S. Mariooryad, D. Stanton, D. Kao, M. Shannon, and T. Bagby, "Location-relative attention mechanisms for robust long-form speech synthesis," 2020. [Online]. Available: https://arxiv.org/abs/1910.10288

[26] J. Ye, X.-C. Wen, Y. Wei, Y. Xu, K. Liu, and H. Shan, "Temporal modeling matters: A novel temporal emotional modeling approach for speech emotion recognition," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[27] J. S. Chung and A. Zisserman, "Out of time: automated lip sync in the wild," in *Computer Vision–ACCV 2016 Workshops: ACCV 2016 International Workshops, Taipei, Taiwan, November 20-24, 2016, Revised Selected Papers, Part II 13*. Springer, 2017, pp. 251–263.