PRELIMINARY PREPRINT VERSION – DO NOT CITE
This is not the version of record. The final version will appear in the Proceedings of IJCAI-ECAI 2026.

FunCineForge: A Unified Dataset Pipeline and Model for Zero-Shot Movie Dubbing in Diverse Cinematic Scenes

Jiaxuan Liu[1], **Yang Xiang**[2], **Han Zhao**[2], **Xiangang Li**[2] and **Zhenhua Ling**[1*]
[1]National Engineering Research Center of Speech and Language Information Processing, University of Science and Technology of China, Hefei, China
[2]Independent Researcher, Beijing, China
jxliu@mail.ustc.edu.cn, xiangyang3131777@gmail.com, hanzhaousc@gmail.com, lixiangang.speech@hotmail.com, zhling@ustc.edu.cn

Abstract

Movie dubbing is the task of synthesizing speech from scripts conditioned on video scenes, requiring accurate lip sync, faithful timbre transfer, and proper modeling of character identity and emotion. However, existing methods face two major limitations: (1) high-quality multimodal dubbing datasets are limited in scale, suffer from high word error rates, contain sparse annotations, rely on costly manual labeling, and are restricted to monologue scenes, all of which hinder effective model training; (2) existing dubbing models rely solely on the lip region to learn audio-visual alignment, which limits their applicability to complex live-action cinematic scenes, and exhibit suboptimal performance in lip sync, speech quality, and emotional expressiveness. To address these issues, we propose FunCineForge, which comprises an end-to-end production pipeline for large-scale dubbing datasets and an MLLM-based dubbing model designed for diverse cinematic scenes. The pipeline enables the construction of the first television dubbing dataset, CineDub, which serves as a high-quality foundation for training and evaluation. Building on this, our dubbing model effectively captures multimodal cues and supports complex dubbing scenarios, including monologue, narration, dialogue, and multi-speaker settings. Experiments demonstrate that our approach consistently outperforms state-of-the-art methods in audio quality, word error rate, lip sync, temporal alignment, timbre transfer, and instruction following. Code and demos are available at https://funcineforge.github.io/.

Figure 1: The overview of FunCineForge. The dataset pipeline automatically transforms raw film and television sources into structured multimodal data, which are used to train and evaluate the dubbing model. During inference, given a silent video clip, dubbing script, segment-level timestamps, an instruction, and reference speech, the model synthesizes scene-consistent speech.

1 Introduction

Movie Dubbing [Chen *et al.*, 2022], also known as Visual-Voice-Cloning (V2C), is a task that synthesizes speech for video clips from scripts using target timbres. This technique has significant potential in television and film production, digital media, and video editing. Unlike Text-to-Speech (TTS), movie dubbing requires incorporating visual information from character faces to precisely control the duration, emotion, and prosody of the synthesized speech.

Existing movie dubbing methods can be broadly categorized into three research paradigms. The first category focuses on improving the synchronization between the synthesized speech and lip movements by learning visual features from the speaker's lip region. For example, a text-video aligner is introduced in NeuralDubber [Hu *et al.*, 2021], a duration aligner is introduced in HPMDubber [Cong *et al.*, 2023], and a phoneme-guided lip aligner is introduced in StyleDubber [Cong *et al.*, 2024b]. These methods aim to explicitly control phoneme-level speech duration by mapping textual phonemes to video frames. However, their align-

*Corresponding author.

IJCAI-ECAI 2026 PREPRINT

PRELIMINARY PREPRINT VERSION – DO NOT CITE
This is not the version of record. The final version will appear in the Proceedings of IJCAI-ECAI 2026.

ment performance heavily relies on the presence of clear and visible lip regions, making them effective only for simple, frontal, and high-resolution monologue scenes. In live-action television, complex cinematic scenes such as facial occlusion, frequent shot changes, multiple coexisting speakers, large speaker-camera distances, and low-resolution frames are common, which prevents reliable audio-visual alignment.

The second category focuses on enhancing dubbing quality and speaker similarity. For example, Speaker2Dubber [Zhang *et al.*, 2024] improves synthesis quality by via prosody-enhanced acoustic pretraining. FlowDubber [Cong *et al.*, 2025] introduces dual contrastive learning and LLM-based flow matching guidance, enhancing speech quality. ProDubber [Zhang *et al.*, 2025b] proposes prosody-enhanced acoustic pretraining and acoustic-disentangled prosody adaptation to improve prosody modeling and expressiveness. EmoDubber [Cong *et al.*, 2024a] employs a flow-based user emotion controlling with positive and negative guidance mechanisms to manipulate dubbing emotions. Authentic-Dubber [Liu *et al.*, 2025] enhances emotional retrieval and matching by constructing a multimodal reference footage library and employing emotion-similarity-based retrieval augmentation. To better clone the reference speaker timbre, most above works adopt acoustic encoders to extract speaker embeddings, while Lee et al. [2023] learn speaker identity from facial features to capture speaker paralinguistic information. However, these methods heavily rely on clear, single-speaker faces and consistent timbre. In the presence of shot-level facial transitions, speaker timbre switching, or intermediate silent segments, alignment becomes difficult, and both dubbing quality and speaker similarity degrade substantially.

The third category leverages Multimodal Large Language Models (MLLMs) to fuse information from multiple modalities for movie dubbing. DeepDubber-V1 [Zheng *et al.*, 2025] utilizes multimodal Chain-of-Thought (CoT) reasoning on visual inputs to infer dubbing styles and speaker attributes, generating visual-derived clues that serve as instructions for the dubbing MLLM. InstructDubber [Zhang *et al.*, 2025a] enables controllable dubbing by conditioning the model on textual prompts that describe speaker identity and emotional clues as an additional modality. These methods primarily rely on natural language instructions to control speech style and emotion, offering a more intuitive and interpretable paradigm for movie dubbing. However, the effectiveness of MLLMs critically depends on large-scale datasets, whereas existing datasets are limited in scale and lack rich annotations, which constrains their effectiveness in real-world scenarios.

To address these issues, we propose FunCineForge, a unified framework for zero-shot movie dubbing that integrates an automated data production pipeline with an MLLM-based dubbing model for diverse cinematic scenes, as illustrated in Fig. 1. The pipeline combines the advantages of lightweight specialized models and general-purpose MLLMs, innovatively adopting a **"specialized model prediction + multimodal CoT correction"** strategy to refine both transcription and speaker diarization, significantly reducing word error rates and speaker identification errors. Based on this pipeline, we construct **CineDub**, a large-scale dubbing dataset comprising two language subsets, CineDub-CN and CineDub-EN, with over 9,000 hours of speech in total. Building upon CineDub, we develop an MLLM-based dubbing model that, for the first time, explicitly introduces a **temporal modality** via a **Timestamp-Speaker tokenizer**, enabling precise control over temporal alignment. Specifically, the model jointly models **when speech occurs, who is speaking, and what is being spoken**, while striving for **fine-grained lip-speech alignment** within speech-active intervals. It is further equipped with an improved flow matching module that **supports flexible speaker switching**. As a result, the proposed dubbing model synthesizes highly natural speech with precise temporal alignment and lip sync, effectively handling complex cinematic scenes and speaker switching. We provide a **project website**[1] to showcase synthesized demos and CineDub dataset samples, and the code repository[2] has been open-sourced.

2 Related Works

2.1 Dubbing Datasets

Dubbing datasets used for model training, including V2C-Animation [Chen *et al.*, 2022], Chem [R *et al.*, 2020], and GRID [Cooke *et al.*, 2006], are limited in scale. Even the largest one, V2C-Animation, contains only 10,217 video-audio-text clips with an average duration of 2.4 seconds, totaling 6.8 hours of data. These datasets heavily rely on manual annotation or crowd-sourced services, are difficult to scale, and are limited to monologue scene due to excessively short clips. The scarcity of data and limited, inconsistent annotations also hinder the effectiveness of MLLMs in complex cinematic scenes, highlighting the urgent need for a fully automated pipeline to construct large-scale dubbing datasets.

2.2 LLM-based Text-to-Speech

Recent advances in large language models (LLMs) have reshaped the field of TTS. Most LLM-based TTS methods adopt a two-stage paradigm, where an LLM first generates speech semantic tokens from text and an acoustic decoder converts them into acoustic features such as Mel-spectrograms, achieving significant advantages in synthesizing expressive speech. Representative models include CosyVoice 3 [Du *et al.*, 2025], IndexTTS 2 [Zhou *et al.*, 2025], FireRedTTS 2 [Xie *et al.*, 2025], VibeVoice [Peng *et al.*, 2025] and Spark-TTS [Wang *et al.*, 2025], In parallel, flow matching has emerged as an effective acoustic decoding strategy. Matcha-TTS [Mehta *et al.*, 2023] adopts optimal-transport conditional flow matching for single-speaker TTS, while CosyVoice 2 [Du *et al.*, 2024] demonstrates the effectiveness of integrating flow matching with LLMs. Despite their success, these methods cannot provide the speech with desired emotion and audio-visual sync for movie dubbing.

2.3 Speaker Diarization for Cinematic Scenes

Speaker Diarization task [Hsu *et al.*, 2017] aims to segment an audio recording into regions corresponding to different speakers by estimating both the number of speakers and the

[1]https://funcineforge.github.io/

[2]https://github.com/FunAudioLLM/FunCineForge/

IJCAI-ECAI 2026 PREPRINT

PRELIMINARY PREPRINT VERSION – DO NOT CITE
This is not the version of record. The final version will appear in the Proceedings of IJCAI-ECAI 2026.

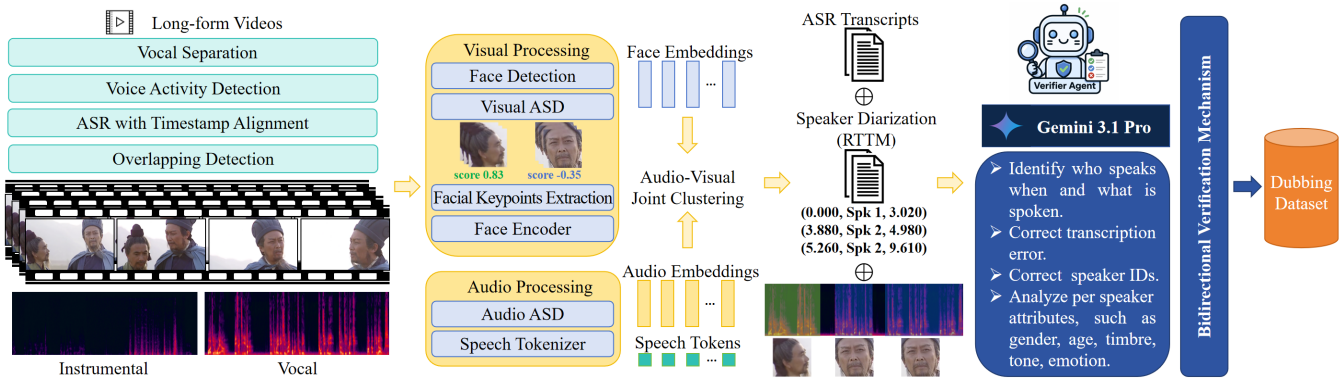

Figure 2: Overview of the FunCineForge dataset pipeline for cinematic dubbing data construction. The pipeline consists of four stages: video preprocessing and segmentation, timestamped transcription and alignment, audio-visual speaker diarization, and multimodal CoT correction.

start-end timestamps of their speech segments. In complex cinematic scenes with multi-speaker interactions, accurate movie dubbing requires explicit knowledge of which speaker is active at each temporal interval, making robust speaker diarization a crucial prerequisite. Most speaker diarization methods [Zheng *et al.*, 2020; Li *et al.*, 2023] follow a feature extraction and unsupervised clustering pipeline, consisting of: (1) voice activity detection (VAD) to remove non-speech regions; (2) sliding-window extraction of speaker embeddings using models such as Res2Net [Gao *et al.*, 2019] and CAM++ [Wang *et al.*, 2023]; and (3) unsupervised clustering of speech segments. 3D-Speaker [Zheng *et al.*, 2023] further incorporates visual embeddings for visually enhanced clustering, achieving notable performance gains. Despite these advances, speaker diarization in long-form videos remains challenging due to: (1) clustering errors caused by a large number of speakers in long-form videos; (2) failures under overlapping multi-speaker speech; and (3) speaker confusion induced by similar tones. Consequently, effective segmentation plus speaker diarization for long-form cinematic scenes remain open problems.

3 FunCineForge

3.1 Dataset Pipeline

To enable training of dubbing models in diverse cinematic scenes, FunCineForge dataset pipeline transforms raw long-form television sources into structured multimodal data. It is specifically designed to address challenges such as high word error rates in transcription and inaccurate speaker diarization. As shown in Fig. 2, the pipeline comprises several coordinated workflows, summarized in four stages.

Video Preprocessing and Segmentation

For long-form videos, we first perform audio-visual preprocessing and segmentation. To improve vocal clarity and suppress background music, thereby reducing word error rates and diarization error rates, we employ a Mel-RoFormer [Wang *et al.*, 2024] module to separate vocal and instrumental components, yielding clean vocal tracks. Next, an FSMN-Monophone VAD [Zuo *et al.*, 2023] model optimized for long-duration speech is applied to extract speech-active segments. Based on the timestamps, the original video is subsequently segmented into shorter video-audio clips. Subsequently, an overlapped speech detection module is applied to the vocal tracks to identify and discard clips containing multiple simultaneous speakers, ensuring the reliability of downstream speaker diarization.

Timestamped Transcription and Alignment

The vocal tracks of the segmented clips are transcribed into text with punctuation using automatic speech recognition (ASR) models, and character-level timestamps are predicted via forced alignment. Specifically, we employ FunASR [Gao *et al.*, 2023] for Chinese and Qwen3-ASR [Shi *et al.*, 2026] for English, which exhibit slightly different performance in terms of word error rate and temporal alignment accuracy. To further reduce temporal discrepancies, we apply an additional alignment step using MMS to refine the audio-text sync.

Audio-Visual Speaker Diarization

A visually enhanced speaker diarization framework is adopted to integrate audio and visual embeddings for robust speaker attribution. In the audio modality, CAM++ [Wang *et al.*, 2023] is employed to extract speaker embeddings from clean vocal tracks, which implicitly capture speaker activity information, while the CosyVoice 3 speech tokenizer [Du *et al.*, 2025] encodes audio into speech tokens at 25 Hz. In the visual modality, frame sampling is first performed by selecting one frame every five frames for 25 fps videos. All faces in these sampled frames are detected using a lightweight face detection module[3] to obtain candidate active frames. The detected faces are then scored by TalkNet-ASD[4] module, and the face with the highest score in each active frame is selected as the active speaker. For the selected face, a FAN-based [Bulat and Tzimiropoulos, 2017] module is applied to detect 2D facial keypoints, from which lip coordinates are obtained and raw face and mouth regions are extracted. An adaptive face encoder [Huang *et al.*, 2020] is used to extract face embeddings of the active speaker, followed by normalization to suppress expression-related variations and retain identity-related representations. Finally, a audio-visual joint

[3]https://github.com/Linzaer/Ultra-Light-Fast-Generic-Face-Detector-1MB

[4]https://github.com/TaoRuijie/TalkNet-ASD

IJCAI-ECAI 2026 PREPRINT

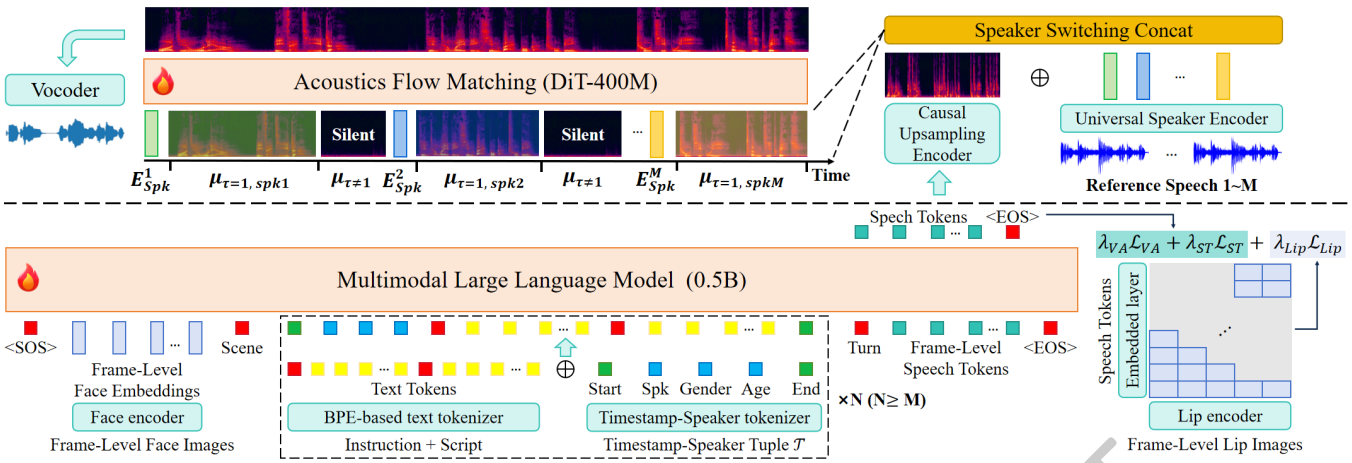

Figure 3: Overview of the FunCineForge dubbing model. The model integrates a multimodal large language model (MLLM) with a flow-matching acoustic decoder.

unsupervised clustering strategy [Zheng *et al.*, 2023] is applied, where audio speaker embeddings serve as the primary modality and normalized visual face embeddings act as auxiliary cues, producing robust speaker clustering results formatted as Rich Transcription Time Marked (RTTM) files.

Multimodal CoT Correction

To enrich each clip with paralinguistic information for controllable dubbing, the pipeline augments video-audio clips with structured character and emotion annotations, which are traditionally costly to obtain manually. Meanwhile, ASR transcripts often contain lexical and punctuation errors, and speaker diarization may suffer from missing or duplicated speaker identities. To address these issues, a multimodal Chain-of-Thought correction strategy is introduced.

Specifically, Gemini-3.1-Pro is employed as a multimodal **verifier agent** to perform cross-modal reasoning over clean vocal tracks, ASR transcripts, and speaker diarization tuples *(start time, speaker ID, end time)*. It first identifies active speakers over time and corrects transcription errors, and then infers speaker attributes, including gender, age group, identity, timbre, tone, speech rate and emotion. These results are further summarized into structured instructions. To ensure reliability, strict prompt templates are adopted, and a bidirectional verification mechanism is introduced to address occasional hallucination in MLLM outputs. Samples with large discrepancies (e.g., Levenshtein distance $> 72\%$) between corrected and original transcripts are discarded, while low-discrepancy samples retain the corrected transcripts. The threshold is selected based on the empirical distribution of transcript discrepancies. Similarly, samples with inconsistent speaker identities are filtered out, and uncertain attribute predictions are replaced with $<unknown>$ to mask unreliable labels. Finally, the remaining samples are categorized into monologue, narration, dialogue, and multi-speaker scenarios based on facial evidence and the number of active speakers.

3.2 Dubbing Model

The architecture of the dubbing model is illustrated in Fig. 3. Given an input video clip, frame-level facial sequences $Face \in \mathbb{R}^T$ of the active speaker are extracted, where T denotes the number of frames. For the video clip containing N speech segments, each segment is associated with a dubbing script $Text$, an instruction $Instruct$, and a timestamp-speaker tuple $\mathcal{T} = (start, spk, gender, age, end)$. In addition, reference speech samples ref^M from M speakers ($M \leq N$) are provided for timbre cloning. The model synthesizes Mel-spectrograms that are temporally aligned with the silent video and exhibit precise lip sync,

$$\hat{Y} = \text{Model}(Face, (Text, Instruct, \mathcal{T})^N, ref^M). \quad (1)$$

The dubbing model is trained in two stages: MLLM training and flow matching training. After flow matching, the synthesized Mel-spectrograms $\hat{Y}$ are converted into waveforms using the widely used HiFiGAN vocoder.

MLLM with Multimodal Alignment

First, the instruction and script are concatenated as "$\langle|\text{startofprompt}|\rangle Instruct \langle|\text{startofprompt}|\rangle Text$", and tokenized using a byte pair encoding (BPE)-based tokenizer to obtain the combined text tokens X_{Text}.

Face and lip images in $Face$ are extracted from 25-fps videos and uniformly sampled every five frames, resulting in 5-fps visual sequences. These downsampled face and lip frames are efficiently encoded into facial and lip representations using the face encoder [Huang *et al.*, 2020] and a lip encoder [Cong *et al.*, 2023], respectively, without normalization, thereby preserving expression-related variations. The resulting features are then upsampled to match the original video frame rate, producing face embeddings E^T_{Face} and lip embeddings E^T_{Lip}. Both E^T_{Face} and E^T_{Lip} are frame-level one-dimensional continuous feature sequences with the same sequence length as the speech tokens X_{Speech}. These features are inherently sparse, as the speaker's face is not consistently visible across all frames in real-world videos.

To handle frequent shot changes, speaker switching and facial occlusion in complex cinematic scenes, and to achieve accurate lip-speech alignment, we design a multimodal alignment mechanism that **combines strong and weak supervi-**

IJCAI-ECAI 2026 PREPRINT

PRELIMINARY PREPRINT VERSION – DO NOT CITE
This is not the version of record. The final version will appear in the Proceedings of IJCAI-ECAI 2026.

sion to jointly model when speech occurs, who is speaking, what is being spoken, and fine-grained lip-speech alignment.

(1) The provided timestamp-speaker tuples $\mathcal{T}^N$ serve as strong supervision, explicitly constraining **when speech occurs and who is speaking**, and enabling precise temporal alignment. A dedicated frame-index codebook is introduced to encode start-end timestamps. Given that each video clip is limited to 60 seconds and sampled at 25 fps, the maximum number of timestamp tokens is 1500. To jointly model temporal and speaker-related information, a Timestamp-Speaker Tokenizer (TST) is designed to map timestamps and speaker attributes into a discrete token sequence X_{TS}. Specifically, each segment is represented by tokens corresponding to speaker identity token X_{Spk}, gender token X_{Gender} for male and female, and X_{Age} corresponding to child, teenager, adult, middle-aged, elderly age groups. Attribute tokens corresponding to $<unknown>$ annotations are masked during training. In addition, a frame-level voice activity indicator is defined as $\tau_t = 1$ if speech is present at frame t, and $\tau_t = 0$ otherwise. Let $\hat{\tau}_t$ denote the predicted voice activity derived from the synthesized speech tokens. A voice activity loss is constructed over X_{Speech} as

$$\mathcal{L}_{VA} = -\sum_{t=1}^{T}[\tau_t \log(\hat{\tau}_t) + (1-\tau_t)\log(1-\hat{\tau}_t)]. \quad (2)$$

(2) For speech tokens X_{Speech}, the supervision focuses on **what is being spoken**. Accordingly, the speech token loss is defined as a cross-entropy loss under conditions $\mathcal{C}_{LM}$,

$$\mathcal{L}_{ST} = -\frac{1}{T+1}\sum_{t=1}^{T+1}\log p(X_{Speech}^{(t)} \mid X_{Speech}^{(<t)}, \mathcal{C}_{LM}). \quad (3)$$

(3) Visual features are introduced to enhance the MLLM's understanding of speaker identity and emotional expression, and to provide weak supervision for **fine-grained lip-speech alignment** when lip movements are clearly observable. Frame-level face embeddings E_{Face}^{T}, temporally aligned with the speech tokens X_{Speech}, are fed into the MLLM. A contrastive learning objective is introduced between frame-level lip embeddings E_{Lip}^{T} and speech token embeddings $E_{ST}^{T} = g_{st}(Emb(X_{Speech}))$,

$$\mathcal{L}_{Lip} = -\sum_{t=1}^{T}\tau_t w_t \log \frac{\exp\left(\langle E_{Lip}^{(t)}, E_{ST}^{(t)}\rangle/\gamma\right)}{\sum_{s=1}^{T}\exp\left(\langle E_{Lip}^{(t)}, E_{ST}^{(s)}\rangle/\gamma\right)}, \quad (4)$$

where τ_t is voice activity indicator activated in speech-active frames, w_t down-weights frames with weak lip motion, and γ is a temperature parameter.

Flow Matching with Speaker Switching

The flow matching module is built upon the CosyVoice 3 [Du *et al.*, 2025] backbone to reconstruct Mel-spectrograms from speech tokens. However, the original design conditions only on a single global speaker embedding, limiting its ability to model speaker transitions in dialogue scenarios. To address this limitation, a speaker switching concatenation strategy is introduced, conditioned on M reference speakers ref^M. A universal speaker encoder [Zheng *et al.*, 2023] is employed to extract audio speaker embeddings E_{Spk}^{M} for ref^M. Given timestamp-speaker tuples $\mathcal{T}^N$, the last silent token preceding each speech segment is identified via bidirectional search. The corresponding speaker embedding $E_{Spk} \in E_{Spk}^{M}$ is then inserted immediately after the last silent token and before the onset of speech tokens, which are encoded into $E_{ST^*}^{T}$ using a causal upsampling encoder. This design enables explicit speaker switching aligned with timestamp boundaries.

The concatenated feature sequence is then fed into a Diffusion Transformer (DiT) [Peebles and Xie, 2022] as conditional context $\mathcal{C}_{flow}$ for flow matching training,

$$\mathcal{L}_{Flow}(\theta) = \mathbb{E}_{u,Y_1}\left\|\mathrm{DiT}_\theta\big(Y_u, u \mid \mathcal{C}_{flow}\big) - (Y_1 - Y_0)\right\|_2^2, \quad (5)$$

where $\mathcal{C}_{flow} = \mathrm{Concat}\big(E_{ST^*}^{T}, E_{Spk}^{M}(\mathcal{T}^N)\big)$, $u \sim \mathcal{U}(0,1)$, $Y_1 \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, and $Y_u = (1-u)Y_0 + uY_1$. Within the DiT module, local attention is adopted to restrict the model's receptive field to speech-active segments. During training, the conditioning $\mathcal{C}_{flow}$ is randomly dropped with a fixed probability to enable classifier-free guidance at inference stage.

4 Experiments

4.1 Implementation Details

Data preprocessing. All videos are standardized to MP4 format, with opening and ending credits removed. Video frames are sampled at 25 fps, and audio is resampled to 16 kHz. The speech tokenizer adopts an FSQ codebook of size 6,560, and the frame-index codebook size is 1,500.

Training setup. The MLLM and flow matching modules are trained separately on the CineDub dataset for over 20 epochs. The CosyVoice3-0.5B checkpoint is used for initialization, providing strong speech naturalness. Training is conducted on 8 NVIDIA A100 GPUs, using the AdamW optimizer with a peak learning rate of $1e{-}4$, 2,000 warm-up steps, and a batch size of 2×10^4 tokens.

Baselines and evaluation. For evaluation, we select several open-source SOTA baselines specifically designed for movie dubbing, and use their official checkpoints trained on the V2C-Animation, Chem, or GRID datasets for inference. For instruction-driven methods, DeepDubber-V1 and InstructDubber, we modify their implementations and retrain them on the CineDub dataset to enable comparison under the same data setting.

Cross-dataset analysis. To evaluate dataset quality, the FunCineForge model is additionally trained on the combined V2C-Animation, Chem, and GRID datasets using only monologue scenes, with empty instruction and timestamp-speaker tuples. This setting enables comparison of how CineDub data types and annotation richness improve model performance.

4.2 Evaluation Metrics

We adopt a comprehensive set of objective and subjective metrics to evaluate dataset quality, dubbing quality, lip sync, temporal alignment, speaker timbre similarity, emotional consistency, and instruction-following consistency.

IJCAI-ECAI 2026 PREPRINT

PRELIMINARY PREPRINT VERSION – DO NOT CITE
This is not the version of record. The final version will appear in the Proceedings of IJCAI-ECAI 2026.

Scene	MCD-DTW↓	MCD-DTW-SL↓	CER(%)↓	WER(%)↓	UTMOS↑	LSE-C↑	LSE-D↓	DER(%)↓	SBFE↓	SSIM(%)↑	ESIM(%)↑	ES-MOS↑
CineDub Dataset (Corrected)												
GT	0.00	0.00	**0.94**	**2.12**	3.92	8.56	4.82	**1.20**	0.000	100.00	100.00	3.99
Monologue	**4.25**	**4.17**	1.49	4.30	**3.97**	**8.92**	**3.60**	7.25	0.073	78.62	**73.20**	**4.04**
Narration	4.30	4.43	1.90	4.53	3.95	-	-	6.37	**0.040**	**78.80**	63.37	3.89
Dialogue	4.73	4.90	2.85	7.98	3.90	8.62	4.73	16.58	0.082	73.35	69.35	3.80
Multi-speaker	4.90	5.13	3.44	12.34	3.82	7.50	8.68	14.31	0.078	64.51	64.42	3.95
CineDub Dataset (Uncorrected)												
GT	0.00	0.00	4.53	9.35	3.85	7.98	6.94	8.38	0.000	100.00	100.00	3.78
Monologue	5.22	5.01	3.52	5.90	3.89	8.30	5.85	9.93	0.098	74.11	70.81	3.84
Narration	5.30	5.35	4.96	6.24	3.82	-	-	8.60	0.075	73.20	60.18	3.80
Dialogue	6.10	6.47	11.08	11.39	3.71	7.82	8.93	23.64	0.147	68.10	64.50	3.68
Multi-speaker	6.24	6.45	12.90	13.40	3.65	7.24	11.24	20.05	0.158	57.29	60.32	3.69
V2C-Animation + Chem + GRID Dataset												
GT	0.00	0.00	-	24.23	3.70	7.05	7.55	-	0.000	100.00	100.00	-
Monologue (Overfitting)	6.58	6.80	-	8.46	3.76	7.50	9.96	-	0.223	68.80	64.25	-

Table 1: Performance of the FunCineForge dubbing model trained on the CineDub dataset (with and without Multimodal CoT Correction) and on the combined V2C-Animation, Chem, and GRID datasets. Results are reported across monologue, narration, dialogue, and multi-speaker scenarios. The ground-truth (GT) rows highlight the impact of Multimodal CoT Correction by comparing the CineDub datasets before and after correction, showing substantial improvements in transcription accuracy (CER/WER) and speaker diarization (DER).

Methods	MCD-DTW↓	MCD-DTW-SL↓	CER(%)↓	WER(%)↓	UTMOS↑	LSE-C↑	LSE-D↓	SBFE↓	SSIM(%)↑	ESIM(%)↑	ES-MOS↑
CineDub Dataset (Corrected)											
FunCineForge Dubbing Model	**4.25**	**4.17**	**1.49**	**4.30**	**3.97**	**8.92**	**3.60**	**0.073**	**78.62**	**73.20**	**4.04**
DeepDubber-V1 [Zheng *et al.*, 2025]	5.25	5.33	6.05	16.90	3.70	7.86	9.24	0.140	71.61	65.57	3.70
InstructDubber [Zhang *et al.*, 2025a]	5.05	5.23	3.84	12.80	3.82	8.08	7.93	0.156	74.53	72.86	3.83
V2C-Animation + Chem + GRID Dataset											
Speaker2Dubber [Zhang *et al.*, 2024]	9.80	10.03	-	32.12	3.42	5.63	12.58	0.307	63.05	46.33	-
StyleDubber [Cong *et al.*, 2024b]	8.52	8.63	-	29.45	3.70	6.03	12.05	0.213	64.80	49.52	-
ProDubber [Zhang *et al.*, 2025b]	6.60	6.72	-	11.02	3.82	5.89	10.45	0.149	70.50	58.19	-
EmoDubber [Cong *et al.*, 2024a]	8.83	8.89	-	22.14	3.84	7.80	7.67	0.160	65.14	68.78	-
DeepDubber-V1 [Zheng *et al.*, 2025]	7.46	7.55	-	24.28	3.68	6.33	10.07	0.208	67.17	60.35	3.32
InstructDubber [Zhang *et al.*, 2025a]	6.69	6.89	-	14.20	3.70	6.50	9.68	0.166	69.55	70.02	3.70

Table 2: Performance comparison of FunCineForge against baselines under the monologue scene.

IJCAI-ECAI 2026 PREPRINT

MCD-DTW & MCD-DTW-SL. Mel Cepstral Distortion with Dynamic Time Warping (MCD-DTW) and its variant MCD-DTW-SL [Chen *et al.*, 2022], which incorporates duration-aware weighting, measure the difference between synthesized speech and ground-truth recordings.

CER & WER. Character Error Rate and Word Error Rate are used to measure pronunciation accuracy, using Whisper-large-v3[5] as the ASR model.

UTMOS. UTMOS[6] is a 5-point speech mean opinion score (MOS) predictor to assess the naturalness of synthesized speech.

LSE-C & LSE-D. To accurately measure lip sync, we use Lip Sync Error Confidence (LSE-C) and Lip Sync Error Distance (LSE-D) [Chung and Zisserman, 2016]. Both metrics are computed based on the pre-trained SyncNet, which is widely adopted in lip-reading task.

DER & SBFE. To evaluate speaker-aware temporal alignment, we adopt Diarization Error Rate (DER) and Speaker Boundary Fidelity Error (SBFE). DER measures the proportion of time that is not correctly attributed to the reference speaker, including missed speech, false alarm, and speaker confusion. Lower DER indicates better overall temporal accuracy.

Given ground-truth time intervals $\mathcal{I}_i = [t_i^{\text{start}}, t_i^{\text{end}})$ from timestamp-speaker tuples $\mathcal{T}^N = \{(t_i^{\text{start}}, s_i, g_i, a_i, t_i^{\text{end}})\}_{i=1}^N$, and predicted intervals $\hat{\mathcal{I}}_i$ extracted from synthesized speech via VAD, SBFE quantifies truncation and leakage via a balanced asymmetric measure:

$$\text{SBFE} = \frac{1}{2} - \frac{1}{2N}\sum_{i=1}^{N}\left(\frac{|\hat{\mathcal{I}}_i \cap \mathcal{I}_i|}{|\mathcal{I}_i|} - \frac{|\hat{\mathcal{I}}_i \setminus \mathcal{I}_i|}{|\hat{\mathcal{I}}_i|}\right), \quad (6)$$

where SBFE $\in [0, 1]$, and lower values indicate more accurate alignment and faithful speaker switching.

SSIM. Speaker cosine similarity is used to evaluate speaker timbre similarity between synthesized speech and reference speech within each speaker's active time interval.

ESIM & ES-MOS. Emotional learning depends on the input face images and instructions. We employ emotion2vec[7] to extract emotion vectors from synthesized speech and ground-truth recordings, and compute the cosine similarity (ESIM) between them. We conduct a subjective 5-point MOS on emotional and style consistency (ES-MOS), where human raters evaluate how well the generated speech matches intended emotion and style in the instruction description.

4.3 CineDub Dataset and Quality Evaluation

To train the dubbing model and evaluate the quality of the dubbing dataset produced by our pipeline, we carefully col-

[5]https://huggingface.co/openai/whisper-large-v3

[6]https://github.com/tarepan/SpeechMOS

[7]https://github.com/ddlBoJack/emotion2vec

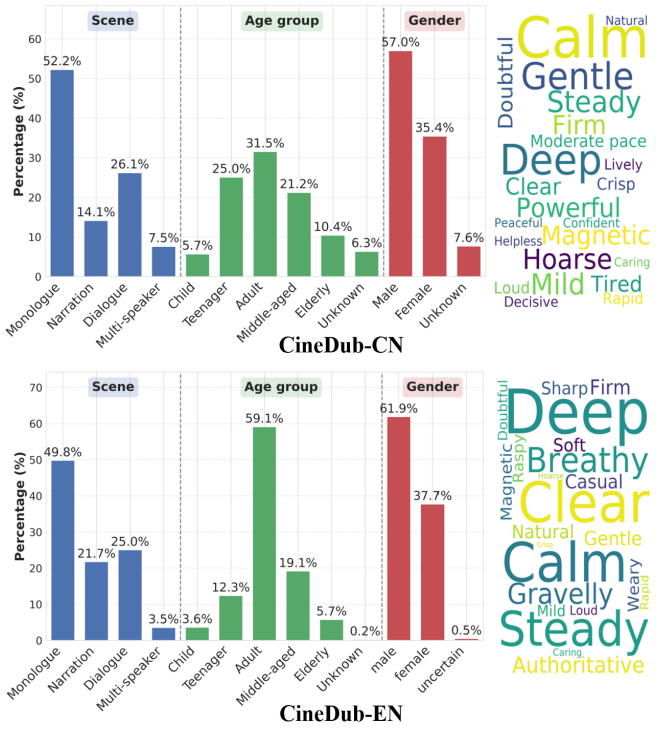

Figure 4: Statistics of the CineDub dataset. The top and bottom panels correspond to the Chinese and English subsets, respectively. Each panel contains: (left) the percentage distribution of scene categories, speaker age groups, and gender; (right) a word cloud of timbre-related keywords extracted from the audio descriptions.

lected over 400 raw television series covering diverse genres, including Chinese dramas, American dramas, and British dramas. Strict selection criteria are applied: non-documentary content, standard pronunciation, prominent vocal tracks, minimal colloquial speech, and unobstructed faces (i.e., no masks or veils). Using the pipeline, over 9,000 hours of effective speech clips are retained, with an average clip length of 10.8 seconds. Non-neutral emotional speech accounts for 38.4% of the dataset, and the average instruction length reaches 55 words with a variance of 16, covering character attributes, tone, speech rate, voice timbre, and emotion, indicating strong diversity and robust multi-scale annotations. Additional dataset statistics are provided in Fig. 4. We select four samples from each television series, corresponding to the four scene categories, to construct the test set. + Table 1 compares the intrinsic quality reported in the GT rows for the CineDub dataset (with and without Multimodal CoT Correction) and the combined V2C-Animation, Chem, and GRID datasets, as well as the performance of the FunCineForge dubbing model trained on these datasets. Results show that models trained on these three small-scale, limited-annotation datasets tend to underfit lip sync, leading to lower alignment accuracy, and produce speech with lower naturalness. In contrast, training on our CineDub dataset significantly improves pronunciation quality and lip sync. Moreover, training on the corrected CineDub dataset leads to clear improvements across all metrics compared with the uncorrected version, especially in pronunciation accuracy and alignment, demonstrating that

Methods	Scene	LSE-C↑	LSE-D↓	SBFE↓	SSIM(%)↑	ES-MOS↑
FunCineForge	Monologue	**8.92**	**3.60**	0.073	78.62	**4.04**
	Narration	-	-	**0.040**	**78.80**	3.89
	Dialogue	8.62	4.73	0.082	73.35	3.80
	Multi-Speaker	7.50	8.68	0.078	64.51	3.95
w/o $\mathcal{T}^N$	Monologue	8.18	6.85	0.132	70.87	3.85
	Narration	-	-	0.154	72.28	3.68
	Dialogue	6.30	10.12	0.285	66.24	3.71
	Multi-Speaker	5.33	14.44	0.389	62.86	3.82
w/o $\mathcal{L}_{Lip}$	Monologue	8.05	4.47	0.084	78.24	3.94
	Narration	-	-	0.059	78.67	3.81
	Dialogue	7.92	6.39	0.090	73.39	3.77
	Multi-Speaker	6.56	10.35	0.088	64.60	3.92
w/o SSC	Monologue	8.87	3.62	0.073	75.20	3.90
	Narration	-	-	0.042	74.92	3.78
	Dialogue	8.56	4.78	0.085	61.53	3.60
	Multi-Speaker	7.44	8.73	0.081	54.64	3.67

Table 3: Results of ablation study on the CineDub dataset.

the proposed bidirectional verification mechanism between lightweight specialized models and general-purpose MLLM significantly reduces word and punctuation errors in the training data, and alleviates duplicated and missing speaker identities in speaker diarization predictions, thereby substantially improving dataset quality.

4.4 Scene Analysis and Baseline Comparison

As shown in Table 1, we evaluate the performance of the FunCineForge dubbing model across four scene categories. The results indicate that our model achieves higher speech naturalness and alignment accuracy in monologue and narration scenes.

Specifically, in monologue scenes, alignment relies on temporal constraints and weakly supervised lip movements; in narration scenes, where no active facial frames are present, the model degrades to a time-controllable TTS model. In dialogue and multi-speaker scenes, which are inherently more challenging, the model maintains strong temporal alignment, accurate lip sync, and effective timbre switching. With longer facial sequences and richer instructions, it further exhibits more accurate emotional expression and improved instruction-following capability. However, as the duration of dubbed clips increases and scene complexity grows, temporal alignment and lip-sync performance degrade, and model robustness decreases.

Furthermore, to verify the effectiveness of FunCineForge dubbing model, as shown in Table 2, we conduct comparisons under a unified monologue setting, where all SOTA methods are evaluated on the same test samples. On three open-source dubbing datasets, our model achieves consistently superior performance in terms of speech quality, WER/CER, and alignment accuracy. In addition, we compare our model with two MLLM-based dubbing models trained on the CineDub dataset and evaluate them on the same monologue test set. The results demonstrate that FunCineForge dubbing model achieves significant improvements across all evaluation metrics, highlighting the effectiveness of the proposed model architecture and design.

4.5 Ablation Studies

To analyze the contribution of the key components in our dubbing model, we conduct ablation studies on the CineDub dataset. Results are summarized in Table 3.

IJCAI-ECAI 2026 PREPRINT

Effectiveness of timestamp-speaker tuples $\mathcal{T}^N$. Removing the timestamp-speaker tokenizer, the tuple inputs $\mathcal{T}^N$ and loss $\mathcal{L}_{VA}$, causes clear degradation in temporal alignment. Both SBFE and LSE-D increase substantially, especially in dialogue and multi-speaker scenes, where speaker truncation and leakage frequently occur. This shows that facial features alone are insufficient for long sequences with frequent shot changes and speaker switching, and explicit temporal speaker supervision is essential for complex cinematic scenes.

Effectiveness of lip contrastive loss $\mathcal{L}_{Lip}$. When removing $\mathcal{L}_{Lip}$ while keeping facial embeddings, LSE-C decreases and LSE-D increases, indicating weaker lip sync. Although facial features in the input visual modality provide coarse lip motion cues, the lack of frame-level contrastive supervision leads to reduced fine-grained alignment accuracy.

Effectiveness of speaker switching concatenation (SSC). We remove the SSC strategy and condition the flow matching on a global reference speaker embedding as in CosyVoice3. Speaker similarity drops slightly in monologue and narration scenes, but degrades significantly in dialogue and multi-speaker scenes, where the synthesized speech converges toward an averaged timbre. Nevertheless, subjective evaluation shows that speaker attributes such as gender and age remain distinguishable, likely due to the multi-task supervision in the CosyVoice3 speech tokenizer, which includes speaker age and gender prediction, leading to partial retention of speaker information in the speech tokens.

5 Conclusion

In this work, we propose an end-to-end dataset pipeline and an MLLM-based dubbing model designed for live-action cinematic dubbing. We construct the first television dubbing dataset CineDub with rich annotations. Multimodal CoT Correction improves dataset quality, while the four complementary modalities provide rich information for synthesizing expressive speech. The frame-index codebook and dedicated MLLM supervision enable accurate alignment in complex scenes, and the improved flow-matching design supports flexible speaker switching. In future work, we will explore the possibility of unifying speech, music, and audio generation through the understanding of visual information, detailed script content and captions.

Ethical Statement

All third-party tools and models used in this work are sourced from open-source projects and comply with their licenses. For the data involved in the CineDub dataset, we have obtained the necessary usage rights from third-party rights holders under formal licensing agreements. Consequently, we are only able to provide sample data from the dataset under the CC-BY-NC 4.0 license to ensure appropriate copyright protection. This work is conducted solely for academic research purposes, and all demos are provided only for demonstration purposes and contain no sensitive or infringing information. All human participants involved in subjective evaluations are compensated under employment agreements. The experiments and datasets pose no ethical issues.

Acknowledgments

This work was funded by the National Nature Science Foundation of China under Grant U23B2053.

References

[Bulat and Tzimiropoulos, 2017] Adrian Bulat and Georgios Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks). *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 1021–1030, 2017.

[Chen *et al.*, 2022] Qi Chen, Mingkui Tan, Yuankai Qi, Jiaqiu Zhou, Yuanqing Li, and Qi Wu. V2c: Visual voice cloning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21242–21251, June 2022.

[Chung and Zisserman, 2016] Joon Son Chung and Andrew Zisserman. Out of time: Automated lip sync in the wild. In Chu-Song Chen, Jiwen Lu, and Kai-Kuang Ma, editors, *Computer Vision - ACCV 2016 Workshops - ACCV 2016 International Workshops, Revised Selected Papers*, volume 10117, pages 251–263. Springer, 2016.

[Cong *et al.*, 2023] Gaoxiang Cong, Liang Li, Yuankai Qi, Zheng-Jun Zha, Qi Wu, Wenyu Wang, Bin Jiang, Ming-Hsuan Yang, and Qingming Huang. Learning to dub movies via hierarchical prosody models. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14687–14697, 2023.

[Cong *et al.*, 2024a] Gaoxiang Cong, Jiadong Pan, Liang Li, Yuankai Qi, Yuxin Peng, Anton van den Hengel, Jian Yang, and Qingming Huang. Emodubber: Towards high quality and emotion controllable movie dubbing. *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15863–15873, 2024.

[Cong *et al.*, 2024b] Gaoxiang Cong, Yuankai Qi, Liang Li, Amin Beheshti, Zhedong Zhang, Anton Hengel, Ming-Hsuan Yang, Chenggang Yan, and Qingming Huang. StyleDubber: Towards multi-scale style learning for movie dubbing. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 6767–6779, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[Cong *et al.*, 2025] Gaoxiang Cong, Liang Li, Jiadong Pan, Zhedong Zhang, Amin Beheshti, Anton van den Hengel, Yuankai Qi, and Qingming Huang. Flowdubber: Movie dubbing with llm-based semantic-aware learning and flow matching based voice enhancing. In *Proceedings of the 33rd ACM International Conference on Multimedia*, MM '25, pages 905–914, New York, NY, USA, 2025. Association for Computing Machinery.

[Cooke *et al.*, 2006] Martin Cooke, Jon Barker, Stuart Cunningham, and Xu Shao. An audio-visual corpus for speech perception and automatic speech recognition. *The Journal of the Acoustical Society of America*, pages 2421–2424, 2006.

IJCAI-ECAI 2026 PREPRINT

[Du *et al.*, 2024] Zhihao Du, Yuxuan Wang, Qian Chen, Xian Shi, Xiang Lv, Tianyu Zhao, Zhifu Gao, Yexin Yang, Changfeng Gao, Hui Wang, Fan Yu, Huadai Liu, Zhengyan Sheng, Yue Gu, Chong Deng, Wen Wang, Shiliang Zhang, Zhijie Yan, and Jing-Ru Zhou. CosyVoice 2: Scalable streaming speech synthesis with large language models. *arXiv preprint arXiv:2412.10117*, 2024.

[Du *et al.*, 2025] Zhihao Du, Changfeng Gao, Yuxuan Wang, Fan Yu, Tianyu Zhao, Hao Wang, Xiang Lv, Hui Wang, Chongjia Ni, Xian Shi, Keyu An, Guanrou Yang, Yabin Li, Yanni Chen, Zhifu Gao, Qian Chen, Yue Gu, Mengzhe Chen, Yafeng Chen, Shiliang Zhang, Wen Wang, and Jieping Ye. CosyVoice 3: Towards in-the-wild speech generation via scaling-up and post-training. *arXiv preprint arXiv:2505.17589*, 2025.

[Gao *et al.*, 2019] Shanghua Gao, Ming-Ming Cheng, Kai Zhao, Xinyu Zhang, Ming-Hsuan Yang, and Philip H. S. Torr. Res2net: A new multi-scale backbone architecture. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43:652–662, 2019.

[Gao *et al.*, 2023] Zhifu Gao, Zerui Li, Jiaming Wang, Haoneng Luo, Xian Shi, Mengzhe Chen, Yabin Li, Lingyun Zuo, Zhihao Du, Zhangyu Xiao, and Shiliang Zhang. Funasr: A fundamental end-to-end speech recognition toolkit, 2023.

[Hsu *et al.*, 2017] Wei-Ning Hsu, Yu Zhang, and James R. Glass. Unsupervised learning of disentangled and interpretable representations from sequential data. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 1878–1889, 2017.

[Hu *et al.*, 2021] Chenxu Hu, Qiao Tian, Tingle Li, Yuping Wang, Yuxuan Wang, and Hang Zhao. Neural dubber: dubbing for videos according to scripts. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, NIPS '21, Red Hook, NY, USA, 2021. Curran Associates Inc.

[Huang *et al.*, 2020] Y. Huang, Yuhan Wang, Ying Tai, Xiaoming Liu, Pengcheng Shen, Shaoxin Li, Jilin Li, and Feiyue Huang. Curricularface: Adaptive curriculum learning loss for deep face recognition. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5900–5909, 2020.

[Lee *et al.*, 2023] Jiyoung Lee, Joon Son Chung, and Soo-Whan Chung. Imaginary voice: Face-styled diffusion model for text-to-speech. *2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2023)*, pages 1–5, 2023.

[Li *et al.*, 2023] Guangxing Li, Wangjin Zhou, Sheng Li, Yi Zhao, Jichen Yang, and Hao Huang. Investigating effective domain adaptation method for speaker verification task. In *Neural Information Processing*, pages 517–527, Singapore, 2023. Springer Nature Singapore.

[Liu *et al.*, 2025] Rui Liu, Yuan Zhao, and Zhenqi Jia. Towards authentic movie dubbing with retrieve-augmented director-actor interaction learning, 2025.

[Mehta *et al.*, 2023] Shivam Mehta, Ruibo Tu, Jonas Beskow, Éva Székely, and Gustav Eje Henter. Matcha-tts: A fast tts architecture with conditional flow matching. *2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2024)*, pages 11341–11345, 2023.

[Peebles and Xie, 2022] William S. Peebles and Saining Xie. Scalable diffusion models with transformers. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4172–4182, 2022.

[Peng *et al.*, 2025] Zhiliang Peng, Jianwei Yu, Wenhui Wang, Yaoyao Chang, Yutao Sun, Li Dong, Yi Zhu, Weijiang Xu, Hangbo Bao, Zehua Wang, Shaohan Huang, Yan Xia, and Furu Wei. Vibevoice technical report. *arXiv preprint arXiv:2508.19205*, 2025.

[R *et al.*, 2020] Prajwal K R, Rudrabha Mukhopadhyay, Vinay P. Namboodiri, and C. V. Jawahar. Learning individual speaking styles for accurate lip to speech synthesis. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13793–13802, 2020.

[Shi *et al.*, 2026] Xian Shi, Xiong Wang, Zhifang Guo, Yongqi Wang, Pei Zhang, Xinyu Zhang, Zishan Guo, Hongkun Hao, Yu Xi, Baosong Yang, Jin Xu, Jingren Zhou, and Junyang Lin. Qwen3-asr technical report, 2026.

[Wang *et al.*, 2023] Haibo Wang, Siqi Zheng, Yafeng Chen, Luyao Cheng, and Qian Chen. Cam++: A fast and efficient network for speaker verification using context-aware masking. *ArXiv*, abs/2303.00332, 2023.

[Wang *et al.*, 2024] Ju-Chiang Wang, Wei-Tsung Lu, and Jitong Chen. Mel-roformer for vocal separation and vocal melody transcription, 2024.

[Wang *et al.*, 2025] Xinsheng Wang, Mingqi Jiang, Ziyang Ma, Ziyu Zhang, Songxiang Liu, Linqin Li, Zheng Liang, Qixi Zheng, Rui Wang, Xiaoqin Feng, Weizhen Bian, Zhen Ye, Sitong Cheng, Ruibin Yuan, Zhixian Zhao, Xinfa Zhu, Jiahao Pan, Liumeng Xue, Pengcheng Zhu, Yunlin Chen, Zhifei Li, Xie Chen, Lei Xie, Yike Guo, and Wei Xue. Spark-TTS: An efficient LLM-based text-to-speech model with single-stream decoupled speech tokens. *arXiv preprint arXiv:2503.01710*, 2025.

[Xie *et al.*, 2025] Kun Xie, Feiyu Shen, Junjie Li, Fenglong Xie, Xu Tang, and Yao Hu. Fireredtts-2:towards long conversational speech generation for podcast and chatbot. *arXiv preprint arXiv:2509.02020*, 2025.

[Zhang *et al.*, 2024] Zhedong Zhang, Liang Li, Gaoxiang Cong, Haibing Yin, Yuhan Gao, Chenggang Yan, Anton van den Hengel, and Yuankai Qi. From speaker to dubber: Movie dubbing with prosody and duration consistency learning. In *Proceedings of the 32nd ACM International Conference on Multimedia, MM 2024, Melbourne, VIC, Australia, 28 October 2024 - 1 November 2024*, pages 7523–7532. ACM, 2024.

IJCAI-ECAI 2026 PREPRINT

[Zhang *et al.*, 2025a] Zhedong Zhang, Liang Li, Gaoxiang Cong, Chunshan Liu, Yuhan Gao, Xiaowan Wang, Tao Gu, and Yuankai Qi. Instructdubber: Instruction-based alignment for zero-shot movie dubbing, 2025.

[Zhang *et al.*, 2025b] Zhedong Zhang, Liang Li, Chenggang Yan, Chunshan Liu, Anton van den Hengel, and Yuankai Qi. Prosody-enhanced acoustic pre-training and acoustic-disentangled prosody adapting for movie dubbing. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 172–182, 2025.

[Zheng *et al.*, 2020] Siqi Zheng, Yun Lei, and Hongbin Suo. Phonetically-aware coupled network for short duration text-independent speaker verification. In *21st Annual Conference of the International Speech Communication Association, Interspeech 2020, Virtual Event, Shanghai, China, October 25-29, 2020*, pages 926–930. ISCA, 2020.

[Zheng *et al.*, 2023] Siqi Zheng, Luyao Cheng, Yafeng Chen, Hui Wang, and Qian Chen. 3d-speaker: A large-scale multi-device, multi-distance, and multi-dialect corpus for speech representation disentanglement, 2023.

[Zheng *et al.*, 2025] Junjie Zheng, Zihao Chen, Chaofan Ding, and Xinhan Di. Deepdubber-v1: Towards high quality and dialogue, narration, monologue adaptive movie dubbing via multi-modal chain-of-thoughts reasoning guidance. *ArXiv*, abs/2503.23660, 2025.

[Zhou *et al.*, 2025] Siyi Zhou, Yiquan Zhou, Yi He, Xun Zhou, Jinchao Wang, Wei Deng, and Jingchen Shu. Indextts2: A breakthrough in emotionally expressive and duration-controlled auto-regressive zero-shot text-to-speech. *arXiv preprint arXiv:2506.21619*, 2025.

[Zuo *et al.*, 2023] Lingyun Zuo, Keyu An, Shiliang Zhang, and Zhijie Yan. Advancing vad systems based on multi-task learning with improved model structures, 2023.

IJCAI-ECAI 2026 PREPRINT