CoSyncDiT: Cognitive Synchronous Diffusion Transformer for Movie Dubbing

Gaoxiang Cong[1,2], Liang Li[1*], Jiaxin Ye[3], Zhedong Zhang[4], Hongming Shan[3], Yuankai Qi[5], and Qingming Huang[2]

[1] Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China
[2] University of Chinese Academy of Sciences, Beijing, China
[3] Fudan University, Shanghai, China
[4] Hangzhou Dianzi University, Hangzhou, China
[5] Macquarie University, Sydney, Australia

Abstract. Movie dubbing aims to synthesize speech that preserves the vocal identity of a reference audio while synchronizing with the lip movements in a target video. Existing methods fail to achieve precise lip-sync and lack naturalness due to explicit alignment at the duration level. While implicit alignment solutions have emerged, they remain susceptible to interference from the reference audio, triggering pronunciation and lip-sync degradation in in-the-wild scenarios. In this paper, we propose a novel flow matching-based movie dubbing framework driven by the **Co**gnitive **Sync**hronous **Di**ffusion **T**ransformer (**CoSyncDiT**), inspired by the cognitive process of professional actors. This architecture progressively guides the noise-to-speech generative trajectory by executing acoustic style adapting, fine-grained visual calibrating, and time-aware context aligning. Furthermore, we design the Joint Semantic and Alignment Regularization (JSAR) mechanism to simultaneously constrain frame-level temporal consistency on the contextual outputs and semantic consistency on the flow hidden states, ensuring robust alignment. Extensive experiments on both standard benchmarks and challenging in-the-wild dubbing benchmarks demonstrate that our method achieves the state-of-the-art performance across multiple metrics. The code is available at https://github.com/GalaxyCong/CoSyncDiT.

Keywords: Movie dubbing · Flow matching · Visual voice cloning

1 Introduction

Movie Dubbing, also known as Visual Voice Cloning (V2C) [3], aims to generate a piece of speech for the character in silent videos based on the specified reference voice and textual scripts (as shown in Fig. 1(a)). It promises significant potential in real-world applications such as film post-production, media production, and personal speech AIGC. Compared to traditional video-to-speech tasks [7,18–20], movie dubbing presents significantly greater challenges. Unlike merely inferring

* Corresponding author.

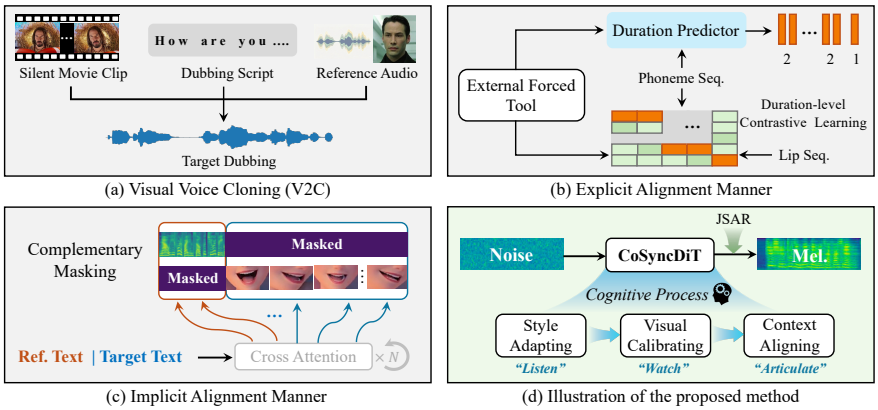

Fig. 1: (a) Illustration of Visual Voice Cloning (V2C) task. (b) Explicit alignment manner requires an external forced tool to compute the ground-truth duration of each phoneme in advance. (c) Implicit alignment eliminates the dependency on external forced tools. (d) We propose a movie dubbing architecture built upon CoSyncDiT, which is structured around a cognitively inspired listen–watch–articulate paradigm to progressively guide the denoising trajectory in flow matching.

speech from visual cues, V2C must faithfully clone the reference acoustic timbre and accurately articulate the textual scripts aligned with dynamic lip motions in the video, while ensuring emotionally expressive and high-fidelity speech quality.

Previous works mainly focus on improving pronunciation clarity and prosody modeling. For instance, Speaker2Dubber [51] introduces a two-stage architecture by pre-training on the clean TTS corpus to improve clarity. Then, ProDubber [53] employs a prosody-enhanced pre-training paradigm on the larger corpus. Recently, InstructDubber [52] integrates pre-trained TTS framework with LLM-driven emotion instructions by predicting pitch and energy to improve emotion expression [47]. However, these methods rely on TTS architecture and predict integer-scaled durations for each phoneme (as shown in Fig. 1(b)), failing to achieve the precise lip-sync required for dubbing. Some dubbing methods [10,12] attempt to introduce duration-level contrastive learning between phoneme and lip sequences to improve alignment. Nevertheless, these methods still depend on an external forced tool (*e.g.*, Montreal Forced Aligner [28]) to extract ground-truth duration boundaries for identifying positive samples during contrastive learning. This rigidly restricts the pronunciation to predefined intervals, damaging the naturalness and expressiveness of the generated speech [17].

Recently, AlignDiT [6] has represented a significant leap forward in the field by eliminating the need for external forced alignment tools and adopting an implicit alignment modeling. As shown in Fig. 1(c), it fuses text (both prefix reference text and target text) with complementary-masked audio (prefix reference) and visual features (target) by adding cross-attention layers in speech DiT [5]. However, this unified cross-attention paradigm makes inference-time alignment

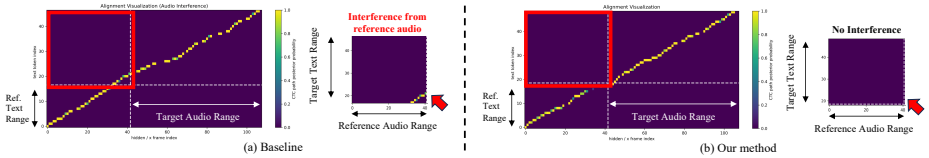

Fig. 2: Alignment comparison between the baseline and our method. The x-axis denotes the DiT hidden frame index, and the y-axis denotes the text-token index. White dashed lines mark the prefix reference and target boundaries for x-axis and y-axis, respectively. Ideally, the alignment path should pass through the dashed intersection. The red box shows that reference-audio regions are associated with target-text tokens in the baseline (left), while our method (right) alleviates this alignment interference.

more difficult, as prefix reference and target visual features must be associated with their correct text tokens. As shown in Fig. 2, this unified paradigm is prone to interference from prefix reference audio. This issue becomes more severe in in-the-wild scenarios, where the prefix reference audio may come from arbitrary speakers. Once the alignment fails, the model may produce pronunciation errors. Moreover, applying this unified cross-attention paradigm in all DiT layers may further increase the alignment difficulty and add computational cost.

In this paper, we propose a novel flow matching-based movie dubbing framework built upon **Co**gnitive **Sync**hronous **Di**ffusion **T**ransformer (**CoSyncDiT**), which progressively guides a noise-to-speech generative trajectory to effectively resolve the demands of high-fidelity style cloning and exact audio-visual synchronization. Inspired by the dubbing workflow of professional actors [1], our method is structured around a sequential cognitive process of listening, watching, and articulating. Rather than treat all transformer layers homogeneously, CoSyncDiT partitions the denoising generation within flow matching into three distinct phases (as shown in Fig. 1(d)). First, we initialize a unified acoustic-semantic prior. Then, the model leverages multi-head self-attention to capture the acoustic style and establish its relationship with the text, without any visual interference. Second, we inject fine-grained visual information through residual connections using a zero-initialized learnable gate. This step calibrates the previously acquired acoustic representations to capture temporal dynamics. Third, we design a time-aware context aligning layer to yield correct articulation by implicitly retrieving the linguistic context based on second phase. Finally, we introduce a Joint Semantic and Alignment Regularization (JSAR) mechanism to further rectify the third phase. This approach innovatively applies frame-level contrastive learning to the queried contextual outputs to enforce strict temporal alignment, while ensuring semantic constraints on the DiT hidden states.

Our key contributions are: **(1)** We propose CoSyncDiT, a novel flow matching-based framework for movie dubbing. By formulating the denoising generation as a cognitively synchronous process, it progressively guides the vector field estimation through three sequential phases: acoustic style adapting, fine-grained visual calibrating, and time-aware context aligning. **(2)** We design the

JSAR mechanism to enforce frame-level temporal consistency on intermediate context outputs and semantic consistency on final flow hidden states, thereby stabilizing the generative trajectory and enhancing context alignment. **(3)** Our method performs favourably against state-of-the-art methods on the challenging datasets, including the in-the-wild movie dubbing scenarios.

2 Related Work

2.1 Visual Voice cloning

Visual Voice Cloning (V2C) [3, 25, 30, 54] has recently attracted substantial attention in multimedia community [31,37,55], as it integrates information across text, speech [48], and visual modalities [14,40,41,56]. By bridging these modalities, V2C shows strong potential for applications in film production [24,38,46,57] and digital media [8, 36, 45]. Early methods widely adopt explicit duration prediction to achieve lip synchronization. For instance, StyleDubber [13] queries lip motions using textual phonemes to construct a phoneme-level visual sequence and predicts the duration of each phoneme. ProDubber [53] and InstructDubber [52] follow a similar paradigm. By enforcing explicit duration boundaries, these approaches typically guarantee high pronunciation clarity. However, they inherently struggle to achieve precise lip synchronization because the predicted durations must be expanded into rigid integer multiples. Recent methods [10,12] introduce duration-level contrastive learning to alleviate this limitation. Nevertheless, they still rely on external forced aligners to extract duration intervals. In this paper, drawing inspiration from the human cognitive workflow, we propose CoSyncDiT, a novel implicit dubbing framework that completely eliminates the dependency on external aligners. It ensures high-fidelity timbre cloning and accurate lip synchronization in challenging in-the-wild dubbing scenarios.

2.2 Flow Matching and Generative Modeling

Flow matching [23] has emerged as a dominant paradigm in generative modeling [2] due to its superior quality and inference speed. It learns a vector field [39] to transport samples from a simple prior distribution to the target data distribution along efficient probability paths. In TTS field, flow matching has found extensive applications [5, 15, 21, 29, 44, 50]. However, these methods cannot parse visual scenes and fail to achieve fine-grained lip-sync, hindering their application in movie dubbing. FlowDubber [10] introduces a dual contrastive learning, which relies on external forced aligners and employs a monotonic search algorithm for sequential expansion. Nevertheless, even minor initial misalignments tend to compound throughout the generation process and degrade the lip-sync quality. AlignDiT [6] implicitly models this alignment by incorporating cross-attention modules across all transformer layers. However, due to its complementary masking strategy, the model struggles to simultaneously align with the reference audio and the target lip motions. This homogeneous injection across all layers further

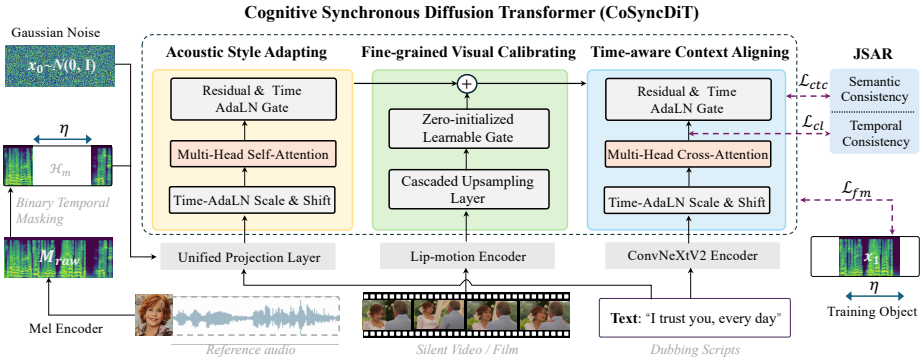

Fig. 3: The overview of architecture

exacerbates alignment instability in complex dubbing scenarios. In this paper, we propose CoSyncDiT to progressively guide the denoising trajectory by executing acoustic style adapting, fine-grained visual calibrating, and time-aware context aligning. Furthermore, we design the JSAR mechanism to learn temporal consistency and semantic consistency.

3 Methodology

As shown in Fig. 3, the overall framework of our method includes the following parts. First, individual encoders extract the foundational representations from the reference audio, silent video, and dubbing scripts. Next, under the Optimal-Transport Conditional Flow Matching (OT-CFM) paradigm, the proposed CoSyncDiT first takes Gaussian noise, masked acoustic conditions, and semantic conditions as priors to execute Acoustic Style Adapting. Sequentially, the Fine-grained Visual Calibrating focuses on capturing the rhythmic dynamics of lip motion. Next, a Time-aware Context Aligning module integrates the textual features to ensure precise audio-visual synchronization. Finally, the designed JSAR mechanism regularizes the alignment stage through semantic and temporal consistency constraints, guiding the generative trajectory toward synchronized and high-fidelity speech.

3.1 Acoustic and Semantic Prior

The reference audio R_a is extracted to a raw mel-spectrogram $m_{raw} \in \mathbb{R}^{F \times L}$, where F is the mel dimension and L is the sequence length. A binary temporal mask $\mathbf{M}_b \in \{0,1\}^L$ is then applied to obscure the target region η. This enables the model to yield the masked acoustic features $\mathcal{H}_m = (1 - \mathbf{M}_b) \odot m_{raw}$. Unlike directly concatenating the masked visual features, we construct a unified sequence by jointly modeling text and $\mathcal{H}_m$. To expand the word-level text sequence to the mel-level, we adopt two expansion strategies: (1) a padding operation to preserve

the complete linguistic content; (2) a cross-attention operation to provide coarse temporal priors. Finally, we concatenate these textual sequences and $\mathcal{H}_m$ into the semantic-acoustic prior, rather than using raw visual signals for early-stage conditioning.

3.2 CoSyncDiT

Traditional Diffusion Transformers (DiTs) struggle to meet the strict demands of movie dubbing. They fail to maintain precise audio-visual synchronization, degrading both timbre fidelity and pronunciation clarity. Drawing inspiration from the human dubbing cognitive process, we propose the Cognitive Synchronous Diffusion Transformer (CoSyncDiT). The model learns to progressively establish the reference speaker's style, perform fine-grained lip calibration, and enforce contextual alignment.

Stage 1: Acoustic Style Adapting. In the scope of OT-CFM, our model predicts the vector field mapping from a standard Gaussian noise distribution $x_0 \sim \mathcal{N}(0, I)$ to the target speech latent x_1, following the time-dependent path $x_t = (1-t)x_0 + tx_1$. In stage 1, the model concatenates the noisy speech x_t and the semantic-acoustic prior by unified projection layer to form the initial sequence $\mathcal{Z}^0$. The Multi-Head Self-Attention (MHSA) models the long-range dependencies in the initial hidden states to capture style information:

$$\mathcal{Z}^l_{style} = \mathcal{Z}^{l-1} + \alpha^l_1(t) \odot \mathrm{MHSA}(\mathrm{AdaLN}_{\beta_1,\gamma_1}(\mathcal{Z}^{l-1}, t)), \tag{1}$$

where l represents the current block level. $\mathrm{AdaLN}_{\beta_1,\gamma_1}$ applies Time-Adaptive Layer Normalization (Time-AdaLN) to modulate the input feature statistics via scale γ_1 and shift β_1 vectors driven by the current time t. The gating term $\alpha_1(t)$ controls the residual contribution. Similarly, a time-adaptive Multilayer Perceptron (MLP) further refines these features via non-linear dimensional transformations to consolidate the relationship between acquired acoustic style and text.

Stage 2: Fine-grained Visual Calibrating. The input silent video is first processed by a lip-motion encoder to extract the raw lip features $\mathcal{X}_{raw}$. Then, $\mathcal{X}_{raw}$ is then passed through a cascaded upsampling layer to yield the refined visual representations $\mathcal{X}_{lip}$, ensuring their temporal resolution strictly matches that of the target mel-spectrogram. Next, $\mathcal{X}_{lip}$ are seamlessly injected into the $\mathcal{Z}^l_{style}$ via a zero-initialized learnable gate:

$$\mathcal{Z}^l_{lip} = \mathcal{Z}^l_{style} + \Lambda^l \odot \mathcal{X}_{lip}, \quad \text{where } \Lambda^l = \mathbf{0} \in \mathbb{R}^d. \tag{2}$$

The zero-initialized learnable gate Λ^l is conditioned on the time step t at each layer, ensuring that visual information is introduced as a subtle rhythm-aware residual for frame-level calibration while preserving the style information.

Stage 3: Time-aware Context Aligning. We place the Multi-Head Cross-Attention (MHCA) at the bottom of the architecture to ensure that text alignment operates on fully mature visual-acoustic representations, rather than forcing a premature and unstable complementary fusion. Furthermore, we introduce

a time-aware mechanism, $\text{AdaLN}_{\beta_3,\gamma_3}$, to modulate the cross-attention process, allowing the alignment dynamics to better adapt to the evolving flow states:

$$\mathcal{Z}^l_{out} = \mathcal{Z}^l_{lip} + \alpha^l_3(t) \odot (\text{MHCA}(\text{AdaLN}_{\beta_3,\gamma_3}(\mathcal{Z}^l_{lip}, t), \mathcal{H}_{text})), \tag{3}$$

where $\alpha^l_3(t)$ controls the residual contribution of the cross-attention output according to the current timestep t. The textual representation $\mathcal{H}_{text}$ extracted by ConvNeXtV2 encoder [43] serves as both the Key and Value, encouraging the generative path to converge toward the aligned content by implicitly retrieving the linguistic context.

3.3 Joint Semantic and Alignment Regularization

While the OT-CFM efficiently constructs the data generation path, the unconstrained vector field estimation often leads to temporal misalignments. In this work, we introduce the Joint Semantic and Alignment Regularization (JSAR) mechanism, constraining both the final flow hidden states and the intermediate context representations.

Alignment Regularization. The intermediate context representations (*i.e.*, the output of the MHCA in Eq. (3), denoted as $\hat{\mathcal{Z}}_{ca}$) need to maintain inherent temporal consistency when queried by the flow hidden features. To implicitly align these context representations in time, we introduce a frame-level contrastive learning by employing the audio branch representations from a pre-trained AV-HuBERT (denoted as $\mathcal{F}_{av}$) as the contrastive ground truth. Both the outputs $\hat{\mathcal{Z}}_{ca}$ and the AV-HuBERT features $\mathcal{F}_{av}$ are then L2-normalized along the feature dimension. Then, it uses the InfoNCE objective to maximize the cosine similarity of matching temporal frames while repelling non-matching frames:

$$\mathcal{L}_{CL} = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{\exp\left(\langle \hat{z}^{(i)}_{ca}, f^{(i)}_{av} \rangle / \tau\right)}{\sum_{j=1}^{N} \exp\left(\langle \hat{z}^{(i)}_{ca}, f^{(j)}_{av} \rangle / \tau\right)}, \tag{4}$$

where $\hat{z}^{(i)}_{ca}$ and $f^{(i)}_{av}$ represent the i-th frame of the flattened $\hat{\mathcal{Z}}_{ca}$ and $\mathcal{F}_{av}$ respectively, $\langle \cdot, \cdot \rangle$ denotes the dot product, and τ is the temperature hyperparameter.

Semantic Regularization. To guarantee pronunciation correctness, we apply a Connectionist Temporal Classification (CTC) $\mathcal{L}_{ctc}$ loss directly to the final hidden features $\mathcal{Z}^l_{out}$ (*i.e.*, the output of Eq. (3)). It encourages the model to retain more linguistic information in predicted hidden states. By jointly optimizing these objectives within the JSAR framework, the model effectively synchronizes the context with the acoustic dynamics and maintains semantic consistency, stably anchoring the flow matching trajectory.

3.4 Optimal-Transport Conditional Flow Matching Objective

The objective of the OT-CFM is to minimize the Mean Squared Error (MSE) between the predicted vector field by CoSyncDiT and the target vector field:

$$\mathcal{L}_{fm} = \mathbb{E}_{t,q(x_1),p(x_0)}[||v_\theta(x_t|t, \mathcal{H}_m, \mathcal{X}_{lip}, \mathcal{H}_{text}) - (x_1 - x_0)||^2], \tag{5}$$

where $t \sim \mathcal{U}[0,1]$ is the timestep, and $x_1 \sim q(x_1)$ denotes the target mel-spectrogram latent. x_t represents the intermediate noise latent. v_θ represents the vector field predicted by the proposed CoSyncDiT with model parameters θ, which is progressively conditioned on the $\mathcal{H}_m$, $\mathcal{X}_{lip}$, and $\mathcal{H}_{text}$.

3.5 Acoustic-Semantic Classifier-Free Guidance

In conditional flow matching, classifier-free guidance (CFG) [16] is typically employed to amplify the influence of the conditioning signals. Driven by acoustic and semantic prior, our CFG focuses on explicitly decoupling these conditions. Let $\mathcal{C}$ denote the joint condition comprising both acoustic and semantic, and $\varnothing$ denote the unconditional prior. The modified vector field is formulated as:

$$v_{t,cfg} = v_t(x_t, \mathcal{C}) + \lambda_a \cdot (v_t(x_t, \mathcal{C}) - v_t(x_t, \mathcal{H}_m)) + \lambda_s \cdot (v_t(x_t, \mathcal{H}_m) - v_t(x_t, \varnothing)), \quad (6)$$

where λ_a and λ_s are the acoustic and semantic guidance scales, enabling a highly controllable dubbing generation.

4 Experiments

4.1 Datasets and Experimental Setup

We conduct experiments on the challenge dubbing datasets, encompassing diverse in-the-wild scenarios (*e.g.*, live action movies, vlogs, and dramas) as well as traditional scenarios. Tab. 1 summarizes the evaluation settings: (1) **Setting 1**: ground-truth speech is used as reference; (2) **Setting 2**: a different clip from the same speaker serves as reference; (3) **Zero-shot (Setting 1&2)**: both unseen voice and unseen video are evaluated by out-of-domain dubbing dataset. The unseen voice can essentially still be categorized into Setting 1 and Setting 2.

Table 1: Overview of evaluation configurations.

Dataset	Speaker	Main Scenes	Experiment Setting
CelebV-Dub [36]	Multi-Speakers	Vlogs, Dramas	Setting 1 & 2
CinePile-Dub [33]	Multi-Speakers	Movies	Zero-shot (Setting 1 & 2)

CelebV-Dub. Built upon the foundations of CelebV-HQ [58] and CelebV-Text [49], this dataset aggregates content from diverse sources such as vlogs and dramas. It serves as a distinctly challenging benchmark by featuring unconstrained real-world settings alongside rich emotional variations. Following the official division [36], this dataset provides 79,933 training samples and 213 testing samples.

CinePile-Dub. Derived from the original CinePile [33] dataset, CinePile-Dub is specifically curated from professional live-action movie productions. It features high emotional intensity and expressive acting-specific prosody characteristic of authentic cinematic environments.

4.2 Evaluation Metrics

We adopt several authority metrics to comprehensively evaluate various aspects of generated dubbing.

WER measures pronunciation clarity by calculating the word error rate derived from an automatic speech recognition model, whisper-large-V3 [32].

EMOSIM measures emotion similarity. We compute the cosine similarity between the synthesized speech and the ground truth by the Emotion2Vec model [27], which is a universal speech emotion representation model.

SPKSIM measures speaker similarity. In contrast to the prior metric in [22], *i.e.*, speaker encoder cosine similarity (SECS) by GE2E model [42], we adopt the WavLM-TDNN model [4], which is widely adopted in speaker verification to ensure a reliable measure [10, 36].

DNSMOS is used to objectively evaluate speech quality by approximating subjective human ratings (Mean Opinion Score, MOS). DNSMOS [34] (Deep Noise Suppression MOS) assesses speech clarity and background noise cleanliness.

MOS-N and MOS-S are subjective evaluation metrics by human ratings. MOS-N measures the naturalness, while MOS-S evaluates the speaker similarity.

Sync-KL also known as duration divergence [52]. Unlike vision-based synchronization metrics (*e.g.*, AVSync [6] and LSE [9]), which can be affected by complex visual variations and perturbations in in-the-wild scenarios, Sync-KL exploits the natural synchronization between the ground-truth speech and video. Please note that this metric remains meaningful under the condition that the generated speech has a duration close to that of the ground-truth speech. When the two total durations differ substantially, Sync-KL mainly reflects relative differences rather than audio-visual synchronization.

LSE-C and LSE-D are vision-based synchronization metrics derived from the SyncNet [9]. We additionally report these metrics to provide a more comprehensive evaluation of audio-visual alignment.

4.3 Implementation Details

The input and output dimensions of the unified projection layer are 712 and 1,024, respectively. ConvPosition is used to inject positional information into the unified projection layer with a kernel size of 31 and 16 groups. CoSyncDiT consists of 22 layers. The multi-head self-attention and multi-head cross-attention layers both have a hidden dimension of 1,024 with 16 attention heads. Lip regions are resized to 96×96 and processed by a pre-trained AV-HuBERT [35] model to extract 1,024-dimensional embeddings. The text encoder comprises a stack of four ConvNeXt V2 blocks with a hidden dimension of 512. In JSAR, the temperature parameter τ is set to 0.07. The CTC projection layer within JSAR includes two temporal downsampling layers with Mish activations to map the 1,024-dimensional features into a 2,547-dimensional space. The final projection layer maps the 1,024-dimensional features to a predicted 100-dimensional vector field. During inference, the predicted vector field is used to solve the ODE

Table 2: Compared with SOTA Dubbing methods on the CelebV-Dub dataset under Setting 1. CelebV-Dub includes diverse video content (vlogs and dramas) with highly expressive speech.

Methods	SPKSIM(%)↑	WER(%)↓	EMOSIM(%)↑	Sync-KL↓	DNSMOS↑
GT	100.00	4.15	100.00	0.000	3.38
HPMDubbing [11] (CVPR'23)	17.26	21.99	73.01	0.441	2.47
StyleDubber [13](ACL'24)	26.39	10.79	79.38	0.415	2.65
EmoDubber [12] (CVPR'25)	22.08	13.44	76.59	0.407	3.26
FlowDubber [10] (MM'25)	22.24	13.95	76.76	0.423	3.27
Produbber [53] (CVPR'25)	26.97	5.18	73.49	0.421	3.12
HD-Dubber [22] (TPAMI'25)	10.96	19.73	66.30	0.410	3.01
AlignDiT [6] (MM'25)	59.71	9.48	84.54	0.402	3.45
InstructDub [52] (AAAI'26)	29.12	6.13	75.57	0.429	3.16
Ours	**65.21**	**4.29**	**84.61**	**0.392**	**3.46**

Table 3: Compared with SOTA methods on the CelebV-Dub dataset under Setting 2.

Methods	SPKSIM(%)↑	WER(%)↓	EMOSIM(%)↑	Sync-KL↓	DNSMOS↑
GT	66.13	4.15	100.00	0.000	3.38
HPMDubbing [11] (CVPR'23)	12.52	23.64	69.25	0.447	2.47
StyleDubber [13] (ACL'24)	19.42	7.03	74.87	0.405	2.63
EmoDubber [12] (CVPR'25)	18.78	16.37	76.30	0.422	3.30
FlowDubber [10] (MM'25)	19.32	14.94	74.18	0.420	3.31
Produbber [53] (CVPR'25)	23.13	6.41	71.38	0.432	3.14
HD-Dubber [22](TPAMI'25)	9.69	16.86	64.71	0.400	3.00
AlignDiT [6] (MM'25)	49.49	13.18	79.70	0.413	**3.47**
InstructDub [52] (AAAI'26)	22.85	**5.64**	74.03	0.434	3.18
Ours	**53.44**	6.39	**80.29**	**0.381**	**3.47**

from Gaussian noise to the target mel-spectrogram using an Euler solver with 32 function evaluations. We optimize the model using the AdamW optimizer [26] with momentum parameters $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The decoupled weight decay is set to 0.01, and the epsilon term is 1×10^{-8}. A random 70% to 100% span of mel-spectrogram frames is masked with span length η.

4.4 Performance Comparison with SOTA Methods

Results on the CelebV-Dub (Setting 1). Results are reported in Tab. 2. Compared with relatively controlled datasets like Chem, CelebV-Dub contains diverse in-the-wild content such as vlogs and dramas with rich emotional variations, posing a distinctly more challenging evaluation. Despite these complexities, our proposed CoSyncDiT successfully achieves state-of-the-art performance across all evaluated metrics. Specifically, it attains 65.21% in SPKSIM, outperforming the strongest baseline AlignDiT by an absolute margin of 5.50%, which confirms the reliability of our acoustic style adapting phase in handling highly

Table 4: Zero-shot main experiment results on CinePile-Dub dataset (Out-of-Domain) under setting 1. All models are trained on the training set of the basic dubbing dataset CelebV-Dub.

Methods	SPKSIM(%)↑	WER(%)↓	EMOSIM(%)↑	Sync-KL↓	DNSMOS↑
GT	100.00	2.56	100.00	0.00	3.15
HPMDubbing [11](CVPR'23)	42.72	99.20	61.22	0.391	3.23
StyleDubber [13](ACL'24)	41.65	74.96	63.88	0.362	2.89
HD-Dubber [22](TPAMI'25)	5.84	42.75	56.32	0.337	2.58
EmoDubber [12](CVPR'25)	23.10	49.57	70.08	0.337	3.05
FlowDubber [10](MM'25)	22.68	54.41	67.39	0.354	3.04
AlignDiT [6](MM'25)	58.90	20.98	77.39	0.342	3.36
ProDubber [53](CVPR'25)	28.86	5.25	70.35	0.335	2.93
InstructDub [52](AAAI'26)	27.61	**4.61**	65.97	0.370	2.95
Ours	**60.04**	5.59	**77.41**	**0.332**	**3.40**

Table 5: Zero-shot main experiment results on CinePile-Dub dataset (Out-of-Domain) under setting 2. Building upon the unseen scenario, the reference audio of this setting is derived from other clips of the same speaker (*i.e.*, Zero-shot+Setting 2).

Methods	SPKSIM(%)↑	WER(%)↓	EMOSIM(%)↑	Sync-KL↓	DNSMOS↑
GT	60.58	2.56	100.00	0.00	3.15
HPMDubbing [11](CVPR'23)	30.75	93.86	52.93	0.382	3.24
StyleDubber [13](ACL'24)	28.67	77.61	57.15	0.372	2.88
HD-Dubber [22](TPAMI'25)	7.04	64.12	47.66	0.330	2.58
EmoDubber [12](CVPR'25)	19.43	57.73	62.46	0.347	3.05
FlowDubber [10](MM'25)	19.86	55.93	63.86	0.346	3.05
AlignDiT [6](MM'25)	43.82	23.92	**72.53**	0.355	3.33
ProDubber [53](CVPR'25)	23.91	8.02	62.06	0.338	2.93
InstructDub [52](AAAI'26)	23.11	7.62	61.14	0.367	2.94
Ours	**47.24**	**7.53**	70.12	**0.319**	**3.39**

expressive speech. Furthermore, our method obtains the highest EMOSIM of 84.61% alongside an exceptionally low WER of 4.29%, closely approaching the Ground Truth limit of 4.15%.

Results on the CelebV-Dub (Setting 2). Finally, we evaluate the models on the CelebV-Dub dataset under Setting 2. As shown in Tab. 3, our CoSyncDiT achieves the highest SPKSIM of 53.44% and EMOSIM of 80.29%, demonstrating robust acoustic style adapting capabilities for complex unseen voices. While InstructDub attains a marginally lower WER of 5.64%, it severely compromises speaker identity preservation by yielding a remarkably low SPKSIM of 22.85%. In contrast, our method maintains a highly competitive WER of 6.39%. Furthermore, achieving the highest DNSMOS score of 3.47 confirms that our dubbing method avoids the pitfall of single-metric over-optimization, providing the most comprehensive and balanced dubbing quality.

Table 6: Results of ablation study on CelebV-Dub dataset under Setting 2.

Methods	SPKSIM (%) ↑	WER (%) ↓	EMOSIM (%) ↑	Sync-KL ↓	DNSMOS ↑
w/o Style Adapting	19.64	6.84	77.24	0.385	3.38
w/o Visual Calibrating	53.25	6.40	80.17	0.419	3.45
w/o Context Aligning	52.75	7.39	80.04	0.446	3.44
w/o JSAR	51.30	8.72	80.14	0.431	3.39
w/o Semantic Consistency	52.34	8.39	80.19	0.392	3.42
w/o Temporal Consistency	51.37	6.58	80.24	0.425	3.44
Full model	**53.44**	**6.39**	**80.29**	**0.381**	**3.47**

Results on the CinePile-Dub (Zero-shot Setting 1). To assess true zero-shot generalization, we conduct an out-of-domain evaluation where models trained exclusively on CelebV-Dub are tested on CinePile-Dub, a highly challenging real-world live-action movie dataset. As presented in Tab. 4, our proposed CoSyncDiT achieves the highest SPKSIM of 60.04% and EMOSIM of 77.41%. This confirms the powerful cross-domain extrapolation capacity of the proposed method when confronted with entirely novel cinematic environments. Although InstructDub attains the lowest WER of 4.61%, it experiences a catastrophic collapse in speaker identity preservation by yielding a poor SPKSIM of 27.61%. Conversely, our method maintains a highly competitive WER of 5.59% while simultaneously establishing the best speaker similarity of 60.04%.

Results on the CinePile-Dub (Zero-shot Setting 2). As reported in Tab. 5, the performance of all evaluated methods experiences a noticeable decline under this highly constrained configuration. Despite these extreme challenges, our proposed method consistently secures the best performance across almost objective metrics. Specifically, our approach achieves the highest speaker similarity of 47.24%. This demonstrates its capability to reliably extract and transfer target vocal characteristics from temporally unaligned reference segments. Furthermore, our model attains the lowest WER of 7.53%. These comprehensive results compellingly confirm the exceptional robustness of our CoSyncDiT. It proves that our progressive cognitive synchronous architecture can reliably maintain high-fidelity speech generation and audio-visual alignment in challenging scenarios.

4.5 Ablation Study

To validate the effectiveness of our core designs, we conduct a comprehensive ablation study on the CelebV-Dub dataset under Setting 2. As reported in Tab. 6, removing the acoustic style adapting phase triggers a severe collapse in speaker similarity, plummeting the SPKSIM from 53.44% to 19.64%. This confirms its indispensable role in capturing and preserving the target timbre. When the fine-grained visual calibrating module is omitted, the Sync-KL metric clearly degrades to 0.419, indicating its absolute necessity for establishing strict lip-sync. Furthermore, we find that the absence of the time-aware context

aligning phase exerts the most profound impact on overall synchronization, as its core function relies on re-retrieving textual content conditioned on previously integrated information to adjust articulation. Finally, we dissect the JSAR mechanism. Eliminating the semantic consistency constraint primarily damages the pronunciation accuracy, dropping the temporal consistency constraint strictly deteriorates the audio-visual synchronization, and discarding the entire JSAR framework compounds both errors.

Table 7: More comparison with recent SOTA methods on two dubbing benchmarks.

Method	CelebV-Dub		CinePile-Dub	
	LSE-C↑	LSE-D↓	LSE-C↑	LSE-D↓
ProDubber (CVPR'25)	2.51	11.95	2.51	11.46
HD-Dubber (TPAMI'25)	3.34	11.08	2.08	11.35
AlignDiT (MM'25)	6.44	7.92	5.05	9.33
InstructDubber (AAAI'26)	2.49	11.88	2.64	11.40
Ours	**6.54**	**7.85**	**5.13**	**9.26**

4.6 Results on the Setting 2 by using LSE-C and LSE-D

Following the suggestion, we further report LSE-C and LSE-D on two dubbing benchmarks. To evaluate model robustness, we follow Setting 2 and compare our method with recent strong baselines. In Setting 2, the reference audio for each character is taken from another clip of the same character, avoiding the use of ground-truth speech as the reference prefix. As shown in Tab. 7, our method achieves the best scores on both benchmarks, with higher LSE-C and lower LSE-D than recent state-of-the-art methods. The gains are also consistent on CinePile-Dub, a more challenging benchmark built from in-the-wild movie clips. These results show that our method effectively improves truly audio-visual synchronization by designing cognitive synchronous diffusion transformer.

4.7 Qualitative Comparison

We visualize the mel-spectrograms of the ground truth and the generated speech from different models in Fig. 4. The blue dashed regions and arrows highlight critical temporal intervals for audio-visual synchronization, while the white dashed boxes emphasize fine-grained pronunciation differences. While most baseline methods perform adequately on the controlled Chem dataset, they degrade sharply on more challenging benchmarks. Specifically in vlog and cinematic scenarios, models like InstructDubber and ProDubber completely fail to maintain proper audio-visual sync. Furthermore, AlignDiT exhibits noticeable temporal shifts and little distortions within the arrow-indicated regions. Conversely,

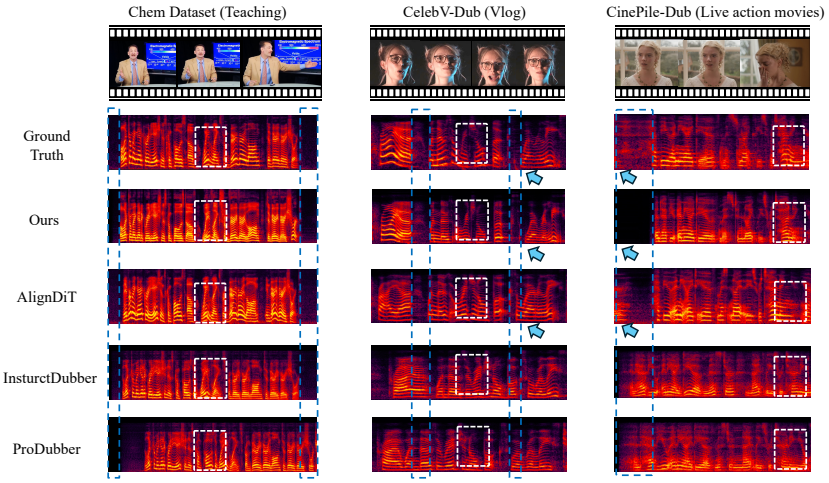

Fig. 4: Visual comparison of the generated mel-spectrograms. The blue arrows highlight specific regions requiring attention for audio-visual synchronization.

our proposed CoSyncDiT synthesizes highly robust mel-spectrograms that most closely match the ground truth across both boundaries and details.

4.8 Alignment Comparison between SOTA Method and CoSyncDiT

To visualize the alignment between DiT hidden frames and text tokens around the reference-target boundary, we use a CTC-based visualization. The CTC head provides an explicit monotonic alignment signal, making boundary leakage easier to inspect than in generic attention maps. Specifically, as shown in Fig. 2, we visualize the frame-level CTC emission probabilities along the non-blank Viterbi path. As mentioned in the introduction, the baseline based on the unified cross-attention paradigm shows clear interference in the red boxed region (Fig. 2(a)), where reference-side audio activations leak into the target-text range. This indicates that the reference audio, used as a prefix prior in the DiT input, can disturb the subsequent alignment process in unified cross-attention layers in DiT block. In contrast, Fig. 2(b) shows that CoSyncDiT produces a more stable boundary. This indicates the effectiveness of the our method by utilizing the cognitive synchronous DiT, alleviating the difficulty of aligning by cross-attention to some extent.

4.9 Details on the Human Evaluation

To comprehensively evaluate the zero-shot dubbing capabilities in real-world cinematic scenarios, we conduct a subjective Human Evaluation. We randomly select 50 video clips from the challenging CinePile-Dub test set. Raters use a 5-point Absolute Category Rating (ACR) scale from 1 (poor) to 5 (excellent),

in 0.5 increments. As reported in Tab. 8, our method achieves the highest scores across all human subjective evaluations. This demonstrates that human listeners perceive our generated speech as the most natural and acoustically similar to the target speaker. In addition to the human MOS results, we incorporate objective MOS metrics. These objective metrics are not scored by human annotators but rely on predictive models to mimic human auditory perception. DNSMOS evaluates speech clarity, and UTMOS assesses the overall speech quality. Consistent with the subjective feedback, our method still achieves the optimal results across these objective metrics in the zero-shot setting. This comprehensive assessment thoroughly validates the high fidelity and robust synthesis quality of our proposed architecture.

Table 8: Zero-shot MOS results on CinePile-Dub dataset (Out-of-Domain). The human subjective MOS metrics comprise MOS-N (MOS-Naturalness) and MOS-S (MOS-Similarity). The objective MOS metrics include DNSMOS and UTMOS.

Methods	DNSMOS ↑	UTMOS ↑	MOS-N ↑	MOS-S ↑
GT	3.15	2.70	4.72 ± 0.08	-
ProDubber [53]	2.93	2.92	3.89 ± 0.14	3.56 ± 0.09
HD-Dubber [22]	2.81	2.12	3.14 ± 0.18	2.95 ± 0.09
AlignDiT [6]	3.36	3.05	4.02 ± 0.11	3.97 ± 0.12
InstructDubber [52]	2.95	3.08	3.98 ± 0.12	3.53 ± 0.14
Ours	**3.40**	**3.24**	**4.04 ± 0.09**	**4.01 ± 0.10**

5 Conclusion

In this paper, we propose CoSyncDiT, a flow matching-based framework for movie dubbing, which sequentially executes acoustic style adapting, fine-grained visual calibrating, and time-aware context aligning to explicitly guide the noise-to-speech generative trajectory. Furthermore, we introduce the Joint Semantic and Alignment Regularization (JSAR) mechanism to simultaneously enforce frame-level temporal consistency on the contextual outputs and semantic consistency on the flow hidden states, guaranteeing reliable pronunciation during the alignment phase. Extensive experiments on standard benchmarks and in-the-wild datasets demonstrate that our method achieves state-of-the-art performance.

Acknowledgements

This work was supported by the National Nature Science Foundation of China (62322211), the "Pioneer" and "Leading Goose" R&D Program of Zhejiang Province (2024C01023). Yuankai Qi is not supported by the aforementioned fundings.

References

1. Alburger, J.R.: The art of voice acting: The craft and business of performing for voiceover. Focal Press (2023)
2. Cao, B., Ye, J., Wei, Y., Shan, H.: RepLDM: Reprogramming pretrained latent diffusion models for high-quality, high-efficiency, high-resolution image generation. In: NeurIPS (2026)
3. Chen, Q., Tan, M., Qi, Y., Zhou, J., Li, Y., Wu, Q.: V2C: visual voice cloning. In: CVPR. pp. 21210–21219 (2022)
4. Chen, S., Wang, C., Chen, Z., Wu, Y., Liu, S., Chen, Z., Li, J., Kanda, N., Yoshioka, T., Xiao, X., Wu, J., Zhou, L., Ren, S., Qian, Y., Qian, Y., Wu, J., Zeng, M., Yu, X., Wei, F.: Wavlm: Large-scale self-supervised pre-training for full stack speech processing. IEEE J. Sel. Top. Signal Process. **16**(6), 1505–1518 (2022)
5. Chen, Y., Niu, Z., Ma, Z., Deng, K., Wang, C., Zhao, J., Yu, K., Chen, X.: F5-tts: A fairytaler that fakes fluent and faithful speech with flow matching. arXiv preprint arXiv:2410.06885 (2024)
6. Choi, J., Kim, J.H., Sung-Bin, K., Oh, T.H., Chung, J.S.: Alignedit: Multimodal aligned diffusion transformer for synchronized speech generation. In: ACM MM. p. 10758–10767 (2025)
7. Choi, J., Kim, M., Ro, Y.M.: Intelligible lip-to-speech synthesis with speech units. In: Annu. Conf. Int. Speech Commun. Assoc. pp. 4349–4353 (2023)
8. Choi, J., Park, S.J., Kim, M., Ro, Y.M.: Av2av: Direct audio-visual speech to audio-visual speech translation with unified audio-visual speech representation. In: CVPR. pp. 27325–27337 (2024)
9. Chung, J.S., Zisserman, A.: Out of time: Automated lip sync in the wild. In: ACCV 2016 Workshops. pp. 251–263 (2016)
10. Cong, G., Li, L., Pan, J., Zhang, Z., Beheshti, A., van den Hengel, A., Qi, Y., Huang, Q.: Flowdubber: Movie dubbing with llm-based semantic-aware learning and flow matching based voice enhancing. In: ACM MM. p. 905–914 (2025)
11. Cong, G., Li, L., Qi, Y., Zha, Z.J., Wu, Q., Wang, W., Jiang, B., Yang, M.H., Huang, Q.: Learning to dub movies via hierarchical prosody models. In: CVPR. pp. 14687–14697 (2023)
12. Cong, G., Pan, J., Li, L., Qi, Y., Peng, Y., Hengel, A.v.d., Yang, J., Huang, Q.: Emodubber: Towards high quality and emotion controllable movie dubbing. arXiv preprint arXiv:2412.08988 (2024)
13. Cong, G., Qi, Y., Li, L., Beheshti, A., Zhang, Z., Hengel, A.v.d., Yang, M.H., Yan, C., Huang, Q.: Styledubber: Towards multi-scale style learning for movie dubbing. arXiv preprint arXiv:2402.12636 (2024)
14. Cui, Y., Li, L., Yin, H., Gao, Y., Sheng, X., Yan, C.: Expert-teacher-student collaborative learning for domain adaptive object detection. In: CVPR (2026)
15. Guo, Y., Du, C., Ma, Z., Chen, X., Yu, K.: Voiceflow: Efficient text-to-speech with rectified flow matching. In: ICASSP. pp. 11121–11125 (2024)
16. Ho, J., Salimans, T.: Classifier-free diffusion guidance. arXiv preprint arXiv:2207.12598 (2022)
17. Jiang, Z., Ren, Y., Li, R., Ji, S., Zhang, B., Ye, Z., Zhang, C., Jionghao, B., Yang, X., Zuo, J., et al.: Megatts 3: Sparse alignment enhanced latent diffusion transformer for zero-shot speech synthesis. arXiv preprint arXiv:2502.18924 (2025)
18. Kim, J.H., Choi, J., Kim, J., Jung, C., Chung, J.S.: From faces to voices: Learning hierarchical representations for high-quality video-to-speech. arXiv preprint arXiv:2503.16956 (2025)

19. Kim, J., Kim, J., Chung, J.S.: Let there be sound: Reconstructing high quality speech from silent videos. In: AAAI. pp. 2759–2767 (2024)
20. Kim, M., Hong, J., Ro, Y.M.: Lip-to-speech synthesis in the wild with multi-task learning. In: ICASSP. pp. 1–5 (2023)
21. Kim, S., Shih, K.J., Badlani, R., Santos, J.F., Bakhturina, E., Desta, M., Valle, R., Yoon, S., Catanzaro, B.: P-flow: A fast and data-efficient zero-shot TTS through speech prompting. In: NeurIPS (2023)
22. Li, L., Cong, G., Qi, Y., Zha, Z.J., Wu, Q., Sheng, Q.Z., Huang, Q., Yang, M.H.: Dubbing movies via hierarchical phoneme modeling and acoustic diffusion denoising. IEEE TPAMI **47**(11), 10361–10377 (2025)
23. Lipman, Y., Chen, R.T., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. arXiv preprint arXiv:2210.02747 (2022)
24. Liu, J., Xiang, Y., Zhao, H., Li, X., Ling, Z.: Funcineforge: A unified dataset toolkit and model for zero-shot movie dubbing in diverse cinematic scenes. arXiv preprint arXiv:2601.14777 (2026)
25. Liu, R., Zhao, Y., Jia, Z.: Towards authentic movie dubbing with retrieve-augmented director-actor interaction learning. arXiv:2511.14249 (2025)
26. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
27. Ma, Z., Zheng, Z., Ye, J., Li, J., Gao, Z., Zhang, S., Chen, X.: emotion2vec: Self-supervised pre-training for speech emotion representation. In: Findings of ACL. pp. 15747–15760 (2024)
28. McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., Sonderegger, M.: Montreal forced aligner: Trainable text-speech alignment using kaldi. In: Interspeech. pp. 498–502 (2017)
29. Mehta, S., Tu, R., Beskow, J., Székely, É., Henter, G.E.: Matcha-tts: A fast tts architecture with conditional flow matching. In: ICASSP. pp. 11341–11345 (2024)
30. Nguyen, N.S., Tran, T.V., Choi, J., Huynh-Nguyen, H.N., Hy, T.S., Nguyen, V.: Diflowdubber: Discrete flow matching for automated video dubbing via cross-modal alignment and synchronization. arXiv preprint arXiv:2603.14267 (2026)
31. Peng, Y., Wang, Z., Zheng, X., Li, G., Yin, S., He, H.: A survey on fine-grained multimodal large language models. Chinese Journal of Electronics (2025)
32. Radford, A., Kim, J.W., Xu, T., Brockman, G., McLeavey, C., Sutskever, I.: Robust speech recognition via large-scale weak supervision. In: ICML (2023)
33. Rawal, R., Saifullah, K., Basri, R., Jacobs, D., Somepalli, G., Goldstein, T.: Cinepile: A long video question answering dataset and benchmark. In: CVPR Workshop (2024)
34. Reddy, C.K., Gopal, V., Cutler, R.: Dnsmos: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors. In: ICASSP. pp. 6493–6497. IEEE (2021)
35. Shi, B., Hsu, W., Lakhotia, K., Mohamed, A.: Learning audio-visual speech representation by masked multimodal cluster prediction. In: ICLR (2022)
36. Sung-Bin, K., Choi, J., Peng, P., Chung, J.S., Oh, T.H., Harwath, D.: Voicecraft-dub: Automated video dubbing with neural codec language models. arXiv preprint arXiv:2504.02386 (2025)
37. Tang, J., Liang, L., Zhang, B., Huang, Q.: Lmda:llm-guided marginal distribution alignment for open-set active learning. Chinese Journal of Electronics (2026)
38. Tian, W., Zhu, X., Liu, H., Zhao, Z., Chen, Z., Ding, C., Di, X., Zheng, J., Xie, L.: Dualdub: Video-to-soundtrack generation via joint speech and background audio synthesis. In: ACM MM. pp. 10671–10680 (2025)

39. Tong, A., Fatras, K., Malkin, N., Huguet, G., Zhang, Y., Rector-Brooks, J., Wolf, G., Bengio, Y.: Improving and generalizing flow-based generative models with minibatch optimal transport. Trans. Mach. Learn. Res. **2024** (2024)
40. Tu, Y., Li, L., Su, L., Yan, C., Huang, Q.: Distractors-immune representation learning with cross-modal contrastive regularization for change captioning. In: ECCV. pp. 311–328 (2024)
41. Tu, Y., Li, L., Su, L., Zha, Z., Huang, Q.: SMART: syntax-calibrated multi-aspect relation transformer for change captioning. IEEE TPAMI **46**(7), 4926–4943 (2024)
42. Wan, L., Wang, Q., Papir, A., López-Moreno, I.: Generalized end-to-end loss for speaker verification. In: ICASSP. pp. 4879–4883 (2018)
43. Woo, S., Debnath, S., Hu, R., Chen, X., Liu, Z., Kweon, I.S., Xie, S.: Convnext v2: Co-designing and scaling convnets with masked autoencoders. In: CVPR. pp. 16133–16142 (2023)
44. Yao, J., Yang, Y., Pan, Y., Ning, Z., Ye, J., Zhou, H., Xie, L.: Stablevc: Style controllable zero-shot voice conversion with conditional flow matching. arXiv preprint arXiv:2412.04724 (2024)
45. Ye, J., Cao, B., Shan, H.: Emotional face-to-speech. In: ICML. vol. 267 (2025)
46. Ye, J., Cong, G., Wang, C., Wen, X.C., Li, Z., Cao, B., Shan, H.: Hierarchical codec diffusion for video-to-speech generation. In: CVPR (2026)
47. Ye, J., Wen, X., Wei, Y., Xu, Y., Liu, K., Shan, H.: Temporal modeling matters: A novel temporal emotional modeling approach for speech emotion recognition. In: ICASSP. pp. 1–5 (2023)
48. Ye, J., Zhang, J., Shan, H.: DepMamba: Progressive fusion mamba for multimodal depression detection. In: IEEE Conf. Acoust. Speech Signal Process. pp. 1–5 (2025)
49. Yu, J., Zhu, H., Jiang, L., Loy, C.C., Cai, W., Wu, W.: Celebv-text: A large-scale facial text-video dataset. In: CVPR. pp. 14805–14814 (2023)
50. Zhang, X., Wang, Y., Wang, C., Li, Z., Chen, Z., Wu, Z.: Advancing zero-shot text-to-speech intelligibility across diverse domains via preference alignment. In: ACL. pp. 12251–12270 (2025)
51. Zhang, Z., Li, L., Cong, G., Haibing, Y., Gao, Y., Yan, C., van den Hengel, A., Qi, Y.: From speaker to dubber: Movie dubbing with prosody and duration consistency learning. In: ACM MM (2024)
52. Zhang, Z., Li, L., Cong, G., Liu, C., Gao, Y., Wang, X., Gu, T., Qi, Y.: Instructdubber: Instruction-based alignment for zero-shot movie dubbing. In: AAAI. pp. 12988–12996 (2026)
53. Zhang, Z., Li, L., Yan, C., Liu, C., van den Hengel, A., Qi, Y.: Prosody-enhanced acoustic pre-training and acoustic-disentangled prosody adapting for movie dubbing. arXiv preprint arXiv:2503.12042 (2025)
54. Zhao, Y., Jia, Z., Liu, R., Hu, D., Bao, F., Gao, G.: Mcdubber: Multimodal context-aware expressive video dubbing. arXiv preprint arXiv:2408.11593 (2024)
55. Zhao, Z., Li, L., Shen, L., Sheng, X., Sun, Y., Kang, F., Yan, C.: Temporal calibrating and distilling for scene-text aware text-video retrieval. In: AAAI (2026)
56. Zhao, Z., Li, L., Zhang, J., Sun, Y., Sheng, X., Yin, H., Jiang, S.: Heterogeneous prompt-guided entity inferring and distilling for scene-text aware cross-modal retrieval. In: AAAI. pp. 10537–10545 (2025)
57. Zheng, J., Chen, Z., Ding, C., Di, X.: Deepdubber-v1: Towards high quality and dialogue, narration, monologue adaptive movie dubbing via multi-modal chain-of-thoughts reasoning guidance. arXiv preprint arXiv:2503.23660 (2025)
58. Zhu, H., Wu, W., Zhu, W., Jiang, L., Tang, S., Zhang, L., Liu, Z., Loy, C.C.: Celebv-hq: A large-scale video facial attributes dataset. In: ECCV. pp. 650–667 (2022)