Hierarchical Codec Diffusion for Video-to-Speech Generation

Jiaxin Ye[1], Gaoxiang Cong[2,3], Chenhui Wang[1],
Xin-Cheng Wen[4], Zhaoyang Li[1], Boyuan Cao[1], Hongming Shan[1†]
[1]Fudan University, [2]Institute of Computing Technology, Chinese Academy of Sciences
[3]University of Chinese Academy of Sciences, [4]Harbin Institute of Technology (Shenzhen)
jxye22@m.fudan.edu.cn, hmshan@fudan.edu.cn

Abstract

*Video-to-Speech (VTS) generation aims to synthesize speech from a silent video without auditory signals, and holds substantial promise for applications such as film dubbing and voice restoration for individuals with aphonia. However, existing VTS methods disregard the hierarchical nature of speech, which spans coarse speaker-aware semantics to fine-grained prosodic details. This oversight hinders direct alignment between visual and speech features at specific hierarchical levels during property matching. In this paper, leveraging the hierarchical structure of Residual Vector Quantization (RVQ)-based codec, we propose **HiCoDiT**, a novel **Hi**erarchical **Co**dec **Di**ffusion **T**ransformer that exploits the inherent hierarchy of discrete speech tokens to achieve strong audio-visual alignment. Specifically, since lower-level tokens encode coarse speaker-aware semantics and higher-level tokens capture fine-grained prosody, HiCoDiT employs low-level and high-level blocks to generate tokens at different levels. The low-level blocks condition on lip-synchronized motion and facial identity to capture speaker-aware content, while the high-level blocks use facial expression to modulate prosodic dynamics. Finally, to enable more effective coarse-to-fine conditioning, we propose a dual-scale adaptive instance layer normalization that jointly captures global vocal style through channel-wise normalization and local prosody dynamics through temporal-wise normalization. Extensive experiments demonstrate that HiCoDiT outperforms baselines in fidelity and expressiveness, highlighting the potential of discrete modelling for VTS. The code and speech demo are both available at https://github.com/Jiaxin-Ye/HiCoDiT.*

†Corresponding author.
Jiaxin Ye and Hongming Shan are with the Institute of Science and Technology for Brain-Inspired Intelligence, MOE Frontiers Center for Brain Science, Key Laboratory of Computational Neuroscience and Brain-Inspired Intelligence, and State Key Laboratory of Brain Function and Disorders, Fudan University, Shanghai, China.

1. Introduction

Video-to-Speech (VTS) [9, 10, 63] generation aims to infer and synthesize speech from visual cues alone. This capability enables transformative applications, including silent film dubbing and assistive communication for aphonic individuals, to seamless interaction in noise-sensitive [13], privacy-critical [41] or embodied [18] environments.

The fundamental challenge in VTS lies in addressing the inherent information asymmetry between visual and acoustic modalities when generating natural and lip-synchronized speech from visual without acoustic input guidance. Specifically, although facial video and speech share consistent content [43], identity [44], and emotional prosody [58, 61, 62], visual features are inherently sparse and insufficient to capture the dense representations of speech, making it difficult to build accurate cross-modal alignment.

Existing approaches predominantly focus on representation alignment for guiding generative models [4, 5, 20, 22], spanning semantic content, vocal identity, and emotional prosody from vision to speech: (*i*) for semantic content alignment, NaturalL2S [29] leverages multimodal self-supervised representations to enhance the alignment between visual semantics and speech content; (*ii*) for vocal identity alignment, Face2Speech [20] aligns features from a face recognition encoder and a speaker recognition encoder to map facial identity to timbre information, while FaceStyleSpeech [23] further incorporates contrastive learning to improve face-to-speech alignment; (*iii*) for emotional prosody alignment, FTV [26] aligns facial emotion embeddings with pitch and energy to enhance prosody expressiveness. However, existing VTS methods typically inject visual features into holistic speech representations while overlooking the hierarchical structure of speech, from coarse speaker-aware semantics to fine-grained prosodic details, which ultimately exacerbates the inherent information asymmetry between visual and acoustic modalities. Therefore, the principal bottleneck in high-quality video-to-speech generation lies in visual conditioning, and how to

exploit the speech hierarchy as a prior to improve generation quality remains unresolved.

In this paper, we propose **HiCoDiT**, a novel **Hi**erarchical **Co**dec **Di**ffusion **T**ransformer that fully leverages the inherent hierarchy of discrete speech tokens to enable more effective vision-speech alignment. To the best of our knowledge, HiCoDiT is the first to introduce an explicit speech hierarchy prior into a discrete diffusion framework for video-to-speech generation. Specifically, leveraging the hierarchical structure of the Residual Vector Quantization (RVQ) codec in Figure 2, the low-level tokens primarily capture rich speaker-aware semantic content, whereas the high-level tokens encode more abstract prosodic details. Therefore, the hierarchy prior dictates that visual features such as lip motion and facial identity should primarily refine low-level speech tokens, while facial emotion features should modulate high-level tokens. Motivated by this prior, we design a hierarchical codec diffusion transformer composed of low-level and high-level blocks, progressively conditioning on speech tokens across different levels. The low-level blocks generate tokens conditioned on synchronized lip-motion representations and facial identity features for semantic and timbre alignment, while the high-level blocks produce tokens guided by facial emotion sequences for prosody alignment. To achieve more effective conditioning in the high-level block, we introduce a dual-scale Adaptive Instance Layer Normalization (AdaLN) that employs channel-wise normalization to model global vocal style, and temporal-wise normalization to capture local prosody dynamics. Extensive experiments demonstrate that HiCoDiT surpasses state-of-the-art baselines in semantic alignment and expressive prosody. Our contributions are summarized as follows.

- To our knowledge, HiCoDiT is the first discrete diffusion framework for VTS to explicitly integrate speech hierarchy prior, bridging the gap between video and speech.
- We propose a novel hierarchical diffusion transformer that models the speech hierarchy while disentangling visual conditioning, and a dual-scale AdaLN to inject global vocal style and local prosody into speech generation, enhancing expressiveness and fidelity.
- Extensive experiments demonstrate superior performance in semantic consistency and speech diversity, highlighting the potential of discrete speech tokens modelling for efficient VTS generation.

2. Related Work

Video-to-Speech (VTS) generation. Video-to-speech (VTS) seeks to generate speech that accurately reflects both linguistic content and speaker identity from visual cues alone [5, 27, 64]. Current approaches typically enforce alignment through auxiliary objectives: some predict text or mel-spectrograms jointly with visual input [27], others condition speaker embeddings on lip motion [5] or minimize cross-modal embedding distances [20]. While effective in isolation, these methods treat speech as a flat sequence without hierarchy and impose multiple supervision signals, leading to suboptimal alignment.

Recent advances have explored more sophisticated generative frameworks. For example, FTV [26] employs flow matching with a hierarchical visual encoder to gradually inject visual features into continuous mel-spectrogram space, while VoiceCraft-Dub [50] adapts pretrained autoregressive discrete text-to-speech models [39] to incorporate visual context. However, both obscure the inherent hierarchical structure of speech representation, in which coarse linguistic content emerges at early token levels and fine prosodic detail is resolved later. In contrast, HiCoDiT introduces the first discrete diffusion model for VTS trained from scratch, which explicitly integrates the speech hierarchy prior, bridging the gap between video and speech.

Hierarchical speech generation. Given the intrinsic hierarchy of speech, extensive research has focused on hierarchical representation modelling to achieve high-quality speech generation. In text-to-speech (TTS), Lee *et al*. [28] propose a hierarchical conditional variational autoencoder (VAE) that leverages self-supervised speech representations to bridge the information gap between text and speech, and Hsu *et al*. [21] likewise propose a conditional VAE with two levels of hierarchical latent variables, which captures coarse acoustic information and refines specific attribute configurations. Similarly, in video-to-speech, Kim *et al*. [26] develop a hierarchical visual encoder that learns a conditional representation by progressively aligning content, timbre, and prosody for a flow-matching decoder. In contrast to prior works that design entangled conditioning for hierarchical speech attributes, we exploit the inherent hierarchy of speech tokens themselves, modelling speech tokens from coarse semantics at lower levels to fine-grained acoustic details at higher levels, enabling disentangled conditioning and improving the fidelity of speech generation.

3. Preliminary: Discrete Diffusion Models

Recently, continuous diffusion models (CDM) [3, 5, 16, 32, 53, 54, 57, 65] have achieved state-of-the-art results in multimedia generation, while they are limited by computational inefficiency, frustrating practical application. An intuitive solution is to utilize discrete speech tokens [14, 56, 68] to build discrete diffusion models (DDMs), which have shown potential in language modeling [2, 31, 36] and speech generation [59, 60]. In this paper, we introduce a masked-based DDM to generate speech tokens under cross-modal guidance and outline below the forward and reverse processes of the DDM, along with its training objective.

Forward diffusion process. Given a token sequence $\boldsymbol{x} = [x^1, \ldots, x^d]$ with length d, where each token belongs to a discrete state sapce $\mathcal{X} = \{1, \ldots, n\}$. The diffusion pro-

cess can be modelled as a continuous-time discrete Markov chain, parameterized by the diffusion matrix $\boldsymbol{Q}_t \in \mathbb{R}^{n^d \times n^d}$, also known as the transition rate matrix at time t, as follows:

$$p(x^i_{t+\Delta t}|x^i_t) = \delta_{x^i_{t+\Delta t} x^i_t} + \boldsymbol{Q}_t(x^i_{t+\Delta t}, x^i_t)\Delta t + o(\Delta t), \quad (1)$$

where δ is Kronecker delta, x^i_t denotes i-th element of $\boldsymbol{x}_t$, and $\boldsymbol{Q}_t(x^i_{t+\Delta t}, x^i_t)$ is the $(x^i_{t+\Delta t}, x^i_t)$ element of $\boldsymbol{Q}_t$, which represents the transition rate from state x^i_t to state $x^i_{t+\Delta t}$ at time t. To further achieve efficient computation, existing methods [31, 37] adopt the assumption of dimensional independence, conducting a one-dimensional diffusion process for each dimension with the same token-level diffusion matrix $\boldsymbol{Q}^{\text{tok}}_t = \sigma(t)\boldsymbol{Q}^{\text{tok}} \in \mathbb{R}^{n \times n}$, where $\sigma(t)$ is the noise schedule and $\boldsymbol{Q}^{\text{tok}}$ is designed to diffuse towards an masked state `[MASK]`. Now, the forward equation can be formulated as $\boldsymbol{P}(x^i_t, x^i_0) = \exp\left(\bar{\sigma}(t)\boldsymbol{Q}^{\text{tok}}(x^i_t, x^i_0)\right)$, where transition probability matrix $\boldsymbol{P}(x^i_t, x^i_0) := p(x^i_t|x_0)$, and cumulative noise $\bar{\sigma}(t) = \int_0^t \sigma(s)ds$. There are two probabilities in the $\boldsymbol{P}_{t|0}$: $1 - e^{-\bar{\sigma}(t)}$ for replacing the current tokens with `[MASK]`, $e^{-\bar{\sigma}(t)}$ for remaining unchanged. Finally, the corrupted sequence $\boldsymbol{x}_t$ can be sampled from $\boldsymbol{x}_0$ in one step.

Reverse unmasking process. Given the diffusion matrix $\boldsymbol{Q}^{\text{tok}}_t$, we need a reverse transition rate matrix $\bar{\boldsymbol{Q}}_t$ [24, 49] to formulate reverse process, where $\bar{\boldsymbol{Q}}_t(x^i_{t-\Delta t}, x^i_t) = \frac{p(x^i_{t-\Delta t})}{p(x^i_t)}\boldsymbol{Q}^{\text{tok}}_t(x^i_t, x^i_{t-\Delta t})$ and $x^i_{t-\Delta t} \neq x^i_t$, or $\bar{\boldsymbol{Q}}_t(x^i_{t-\Delta t}, x^i_t) = -\sum_{z \neq x_t} \bar{\boldsymbol{Q}}_t(z, x^i_t)$. The reverse equation is formulated as follows:

$$p(x^i_{t-\Delta t}|x^i_t) = \delta_{x^i_{t-\Delta t} x^i_t} + \bar{\boldsymbol{Q}}_t(x^i_{t-\Delta t}, x^i_t)\Delta t + o(\Delta t). \quad (2)$$

The core of the reverse unmasking process is to estimate the concrete score $c_{x^i_{t-\Delta t} x^i_t} = \frac{p(x^i_{t-\Delta t})}{p(x^i_t)}$ of $\bar{\boldsymbol{Q}}_t$, representing to measure the *transition probability or closeness* from a state x^i at time t to a state $\hat{x}^i$ at time $t - \Delta t$. We can introduce a score network $s_\theta(x^i_t, t)_{x^i_{t-\Delta t}} \approx [\frac{p(x^i_{t-\Delta t})}{p(x^i_t)}]_{x^i_t \neq x^i_{t-\Delta t}}$ to learn the score, so that the reverse matrix is parameterized to model the reverse process $q_\theta(x^i_{t-\Delta t}|x^i_t)$ (*i.e.*, parameterize the concrete score).

Training objective. Denoising score entropy (DSE) [31] is introduced to train the score network s_θ:

$$\int_0^T \mathbb{E}_{\boldsymbol{x}_t \sim p(\boldsymbol{x}_t|\boldsymbol{x}_0)} \sum_{\hat{\boldsymbol{x}}_t \neq \boldsymbol{x}_t} \boldsymbol{Q}_t\left(\hat{x}^i_t, x^i_t\right) \Big[s_\theta\left(x^i_t, t\right)_{\hat{x}^i_t} - c_{\hat{x}^i_t x^i_t} \log s_\theta\left(x^i_t, t\right)_{\hat{x}^i_t} + \mathrm{N}(c_{\hat{x}^i_t x^i_t}) \Big] dt, \quad (3)$$

where the concrete score $c_{\hat{x}^i_t x^i_t} = \frac{p(\hat{x}^i_t|x^i_0)}{p(x^i_t|x^i_0)}$ and a normalizing constant function $\mathrm{N}(c) := c \log c - c$ that ensures loss non-negative. During sampling, we start from $\boldsymbol{x}_T$ filled with masked token `[MASK]`, and iteratively sample new set of tokens $\boldsymbol{x}_{t-1}$ from $p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)$ by replacing the concrete score with the trained score network on Eq. (2).

4. Methodology

4.1. Overview of HiCoDiT

Given a silent video $\mathcal{V}$, the goal of VTS system is to synthesize high-fidelity speech that aligns with the extracted visual features from the input video, including lip motion $\boldsymbol{c}_{\text{lip}}$, identity $\boldsymbol{c}_{\text{id}}$, and emotional expression $\boldsymbol{c}_{\text{emo}}$. To explicitly integrate speech hierarchy prior, we formulate VTS as a hierarchical masked token prediction task, employing an RVQ codec to tokenize speech for high-fidelity generation [14] and a discrete diffusion model to decode masked tokens for strong in-context perception [31]. Specifically, as shown in Figure. 1, HiCoDiT takes masked speech token sequence $\boldsymbol{x}_t$ as input and decomposes it into low-level component $\boldsymbol{x}^{\text{low}}_t = \boldsymbol{x}^{r_1:r_2}_t$ and high-level component $\boldsymbol{x}^{\text{high}}_t = \boldsymbol{x}^{r_3:r_{12}}_t$. Then, according to the inherent hierarchy of speech tokens, we disentangle the visual features extraction from the input video $\mathcal{V}$ and inject them into HiCoDiT. The lip motion features $\boldsymbol{c}_{\text{lip}}$ and identity features $\boldsymbol{c}_{\text{id}}$ are embedded into the low-level blocks to refine the generation of content- and timbre-centric tokens, while the emotional expression features $\boldsymbol{c}_{\text{emo}}$ are injected into the high-level blocks for enhancing prosody-related tokens generation. Finally, HiCoDiT outputs concrete scores for the reverse diffusion process to recover the masked tokens, which are decoded by the codec to synthesize high-fidelity speech.

4.2. Disentangled Visual Conditioning

Lip adapter for content modelling. Due to the strong temporal alignment between lip motion and speech content [12], we extract visual features using AV-HuBERT [47], taking the last-layer hidden states as they encode the most discriminative audio-visual semantics. These features are projected via a Multilayer Perceptron (MLP) to obtain $\boldsymbol{c}_{\text{lip}} \in \mathbb{R}^{L \times C}$, where L and C denote the sequence length and channel dimension, respectively, matching those of the masked low-level speech tokens $\boldsymbol{m}^{\text{low}}_t$.

Identity adapter for timbre modelling. Since both speech timbre and facial appearance encode speaker identity—despite lacking direct correspondence, we align their representation through cross-modal identity modelling. Specifically, visual identity features are extracted from facial images using ArcFace [15] and projected via an MLP into $\boldsymbol{c}_{\text{id}} \in \mathbb{R}^{L \times C_{\text{ge2e}}}$, where C_{ge2e} matches the channel dimension of acoustic identity features extracted by the GE2E model [52]. The two modalities are aligned by minimizing the ℓ_1 distance between their embeddings. The $\boldsymbol{c}_{\text{id}}$ is then fed into an MLP to generate modulation parameters for our dual-level AdaLN for timbre conditioning.

Emotion adapter for prosody modelling. Prosody refers to the non-lexical acoustic properties that convey the emotion of the speaker [19]. We leverage facial expression as a proxy signal by employing Poster2 [35], which

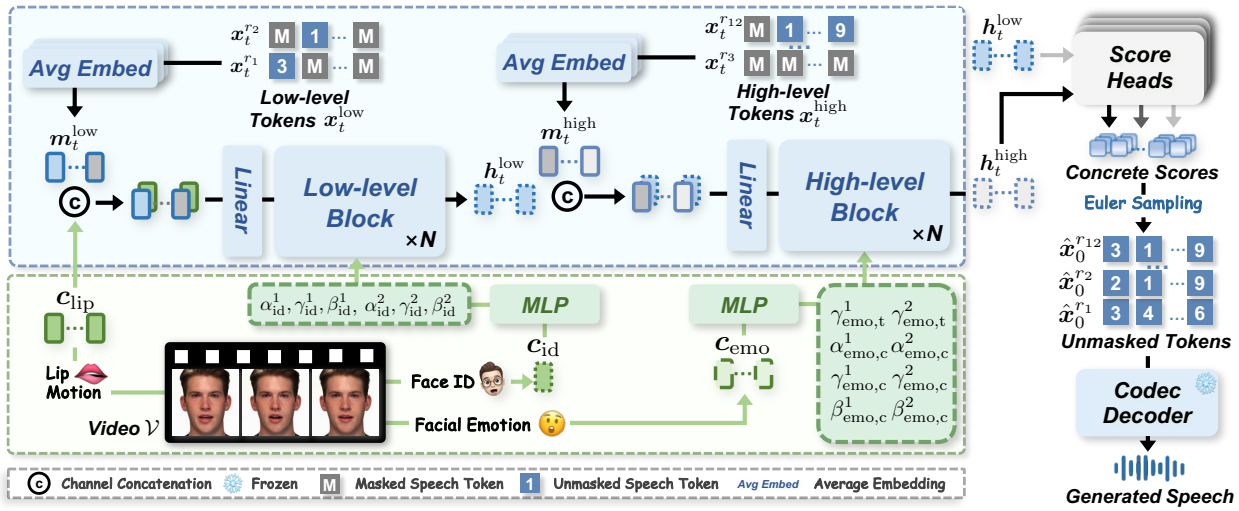

Figure 1. **Overall framework of HiCoDiT**. We formulate video-to-speech generation as a hierarchical masked token prediction task. Speech is tokenized using an RVQ codec and split into low-level components $\boldsymbol{x}_t^{r_1:r_2}$ and high-level components $\boldsymbol{x}_t^{r_3:r_{12}}$, reflecting the intrinsic hierarchy of speech tokens. Guided by this structure, we disentangle visual features from the input video $\mathcal{V}$ into lip motion $\boldsymbol{c}_{\text{lip}}$, identity $\boldsymbol{c}_{\text{id}}$, and emotion $\boldsymbol{c}_{\text{emo}}$, and inject them into the corresponding diffusion blocks. Finally, score heads take output features $\boldsymbol{h}_t^{\text{low}}$ and $\boldsymbol{h}_t^{\text{high}}$ from both blocks to predict concrete scores of all level tokens for unmasking.

is a strong video facial expression recognition model. To suppress identity-biased fluctuations, we only predict emotional class over all frames and apply temporal smoothing over 0.5-second windows, reducing the sequence to length L^{emo}. A learnable embedding layer then maps the smoothed class sequence to emotional features $\boldsymbol{c}_{\text{emo}} \in \mathbb{R}^{L_{\text{emo}} \times C}$, conditioning high-level speech tokens to modulate prosody.

4.3. Hierarchical Masked Token Prediction

Hierarchical speech tokenization and diffusion. Given a single-channel speech signal, we utilize the RVQ-based codec [56] to compresses it into tokens represented as $\boldsymbol{x}^{r_1:r_{12}} = \{1, \ldots, C_{\text{code}}\}^{12 \times L}$, where r_i is the i-th level of token, L is the length of the token sequence, respectively. The number of RVQ layers is 12 with a codebook size $C_{\text{code}} = 1,024$ in each level. We partition RVQ tokens into low-level $\boldsymbol{x}_t^{\text{low}} = \boldsymbol{x}_t^{r_1:r_2}$ and high-level $\boldsymbol{x}_t^{\text{high}} = \boldsymbol{x}_t^{r_3:r_{12}}$, reflecting the hierarchical structure of speech and consistent with the hierarchy analysis in Figure 2. The tokens are then masked via the discrete diffusion process of SEDD [31], as formalized in Eq. (1), yielding $\boldsymbol{x}_t^{\text{low}}$ and $\boldsymbol{x}_t^{\text{high}}$ at step t.

Hierarchical codec diffusion transformer. The proposed HiCoDiT serves as the score network in Eq. (3), predicting concrete scores for masked speech tokens that parametrize the transition rate from the masked state to each valid token. To align visual cues with the hierarchical structure of speech, we employ two complementary conditioning mechanisms: (i) direct concatenation for fine-grained and frame-synchronized signals such as lip motion, and (ii) dual-scale AdaLN for class-like attributes like speaker identity and emotion. For the content conditioning, the masked features $\boldsymbol{m}_t^{\text{low}} \in \mathbb{R}^{L \times C}$ are first concatenated with lip motion features $\boldsymbol{c}_{\text{lip}}$ along the channel dimension, followed by a linear layer to enhance temporally synchronized fusion. For the timbre conditioning, we utilize a MLP predicts channel-level scale and shift parameters $\alpha_{\text{id}}^1, \gamma_{\text{id}}^1, \beta_{\text{id}}^1, \alpha_{\text{id}}^2, \gamma_{\text{id}}^2, \beta_{\text{id}}^2 \in \mathbb{R}^C$ based on both identity features $\boldsymbol{c}_{\text{id}}$ and time t features. We can formulate the identity conditioning of the single-scale AdaLN as:

$$(1 + \gamma_{\text{id}}^i) \cdot \frac{\boldsymbol{h}_t - \mu(\boldsymbol{h}_t)}{\sigma(\boldsymbol{h}_t)} + \beta_{\text{id}}^i, \tag{4}$$

where $i = \{1, 2\}$ for multi-head attention and feed-forward network, and $\boldsymbol{h}_t \in \mathbb{R}^{L \times C}$ is the hidden embedding after layer normalization in low-level blocs. $\mu(\cdot)$ and $\sigma(\cdot)$ are the mean and standard deviation for $\boldsymbol{h}_t$ across the channel dimension. Furthermore, for the prosody conditioning, we introduce a temporal MLP to predict **t**emporal-level scale parameters $\gamma_{\text{emo,t}}^1, \gamma_{\text{emo,t}}^2 \in \mathbb{R}^{L^{\text{emo}}}$ using emotion features $\boldsymbol{c}_{\text{emo}}$ and time t features, and a channel MLP to predict **c**hannel-level scale and shift parameters $\alpha_{\text{emo,c}}^1, \gamma_{\text{emo,c}}^1, \beta_{\text{emo,c}}^1, \alpha_{\text{emo,c}}^2, \gamma_{\text{emo,c}}^2, \beta_{\text{emo,c}}^2 \in \mathbb{R}^C$ using pooling emotion features and time features. We can formulate the prosody conditioning of the dual-scale AdaLN as:

$$\underbrace{\gamma_{\text{emo,t}}^i \otimes \mathbf{1}_{25}}_{\text{Temporal-level}} \cdot \underbrace{\left((1 + \gamma_{\text{emo,c}}^i) \cdot \frac{\boldsymbol{h}_t - \mu(\boldsymbol{h}_t)}{\sigma(\boldsymbol{h}_t)} + \beta_{\text{emo,c}}^i \right)}_{\text{Channel-level}}, \tag{5}$$

where $i = \{1, 2\}$, $\otimes$ denotes Kronecker product, and $\mathbf{1}_{25} \in \mathbb{R}^{25}$ is an all-ones vector to up-sample γ_{te}^i with $L_{\text{emo}} = \frac{L}{25}$

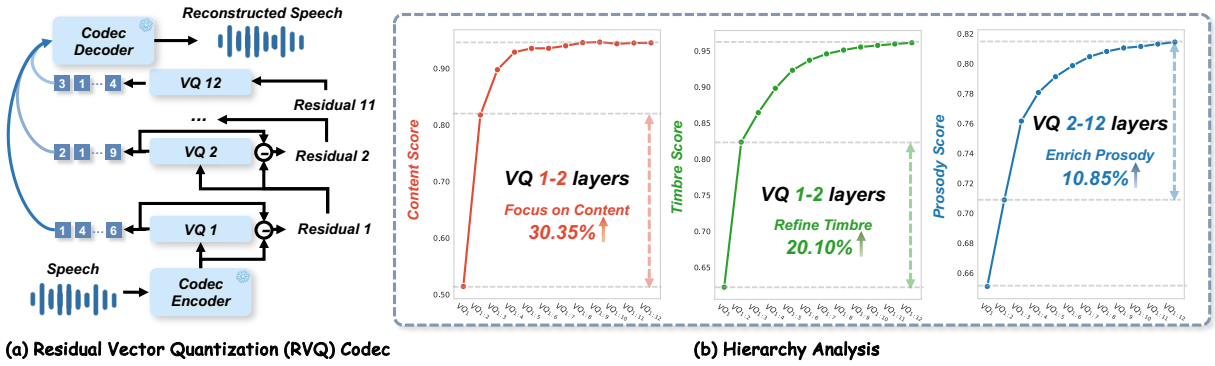

Figure 2. **Hierarchy analysis of speech token.** (a) RVQ codec encodes and decodes speech through multiple VQ layers. (b) The x-axis denotes cumulative decoding across token levels, and the y-axis reports scores for semantic fidelity, timbre similarity, and prosody quality. It can be observed that speaker-aware semantics improvements concentrate in lower layers, while prosody gains emerge in higher layers.

parameters to align with the hidden embedding with 50 Hz sampling rate. Finally, for the output, we incorporate 12 linear score heads to predict concrete scores for each level. These conditioning mechanisms enable HiCoDiT to faithfully modulate speech generation according to the cross-modal prior, bridging the gap between video and speech.

4.4. Training and Inference

Training. HiCoDiT is optimized by multi-level DSE loss based on Eq. (3) with the sum across all 12 RVQ levels as $\mathcal{L}_{\text{score}} = \sum_{i=1}^{12} \mathcal{L}_{\text{DSE}}(\boldsymbol{x}^{r_i}, t, \boldsymbol{c})$. For conducting predictor-free guidance, we randomly set $\varnothing$ with 10% probability for each condition and enforce all conditions set to $\varnothing$ for 10% samples. An additional loss $\mathcal{L}_{\text{id}} = \ell_1(\boldsymbol{c}_{\text{id}}, \boldsymbol{c}_{\text{GE2E}})$ aligns the visual identity embedding $\boldsymbol{c}_{\text{id}}$ with the GE2E speech embedding $\boldsymbol{c}_{\text{GE2E}}$ to reinforce speaker consistency. To summarize, the total loss function $\mathcal{L}_{\text{total}}$ is defined as follows:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{score}} + \lambda \mathcal{L}_{\text{id}}, \tag{6}$$

where λ is set to 100.0 in our experiments.

Inference. Following Eq. (2), the reverse process is executed with Euler sampling [31] and enhanced predictor-free guidance [63] with 64 sampling steps. Notably, to ensure training stability, we utilize ground truth acoustic features to replace $\boldsymbol{c}_{\text{id}}$ and $\boldsymbol{c}_{\text{emo}}$ during training, whereas only visual features are used during inference.

5. Experimental Results

5.1. Experimental Setups

Datasets. Our HiCoDiT is trained on the VoxCeleb2 [8] dataset, which provides large-scale speaker-diverse audio-visual recordings. To ensure well-aligned data, we perform a multi-stage data preprocessing pipeline. We first resample all audio to 16 kHz and employ a speech language identification model [44, 51] to filter out non-English utterances. We then apply a speaker diarization model [40] to remove multi-speaker segments, followed by the ClearerVoice speech separation model [66] to enhance signal-noise ratio. Finally, we leverage [43] to discard misaligned text–speech pairs. Finally, the pre-processing dataset comprises 261.5 hours of audio recordings with 169k utterances across 7 basic emotions and 3,438 speakers. For evaluation, we test our models on two in-the-wild datasets without any specific training, LRS2 [48] and LRS3 [1].

Evaluation metrics. The generation performance is evaluated using both subjective and objective metrics. For subjective assessment, we conduct a Mean Opinion Score (MOS) and A/B testing. For objective assessment, we first quantify spectral differences with Mel Cepstral Distortion (MCD) [4], DNSMOS [45], and UTMOS [46], which are widely used networks to estimate perceptual audio quality. We also calculate the Word Error Rate (WER) [43, 55] to gauge intelligibility. For the synchronization, we report the distance and confidence scores of lip sync errors (LSE-C and LSE-D) between speech and video using the pre-trained SyncNet [7]. For the expressiveness, we calculate cosine similarity metrics based on ECAPA-TDNN [17] to obtain speaker identity similarity (SpkSim). Additionally, we evaluate emotion accuracy (EmoAcc) using a strong speech emotion recognition model [33, 34].

Implementation details. For the speech tokenization, we employ a pre-trained RVQ-based codec from MaskGCT [56], and adopt a log-linear noise schedule $\sigma(t)$ [31] for the diffusion process, where the expectation of the number of masked tokens is linear with time t. For the disentangled visual conditioning, AV-HuBERT-Large, ArcFace, and Poster2 are used for lip, identity, and emotion feature extraction, respectively. For the transformer, the numbers of low- and high-level blocks are 8 and 8, respectively. The channel dimension C is set to 768 with 12 attention heads. During training, we use the AdamW optimizer [30] with a learning rate of 1e-4, batch size 32. The

Methods	Venue	A	V	*Naturalness*			*Synchronization*			*Expressiveness*	
				WER↓	DNSMOS↑	UTMOS↑	MCD↓	LSE-C↑	LSE-D↓	EmoAcc↑	SpkSim↑
Ground Truth	-	-	-	2.29	3.29	3.57	0.00	6.66	6.89	100.00	1.0000
Lip2Wav† [42]	CVPR'20	✓	✓	98.68	2.47	1.29	13.43	3.37	9.85	63.11	0.4785
MTL [27]	ICASSP'23	✓	✓	76.61	2.42	1.28	9.84	5.87	7.51	61.24	0.3347
EmoDubber† [11]	CVPR'25	✓	✓	41.52	2.95	2.83	9.25	6.88	6.85	72.01	0.6052
DiffV2S [5]	ICCV'23	✗	✓	41.07	2.56	3.06	-	-	-	-	-
LTBS† [25]	AAAI'24	✗	✓	84.00	2.36	2.42	-	-	-	-	-
AlignDiT [6]	ACM MM'25	✗	✓	31.37	3.24	3.76	10.02	6.95	6.82	76.11	0.5597
FTV [26]	CVPR'25	✗	✓	30.37	3.22	**3.99**	10.54	7.08	6.66	73.19	0.5981
HiCoDiT†(ours)	-	✗	✓	29.41	**3.50**	**3.84**	9.62	**7.15**	**6.58**	**79.41**	0.5678
		✓	✓	**28.98**	3.44	3.80	**8.69**	7.10	6.61	77.08	**0.6715**

Table 1. **Quantitative results on LRS3.** A/V indicate use of audio/video guidance (✓/✗). The superscript † indicates that the model is not trained on LRS3. ↑ (↓) indicates that higher (lower) is better. Best results are highlighted in **deeper blue**, second-best in lighter blue.

Methods	Venue	A	V	*Naturalness*			*Synchronization*			*Expressiveness*	
				WER↓	DNSMOS↑	UTMOS↑	MCD↓	LSE-C↑	LSE-D↓	EmoAcc↑	SpkSim↑
Ground Truth	-	-	-	8.93	3.14	3.05	0.00	7.20	6.67	100.00	1.0000
Lip2Wav† [42]	CVPR'20	✓	✓	100.05	2.47	1.31	14.09	3.83	9.80	54.38	0.4438
MTL [27]	ICASSP'23	✓	✓	58.03	2.42	1.30	10.71	6.58	7.16	63.89	0.3556
EmoDubber [11]	CVPR'25	✓	✓	47.60	2.84	2.77	**7.02**	7.42	6.60	66.76	0.5252
DiffV2S [5]	ICCV'23	✗	✓	54.86	2.36	2.95	-	-	-	-	-
LTBS† [25]	AAAI'24	✗	✓	94.30	2.17	2.29	-	-	-	-	-
AlignDiT† [6]	ACM MM'25	✗	✓	42.26	3.13	3.65	8.46	7.50	6.58	67.01	0.5187
FTV [26]	CVPR'25	✗	✓	**38.09**	3.11	**3.88**	12.91	7.71	6.35	67.84	0.5368
HiCoDiT†(ours)	-	✗	✓	39.99	**3.35**	3.68	8.74	**7.95**	**6.17**	**68.21**	0.5222
		✓	✓	40.75	3.27	3.38	8.36	7.83	6.24	65.65	**0.5954**

Table 2. **Quantitative results on LRS2.** A/V indicate use of audio/video guidance (✓/✗). The superscript † indicates that the model is not trained on LRS2. ↑ (↓) indicates that higher (lower) is better. Best results are highlighted in **deeper blue**, second-best in lighter blue.

total number of iterations is 200k. During inference, we employ an Euler sampler to perform the reverse process in 64 steps. For multi-conditional guidance, we adopt the enhanced predictor-free guidance [63], setting the joint guidance scale to $w_{\text{all}} = 2.5/2.25$ and the compositional scales to $w_{\text{id}} = 1.25/1.25$, $w_{\text{id}} = 1.5/1.5$, and $w_{\text{lip}} = 2.0$ for the LRS3 and LRS2 datasets, respectively.

Baseline models. Our method is compared with several state-of-the-art approaches: FTV [26], AlignDiT [6], EmoDubber [11], MTL [27], Lip2Wav [42], LTBS [25], and DiffV2S [5]. For FTV, the test samples on both LRS2 and LRS3 are provided by the authors. For AlignDiT, MTL, and Lip2Wav, we use the publicly released models for inference. For EmoDubber, we reproduce results using the official training code. Since no public models or test samples are available, we report results as cited in their original publications for both LTBS and DiffV2S.

5.2. Quantitative Evaluation

Objective evaluation. Tables 1 and 2 summarize the objective evaluation results on the LRS3 and LRS2 datasets, respectively. Although our HiCoDiT is not trained on either LRS3 or LRS2, it achieves leading performance on key metrics, including overall speech quality (UTMOS, DNSMOS), intelligibility (WER), and lip synchronization (LSE-C). While EmoDubber achieves the best spectral clarity on MCD by directly optimizing spectrograms, our method, which focuses on discrete speech token generation, achieves the second-best performance. Furthermore, our method exhibits a degradation in speaker similarity relative to FTV, reflecting the limited diversity of our training data. However, when a speech signal is introduced as identity guidance, our method achieves the highest score on this metric, showing great voice cloning ability. Similar trends are observed on LRS2. Overall, these results demonstrate the effectiveness of our hierarchical masked token prediction for VTS.

Subjective evaluation. We further conduct the subjective evaluation with 20 participants, to compare our HiCoDiT with SOTA methods. Specifically, we introduce five MOS with rating scores from 1 to 5 in 0.5 increments, including MOS_{nat}, MOS_{exp}, MOS_{syn} for speech naturalness, expressiveness, and lip-synchronization. We randomly generate 30 samples from the test set. The scoring results of the user study are presented in Table 3, demonstrating that HiCoDiT outperforms SOTA methods across nearly all metrics, particularly surpassing 2.94%in MOS_{syn}

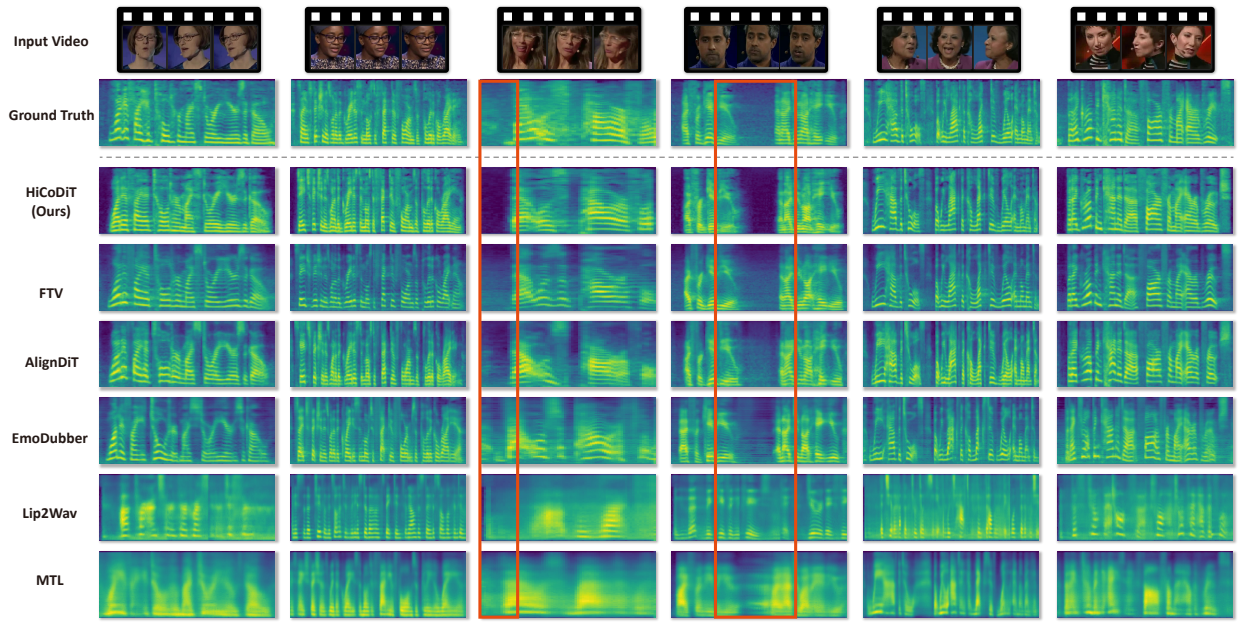

Figure 3. **The visualization of the mel-spectrograms of ground truth (GT) and synthesized speech obtained by different models.** As highlighted in the red boxes, the spectrograms generated by our method exhibit higher clarity with improved signal-to-noise ratio.

with the ground truth. Tables 1 and 2 summarize the objective evaluation results on the LRS3 and LRS2 datasets, respectively. Although our HiCoDiT is not trained on either LRS3 or LRS2, it achieves leading performance on key metrics, including overall speech quality (UTMOS, DNS-MOS), intelligibility (WER), and lip synchronization (LSE-C). While EmoDubber achieves the best spectral clarity on MCD by directly optimizing spectrograms, our method, which focuses on discrete speech token generation, achieves the second-best performance. Furthermore, our method exhibits a degradation in speaker similarity relative to FTV, reflecting the limited diversity of our training data. However, when a speech signal is introduced as identity guidance, our method achieves the highest score on this metric, showing great voice cloning ability. Similar trends are observed on LRS2. Overall, these results demonstrate the effectiveness of our hierarchical masked token prediction for VTS. Furthermore, the proposed model HiCoDiT achieves the highest MOS_{nat} (3.17) and MOS_{sync} (3.50), indicating superior naturalness and synchronization compared to existing methods like AlignDiT and FTV. Although the expressiveness is slightly lower than FTV, indicating that a more diverse speaker dataset can enhance expressiveness.

In addition, the Table 4 compares A/B test preferences for synthesized speech. Our method demonstrates clear superiority over AlignDiT, achieving a 57.0% preference, and also outperforms FTV with 52.1% preference. Additionally, ground-truth speech is preferred over FTV (51.5%) and is outperformed by our method with a 53.9% preference, showing the strength of our model in generating high-quality speech nearly indistinguishable from real speech.

Methods	$\text{MOS}_{\text{nat}} \uparrow$	$\text{MOS}_{\text{exp}} \uparrow$	$\text{MOS}_{\text{syn}} \uparrow$
Ground Truth	3.07±1.02	3.30±1.19	3.40±0.93
AlignDiT [6]	2.47±1.19	2.63±1.30	3.13±0.75
FTV [26]	2.80±1.03	**2.90±1.45**	3.48±1.02
HiCoDiT (**ours**)	**3.17±1.31**	2.88±1.53	**3.50±0.86**

Table 3. **Subjective evaluation** on speech naturalness, expressiveness, and synchronization, compared with other SOTA methods.

A vs. B	A wins (%)	Neutral	B wins (%)
Ours vs. AlignDiT [6]	**57.0**	4.9	38.1
Ours vs. FTV [26]	**52.1**	6.1	41.8
GT vs. FTV [26]	**51.5**	14.0	34.5
GT vs. Ours	45.5	0.6	**53.9**

Table 4. **A/B testing results.** We report the preferences (%) between A and B across various aspects of synthesized speech.

5.3. Qualitative Results

Qualitative spectrogram comparisons. As shown in Figure 3, we compare generated mel-spectrograms with other methods. For Lip2Wav and MTL, we observe severe over-smoothing or acoustic artifacts, resulting in significant degradation of speech quality and limiting their practical utility, which may be attributed to the insufficient probabilistic modeling capacity of the generative models

Datasets	Ablations	WER↓	DNSMOS↑	UTMOS↑	MCD↓	LSE-C↑	LSE-D↓	EmoAcc↑	SpkSim↑
LRS3	w/o Hierarchical Modeling	30.65	3.36	3.73	10.07	7.02	6.75	76.98	0.5652
	w/o Dual Scale AdaLN	29.60	3.45	**3.92**	9.75	7.12	6.60	78.55	0.5621
	HiCoDiT (full)	**29.41**	**3.50**	3.84	**9.62**	**7.15**	**6.58**	**79.41**	**0.5678**
LRS2	w/o Hierarchical Modeling	44.57	3.18	3.48	9.43	7.66	6.47	64.69	0.4946
	w/o Dual Scale AdaLN	41.01	3.30	**3.75**	9.33	7.88	6.22	**68.61**	0.5155
	HiCoDiT (full)	**39.99**	**3.35**	3.68	**8.74**	**7.95**	**6.17**	68.21	**0.5222**

Table 5. **Ablation study on LRS3 and LRS2.** Best results are highlighted in **Bold**.

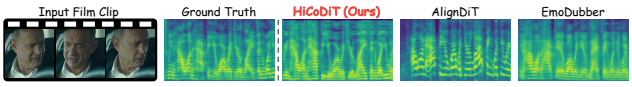

Figure 4. Comparison of generated Mels on real-world film data.

Method	WER↓	MCD↓	DNSMOS↑	Emo↑	Spk↑	LSE-D↓
EmoDubber	88.3	9.9	2.8	76.5	45.1	7.72
AlignDiT	80.8	11.4	3.2	75.2	**58.5**	8.23
HiCoDiT	**58.7**	**9.8**	**3.5**	**82.0**	50.1	**7.60**

Table 6. Quantitative comparison of real-world OOD film data.

used. Methods based on powerful diffusion models, produce high-quality speech. However, their mel-spectrograms still exhibit noise in the silent clip. In contrast, our method generates clarity mel-spectrograms with richer acoustic details and precise lip-synchronization, benefiting from the strong speech reconstruction capability of Codec.

5.4. Ablation Studies

Ablation on hierarchical modeling. We explore the impact of hierarchical modeling on video-to-speech generation in Table 5. The removal of hierarchical modeling collapses the multi-level speech representation into a single uniform module, while simultaneously forcing visual conditioning to be injected across all tokens. Experimental results demonstrate that performance degrades significantly across all metrics, underscoring the validity of our proposed speech hierarchy prior. This further indicates that visual features corresponding to specific attributes should align with speech tokens carrying matching content.

Ablation on dual scale AdaLN. To demonstrate the effectiveness of the proposed Dual-scale AdaLN, we utilize the vanilla adaLN of DiT [38] and replace the temporal embedding with utterance-level emotional embedding combined with global style as an acoustic guidance. As shown in Table 5, pooling dynamic emotions struggles to model prosody dynamics, with a negligible decrease in terms of EmoAcc, while other metrics gain a lot. The results highlight the effectiveness of our dual-scale AdaLN mechanism.

Ablation on out-of-domain data. To assess generalization in complex, real-world environments, we curate an authentic film benchmark comprising 160 utterances across 56 speakers from *CinePile* to ensure realistic audio-visual complexity. We compared HiCoDiT against primary open-source SOTA methods EmoDubber and AlignDiT. Table 6 and Figure 4 demonstrate that our method achieves robust intelligibility and lip-synchronization on this challenging OOD data, underscoring HiCoDiT's robustness and adaptability to authentic scenarios.

Ablations	WER↓	MCD↓	DNSMOS↑	Emo↑	Spk↑	LSE-D↓
(a) wo GE2E $\mathcal{L}_{\text{id}}$	**29.38**	10.18	3.41	74.47	34.10	6.71
(b) wo Poster2	29.41	9.68	3.50	76.29	55.28	6.67
HiCoDiT	29.41	**9.62**	**3.50**	**79.41**	**56.78**	**6.58**

Table 7. Ablation study of visual conditioning on LRS3 test set.

Ablation on visual conditioning. To further explore our visual conditioning, we conduct ablation studies on the LRS3 benchmark in Table 7. We evaluate the impact of the GE2E loss $\mathcal{L}_{\text{id}}$ by removing it from the training objective. The results reveal a substantial decline in speaker similarity from 56.78% to 34.10%, while WER remain unaffected. This confirms that the GE2E loss is indispensable for identity preservation, effectively guiding the model to extract implicit vocal timbre from facial cues. Second, we assess the Poster2 [35] encoder by replacing it with Poster [67]. This substitution leads to a noticeable drop in emotion accuracy from 79.41% to 76.29%, validating the superiority of Poster2 in capturing fine-grained affective information.

6. Conclusion

We present HiCoDiT, a Hierarchical Codec Diffusion Transformer that redefines how visual features and speech tokens are aligned in VTS generation. By leveraging the hierarchy of discrete speech tokens, HiCoDiT enables precise synchronization of lip motion and identity at lower levels, while capturing expressive emotional and prosodic dynamics at higher levels. We also design a dual-scale AdaLN, which effectively captures global vocal style and local prosody dynamics. Extensive experiments conducted on benchmark datasets, including LRS2 and LRS3, demonstrate the superiority of HiCoDiT over state-of-the-art methods in terms of naturalness, expressiveness, and synchronization fidelity, establishing HiCoDiT as a promising solution for real-world VTS applications.

Acknowledgment

This work was supported in part by National Natural Science Foundation of China (No. 62471148), STI2030-Major Projects (No. 2021ZD0200204), and Shanghai Center for Brain Science and Brain-inspired Technology.

References

[1] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. LRS3-TED: A large-scale dataset for visual speech recognition. *arXiv preprint:1809.00496*, 2018. 5

[2] Jacob Austin, Daniel D. Johnson, Jonathan Ho, Daniel Tarlow, and Rianne van den Berg. Structured denoising diffusion models in discrete state-spaces. In *Adv. Neural Inform. Process. Syst.*, pages 17981–17993, 2021. 2

[3] Boyuan Cao, Jiaxin Ye, Yujie Wei, and Hongming Shan. RepLDM: Reprogramming pretrained latent diffusion models for high-quality, high-efficiency, high-resolution image generation. In *Adv. Neural Inform. Process. Syst.* 2

[4] Qi Chen, Mingkui Tan, Yuankai Qi, Jiaqiu Zhou, Yuanqing Li, and Qi Wu. V2C: Visual voice cloning. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 21210–21219, 2022. 1, 5

[5] Jeongsoo Choi, Joanna Hong, and Yong Man Ro. DiffV2S: Diffusion-based video-to-speech synthesis with vision-guided speaker embedding. In *Int. Conf. Comput. Vis.*, pages 7778–7787, 2023. 1, 2, 6

[6] Jeongsoo Choi, Ji-Hoon Kim, Kim Sung-Bin, Tae-Hyun Oh, and Joon Son Chung. AlignDiT: Multimodal aligned diffusion transformer for synchronized speech generation. In *ACM Int. Conf. Multimedia*, 2025. 6, 7

[7] Joon Son Chung and Andrew Zisserman. Out of time: Automated lip sync in the wild. In *ACCV. Int. Worksh.* , pages 251–263, 2016. 5

[8] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. Voxceleb2: Deep speaker recognition. *arXiv preprint:1806.05622*, 2018. 5

[9] 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 *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 14687–14697, 2023. 1

[10] 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 *ACM Int. Conf. Multimedia*, pages 905–914, 2025. 1

[11] 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. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 15863–15873, 2025. 6

[12] Yusheng Dai, Hang Chen, Jun Du, Ruoyu Wang, Shihao Chen, Haotian Wang, and Chin-Hui Lee. A study of dropout-induced modality bias on robustness to missing video frames for audio-visual speech recognition. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 27435–27445, 2024. 3

[13] Balaji Darur and Karan Singla. Visual-aware speech recognition for noisy scenarios. In *Proc. Conf. Empir. Methods Natural Lang. Process.*, pages 16709–16717, 2025. 1

[14] Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. High fidelity neural audio compression. *Trans. Mach. Learn. Res.*, 2023, 2023. 2, 3

[15] Jiankang Deng, Jia Guo, Jing Yang, Niannan Xue, Irene Kotsia, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(10):5962–5979, 2022. 3

[16] Yan Deng, Ning Wu, Chengjun Qiu, Yangyang Luo, and Yan Chen. MixGAN-TTS: Efficient and stable speech synthesis based on diffusion model. *IEEE Access*, 11:57674–57682, 2023. 2

[17] Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck. ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in TDNN based speaker verification. In *Annu. Conf. Int. Speech Commun. Assoc.*, pages 3830–3834, 2020. 5

[18] Tongtong Feng, Xin Wang, and Wenwu Zhu. Self-evolving embodied ai. *arXiv preprint:2602.04411*, 2026. 1

[19] Robert W Frick. Communicating emotion: The role of prosodic features. *Psychological bulletin*, 97(3):412, 1985. 3

[20] Shunsuke Goto, Kotaro Onishi, Yuki Saito, Kentaro Tachibana, and Koichiro Mori. Face2Speech: Towards multi-speaker text-to-speech synthesis using an embedding vector predicted from a face image. In *Annu. Conf. Int. Speech Commun. Assoc.*, pages 1321–1325, 2020. 1, 2

[21] Wei-Ning Hsu, Yu Zhang, Ron J. Weiss, Heiga Zen, Yonghui Wu, Yuxuan Wang, Yuan Cao, Ye Jia, Zhifeng Chen, Jonathan Shen, Patrick Nguyen, and Ruoming Pang. Hierarchical generative modeling for controllable speech synthesis. In *Int. Conf. Learn. Represent.*, 2019. 2

[22] Youngjoon Jang, Ji-Hoon Kim, Junseok Ahn, Doyeop Kwak, et al. Faces that speak: Jointly synthesising talking face and speech from text. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 8818–8828, 2024. 1

[23] Minki Kang, Wooseok Han, and Eunho Yang. Face-StyleSpeech: Improved face-to-voice latent mapping for natural zero-shot speech synthesis from a face image. *arXiv preprint:2311.05844*, 2023. 1

[24] Frank P Kelly. *Reversibility and stochastic networks*. Cambridge University Press, 2011. 3

[25] Ji-Hoon Kim, Jaehun Kim, and Joon Son Chung. Let there be sound: Reconstructing high quality speech from silent videos. In *AAAI Conf. Artif. Intell.*, pages 2759–2767, 2024. 6

[26] Ji-Hoon Kim, Jeongsoo Choi, Jaehun Kim, Chaeyoung Jung, and Joon Son Chung. From faces to voices: Learning hierarchical representations for high-quality video-to-speech. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 15874–15884, 2025. 1, 2, 6, 7

[27] Minsu Kim, Joanna Hong, and Yong Man Ro. Lip-to-speech synthesis in the wild with multi-task learning. In *IEEE Conf. Acoust. Speech Signal Process.*, pages 1–5, 2023. 2, 6

[28] Sang-Hoon Lee, Seung-Bin Kim, Ji-Hyun Lee, et al. HierSpeech: Bridging the gap between text and speech by hierarchical variational inference using self-supervised representations for speech synthesis. In *Adv. Neural Inform. Process. Syst.*, 2022. 2

[29] Yifan Liang, Fangkun Liu, Andong Li, Xiaodong Li, Chengyou Lei, and Chengshi Zheng. NaturalL2S: End-to-end high-quality multispeaker lip-to-speech synthesis with differential digital signal processing. *Neural Networks*, 194: 108163, 2026. 1

[30] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *Int. Conf. Learn. Represent.*, 2019. 5

[31] Aaron Lou, Chenlin Meng, and Stefano Ermon. Discrete diffusion modeling by estimating the ratios of the data distribution. In *Int. Conf. on Mach. Learn.*, 2024. 2, 3, 4, 5

[32] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. DPM-Solver++: Fast solver for guided sampling of diffusion probabilistic models. *Mach. Intell. Res.*, 22 (4):730–751, 2025. 2

[33] Ziyang Ma, Mingjie Chen, Hezhao Zhang, Zhisheng Zheng, Wenxi Chen, Xiquan Li, Jiaxin Ye, Xie Chen, and Thomas Hain. EmoBox: Multilingual multi-corpus speech emotion recognition toolkit and benchmark. In *Annu. Conf. Int. Speech Commun. Assoc.*, 2024. 5

[34] Ziyang Ma, Zhisheng Zheng, Jiaxin Ye, Jinchao Li, Zhifu Gao, Shiliang Zhang, and Xie Chen. emotion2vec: Self-supervised pre-training for speech emotion representation. In *Findings Proc. Annu. Meeting Assoc. Comput. Linguistics*, pages 15747–15760, 2024. 5

[35] Jiawei Mao, Rui Xu, Xuesong Yin, Yuanqi Chang, Binling Nie, and Aibin Huang. POSTER V2: A simpler and stronger facial expression recognition network. *arXiv preprint:2301.12149*, 2023. 3, 8

[36] Chenlin Meng, Kristy Choi, Jiaming Song, and Stefano Ermon. Concrete score matching: Generalized score matching for discrete data. In *Adv. Neural Inform. Process. Syst.*, 2022. 2

[37] Jingyang Ou, Shen Nie, Kaiwen Xue, Fengqi Zhu, Jiacheng Sun, Zhenguo Li, and Chongxuan Li. Your absorbing discrete diffusion secretly models the conditional distributions of clean data. *CoRR*, abs/2406.03736, 2024. 3

[38] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Int. Conf. Comput. Vis.*, pages 4172–4182, 2023. 8

[39] Puyuan Peng, Po-Yao Huang, Shang-Wen Li, Abdelrahman Mohamed, and David Harwath. VoiceCraft: Zero-shot speech editing and text-to-speech in the wild. In *Proc. Annu. Meeting Assoc. Comput. Linguistics*, pages 12442–12462, 2024. 2

[40] Alexis Plaquet and Hervé Bredin. Powerset multi-class cross entropy loss for neural speaker diarization. In *Annu. Conf. Int. Speech Commun. Assoc.*, 2023. 5

[41] Gerasimos Potamianos, Chalapathy Neti, Juergen Luettin, Iain Matthews, et al. Audio-visual automatic speech recognition: An overview. *Issues in visual and audio-visual speech processing*, 22:23, 2004. 1

[42] K. R. Prajwal, Rudrabha Mukhopadhyay, Vinay P. Namboodiri, and C. V. Jawahar. Learning individual speaking styles for accurate lip to speech synthesis. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 13793–13802, 2020. 6

[43] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *Int. Conf. on Mach. Learn.*, pages 28492–28518, 2023. 1, 5

[44] Mirco Ravanelli, Titouan Parcollet, Peter Plantinga, Aku Rouhe, et al. SpeechBrain: A general-purpose speech toolkit. *arXiv preprint:2106.04624*, 2021. 1, 5

[45] Chandan K. A. Reddy, Vishak Gopal, and Ross Cutler. Dnsmos: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors. In *IEEE Conf. Acoust. Speech Signal Process.*, pages 6493–6497, 2021. 5

[46] Takaaki Saeki, Detai Xin, Wataru Nakata, Tomoki Koriyama, Shinnosuke Takamichi, and Hiroshi Saruwatari. Utmos: Utokyo-sarulab system for voicemos challenge 2022. *arXiv preprint:2204.02152*, 2022. 5

[47] Bowen Shi, Wei-Ning Hsu, Kushal Lakhotia, and Abdelrahman Mohamed. Learning audio-visual speech representation by masked multimodal cluster prediction. In *Int. Conf. Learn. Represent.*, 2022. 3

[48] Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Zisserman. Lip reading sentences in the wild. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 6447–6456, 2017. 5

[49] Haoran Sun, Lijun Yu, Bo Dai, Dale Schuurmans, and Hanjun Dai. Score-based continuous-time discrete diffusion models. In *Int. Conf. Learn. Represent.*, 2023. 3

[50] Kim Sung-Bin, Jeongsoo Choi, Puyuan Peng, Joon Son Chung, Tae-Hyun Oh, and David Harwath. VoiceCraft-Dub: Automated video dubbing with neural codec language models. In *Int. Conf. Comput. Vis.*, 2025. 2

[51] Jörgen Valk and Tanel Alumäe. VoxLingua107: A dataset for spoken language recognition. In *Proc. IEEE SLT Workshop*, 2021. 5

[52] Li Wan, Quan Wang, Alan Papir, and Ignacio López-Moreno. Generalized end-to-end loss for speaker verification. In *IEEE Conf. Acoust. Speech Signal Process.*, pages 4879–4883, 2018. 3

[53] Chenhui Wang, Tao Chen, Zhihao Chen, Zhizhong Huang, Taoran Jiang, Qi Wang, and Hongming Shan. FLDM-VTON: Faithful latent diffusion model for virtual try-on. In *Proc. Int. Joint Conf. Artif. Intell.*, pages 1362–1370, 2024. 2

[54] Chenhui Wang, Boyun Zheng, Liuxin Bao, Zhihao Peng, Peter YM Woo, Hongming Shan, and Yixuan Yuan. Brain-WM: Brain glioblastoma world model. *arXiv preprint:2603.07562*, 2026. 2

[55] Yuxuan Wang, Daisy Stanton, Yu Zhang, R. J. Skerry-Ryan, Eric Battenberg, Joel Shor, Ying Xiao, Ye Jia, Fei Ren, and Rif A. Saurous. Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis. In *Int. Conf. on Mach. Learn.*, pages 5167–5176, 2018. 5

[56] Yuancheng Wang, Haoyue Zhan, Liwei Liu, Ruihong Zeng, Haotian Guo, Jiachen Zheng, Qiang Zhang, Shunsi Zhang, and Zhizheng Wu. MaskGCT: Zero-shot text-to-speech with masked generative codec transformer. *arXiv preprint:2409.00750*, 2024. 2, 4, 5

[57] Yujie Wei, Shiwei Zhang, Hangjie Yuan, Xiang Wang, Haonan Qiu, Rui Zhao, Yutong Feng, Feng Liu, Zhizhong Huang, Jiaxin Ye, Yingya Zhang, and Hongming Shan. DreamVideo-2: Zero-shot subject-driven video customization with precise motion control. *arXiv preprint:2410.13830*, 2024. 2

[58] Xin-Cheng Wen, Jiaxin Ye, Yan Luo, Yong Xu, Xuan-Ze Wang, Chang-Li Wu, and Kun-Hong Liu. CTL-MTNet: A novel capsnet and transfer learning-based mixed task net for single-corpus and cross-corpus speech emotion recognition. In *Proc. Int. Joint Conf. Artif. Intell.*, pages 2305–2311, 2022. 1

[59] Zhichao Wu, Qiulin Li, Sixing Liu, and Qun Yang. DCTTS: Discrete diffusion model with contrastive learning for text-to-speech generation. In *IEEE Conf. Acoust. Speech Signal Process.*, pages 11336–11340. IEEE, 2024. 2

[60] Dongchao Yang, Jianwei Yu, Helin Wang, Wen Wang, Chao Weng, Yuexian Zou, and Dong Yu. Diffsound: Discrete diffusion model for text-to-sound generation. *IEEE ACM Trans. Audio Speech Lang. Process.*, 31:1720–1733, 2023. 2

[61] Jiaxin Ye, Yujie Wei, Xin-Cheng Wen, Chenglong Ma, Zhizhong Huang, Kunhong Liu, and Hongming Shan. Emo-DNA: Emotion decoupling and alignment learning for cross-corpus speech emotion recognition. In *ACM Int. Conf. Multimedia*, pages 5956–5965, 2023. 1

[62] Jiaxin Ye, Xin-Cheng Wen, Yujie Wei, Yong Xu, Kunhong Liu, and Hongming Shan. Temporal modeling matters: A novel temporal emotional modeling approach for speech emotion recognition. In *IEEE Conf. Acoust. Speech Signal Process.,*, pages 1–5, 2023. 1

[63] Jiaxin Ye, Boyuan Cao, and Hongming Shan. Emotional face-to-speech. In *Int. Conf. on Mach. Learn.*, 2025. 1, 5, 6

[64] Yochai Yemini, Aviv Shamsian, Lior Bracha, Sharon Gannot, and Ethan Fetaya. LipVoicer: Generating speech from silent videos guided by lip reading. In *Int. Conf. Learn. Represent.*, 2024. 2

[65] Xulu Zhang, Xiaoyong Wei, Wentao Hu, Jinlin Wu, Jiaxin Wu, Wengyu Zhang, Zhaoxiang Zhang, Zhen Lei, and Qing Li. A survey on personalized content synthesis with diffusion models. *Mach. Intell. Res.*, 22(5):817–848, 2025. 2

[66] Shengkui Zhao, Zexu Pan, and Bin Ma. Clearervoice-studio: Bridging advanced speech processing research and practical deployment. *arXiv preprint:2506.19398*, 2025. 5

[67] Ce Zheng, Matías Mendieta, and Chen Chen. POSTER: A pyramid cross-fusion transformer network for facial expression recognition. In *Int. Conf. Comput. Vis. Workshop*, pages 3138–3147, 2023. 8

[68] Youqiang Zheng, Weiping Tu, Li Xiao, and Xinmeng Xu. Srcodec: Split-residual vector quantization for neural speech codec. In *IEEE Conf. Acoust. Speech Signal Process.*, pages 451–455, 2024. 2