AlignDiT: Multimodal Aligned Diffusion Transformer for Synchronized Speech Generation

Jeongsoo Choi
Korea Advanced Institute of Science and Technology
Daejeon, Republic of Korea
jeongsoo.choi@kaist.ac.kr

Ji-Hoon Kim
Korea Advanced Institute of Science and Technology
Daejeon, Republic of Korea
jihoon@mm.kaist.ac.kr

Kim Sung-Bin
Pohang University of Science and Technology
Pohang, Republic of Korea
sungbin@postech.ac.kr

Tae-Hyun Oh
Korea Advanced Institute of Science and Technology
Daejeon, Republic of Korea
taehyun.oh@kaist.ac.kr

Joon Son Chung
Korea Advanced Institute of Science and Technology
Daejeon, Republic of Korea
joonson@kaist.ac.kr

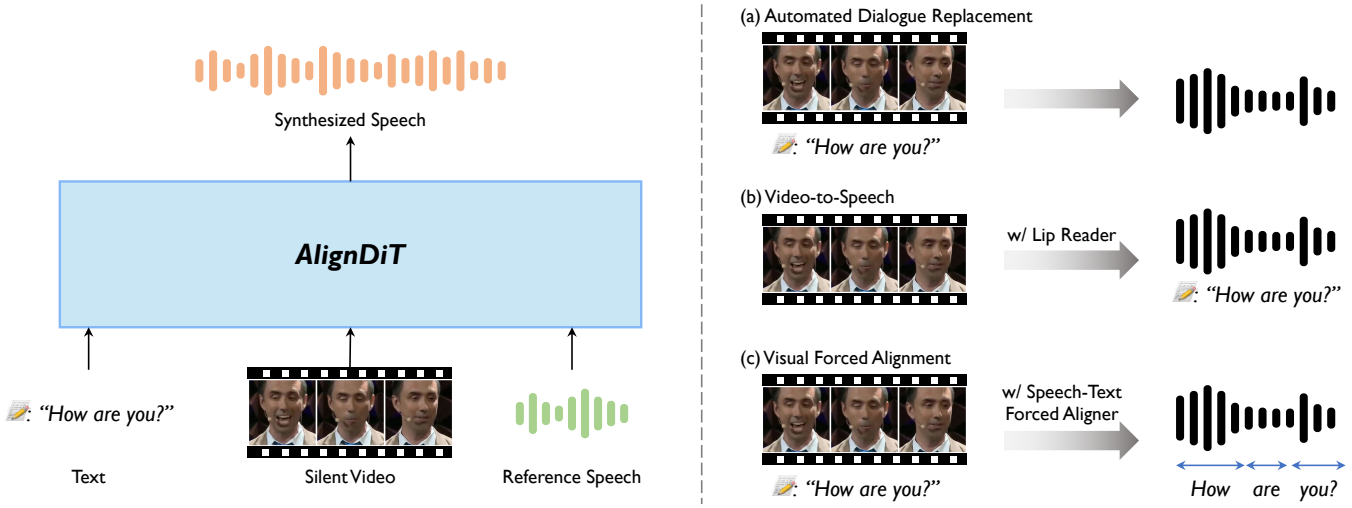


Figure 1: AlignDiT aims to generate natural speech that is (1) accurately aligned with the given text, (2) temporally synchronized with the input video, and (3) acoustically consistent with the reference speech. This versatile system has a broad range of applications, including (a) Automated dialogue replacement, (b) Video-to-Speech synthesis, and (c) Visual Forced Alignment.

Abstract

In this paper, we address the task of multimodal-to-speech generation, which aims to synthesize high-quality speech from multiple input modalities: text, video, and reference audio. This task has gained increasing attention due to its wide range of applications, such as film production, dubbing, and virtual avatars. Despite recent progress, existing methods still suffer from limitations in speech intelligibility, audio-video synchronization, speech naturalness, and voice similarity to the reference speaker. To address these challenges, we propose AlignDiT, a multimodal Aligned Diffusion Transformer that generates accurate, synchronized, and natural-sounding speech from aligned multimodal inputs. Built upon the in-context learning capability of the DiT architecture, AlignDiT explores three effective strategies to align multimodal representations. Furthermore, we introduce a novel multimodal classifier-free guidance mechanism that allows the model to adaptively balance information from each modality during speech synthesis. Extensive experiments demonstrate that AlignDiT significantly outperforms existing methods across multiple benchmarks in terms of quality, synchronization, and speaker similarity. Moreover, AlignDiT exhibits strong generalization capability across various multimodal tasks, such as video-to-speech synthesis and visual forced alignment, consistently achieving state-of-the-art performance. The demo page is available at https://mm.kaist.ac.kr/projects/AlignDiT.

MM '25, Dublin, Ireland

ACM ISBN 979-8-4007-2035-2/2025/10
https://doi.org/10.1145/3746027.3754577

CCS Concepts

• **Information systems** → **Multimedia content creation**.

Keywords

Diffusion Transformer, Speech Generation, Multimodal Alignment, Automated Dialogue Replacement

ACM Reference Format:

Jeongsoo Choi, Ji-Hoon Kim, Kim Sung-Bin, Tae-Hyun Oh, and Joon Son Chung. 2025. AlignDiT: Multimodal Aligned Diffusion Transformer for Synchronized Speech Generation. In *Proceedings of the 33rd ACM International Conference on Multimedia (MM '25), October 27–31, 2025, Dublin, Ireland*. ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3746027.3754577

1 Introduction

Human communication involves the exchange of information through spoken language, which serves as a fundamental means of interaction between individuals [19, 28, 50]. It naturally generates multimodal signals that convey both linguistic content and paralinguistic cues. Among them, audio, video, and text are three key modalities widely utilized in speech and language processing. The audio modality delivers the sound signal that carries phonetic content, prosodic features such as intonation and rhythm, as well as speaker-specific characteristics. The video modality captures lip movements, facial expressions, and non-verbal signals that are synchronized with the audio. Text, while not directly observable in the physical world, serves as a symbolic and discrete representation of the linguistic content, monotonically aligned with audio and video over time [58, 64]. Together, these modalities complement one another, offering a comprehensive view of human verbal communication.

Cross-modal generation tasks, where one modality is synthesized or inferred from others, have been extensively studied to support accessibility and multimodal interaction. Typical examples include text-to-speech synthesis (TTS) [6, 8, 51, 63], automatic speech recognition (ASR) [4, 23, 62, 65], lip reading [46, 47, 67, 76], video-to-speech synthesis [10, 34, 52, 75], and talking face generation [56, 61, 72]. While many of these tasks focus on converting a single input modality into another, incorporating multiple input modalities can lead to more accurate and robust generation, as different modalities provide complementary information. For example, audio-visual speech recognition (AVSR) [1, 13, 29] improves transcription accuracy by leveraging both audio and video inputs, especially under noisy conditions.

This motivates generating speech audio conditioned on both video and text inputs, where visual cues from lip movements and explicit linguistic content complement each other. A representative application of this task is Automated Dialogue Replacement (ADR) [15, 16], widely used in film and television post-production. During this process, actors' lines are re-recorded to improve audio quality while ensuring synchronization with the video, especially when original recordings are affected by background noise or challenging recording conditions. Building on recent advances in deep learning based speech generative models, prior studies have explored generating speech from silent video, transcriptions, and short reference audio clips. Despite their potential, existing multimodal-to-speech generation approaches face three key challenges. First, they often struggle to generate natural and intelligible speech due to limited modeling capacity and reliance on limited datasets. Second, they are not robust when one or more input modalities are missing or corrupted, as they lack mechanisms to adjust the importance of each modality. Third, many methods depend on external forced aligners or duration predictors for synchronization, increasing supervision costs and risks propagating alignment errors.

To address these challenges, we propose AlignDiT, a multimodal aligned diffusion transformer architecture designed for natural and synchronized speech generation. AlignDiT jointly models video, text, and reference audio within a unified framework, implicitly learning cross-modal alignments without relying on explicit duration predictors or external forced aligners. By framing speech synthesis as a conditional generative diffusion process, AlignDiT naturally aligns the generated speech with visual lip movements, linguistic content, and speaker-specific voice characteristics. We conduct extensive experiments using both subjective and objective evaluation metrics. The results clearly show that AlignDiT significantly outperforms existing ADR methods across all criteria, including speech intelligibility, synchronization accuracy, and speaker similarity. Furthermore, AlignDiT effectively generalizes to related multimodal tasks, such as video-to-speech synthesis and visual forced alignment, highlighting its robustness and flexibility across diverse multimodal scenarios, as demonstrated in Fig. 1.

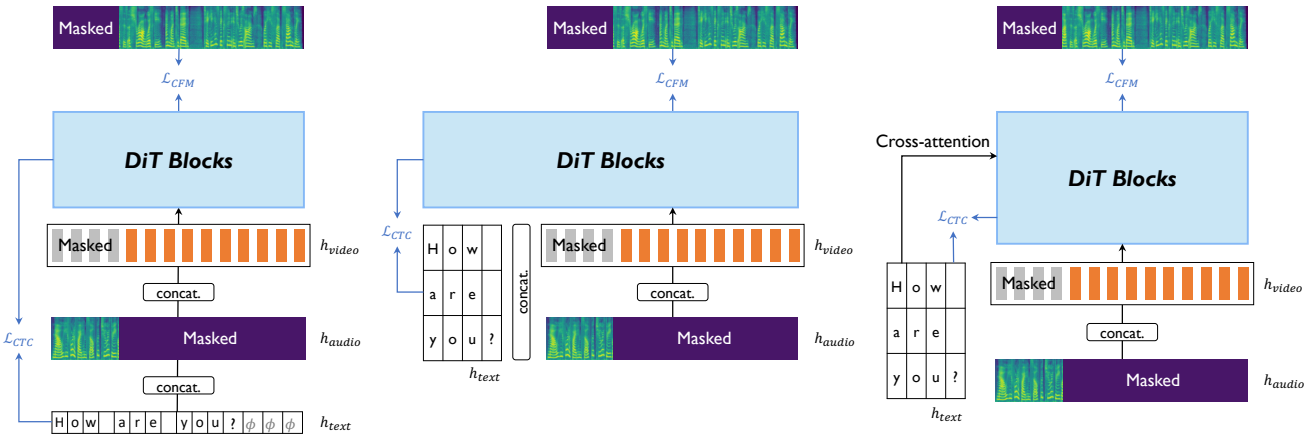
Our major contributions can be summarized as follows:

- We propose AlignDiT, a model that jointly leverages video, text, and reference audio to synthesize accurate, high-quality, and synchronized speech for the ADR task.
- We conduct extensive analyses and experiments to explore various settings and identify the most effective approach for multimodal alignment.
- We demonstrate the versatility of AlignDiT by successfully adapting it to related multimodal tasks, such as video-to-speech and visual forced alignment.

2 Related Works

2.1 Multimodal Speech Tasks

Automated dialogue replacement (ADR). Early efforts in ADR, also known as automated video dubbing, framed it as a multimodal text-to-speech (TTS) problem. Neural Dubber [31] pioneered this direction by generating speech from text conditioned on video lip movements. Subsequent work, VisualTTS [45], integrated visual information more explicitly into the TTS pipeline by introducing a textual-visual attention mechanism to learn alignments between phonemes and lip frames, as well as visual feature fusion during acoustic decoding. VDTTS [24] extended this idea to unconstrained, multi-speaker settings, predicting finer prosodic elements for a more natural dubbing. More recently, HPMDubbing [15] presented a unified architecture employing hierarchical prosody modeling. It extracts visual features at the lip, face, and scene levels to control different aspects of speech (timing, energy/pitch, and global emotion), synthesizing natural and emotional speech and setting a strong state-of-the-art benchmark for ADR. StyleDubber [16] further focused on speaker style and pronunciation habits by introducing two-level style learning (phoneme-level and utterance-level) using reference audio. Both HPMDubbing and StyleDubber require explicit alignment through a forced aligner during training to ensure accurate synchronization between video and speech.



(a) Early Fusion & Self-attention (b) Prefix & Self-attention (c) Multimodal Cross-attention

Figure 2: Various methods for conditioning multimodal inputs to DiT blocks: (a) channel-wise concatenation of text, reference speech, and visual features; (b) providing the text as a prefix; and (c) conditioning text inputs through cross-attention. In (a), ϕ represents filler tokens.

Video-to-speech. Distinct from ADR, video-to-speech generates speech directly from silent videos without transcripts. Initial methods approached this task using CNN-based [20, 38] or sequence-to-sequence models [60] to capture lip movements and generate speech. Recent approaches have further improved speech generation quality by incorporating advanced techniques, such as GANs [35, 53], normalizing flows [25, 34], and diffusion models [9, 75]. Additionally, to capture speaker characteristics, most works [11, 52] utilize speaker embeddings derived from reference audio. As obtaining reference audio during inference is not always feasible, several studies [9, 33, 34, 75] extract speaker information directly from the given video to facilitate speaker-aware speech synthesis.
Visual forced alignment (VFA). VFA is a task that identifies the timeline—specifically, the start and end times—for each word or phoneme in silent videos based on corresponding textual content. VFA requires accurately aligning lip movements with text. One approach employs visual keyword spotting [54, 59], a technique identifying durations of individual words. Repeating this across an entire video achieves visual forced alignment, but involves substantial computational demands and potential inaccuracies due to overlaps between adjacent words. Another common approach [39] utilizes Connectionist Temporal Classification (CTC) [22], frequently employed in visual speech recognition (VSR) for video-text alignment. DVFA [36] represents the first method explicitly designed for VFA, leveraging a multimodal attention mechanism to effectively align textual transcriptions with corresponding lip movements.

Our approach builds on insights from these three different lines of work, aiming to achieve the best of each by learning cross-modal alignment across speech, video, and text. Crucially, we avoid dedicated speech-text aligners, such as those required by HPMDubbing and StyleDubber, by learning implicit multimodal alignment. This results in improved content accuracy and speaker similarity. Furthermore, the flexibility of our multimodal alignment approach allows it to generalize beyond ADR, effectively tackling video-to-speech and VFA, thus demonstrating significant versatility and broad applicability.

2.2 Text-to-speech (TTS) Synthesis

Recent research in Text-to-Speech (TTS) synthesis has achieved significant advancements. One notable approach, autoregressive (AR)-based TTS models [6, 32, 70], combines powerful speech tokenizers [17, 78] with next-token prediction language modeling and has demonstrated promising results. Despite their high quality, non-autoregressive (NAR) models benefit from fast inference due to parallel processing, effectively balancing quality and latency. Specifically, diffusion models [26, 68] have significantly contributed to the success of current NAR approaches. For example, Matcha-TTS [51] adopts conditional flow matching with optimal transport paths (OT-CFM) [42] for training and relies on a phoneme-level duration model for speech synthesis. DiTTo-TTS [41] improves alignment by utilizing a Diffusion Transformer (DiT) [57] with cross-attention conditioned on encoded text from a pretrained language model. E2 TTS [21] removes phoneme and duration predictors, directly using characters padded with filler tokens to match the length of mel spectrograms. Additionally, F5-TTS [8] enhances text-speech alignment by integrating ConvNeXt V2 [71] into a diffusion transformer framework. In this work, we aim to extend NAR TTS models, specifically DiT, to multimodal TTS synthesis. By leveraging classifier-free guidance and carefully designed conditioning, our approach flexibly handles varying inputs—including text-only, video-only, and multimodal scenarios—for high-quality speech synthesis.

3 AlignDiT

Our goal is to generate a speech waveform that matches the lip movements in a video, conveys the content of a given text script, and resembles the voice characteristics of a target speaker indicated by reference speech. We employ an in-context learning-based speech synthesis approach to handle multimodal inputs. For text-to-speech synthesis task, various methods [8, 21, 41, 74] have explored diverse strategies for conditioning text inputs. However, to the best of our knowledge, no existing method simultaneously handles video input alongside text and reference speech to achieve multimodal

alignment for speech synthesis within in-context learning-based generative models. We aim to explore various settings and identify the most suitable method for multimodal alignment, capable of flexibly handling both unimodal and multimodal inputs to synthesize high-quality and accurate speech.

3.1 Model Architecture

We utilize the Diffusion Transformer (DiT) [57] for multimodal-to-speech generation, as it has been shown to be effective for speech generation task. Following prior works about text-to-speech [8, 40], we train the model using a flow matching objective and generate mel-spectrogram through an iterative inference process starting from random noise. The model consists of multiple blocks, each following the standard Transformer [69] architecture with self-attention, along with additional parameters for conditioning diffusion timestep information. For multimodal-to-speech generation, we guide the generative process using a fused multimodal representation as a condition, allowing the model to progressively refine the noisy mel-spectrogram into a representation that aligns with the given multimodal inputs.

3.2 Multimodal Conditioning

Audio-Video Fusion. Audio and video are naturally synchronized modalities. When a person speaks, their lip movements correspond closely to the acoustic features at that same moment, making frame-by-frame fusion straightforward. However, since the frame rates of the two modalities typically differ, we extract and encode each modality to have a common temporal resolution. Specifically, audio is converted into mel-spectrogram sequence h_{audio} at 100 fps. For video, we first extract a sequence of video features h_{video} using a pretrained video encoder specialized in lip motion. To match the frame rate, we upsample the 25 fps video features to 100 fps via lightweight transposed convolutional layers. We then apply Conformer [23] encoder for better capturing contextual information. Once aligned temporally, the audio and video features are concatenated channel-wise to form a unified multimodal representation, which effectively encodes when and how to speak.

To enable our model to generate speech that follows the voice characteristics of a reference speech, we apply a binary temporal mask M to the mel-spectrogram and train the model to inpaint the masked regions, using $(1-M)\odot h_{audio}$ as input. The masked spans are randomly selected to enhance the in-context learning ability of the model. Since requiring paired audio-video data of reference utterance during inference is critical, we train the model to operate with reference speech alone by applying complementary masking to the video features. Specifically, the input becomes $M \odot h_{video}$, allowing the model to rely on audio while ignoring the masked video during training. This audio-video fusion can be formulated as follows:

$$h_{av} = [(1-M)\odot h_{audio}; M \odot h_{video}] \in \mathbb{R}^{T\times 2D}, \tag{1}$$

where T and D denote the temporal length and the hidden dimension, respectively.

Audio-Video-Text Fusion. While the audio-video streams are inherently time-aligned, text is only monotonically aligned with them and lacks strict frame-level synchronization. Unlike previous methods that rely on text token-level duration information from external forced aligners [49] and duration predictors [63], we aim to train the generative model fuse the modalities naturally, without explicit duration constraints. This simplifies the data preprocessing pipeline, making it easier to scale the dataset, and avoids potential biases introduced by forced alignments, thereby facilitating a more natural audio-video-text alignment. To investigate feasible modality fusion methods in the DiT blocks, we explore three conditioning strategies as illustrated in Fig. 2. In all cases, for text encoding, the character sequence is first embedded through a lookup table and then refined by a convolutional encoder to be h_{text} which has the length of L.

(a) Early Fusion & Self-Attention: A naive approach for fusion is concatenating all conditioning modalities along the channel axis. To match the total length, we add filler tokens at the end of the text sequence inspired by E2 TTS [21]. After this early fusion, the self-attention layers in DiT blocks learn alignment naturally to predict the masked part of mel-spectrogram. This can be expressed as follows:

$$h' = [h_{av}; h_{text}] \in \mathbb{R}^{T\times 3D}, \tag{2}$$

$$h = h'W_a + b_a \in \mathbb{R}^{T\times D}, \tag{3}$$

where W_a and b_a are the parameters of a fully connected layer to transform the channel dimension to D.

(b) Prefix & Self-Attention: Another approach for fusion is frame-wise concatenation between text feature and audio-video feature, treating the text as a prefix. This leverages in-context conditioning [70, 74], allowing the text prefix to guide the model within the self-attention layers of DiT blocks. After the model outputs the full sequence, the prefix part corresponding to the text length is discarded, yielding the predicted mel-spectrogram. This can be formulated as follows:

$$h' = \mathrm{Concat}(h_{text}, h_{av}) \in \mathbb{R}^{(L+T)\times 2D}, \tag{4}$$

$$h = h'W_b + b_b \in \mathbb{R}^{(L+T)\times D}, \tag{5}$$

where W_b and b_b are the parameters of a fully connected layer.

(c) Multimodal Cross-Attention (AlignDiT): Lastly, instead of fusing the audio-video and text features at the input level, we design the model to gradually incorporate text information while preserving the natural synchronization between audio and video. To this end, we revise each DiT block by inserting a cross-attention layer in addition to self-attention. In this setup, the audio-video representation h_{av} is used as the query, while the text embedding h_{text} serves as the key and value in a multi-head cross-attention mechanism as follows:

$$h = \mathrm{MHCA}(h_{av}W_Q, h_{text}W_K, h_{text}W_V) \in \mathbb{R}^{T\times D}, \tag{6}$$

where W_Q, W_K, and W_V are learnable projection matrices. By analyzing each approach through (a)-(c) in Section 5.1, we observe that (c) naturally aligns audio-visual features while effectively incorporating text features, achieving the best multimodal alignment. Therefore, we adopt this variant in our AlignDiT model.

3.3 Training Objective

Based on the fused multimodal representations, we train our AlignDiT with multi-task learning. First, we adopt the conditional flow

matching (CFM) training objective, which has been proven its effectiveness in generating high-quality data samples in an efficient manner [42]. The CFM seeks to match a probability path p_t from a tractable distribution p_0 to p_1 approximating the target distribution. Given the fused multimodal representation h and a noisy mel-spectrogram x_t ($t \in [0, 1]$), AlignDiT is trained to regress the vector field u_t with CFM objectives:

$$\mathcal{L}_{CFM} = \mathbb{E}_{t,p_t} \|v_t(x_t|h, \theta) - u_t(x_t)\|^2, \tag{7}$$

where θ defines the DiT blocks and $v_t(x_t|h, \theta)$ denotes the estimated vector fields with $x_t \sim p_t(x)$.

Second, while the CFM loss encourages natural mel-spectrogram generation, relying solely on the CFM loss can be insufficient for modality alignment, as it provides only an indirect learning signal [12]. To address this, we introduce a Connectionist Temporal Classification (CTC) [22] loss to guide the intermediate representations of DiT blocks to align more directly with the textual content. Specifically, we attach lightweight projection heads to several intermediate blocks to predict the text sequence from their hidden representations. The loss can be expressed as:

$$\mathcal{L}_{\text{CTC}} = -\sum_{i \in \mathcal{I}} \log p_{\text{CTC}}(\text{text} \mid h^i), \tag{8}$$

where h^i is the output of the i-th DiT block and $\mathcal{I}$ denotes the selected layers. The CTC loss encourages the model to retain more linguistic information, without employing external alignment modules. The total multi-task loss is defined as follows:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{CFM}} + \lambda_{\text{CTC}} \mathcal{L}_{\text{CTC}}, \tag{9}$$

where λ_{CTC} is a balancing hyperparameter.

3.4 Audio-only Pretraining

Generating high-quality speech and learning multimodal alignment simultaneously can be challenging. To provide a better initialization for AlignDiT, we employ an audio-only pretraining phase before conditioning on video and text. The model learns to predict masked regions of the input mel-spectrogram from unmasked context during pretraining, and this has demonstrated its effectiveness in text-to-speech task [43]. In addition, we follow the approach [77] that improves convergence speed by distilling rich features into the intermediate layers of DiT, using a self-supervised speech model [30] as the teacher. By allowing the model to acquire unconditional speech generation abilities prior to multimodal conditioning, this pretraining phase simplifies the subsequent audio-video-text fusion. Additionally, since audio-only data is abundant compared to audio-video-text paired data, our audio-only pretraining strategy is readily scalable and offers a practical way to improve the performance.

3.5 Multimodal Classifier-Free Guidance

In diffusion-based generative models, classifier-free guidance [27] is well-explored to strengthen the influence of conditioning signals during inference. This is achieved by using both conditional and unconditional predictions from the same model to guide the generation process as follows:

$$v_{t,CFG} = v_t(x_t, h) + s \cdot (v_t(x_t, h) - v_t(x_t, \emptyset)), \tag{10}$$

where s is guidance scale.

Since each modality exhibits different characteristics, we hypothesize that using a single guidance scale for all modalities may be sub-optimal. To allow better control over each modality during inference, we propose multimodal classifier-free guidance by assigning modality-specific guidance scales:

$$\begin{aligned} v_{t,CFG} = v_t(x_t, h_{text}, h_{video}) \\ + s_{video} \cdot (v_t(x_t, h_{text}, h_{video}) - v_t(x_t, h_{text}, \emptyset)) \\ + s_{text} \cdot (v_t(x_t, h_{text}, \emptyset) - v_t(x_t, \emptyset, \emptyset)), \end{aligned} \tag{11}$$

where s_{text} is guidance scale for text modality and s_{video} for video. By adjusting s_{text} and s_{video}, we can adaptively control the focus between modalities. Higher s_{text} encourages the model to follow the text more closely, improving intelligibility, while higher s_{video} leads to better lip synchronizations.

To support CFG, we apply modality dropout during training by randomly dropping text, video, or both. This not only enables multimodal CFG but also improves robustness in cases where a modality may be missing.

4 Experiments

4.1 Datasets

We train and evaluate our proposed AlignDiT on the large-scale audio-visual dataset LRS3 [2], which contains 439 hours of English sentence-level data sourced from TED and TEDx talk videos from thousands of speakers. Each video segment contains unconstrained audio-visual speech paired with an accurate transcript, making the dataset suitable for our multimodal setting. Around 131,000 utterances are utilized for training. For evaluation, we construct an LRS3-cross test set consisting of triplets of {reference speech, text, silent video}, where the reference speech is from a different utterance by the same speaker, rather than using the ground-truth speech as the reference. This prevents the model from accessing ground-truth speech during testing, ensuring a rigorous evaluation without information leakage. For video-to-speech, instead of using the provided text from the dataset, we utilize an off-the-shelf lip reading model to extract transcripts directly from the silent videos.

4.2 Evaluation Metrics

Subjective metrics. Since our primary focus is generating natural and synchronized speech, subjective evaluation is essential for accurately assessing model performance [3, 48]. Thus, we conduct human evaluations using Mean Opinion Scores (MOS) for the ADR task along two criteria: naturalness, evaluating the overall quality of the speech; and similarity, assessing speaker similarity between the reference speech and synthesized speech. 20 participants rate the randomly sampled 30 utterances using a 5-point Likert scale, where 1 indicates "very poor" (or "very different") and 5 indicates "very good" (or "very similar").

Objective metrics. We employ comprehensive objective metrics to evaluate the synthesized speech. For both ADR and video-to-speech tasks, we report Word Error Rate (WER) to assess content accuracy, using Whisper-large-v3 [62] to transcribe the synthesized speech. Speaker similarity (spkSIM) is measured by computing the cosine similarity between speaker embeddings extracted from synthesized and reference speech using a WavLM-large-based speaker verification model [7]. Lip-sync accuracy (AVSync) between the

Table 1: Ablation study of modality conditioning method.

Method	$\mathcal{L}_{CTC}$	WER ↓	AVSync ↑	spkSIM ↑
(a)	✗	28.828	0.666	0.530
	✓	2.236	0.748	0.511
(b)	✗	5.456	0.726	0.536
	✓	1.917	0.745	0.519
(c)	✗	2.507	0.745	**0.543**
	✓	**1.401**	**0.751**	0.515

Table 2: Ablation study of audio-only pretraining.

Pretraining data.	WER ↓	AVSync ↑	spkSIM ↑
✗	2.040	0.746	0.396
LRS3	1.720	0.750	0.503
LibriSpeech	**1.401**	**0.751**	**0.515**

video and synthesized speech is measured using AVHubert [66]. Specifically, we compute the cosine similarity between AVHubert features extracted from video paired with ground-truth speech and features extracted from video paired with synthesized speech. This assessment has been shown in [73] to be more robust and effective for validating audio-visual synchronization compared to conventional lip-sync accuracy metrics based on SyncNet [14]. For the alignment task, we evaluate alignment accuracy by comparing word-level timestamps obtained from our method with ground truth timestamps derived from original audio alignment. We report the average absolute time error per word (in milliseconds).

4.3 Implementation Details

Data preprocessing. For audio features, we use 16 kHz mono audio and convert it into 80 bins mel-spectrogram using a filter size of 640, a hop size of 160, resulting in a frame rate of 100 Hz. We utilize 25 fps video and extract visual features as follows. Face detection is performed using RetinaFace [18], followed by facial landmark extraction using FAN [5]. We crop the lip-centered region based on the detected landmarks and resize it into 88 × 88. The pretrained AV-HuBERT (Large) model [66] is utilized to extract 25 Hz visual representations, which result in a fixed 1:4 length ratio to the audio features. For text, we represent input as a sequence of characters. Compared to the audio features, the character sequence generally has a shorter temporal length [64].

Architecture. The visual feature encoder is composed of two transposed convolution layers with stride 2, followed by two Conformer [23] encoder layers, each with embedding size of 512, 4 attention heads, and a 1024-dimensional feed-forward layer. For the text encoder, we adopt 4 layers of ConvNeXt v2 [71] with 512-dimensional hidden embeddings. After concatenating multimodal features, we apply a linear layer to project the fused representation. This representation is then processed by 18 DiT blocks, each with 768-dimensional embedding size, 12 attention heads, and a 3072-dimensional feed-forward layer.

Training. We train AlignDiT using AdamW [44] optimizer with a warmup of 20k steps to a peak learning rate of 7.5×10^{-5}, followed by linearly decay. Pretraining is conducted for 500k steps on audio-only data with a total batch size of 0.3 hours. We finetune the model for

Table 3: Ablation study of multimodal CFG.

s_{text}	s_{video}	WER ↓	AVSync ↑	spkSIM ↑
0	0	3.785	0.716	0.391
2	2	2.310	0.758	0.497
5	2	**1.401**	0.751	**0.515**
5	5	2.507	**0.760**	0.501

Table 4: Ablation study of input modalities.

Input mod. text	Input mod. video	WER ↓	AVSync ↑	spkSIM ↑
✗	✓	27.820	0.674	0.486
✓	✗	1.769	0.270	0.501
✓	✓	**1.401**	**0.751**	**0.515**

400k steps using paired audio, video, and text data with a total batch size of 0.1 hours. λ_{CTC} is set to 0.1 to balance the CTC loss against the CFM loss in initial stage. The modality dropout probability is set to 0.2 for text, video, and all modalities, respectively.

Inference. During inference, we apply Exponential Moving Averaged (EMA) weights with a decay rate of 0.999 to stabilize the model prediction. AlignDiT takes a reference speech along with its corresponding transcript as inputs, and the total duration is determined by the input video length. For sampling, we use the Euler ODE solver with timestep scaling based on sway sampling strategy with a coefficient of -1, following F5-TTS [8]. To convert generated mel spectrograms into waveforms, we employ a HiFi-GAN [37] model trained on the LRS3 dataset.

4.4 Baseline Models

We compare AlignDiT with state-of-the-art open-source ADR systems, HPMDubbing [15] and StyleDubber [16]. To ensure a fair comparison and improve their generalization capability, we train these models on the LRS3 dataset [2], which is significantly larger than the datasets originally used in their training. Additionally, we replace the lip feature extractor in each model with AV-HuBERT (large) [66], unifying the lip feature extractor across AlignDiT, HPMDubbing, and StyleDubber to purely evaluate and compare their effectiveness under identical settings. Note that both baseline models require an explicit duration aligner during training. For video-to-speech, we compare our method with DiffV2S [9], Intelligible [11], and LipVoicer [75], which are specifically designed for this task. For visual forced alignment, we compare against visual keyword spotting methods such as KWS-Net [54] and Transpotter [59], as well as CTC-based methods [39] and DVFA [36].

5 Results

5.1 Ablation Studies

We analyze the contribution of each AlignDiT component through in-depth ablation studies. We conduct a series of experiments using various objective metrics, *i.e.* WER, AVSync, and spkSIM, and present comprehensive findings on modality conditioning method, audio-only pretraining, multimodal CFG, and input modalities.

Modality conditioning method. To validate the effectiveness of the conditioning strategy of AlignDiT, we carry out a detailed analysis of alternative modality conditioning strategies. We compare

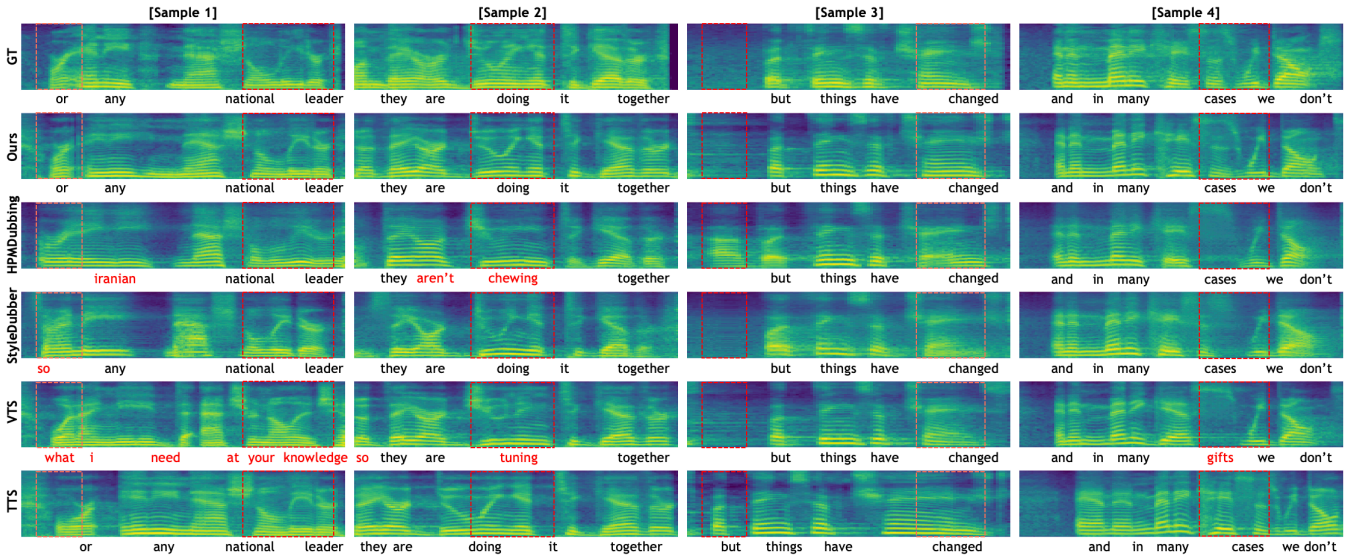


Figure 3: We qualitatively compare mel-spectrogram visualizations of ground-truth speech and synthesized speech from the ADR task using AlignDiT (ours) and existing methods, HPMDubbing and StyleDubber. We also provide results from unimodal inputs, either video-only (VTS) or text-only (TTS). The text below each mel-spectrogram represents time-aligned speech extracted using Whisper-large-v3, with red text indicating incorrectly synthesized word compared to the ground truth.

Table 5: Experimental results on LRS3-cross set. Subjective MOS results are presented with 95% confidence interval. ↑ represents that a higher score is better, and ↓ denotes that a lower score is better. The best-performing result is shown in bold.

Method	Subjective		Objective		
	Naturalness ↑	Speaker Similarity ↑	WER ↓	AVSync ↑	spkSIM ↑
GT	4.13±0.13	3.96±0.14	2.335	1.000	0.562
HPMDubbing [15]	2.37±0.14	2.43±0.16	5.382	0.750	0.287
StyleDubber [16]	1.79±0.12	2.35±0.18	3.170	0.586	0.349
Ours	**3.79±0.16**	**3.96±0.13**	**1.401**	**0.751**	**0.515**

three conditioning variants—(a) early fusion & self-attention, (b) prefix & self-attention, and (c) multimodal cross-attention—and analyze the impact of the CTC loss ($\mathcal{L}_{\text{CTC}}$) in each case. The results in Table 1 confirm the strength of the proposed conditioning method (multimodal cross-attention) which achieves the best performance in WER, AVSync, and spkSIM.

In addition, the absence of $\mathcal{L}_{CTC}$ leads to consistent quality degradation across all variants and metrics, except for a slight deviance in spkSIM of the (c) multimodal cross-attention. In particular, the WER is significantly worsened in every variant, indicating that applying $\mathcal{L}_{CTC}$ facilitates more accurate alignment learning between the input text and output speech within the DiT blocks.

Audio-only pretraining. Table 2 illustrates the benefits of audio-only pretraining. Pretraining AlignDiT on the LRS3 dataset, which is also used during the main training phase, shows consistent quality improvements across all evaluation metrics. Even greater improvements are achieved when using LibriSpeech [55], a dataset not seen during the subsequent training. These results underscore the benefits of leveraging diverse audio data during pretraining, and demonstrate that the performance of AlignDiT can be readily enhanced using easily accessible audio resources.

Multimodal CFG. In order to explore how AlignDiT benefits from its multimodal CFG, we experiment with different values of s_{text} and s_{video}. From the results in Table 3, we derive two key findings. First, applying CFG ($s_{text} = s_{video} \in \{2, 5\}$) consistently improves performance compared to not using CFG ($s_{text} = s_{video} = 0$), aligning with observations from previous works [27, 40]. More importantly, leveraging the proposed multimodal CFG, i.e., adjusting s_{text} and s_{video} to balance information from each modality, further enhances overall performance, demonstrating its effectiveness in synthesizing more natural speech. Through analysis of optimal guidance scales for each modality, we set $s_{text} = 5$ and $s_{text} = 2$, which achieves the best quality in terms of WER and spkSIM.

Input modalities. To examine the effect of multimodal inputs, we compare models trained with video only, text only, and both text and video modalities (Table 4). The WER significantly increases when text input is omitted, highlighting the essential role of textual information in generating accurate and intelligible speech. Excluding the video modality also degrades overall performance, particularly in terms of AVSync. In contrast, utilizing both modalities yields speech that is not only intelligible but also well-synchronized, underscoring the advantages of multimodal-to-speech generation.

Table 6: Video-to-Speech benchmark.

Method	WER ↓	AVSync ↑	spkSIM ↑
DiffV2S [9]	35.210	0.608	0.115
Intelligible [11]	27.432	0.675	0.316
LipVoicer [75]	21.164	0.524	0.094
Ours	**19.513**	**0.688**	**0.508**

Table 7: Visual forced alignment benchmark.

Method	MAE ↓	ACC ↑
KWS-Net [54]	262.9ms	42.6%
CTC-based [39]	124.5ms	60.6%
Transpotter [59]	167.3ms	61.8%
DVFA [36]	97.7ms	80.2%
Ours	**41.5ms**	**83.7%**

5.2 Quantitative Comparison

Table 5 shows both subjective and objective evaluation results of AlignDiT compared to baseline systems. As shown, AlignDiT consistently outperforms all baselines across all evaluation metrics. To be specific, in the subjective evaluation, our method achieves a naturalness score of 3.79 and a speaker similarity score of 3.96, substantially surpassing the baselines by a large margin. These results suggest that AlignDiT generates fluent speech that is perceptually superior in both naturalness and speaker similarity.

The objective evaluation further supports the effectiveness of AlignDiT. It achieves the lowest WER of 1.401, demonstrating its ability to accurately synthesize the input text. It also obtains the highest AVSync score of 0.751. This indicates AlignDiT effectively leans audio-video temporal alignment, without depending on additional aligners or duration predictors which commonly used in the baselines. In terms of speaker consistency, AlignDiT also achieves the best spkSIM score, demonstrating its ability to effectively mimic the voice characteristics of the target speaker.

5.3 Qualitative Comparison

We visually compare the mel-spectrograms converted from synthesized speech produced by existing ADR methods [15, 16] and our AlignDiT with multimodal input, alongside those from ground-truth speech, in Fig. 3. We also include results synthesized from unimodal inputs: video-only (VTS) and text-only (TTS). Focusing on the orange and red boxes in columns 1 and 2, as well as the transcribed speech below each mel-spectrogram, we observe that existing methods frequently generate incorrect speech and blurry spectrograms. In columns 3 and 4, although existing methods produce accurate speech content, they occasionally generate unintended sounds or produce mel-spectrograms that lack details and appear blurry. Regarding VTS and TTS, we observe consistent results in Sec. 5.1. For VTS across all samples, since no text input is provided, content accuracy becomes significantly unstable, leading to incorrect speech. Conversely, TTS generates accurate speech content across all examples due to the text input, yet the absence of video input results in temporal misalignment compared to the ground-truth. In contrast, our model consistently synthesizes accurate speech content, exhibiting fine details in the mel-spectrogram that closely match the ground-truth across all examples. These results clearly demonstrate that our model effectively synthesizes high-quality, accurate speech comparable to the ground-truth.

5.4 Applications

Our proposed AlignDiT is robust and flexible across diverse multimodal scenarios. We demonstrate the versatility of our method by showing its effective generalization to related multimodal tasks, such as video-to-speech synthesis and visual forced alignment.

Video-to-speech. Unlike ADR, video-to-speech takes silent video (without text) as input to read lip movements and synthesize corresponding speech. Some prior works [9, 11] operate in textless manner, while others, such as LipVoicer [75], leverage off-the-shelf lip reading models to guide speech synthesis. For a fair comparison with LipVoicer, we adopt the same lip reading model [46] to obtain pseudo-text labels. Table 6 summarizes the performance comparison between our approach and existing methods. Interestingly, our model significantly outperforms existing methods across all evaluation metrics, including LipVoicer, which is specifically designed for this task and also utilizes an expert lip reading model. This highlights the robustness of our multimodal alignment approach, demonstrating that it generalizes effectively and achieves superior performance even on tasks beyond its primary design.

Visual forced alignment (VFA). Conventionally, the VFA task involves identifying timestamps for each word or phoneme in silent videos. Since our model can synthesize speech given silent video and text input, we bypass direct comparison of silent video and text, and instead leverage these inputs to generate speech signals. We then apply the Montreal Forced Aligner [49] to align the synthesized speech with the corresponding text, thus determining timestamps of each word for VFA task. It is worth noting that we use a single canonical reference speech across all test samples, removing the need for speaker-specific references. Table 7 presents a comparison of alignment performance on the LRS3 dataset. The synthesized speech from our proposed AlignDiT is highly synchronized with the input video, enabling precise forced alignment with text, resulting in significantly better performance compared to existing methods [36, 39, 54, 59] specifically designed for this task. These results support that our model generates highly accurate and temporally synchronized speech from multimodal inputs.

6 Conclusion

We introduced AlignDiT, a unified framework for generating accurate, natural, and synchronized speech from text, video, and reference audio. Through extensive analysis, we explore various configurations and identify the most effective strategy for aligning multiple modalities, without the need for explicit duration modeling. We also proposed a multimodal classifier-free guidance mechanism that adaptively balances information across modalities. AlignDiT achieves state-of-the-art performance across several benchmarks and demonstrates its effectiveness in key multimodal tasks, including video-to-speech synthesis and visual forced alignment. We believe our findings offer valuable insights for future research in multimodal alignment and generation.

Acknowledgments

This work was supported by IITP grants funded by the Korean government (MSIT, RS-2025-02263169, Detection and Prediction of Emerging and Undiscovered Voice Phishing and RS-2024-00457882, National AI Research Lab Project).

References

[1] Triantafyllos Afouras, Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Zisserman. 2018. Deep audio-visual speech recognition. *IEEE Trans. on Pattern Analysis and Machine Intelligence* (2018).

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

[3] Junseok Ahn, Youkyum Kim, Yeunju Choi, Doyeop Kwak, Ji-Hoon Kim, Seongkyu Mun, and Joon Son Chung. 2024. VoxSim: A perceptual voice similarity dataset. In *Proc. Interspeech*.

[4] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Proc. NeurIPS*.

[5] Adrian Bulat and Georgios Tzimiropoulos. 2017. How Far are We from Solving the 2D & 3D Face Alignment Problem?(and a Dataset of 230,000 3D Facial Landmarks). In *Proc. ICCV*.

[6] Sanyuan Chen, Shujie Liu, Long Zhou, Yanqing Liu, Xu Tan, Jinyu Li, Sheng Zhao, Yao Qian, and Furu Wei. 2024. Vall-e 2: Neural codec language models are human parity zero-shot text to speech synthesizers. *arXiv preprint arXiv:2406.05370* (2024).

[7] Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al. 2022. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing* (2022).

[8] Yushen Chen, Zhikang Niu, Ziyang Ma, Keqi Deng, Chunhui Wang, Jian Zhao, Kai Yu, and Xie Chen. 2025. F5-tts: A fairytaler that fakes fluent and faithful speech with flow matching. In *Proc. ACL*.

[9] Jeongsoo Choi, Joanna Hong, and Yong Man Ro. 2023. DiffV2S: Diffusion-based video-to-speech synthesis with vision-guided speaker embedding. In *Proc. ICCV*.

[10] Jeongsoo Choi, Ji-Hoon Kim, Jinyu Li, Joon Son Chung, and Shujie Liu. 2025. V2SFlow: Video-to-Speech Generation with Speech Decomposition and Rectified Flow. In *Proc. ICASSP*.

[11] Jeongsoo Choi, Minsu Kim, and Yong Man Ro. 2023. Intelligible Lip-to-Speech Synthesis with Speech Units. In *Proc. Interspeech*.

[12] Jeongsoo Choi, Zhikang Niu, Ji-Hoon Kim, Chunhui Wang, Joon Son Chung, and Xie Chen. 2025. Accelerating Diffusion-based Text-to-Speech Model Training with Dual Modality Alignment. In *Proc. Interspeech*.

[13] Jeongsoo Choi, Se Jin Park, Minsu Kim, and Yong Man Ro. 2024. Av2av: Direct audio-visual speech to audio-visual speech translation with unified audio-visual speech representation. In *Proc. CVPR*.

[14] Joon Son Chung and Andrew Zisserman. 2016. Out of time: automated lip sync in the wild. In *ACCV workshop*.

[15] Gaoxiang Cong, Liang Li, Yuankai Qi, Zheng-Jun Zha, Qi Wu, Wenyu Wang, Bin Jiang, Ming-Hsuan Yang, and Qingming Huang. 2023. Learning to dub movies via hierarchical prosody models. In *Proc. CVPR*.

[16] Gaoxiang Cong, Yuankai Qi, Liang Li, Amin Beheshti, Zhedong Zhang, Anton van den Hengel, Ming-Hsuan Yang, Chenggang Yan, and Qingming Huang. 2024. Styledubber: towards multi-scale style learning for movie dubbing. In *Findings of ACL*.

[17] Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. 2023. High Fidelity Neural Audio Compression. *Trans. on Machine Learning Research* (2023).

[18] Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. 2020. Retinaface: Single-shot multi-level face localisation in the wild. In *Proc. CVPR*.

[19] Linda Drijvers and Judith Holler. 2023. The multimodal facilitation effect in human communication. *Psychonomic Bulletin & Review* (2023).

[20] Ariel Ephrat and Shmuel Peleg. 2017. Vid2speech: speech reconstruction from silent video. In *Proc. ICASSP*.

[21] Sefik Emre Eskimez, Xiaofei Wang, Manthan Thakker, Canrun Li, Chung-Hsien Tsai, Zhen Xiao, Hemin Yang, Zirun Zhu, Min Tang, Xu Tan, et al. 2024. E2 tts: Embarrassingly easy fully non-autoregressive zero-shot tts. In *IEEE Spoken Language Technology workshop*.

[22] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. 2006. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proc. ICML*.

[23] Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, et al. 2020. Conformer: Convolution-augmented transformer for speech recognition. In *Proc. Interspeech*.

[24] Michael Hassid, Michelle Tadmor Ramanovich, Brendan Shillingford, Miaosen Wang, Ye Jia, and Tal Remez. 2022. More than words: In-the-wild visually-driven prosody for text-to-speech. In *Proc. CVPR*.

[25] Jinzheng He, Zhou Zhao, Yi Ren, Jinglin Liu, Baoxing Huai, and Nicholas Yuan. 2022. Flow-based unconstrained lip to speech generation. In *Proc. AAAI*.

[26] Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Proc. NeurIPS* (2020).

[27] Jonathan Ho and Tim Salimans. 2021. Classifier-Free Diffusion Guidance. In *NeurIPS workshop*.

[28] Judith Holler and Stephen C Levinson. 2019. Multimodal language processing in human communication. *Trends in Cognitive Sciences* (2019).

[29] Joanna Hong, Minsu Kim, Jeongsoo Choi, and Yong Man Ro. 2023. Watch or listen: Robust audio-visual speech recognition with visual corruption modeling and reliability scoring. In *Proc. CVPR*.

[30] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Trans. on Audio, Speech, and Language Processing* (2021).

[31] Chenxu Hu, Qiao Tian, Tingle Li, Wang Yuping, Yuxuan Wang, and Hang Zhao. 2021. Neural dubber: Dubbing for videos according to scripts. In *Proc. NeurIPS*.

[32] Eugene Kharitonov, Damien Vincent, Zalán Borsos, Raphaël Marinier, Sertan Girgin, Olivier Pietquin, Matt Sharifi, Marco Tagliasacchi, and Neil Zeghidour. 2023. Speak, read and prompt: High-fidelity text-to-speech with minimal supervision. *Trans. of the Association for Computational Linguistics* (2023).

[33] Ji-Hoon Kim, Jeongsoo Choi, Jaehun Kim, Chaeyoung Jung, and Joon Son Chung. 2025. From Faces to Voices: Learning Hierarchical Representations for High-quality Video-to-Speech. In *Proc. CVPR*.

[34] Ji-Hoon Kim, Jaehun Kim, and Joon Son Chung. 2024. Let there be sound: reconstructing high quality speech from silent videos. In *Proc. AAAI*.

[35] Minsu Kim, Joanna Hong, and Yong Man Ro. 2021. Lip to speech synthesis with visual context attentional GAN. In *Proc. NeurIPS*.

[36] Minsu Kim, Chae Won Kim, and Yong Man Ro. 2023. Deep visual forced alignment: learning to align transcription with talking face video. In *Proc. AAAI*.

[37] Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. 2020. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. In *Proc. NeurIPS*.

[38] Yaman Kumar, Rohit Jain, Khwaja Mohd Salik, Rajiv Ratn Shah, Yifang Yin, and Roger Zimmermann. 2019. Lipper: Synthesizing thy speech using multi-view lipreading. In *Proc. AAAI*.

[39] Ludwig Kürzinger, Dominik Winkelbauer, Lujun Li, Tobias Watzel, and Gerhard Rigoll. 2020. Ctc-segmentation of large corpora for german end-to-end speech recognition. In *International Conference on Speech and Computer*.

[40] Matthew Le, Apoorv Vyas, Bowen Shi, Brian Karrer, Leda Sari, Rashel Moritz, Mary Williamson, Vimal Manohar, Yossi Adi, Jay Mahadeokar, et al. 2023. Voicebox: Text-guided multilingual universal speech generation at scale. In *Proc. NeurIPS*.

[41] Keon Lee, Dong Won Kim, Jaehyeon Kim, and Jaewoong Cho. 2025. Ditto-tts: Efficient and scalable zero-shot text-to-speech with diffusion transformer. In *Proc. ICLR*.

[42] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. 2023. Flow matching for generative modeling. In *Proc. ICLR*.

[43] Alexander H Liu, Matt Le, Apoorv Vyas, Bowen Shi, Andros Tjandra, and Wei-Ning Hsu. 2024. Generative pre-training for speech with flow matching. In *Proc. ICLR*.

[44] Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. In *Proc. ICLR*.

[45] Junchen Lu, Berrak Sisman, Rui Liu, Mingyang Zhang, and Haizhou Li. 2022. Visualtts: Tts with accurate lip-speech synchronization for automatic voice over. In *Proc. ICASSP*.

[46] Pingchuan Ma, Alexandros Haliassos, Adriana Fernandez-Lopez, Honglie Chen, Stavros Petridis, and Maja Pantic. 2023. Auto-avsr: Audio-visual speech recognition with automatic labels. In *Proc. ICASSP*.

[47] Pingchuan Ma, Yujiang Wang, Jie Shen, Stavros Petridis, and Maja Pantic. 2021. Lip-reading with densely connected temporal convolutional networks. In *Proc. WACV*.

[48] Soumi Maiti, Yifan Peng, Takaaki Saeki, and Shinji Watanabe. 2023. Speechlmscore: Evaluating speech generation using speech language model. In *Proc. ICASSP*.

[49] Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger. 2017. Montreal forced aligner: Trainable text-speech alignment using kaldi.. In *Proc. Interspeech*.

[50] Harry McGurk and John MacDonald. 1976. Hearing lips and seeing voices. *Nature* (1976).

[51] Shivam Mehta, Ruibo Tu, Jonas Beskow, Éva Székely, and Gustav Eje Henter. 2024. Matcha-TTS: A fast TTS architecture with conditional flow matching. In *Proc. ICASSP*.

[52] Rodrigo Mira, Alexandros Haliassos, Stavros Petridis, Björn W Schuller, and Maja Pantic. 2022. SVTS: Scalable Video-to-Speech Synthesis. In *Proc. Interspeech*.

[53] Rodrigo Mira, Konstantinos Vougioukas, Pingchuan Ma, Stavros Petridis, Björn W Schuller, and Maja Pantic. 2022. End-to-end video-to-speech synthesis using generative adversarial networks. *IEEE Trans. on Cybernetics* (2022).

[54] Liliane Momeni, Triantafyllos Afouras, Themos Stafylakis, Samuel Albanie, and Andrew Zisserman. 2020. Seeing wake words: Audio-visual keyword spotting. In *Proc. BMVC*.

[55] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: an asr corpus based on public domain audio books. In *Proc. ICASSP*.

[56] Se Jin Park, Minsu Kim, Joanna Hong, Jeongsoo Choi, and Yong Man Ro. 2022. Synctalkface: Talking face generation with precise lip-syncing via audio-lip memory. In *Proc. AAAI*.

[57] William Peebles and Saining Xie. 2023. Scalable diffusion models with transformers. In *Proc. CVPR*.

[58] Stavros Petridis, Themos Stafylakis, Pingchuan Ma, Georgios Tzimiropoulos, and Maja Pantic. 2018. Audio-visual speech recognition with a hybrid ctc/attention architecture. In *IEEE Spoken Language Technology workshop*.

[59] KR Prajwal, Liliane Momeni, Triantafyllos Afouras, and Andrew Zisserman. 2021. Visual keyword spotting with attention. In *Proc. BMVC*.

[60] KR Prajwal, Rudrabha Mukhopadhyay, Vinay P Namboodiri, and CV Jawahar. 2020. Learning individual speaking styles for accurate lip to speech synthesis. In *Proc. CVPR*.

[61] KR Prajwal, Rudrabha Mukhopadhyay, Vinay P Namboodiri, and CV Jawahar. 2020. A lip sync expert is all you need for speech to lip generation in the wild. In *Proc. ACM MM*.

[62] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proc. ICML*.

[63] Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. 2021. Fastspeech 2: Fast and high-quality end-to-end text to speech. *Proc. ICLR* (2021).

[64] Yi Ren, Yangjun Ruan, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. 2019. Fastspeech: Fast, robust and controllable text to speech. In *Proc. NeurIPS*.

[65] Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. 2019. wav2vec: Unsupervised pre-training for speech recognition. In *Proc. Interspeech*.

[66] Bowen Shi, Wei-Ning Hsu, Kushal Lakhotia, and Abdelrahman Mohamed. 2022. Learning audio-visual speech representation by masked multimodal cluster prediction. In *Proc. ICLR*.

[67] Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Zisserman. 2017. Lip reading sentences in the wild. In *Proc. CVPR*.

[68] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. 2020. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456* (2020).

[69] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proc. NeurIPS*.

[70] Chengyi Wang, Sanyuan Chen, Yu Wu, Ziqiang Zhang, Long Zhou, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, et al. 2023. Neural codec language models are zero-shot text to speech synthesizers. *arXiv preprint arXiv:2301.02111* (2023).

[71] Sanghyun Woo, Shoubhik Debnath, Ronghang Hu, Xinlei Chen, Zhuang Liu, In So Kweon, and Saining Xie. 2023. Convnext v2: Co-designing and scaling convnets with masked autoencoders. In *Proc. CVPR*.

[72] Sicheng Xu, Guojun Chen, Yu-Xiao Guo, Jiaolong Yang, Chong Li, Zhenyu Zang, Yizhong Zhang, Xin Tong, and Baining Guo. 2024. Vasa-1: Lifelike audio-driven talking faces generated in real time. In *Proc. NeurIPS*.

[73] Dogucan Yaman, Fevziye Irem Eyiokur, Leonard Bärmann, Seymanur Akti, Hazım Kemal Ekenel, and Alexander Waibel. 2024. Audio-Visual Speech Representation Expert for Enhanced Talking Face Video Generation and Evaluation. In *CVPR workshop*.

[74] Dongchao Yang, Dingdong Wang, Haohan Guo, Xueyuan Chen, Xixin Wu, and Helen Meng. 2024. Simplespeech: Towards simple and efficient text-to-speech with scalar latent transformer diffusion models. In *Proc. Interspeech*.

[75] Yochai Yemini, Aviv Shamsian, Lior Bracha, Sharon Gannot, and Ethan Fetaya. 2024. LipVoicer: Generating Speech from Silent Videos Guided by Lip Reading. In *Proc. ICLR*.

[76] Jeong Hun Yeo, Minsu Kim, Jeongsoo Choi, Dae Hoe Kim, and Yong Man Ro. 2024. Akvsr: Audio knowledge empowered visual speech recognition by compressing audio knowledge of a pretrained model. *IEEE Trans. on Multimedia* (2024).

[77] Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin, and Saining Xie. 2025. Representation Alignment for Generation: Training Diffusion Transformers Is Easier Than You Think. In *Proc. ICLR*.

[78] Neil Zeghidour, Alejandro Luebs, Ahmed Omran, Jan Skoglund, and Marco Tagliasacchi. 2021. Soundstream: An end-to-end neural audio codec. *IEEE/ACM Trans. on Audio, Speech, and Language Processing* (2021).