

UniCUE: Unified Recognition and Generation Framework for Chinese Cued Speech Video-to-Speech Generation

Jinting Wang¹, Shan Yang², Chenxing Li², Dong Yu², Li Liu^{1*}

¹The Hong Kong University of Science and Technology (Guangzhou)

²Tencent AI Lab

Abstract

Cued Speech (CS) enhances lipreading via hand coding, offering visual phonemic cues that support precise speech perception for the hearing-impaired. The task of **CS Video-to-Speech** generation (**CSV2S**) aims to convert CS videos into intelligible speech signals. Most existing research focuses on **CS Recognition (CSR)**, which transcribes video content into text. Consequently, a common solution for CSV2S is to integrate CSR with a text-to-speech (TTS) system. However, this pipeline relies on text as an intermediate medium, which may lead to error propagation and temporal misalignment between speech and CS video dynamics. In contrast, directly generating audio speech from CS video (direct CSV2S) often suffer from the inherent multimodal complexity and the limited availability of CS data. To address these challenges, we propose **UniCUE**, the first unified framework for CSV2S that **directly** generates speech from CS videos without relying on intermediate text. The core innovation of UniCUE lies in integrating a understanding task (CSR) that provides fine-grained CS visual-semantic cues to to guide the speech generation. Specifically, UniCUE incorporates a pose-aware visual processor, a semantic alignment pool that enables precise visual-semantic mapping, and a VisioPhonetic adapter to bridge the understanding and generation tasks within a unified architecture. To support this framework, we construct UniCUE-HI, a large-scale Mandarin CS dataset containing 11,282 videos from 14 cued speakers, including both hearing-impaired and normal-hearing individuals. Extensive experiments conducted on this dataset demonstrate that UniCUE achieves state-of-the-art (SOTA) performance across multiple evaluation metrics.

Project — <https://beria-moon.github.io/UniCUE/>

Extended Version — <https://arxiv.org/abs/2506.04134>

Introduction

Cued Speech (CS) is an visual phonetic encoding system that utilizes specific hand shapes and positions to enhance lip reading, providing an accurate visual representation of all phonemes in spoken language (Cornett 1967; Liu and Feng 2019; Leybaert and LaSasso 2010). CS maintains a high level of consistency with spoken language in terms of phonemes and speech patterns, enabling hearing-impaired

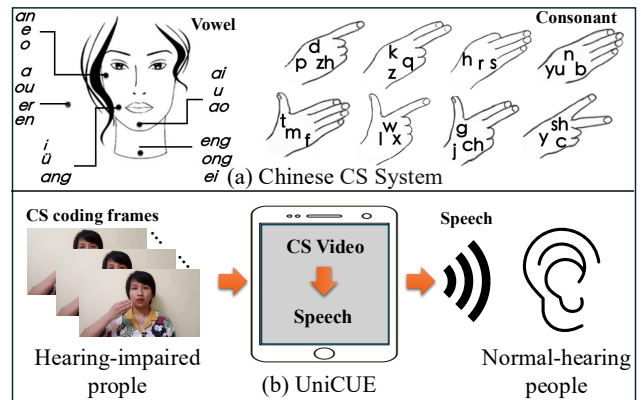


Figure 1: Illustration of the rules of the Chinese CS system and the proposed framework (UniCUE). (a) The chart for Mandarin Chinese CS (figure from (Liu and Feng 2019)), where five distinct hand positions are used to encode vowels, and eight finger shapes are employed to represent consonants in Mandarin Chinese. (b) Our framework enables the direct generation of synchronized natural speech from video.

individuals to better integrate into speech-dominant social and educational environments (Cornett 1967; Leybaert and LaSasso 2010; Leybaert, Aparicio, and Alegria 2010). In Mandarin Chinese, CS employs 8 hand shapes and 5 positions to encode consonants and vowels (as illustrated in Figure 1(a)), addressing challenges such as the phonemes with similar lip shapes (Liu and Feng 2019).

CS Video-to-Speech generation (CSV2S) task aims to convert CS videos of into comprehensible speech signals. However, directly constructing an end-to-end CSV2S model faces several challenges. **Firstly**, this task involves complex multimodal semantic correlations, requiring precise mapping from visual cues (lip movements and hand coding) to acoustic speech, while the limited scale of existing CS datasets further constrains model capacity. **Secondly**, fine-grained spatiotemporal modeling of visual information is essential to resolve the intrinsic asynchrony, *i.e.*, the hand-preceding phenomenon, where hand cues precede corresponding lip movements (Liu et al. 2020a). **To the best of our knowledge, the CSV2S task has not been explicitly studied in prior literature.**

*Corresponds to Li Liu (avrillliu@hkust-gz.edu.cn)

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

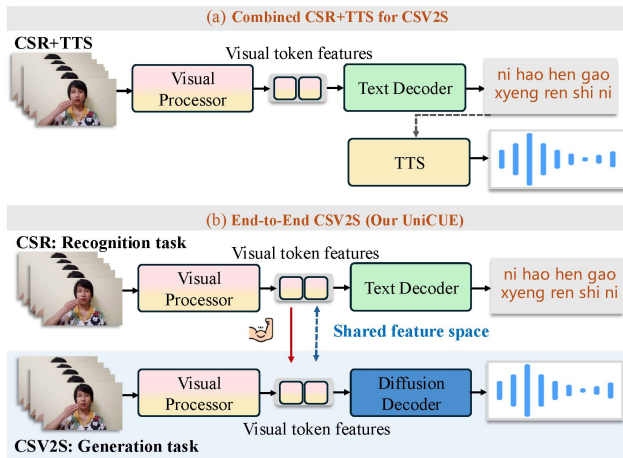


Figure 2: (a) The combined CSV2S architecture combines separately trained CSR and TTS models. (b) Our unified framework (UniCUE) that transfers understanding capabilities of CSR into speech generation by integrating the visual processor of CSR into CSV2S.

Existing research primarily focuses on **CS Recognition (CSR)** that converts CS videos into phoneme-level text (Liu et al. 2020a; Liu, Liu, and Li 2024; Liu et al. 2018, 2019), neglecting the critical need for natural speech generation. This limitation significantly impairs real-time communication between hearing-impaired and normal-hearing individuals, especially in educational and social scenarios. For instance, in group conversations, normal-hearing participants must quickly comprehend and respond to questions posed by their hearing-impaired peers. Textual output from CSR systems is often insufficient for such natural and smooth interactions. Additionally, recent lipreading-based video-to-speech models such as LipVoicer (Yemini et al. 2024) rely solely on lip movements, failing to capture the complementary hand-coded information in CS that conveys critical phonemic distinctions. These shortcomings underscore the need for a more comprehensive approach. Motivated by this, we aim **to develop the first Chinese CSV2S system that directly decodes CS videos into intelligible speech**, as illustrated in Figure 1(b).

A straightforward solution, shown in Figure 2(a), is to combine a CSR model with a Text-to-Speech (TTS) system. However, this combined pipeline suffers from two key drawbacks. Firstly, the intermediate textual representation introduces error propagation, as misrecognitions in the CSR stage lead to incorrect speech output. Secondly, the textual intermediate discards fine-grained spatiotemporal cues in the CS video, resulting in synthesized speech that lacks temporal alignment with the visual input.

To overcome these challenges, we draw inspiration from recent advances in multimodal learning, where semantic reasoning from vision-language models (VLMs) has shown strong promise in tasks like text-guided image synthesis with interleaved control (Mi et al. 2025; Chen et al. 2025a). We hypothesize that the multimodal visual understanding inher-

ent in CSR can serve as a semantic bridge to support more accurate and controllable speech generation in CSV2S. As depicted in Figure 2(b), we introduce a unified framework that leverages a shared visual processor to bridge CSR (understanding task) and CSV2S (generation task). This processor serves as a two-way translator: during CSR, it extracts linguistic semantics from fine-grained lip-hand motion patterns; in CSV2S, it utilizes these semantics to guide speech generation. The core innovation of our framework lies in modeling a semantic compensation flow, where phoneme-level supervision from CSR reduces ambiguity in speech synthesis, enabling more faithful and coherent voice generation under complex multimodal conditions.

Building upon this semantic compensation paradigm, in this work, we propose **UniCUE**, the first unified framework that bridges CSR and CSV2S tasks through three specific components: **Firstly**, unlike prior CSR methods (Liu, Liu, and Li 2024; Liu and Liu 2023) that process lip and hand modalities independently and rely on raw video embeddings, UniCUE employs a pose-aware processor that fuses video and pose streams into a mixed representation. This enables fine-grained spatiotemporal modeling of the hand-preceding phenomenon and improves generalization to cuer-specific expressive styles. **Secondly**, to enhance the alignment between visual and linguistic semantics, we introduce a semantic alignment pool to map the video and pose latent spaces into a shared textual space using contrastive learning. This facilitates cross-modal correlation modeling and improves semantic consistency in the generated speech. **Thirdly**, to unify the understanding and generation tasks, we reuse the CSR visual encoder within our diffusion-based CSV2S decoder and introduce a VisioPhonic Adapter (VPA) that transforms the visual representations into diffusion-compatible codes. This design enables the decoder to effectively incorporate fine-grained semantic information derived from multimodal visual inputs

To evaluate UniCUE on hearing-impaired individuals, we extend the MCCS dataset (Lei, Liu, and Wang 2024) by adding data from 8 hearing-impaired and 2 normal-hearing cuers¹, forming the Unified-HI Corpus with 14 cuers.

Experimental results on this dataset demonstrate that UniCUE not only produces accurate and intelligible speech, but also maintains temporal synchronization with the CS video.

The main contributions of this work can be summarized as:

- We propose the first CSV2S framework by constructing a unified multimodal system that integrates CSR capabilities to enhance speech generation.
- We propose a pose-aware visual processor and a semantic alignment pool to enhance fine-grained, semantically aligned visual representations, and introduce an VPA module to convert fine-grained semantic information into understandable coding for the speech synthesis model.
- We construct a new Mandarin Chinese CS dataset comprising both hearing-impaired and normal-hearing cuers.

¹Cuer means the people who perform CS.

Experimental results demonstrate that our UniCUE outperforms the state-of-the-art (SOTA) methods in terms of speech accuracy, consistency, and quality.

Related Work

Video-to-Speech Generation

V2S aims to synthesize natural speech aligned with silent talking videos, but is challenged by limited data. UniDubbing (Lei et al. 2024) addresses this via modality-aligned pre-training on multimodal data and fine-tuning with both multimodal and audio-only inputs. Similarly, Kefalas *et al.* (Kefalas, Panagakis, and Pantic 2024) pre-train on large audio-only corpora before tuning on paired data. Some studies (Kim, Hong, and Ro 2023; Yemini et al. 2024; Gupta et al. 2024) incorporate transcripts to enhance generation. Kim *et al.* (Kim, Hong, and Ro 2023) use text-speech supervision to improve word-level representation via multi-task learning. Existing V2S methods primarily focus on lip reading. However, CS conveys phonemic information through both lip and hand movements. Ignoring hand cues results in incomplete visual representations and degraded speech synthesis quality, limiting the applicability of these methods to CS. Notably, no prior work has addressed the CSV2S task.

Cued Speech Recognition

CS augments lip reading with hand coding to support the hearing-impaired. The CSR task aims to transcribe CS videos into text by leveraging lips and hands as complementary modalities (Liu and Feng 2019; Papadimitriou and Potamianos 2021). Most CSR methods extract lip and hand features separately and fuse them for recognition (Liu et al. 2020a; Liu, Liu, and Li 2024; Liu and Liu 2023; Zhang, Liu, and Liu 2023). Due to the asynchronous nature of these modalities, effective fusion remains challenging. Liu *et al.* (Liu et al. 2020a) proposed re-synchronization to align hand with lip features, while transformer-based mutual learning (Liu and Liu 2023; Liu, Liu, and Li 2024) improves multimodal interaction. Zhang *et al.* (Zhang, Liu, and Liu 2023) addressed privacy concerns via federated learning. In contrast, we directly model lip and hand cues from whole frames, avoiding explicit fusion. A pose-aware visual processor is introduced further to enhance cross-modal representation and improve performance.

Unified Understanding and Generation

Recent advances in unifying understanding and generation tasks fall into two main paradigms. The first integrates visual-language understanding with external generative models (*e.g.*, diffusion models) for multimodal generation (Wu et al. 2024a; Dong et al. 2024; Jin et al. 2024; Li et al. 2024; Sun et al. 2024; Ge et al. 2024, 2025). For example, (Jin et al. 2024; Li et al. 2024) utilize large language models (LLMs) for semantic understanding and diffusion models (Rombach et al. 2022; Podell et al. 2023) for high-fidelity image synthesis. The second paradigm trains LLM-based foundation models via next-token prediction for both vision understanding and generation (Yu et al. 2023; Sun et al. 2023; Zhou et al. 2024; Wu et al. 2024b; Fang

et al. 2024; Wu et al. 2025; Chen et al. 2025b). Transfusion (Zhou et al. 2024), for instance, unifies image understanding and generation within a single transformer, enabling controllable text-to-image synthesis by preserving visual details. However, existing approaches mainly focus on visual-text settings, leaving visual-to-speech generation underexplored. In this work, we introduce the first unified framework that bridges visual understanding and speech generation.

Method

Overview of UniCUE

To achieve accurate CSV2S generation, the proposed method needs to simultaneously address two critical challenges: (1) **semantic understanding** of the linguistic correlations between visual cues and speech content, and (2) **speech synthesis** that preserves cue-specific characteristics and temporal alignment. Inspired by the auxiliary benefits of unified understanding and generation for multi-modal controllable image synthesis (Wu et al. 2024a; Dong et al. 2024), we design a unified architecture that integrates CSR and CSV2S, enabling CSV2S with understanding capability improvement through shared visual feature representations. As illustrated in Figure 3, the framework operates via two pathways.

CSR: Fine-grained Visual Cues Understanding. As the recognition pathway, CSR models fine-grained spatiotemporal visual semantics to transcribe CS videos into linguistic sequences. Given a CS video I_v and its corresponding pose maps I_p (extracted via OpenPose (Cao et al. 2019)), we first utilize a pose-aware visual processor to extract multi-modal embeddings Z_{mv} , which capture lip and hand motion cues. And then Z_{mv} is fed into a auto-regressive Transformer-based text decoder D_T , which models long-range dependencies and contextual interactions across the sequence to generate the predicted token sequence: $T_p = D_T(Z_{mv})$, where T_p denotes the predicted token sequence.

Unlike prior approaches relying on Connectionist Temporal Classification (CTC) loss (Graves et al. 2006), which predict each token independently and thus limit the model’s ability to capture cross-token dependencies and coarticulatory effects, our method employs an auto-regressive decoder D_T supervised by cross-entropy loss. This design allows D_T to generate tokens conditioned on previously generated outputs and spatialtemporal visual cues, which is more suited to modeling the asynchronous and dynamic nature of CS.

To further enhance both token-level precision and sequence-level linguistic consistency, we employ a hybrid training objective: a masked language modeling loss $\mathcal{L}_{CE}^{\text{masked}}$ supervises selectively masked ground-truth tokens to enhance contextual understanding; a sequence-level cross-entropy loss $\mathcal{L}_{CE}^{\text{seq}}$ enforces supervision over the full sequence to promote accurate transcription. The final training objective for CSR is:

$$\mathcal{L}_R = \mathcal{L}_{CE}^{\text{masked}}(T_p, T_g) + \mathcal{L}_{CE}^{\text{seq}}(T_p, T_g), \quad (1)$$

where T_g denotes the ground-truth token sequence. This dual-loss strategy enhances token-level accuracy while preserving global sequence semantics, enabling the model to

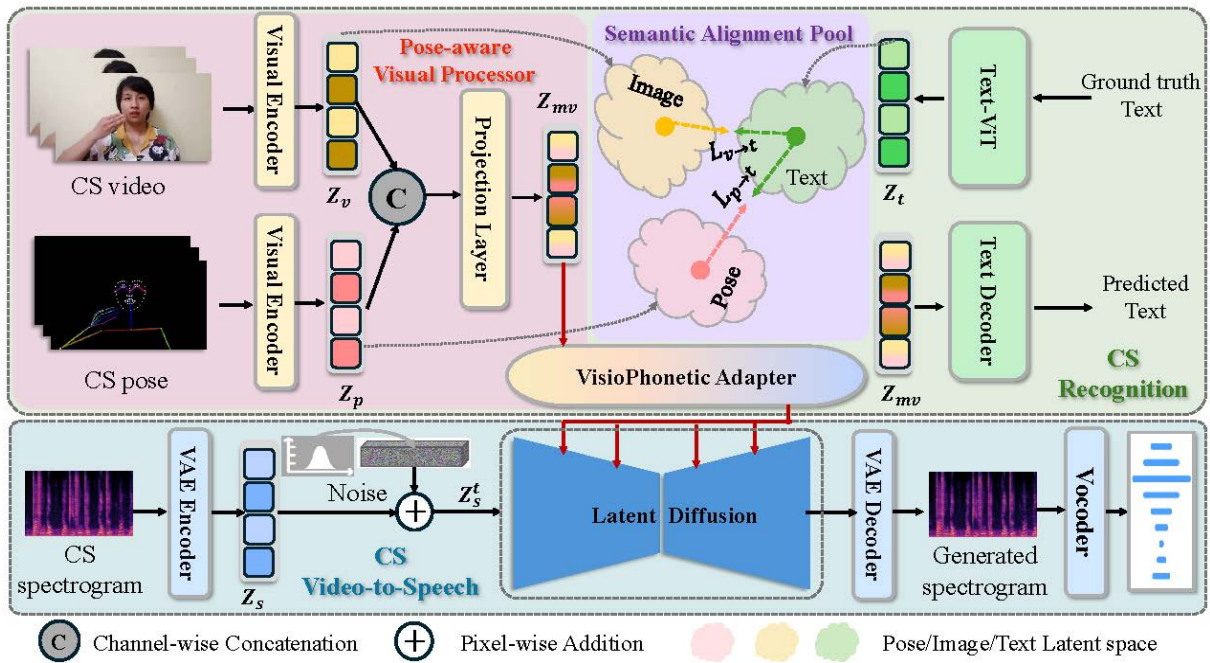


Figure 3: **Overview of our unified framework (UniCUE)**. It achieves direct Chinese CSV2S generation with semantic consistency, temporal alignment, and characteristics coherence by aligning the fine-grained spatiotemporal visual representations of CSR with the diffusion-based speech generator. The framework consists of three core modules: (1) **Pose-Aware Visual Processor**: Integrates video and pose embeddings to perform fine-grained spatiotemporal modeling of lip and hand movements. (2) **Semantic Alignment Pool**: Enhances the semantic mapping between visual features and speech content through video-text and pose-text contrastive learning. (3) **VisioPhonetic Adapter (VPA)**: Converts fine-grained visual representation of CSR into condition encodings compatible with the diffusion-based generator.

capture subtle visual-linguistic cues and temporal dynamics inherent in CS videos, thus improving recognition performance and supporting speech synthesis.

CSV2S: Cuer-specific Speech Synthesis. To directly synthesize intelligible and personalized speech from CS videos, we formulate speech generation as a conditional denoising process within a latent diffusion model (LDM) (Rombach et al. 2022). Since both lip shapes and hand cues in CS convey phonemic content, the speech generation is conditioned a refined visual embedding Z'_{mv} , which is derived by transforming the CSR multimodal feature Z_{mv} via a VisioPhonetic adapter (VPA). Specifically, a pretrained VAE encoder compresses ground-truth mel-spectrograms into latent codes Z_s , which are progressively corrupted with Gaussian noise ϵ over t steps: $Z_s^t := \alpha_t \cdot Z_s + (1 - \alpha_t) \cdot \epsilon$, where α_t denotes the noise level at timestep t . The noisy latent Z_s^t then denoised by the LDM conditioned on Z'_{mv} . The generation objective is defined as:

$$\mathcal{L}_G := \mathbb{E}_{Z_s^t, Z_{mv}, \epsilon, t} \left[\|\epsilon - \mathcal{M}(Z_s^t, Z_{mv}, t)\|_2^2 \right], \quad (2)$$

where \mathcal{M} represents the denoising network. By learning this conditional distribution, our model generates temporally aligned speech that reflects the visual expressions of cuers.

UniCUE: Unified Understanding and Generation. The CSR pathway learns fine-grained multi-modal visual embeddings Z_{mv} through detailed linguistic recognition. To

bridge the architectural gap between the CSR and the diffusion-based speech generator, we introduce a VPA that transforms Z_{mv} into a refined representation Z'_{mv} . These embeddings are subsequently utilized as conditional inputs to the CSV2S pathway, enabling the speech synthesis model to leverage enriched visual understanding for improved generation accuracy. By sharing visual feature representations within this unified framework, our approach effectively reduces information loss and mitigates error propagation that often arises from intermediate text conversions. As a result, CSV2S is capable of generating cue-specific speech that faithfully preserves linguistic fidelity and temporal alignment, producing personalized and intelligible speech outputs tailored to individual cuers.

Pose-aware Visual Processor

Considering the strong spatiotemporal correlation between hand coding, lip movement, and their underlying semantic content, both CSV2S and CSR require accurate modeling of lip and hand motion patterns. This necessitates a visual encoder capable of capturing fine-grained and temporally coherent features. While video frames offer rich appearance information, they often suffer from redundancy and visual ambiguity. In contrast, pose maps provide a compact, structured, and noise-resilient representation of motion dynamics. To leverage the complementary strengths of both modal-

ities, we design a pose-aware visual processor that constructs fused visual representations, as shown in Figure 3.

Specifically, the input to the processor consists of video frames I_v and pose maps I_p , both formatted as tensors of shape $T \times 3 \times H \times W$, where T indicates the frame lengths, and $H \times W$ denotes the spatial resolution. The processor comprises two main components. First, a shared visual encoder E_V extracts spatiotemporal features from both modalities via a sequential architecture: a 2D ResNet backbone extracts frame-wise spatial features, which are stacked along the temporal axis and passed through a 1D temporal convolution to model short-term motion patterns. The resulting sequence is then fed into a Transformer encoder to capture long-range temporal dependencies across frames. This process yields the video features $Z_v = E_V(I_v)$ and pose features $Z_p = E_V(I_p)$, where $Z_v \in \mathbb{R}^{L \times D}$, $Z_p \in \mathbb{R}^{L \times D}$ with D denoting the embedding dimension and $L = T \times N$ being the total number of tokens, where N is the number of spatial patches per frame. Second, the projection layer integrates the two feature streams. The video and pose features are concatenated along the channel dimension and passed through a multi-layer perceptron (MLP), consisting of two linear layers with ReLU activation and LayerNorm, to produce the final mixed visual representation:

$$Z_{mv} = \text{MLP}(\text{Concat}(Z_v, Z_p)). \quad (3)$$

The fused representation Z_{mv} serves as a unified visual embedding that drives both recognition and generation pathways. In the subsequent modules, this representation is semantically aligned with linguistic content and refined for diffusion-based speech synthesis.

Semantic Alignment Pool

To further enhance semantic consistency between visual representation and linguistic content, we introduce a semantic alignment mechanism that aligns video, pose, and textual modalities through contrastive learning. Specifically, a ViT-based text encoder encodes the ground-truth transcript tokens T_g into text embeddings Z_t . The visual features Z_v and pose features Z_p , extracted by the pose-aware visual processor, are projected into a shared latent space via learnable linear layers. The text embedding Z_t is similarly projected. We adopt a contrastive loss across the batch, treating each video-text and pose-text pair from the same sample as a positive pair, and all others as negatives. The loss is denoted as:

$$\mathcal{L}_{v \leftrightarrow t} = 1 - \cos(Z_v, Z_t), \quad \mathcal{L}_{p \leftrightarrow t} = 1 - \cos(Z_p, Z_t), \quad (4)$$

where $\cos(\cdot, \cdot)$ denotes the cosine similarity between normalized embeddings. The total semantic alignment loss is calculated as:

$$\mathcal{L}_S = \mathcal{L}_{v \leftrightarrow t} + \mathcal{L}_{p \leftrightarrow t}. \quad (5)$$

By enforcing this high-level alignment, the model is encouraged to extract complementary and discriminative semantics from visual modalities. These aligned features not only enhance linguistic recognition in CSR, but also offer semantically grounded condition for accurate speech synthesis.

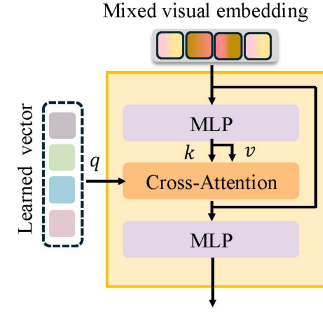


Figure 4: The details of the VisioPhonetic Adapter, which transforms semantic visual embeddings into phonetic-aware features to enable seamless conditioning for diffusion-based speech synthesis.

VisioPhonetic Adapter

While the CSR-derived embeddings capture rich visual-linguistic semantics, they remain mismatched in format and granularity for direct use in diffusion-based speech generation. To bridge this modality gap, we propose the **VisioPhonetic Adapter (VPA)**, which transforms semantically aligned visual features into a phonetic-aware conditioning signal suitable for the LDM. As illustrated in Figure 4, this lightweight module employs a sequential architecture to progressively refine visual-semantic representations into a diffusion-compatible conditioning signal:

$$Z'_{mv} = \text{MLP}\left(\text{CrossAttn}\left(\text{MLP}(Z_{mv})\right)\right), \quad (6)$$

which includes two MLPs and a Q-Former-style (Li et al. 2023) cross-attention layer. We use N_q learnable semantic queries $\mathbf{f} \in \mathbb{R}^{N_q \times D}$, which is initialized by computing the average latent representation from ground-truth mel-spectrograms encoded by the pretrained VAE. This provides a phonetic-aware initialization aligned with the diffusion model's target space. These queries act as phonetic slots to extract and reorganize relevant patterns from Z_{mv} . The cross-attention mechanism operates as: $\mathbf{q} = \mathbf{W}^q \mathbf{f}$, $\mathbf{k} = \mathbf{W}^k Z_{mv}$, $\mathbf{v} = \mathbf{W}^v Z_{mv}$, $\mathbf{a} = \text{Softmax}\left(\frac{\mathbf{qk}^T}{\sqrt{d}}\right) \mathbf{v}$, $Z_{mv}' = \text{MLP}(Z_{mv} + \mathbf{a})$. The adapted features Z'_{mv} serve as the final interface between visual understanding and speech synthesis, ensuring that the generated audio is not only temporally coherent but also linguistically faithful to the video input.

Experiment

Experimental Setting

Dataset. Existing CS datasets are limited to normal-hearing cuers and lack data from hearing-impaired individuals, hindering model generalization to the primary users of assistive communication systems. To bridge this gap, we construct a new dataset, the **Unified-HI Corpus**, which includes CS videos from 8 hearing-impaired and 6 normal-hearing cuers. This diverse composition significantly enriches variations in gesture styles, lip movements, and speech patterns. The expanded coverage introduces more realistic challenges and

Dataset	Cuers	Sentences	Character	Word	Resolution	FPS
French CS (Liu et al. 2018)	1-H	238	12872	-	720 × 576	50
British CS (Liu et al. 2019)	1-H	97	2741	-	720 × 1280	25
MCCS (Lei, Liu, and Wang 2024)	4-H	4000	131608	42256	720 × 1280	30
Unified-HI (Ours)	6-H and 8-HI	11282	350333	112664	720 × 1280	30

Table 1: Comparison between our Chinese Mandarin CS dataset and existing CS dataset. **H** denotes the cuers with normal hearing, while **HI** indicates hearing-impaired cuer. Our newly proposed Unified-HI Corpus is the first large-scale Chinese CS dataset with both hearing-impaired and normal-hearing cuers.

Method	Normal-hearing cuers					Hearing-impaired cuers			
	WER ↓	LSE-C ↑	LSE-D ↓	DNSMOS ↑	STOI ↑	WER ↓	LSE-C ↑	LSE-D ↓	DNSMOS ↑
GT	-	7.274	7.314	2.79	-	-	-	-	-
CMML	0.663	4.135	9.241	1.24	0.11	0.924	2.141	10.132	1.03
EcoCued	0.657	4.327	9.146	1.28	0.12	0.917	2.165	10.079	1.07
CSR (Ours)	0.186	4.874	9.125	2.53	0.57	0.224	3.342	9.315	2.29
Lip2Speech	0.803	4.215	9.367	1.03	0.05	0.989	2.424	10.816	0.02
LipVoicer	0.754	4.361	9.226	1.12	0.08	0.971	2.623	10.517	0.04
CSV2S (Ours)	0.374	<u>6.245</u>	<u>7.962</u>	2.27	0.42	0.422	<u>5.938</u>	<u>8.347</u>	2.04
UniCUE (Ours)	<u>0.205</u>	6.729	7.632	<u>2.46</u>	<u>0.53</u>	<u>0.248</u>	6.491	8.076	<u>2.17</u>

Table 2: Comparison with SOTA methods on test data of normal-hearing cuers and hearing-impaired cuers. **Bold** and underlined results are the best and second-best results. ↑ indicates that larger values are better, while ↓ indicates that smaller values are preferable.

better reflects practical use cases, enabling models to capture cue-specific nuances essential for hearing-impaired users. A comparison with existing CS datasets is shown in Table ??.

Due to the noisy speech data from hearing-impaired cuers, we use CS data from 6 normal-hearing cuers for training. The data from normal-hearing cuers is split by sentence into training and test sets with a 95:5 ratio to ensure effective training and validation. Importantly, **all CS data from the 8 hearing-impaired cuers are used in the test set**, enabling a robust evaluation of model generalization to this group.

Architecture Details. The CSV2S pathway is entirely built upon the AudioLDM (Liu et al. 2023), including its VAE encoder-decoder, latent diffusion model, and vocoder components. For CSR, the Transformer in visual process, tokenizer, text-ViT, and text decoder are initialized from MBart (Liu et al. 2020b).

Evaluation Metrics. We evaluate the synthesized speech from three perspectives: linguistic accuracy, temporal synchronization, and speech quality. Linguistic accuracy is quantified by the Word Error Rate (WER) between the recognized text and ground truth. Temporal synchronization is assessed using SyncNet (Chung and Zisserman 2016), reporting LSE-D (temporal distance) and LSE-C (confidence score). Speech quality is evaluated via STOI (Taal et al. 2010) for intelligibility and DNSMOS (Reddy, Gopal, and Cutler 2021) for naturalness.

Comparison Methods. We evaluate our UniCUE against: (1) **CSV2S (Ours)**: direct speech synthesis without CSR assistance; (2) **CSR (Ours)**: including pose-aware visual processor, text encoder and decoder, and semantic alignment pool; (3) **CSR methods**: CMML (Liu and Liu 2023) and EcoCued (Liu, Liu, and Li 2024); (4) **V2S methods**:

Lip2Speech (Choi, Kim, and Ro 2023) and LipVoicer (Yemini et al. 2024).

Comparison with SOTA Methods

Quantitative Comparison. We compare our framework against SOTA methods, as summarized in Table 2. Our CSR model, empowered by the pose-aware visual processor and semantic alignment pool, achieves significantly lower WERs (0.186 for normal-hearing and 0.224 for hearing-impaired cuers), surpassing previous CSR methods. Building on this strong semantic understanding, UniCUE outperforms V2S methods across LSE-D, LSE-C, DNSMOS, and STOI metrics, demonstrating superior linguistic accuracy, temporal alignment, and speech quality.

Qualitative Comparison. Mel-spectrogram visualizations (**Figure 4 in Appendix**) further highlight the advantages of our method, showcasing improved temporal synchronization and clearer acoustic structures compared to others.

Ablation Studies

To verify the contribution of each component, we conduct ablation studies on both normal-hearing and hearing-impaired test data. Results are summarized in Table 3.

Unified Training Paradigm. Compared to direct CSV2S, UniCUE reduces WER by 45% (0.205 vs. 0.374) on normal-hearing cuers and 41% (0.248 vs. 0.422) on hearing-impaired cuers. These results highlight the benefit of leveraging fine-grained visual semantics from CSR to enhance CSV2S, alleviating the challenge of modeling complex multimodal correlations.

Visual Processor Design. Models that rely solely on raw video features struggle to capture fine-grained motion due

Method	Normal-hearing cuers					Hearing-impaired cuers			
	WER↓	LSE-C↑	LSE-D↓	DNSMOS↑	STOI↑	WER↓	LSE-C↑	LSE-D↓	DNSMOS↑
GT	-	7.274	7.314	2.79	-	-	-	-	-
CSR ^{††}	0.210	4.746	9.129	2.42	0.49	0.250	3.218	9.402	2.19
CSR [‡]	0.204	4.783	9.224	2.46	0.53	0.247	3.234	9.397	2.21
CSR	0.186	4.874	9.125	2.53	0.57	0.224	3.342	9.315	2.29
CSV2S [†]	0.398	6.158	8.122	2.21	0.40	0.398	5.821	8.582	1.96
CSV2S	0.374	6.245	7.962	2.27	0.42	0.422	5.938	8.347	2.04
UniCUE ^{††}	0.239	6.637	7.724	2.30	0.44	0.267	6.419	8.163	2.08
UniCUE [‡]	0.231	<u>6.641</u>	<u>7.716</u>	2.33	0.46	0.276	<u>6.421</u>	<u>8.159</u>	2.10
UniCUE [*]	0.226	6.613	7.731	2.37	0.48	0.271	6.410	8.167	2.12
UniCUE	<u>0.205</u>	6.729	7.632	<u>2.46</u>	<u>0.53</u>	<u>0.248</u>	6.491	8.076	<u>2.17</u>

Table 3: Ablation Studies of model components on test data of normal hearing cuers and hearing-impaired cuers. The notations X^{††}, X[‡], and X^{*} indicate ablated versions of the architecture X, where the pose maps, semantic alignment pool, and VPA module are removed, respectively.

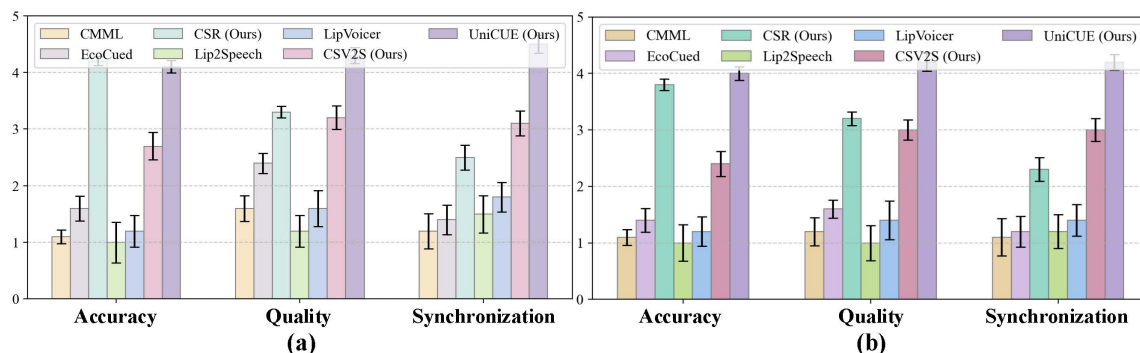


Figure 5: **User study** results for accuracy, quality, and synchronization metrics on normal-hearing (a) and hearing-impaired (b) test samples.

to redundant and noisy visual information, resulting in sub-optimal performance. By incorporating pose cues, our visual processor effectively captures cuer-specific dynamics, leading to significantly improved accuracy and robustness across diverse cuers. **Semantic Alignment Mechanism.** Disabling the Semantic Alignment Pool (SAP) degrades visual-semantic consistency, resulting in higher WERs for both CSR and UniCUE. This underscores the importance of the alignment in enforcing spatiotemporal coherence between visual cues and phonemic representations for accurate semantic modeling.

VisioPhonetic Adapter. Removing the VPA results in noticeable degradation in temporal alignment, demonstrating its crucial role in bridging the representation gap between CSR and CSV2S. By adaptively selecting and refining fine-grained spatiotemporal visual cues through learnable queries, the VPA enables more accurate and temporally coherent speech synthesis.

Impact of Hand Cues. Removing hand cues leads to substantial performance degradation, particularly for hearing-impaired users who often exhibit limited oral articulation and atypical lip shapes (see **Appendix Table 1**). The results highlight the complementary role of hand gestures in enhancing visual phonemic representations for CS.

User Study

To comprehensively assess the perceptual quality of synthesized speech, we conduct a user study involving 20 randomly selected test samples per cuer. Twenty volunteers rate the generated speech on three perceptual dimensions using 5-point Likert scales: **Accuracy** (1: unintelligible, 5: perfectly intelligible), **Quality** (1: artificial, 5: human-like), and **Synchronization** (1: desynchronized, 5: perfectly aligned). As shown in Figure 5, UniCUE consistently achieves significantly higher scores across all metrics, demonstrating statistically meaningful improvements. These findings validate that our unified framework effectively bridges visual understanding and speech generation, delivering superior performance in human perception compared to both modular pipelines and task-specific baselines.

Conclusion

This work introduces UniCUE, the first unified framework for directly generating speech from CS videos. By integrating fine-grained visual understanding with diffusion-based speech synthesis, UniCUE produces intelligible speech with precise temporal alignment. Key components including the pose-aware visual processor, semantic alignment pool, and VisioPhonetic Adapter, enable effective knowledge trans-

fer from CS recognition (CSR) to CS video-to-speech generation (CSV2S), enhancing both linguistic accuracy and temporal synchronization. Additionally, we introduce the UniCUE-HI corpus, a new CS dataset featuring both normal-hearing and hearing-impaired users. Extensive experiments on this dataset demonstrate that UniCUE outperforms SOTA methods across multiple evaluation metrics.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No. 62471420), Guangdong Basic and Applied Basic Research Foundation (2025A1515012296), and 2025 Tencent AI Lab Rhino-Bird Program.

References

- Cao, Z.; Hidalgo Martinez, G.; Simon, T.; Wei, S.; and Sheikh, Y. A. 2019. OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.
- Chen, L.; Bai, S.; Chai, W.; Xie, W.; Zhao, H.; Vinci, L.; Lin, J.; and Chang, B. 2025a. Multimodal Representation Alignment for Image Generation: Text-Image Interleaved Control Is Easier Than You Think. *CoRR*.
- Chen, X.; Wu, Z.; Liu, X.; Pan, Z.; Liu, W.; Xie, Z.; Yu, X.; and Ruan, C. 2025b. Janus-Pro: Unified Multimodal Understanding and Generation with Data and Model Scaling. *CoRR*.
- Choi, J.; Kim, M.; and Ro, Y. M. 2023. Intelligible Lip-to-Speech Synthesis with Speech Units. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, volume 2023, 4349–4353.
- Chung, J. S.; and Zisserman, A. 2016. Out of time: automated lip sync in the wild. In *Workshop on Multi-view Lip-reading, ACCV*.
- Cornett, R. O. 1967. Cued speech. *American annals of the deaf*, 3–13.
- Dong, R.; Han, C.; Peng, Y.; Qi, Z.; Ge, Z.; Yang, J.; Zhao, L.; Sun, J.; Zhou, H.; Wei, H.; et al. 2024. DreamLLM: Synergistic Multimodal Comprehension and Creation. In *ICLR*.
- Fang, R.; Duan, C.; Wang, K.; Li, H.; Tian, H.; Zeng, X.; Zhao, R.; Dai, J.; Li, H.; and Liu, X. 2024. PUMA: Empowering Unified MLLM with Multi-granular Visual Generation. *CoRR*.
- Ge, Y.; Li, Y.; Ge, Y.; and Shan, Y. 2025. Divot: Diffusion powers video tokenizer for comprehension and generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 13606–13617.
- Ge, Y.; Zhao, S.; Zhu, J.; Ge, Y.; Yi, K.; Song, L.; Li, C.; Ding, X.; and Shan, Y. 2024. SEED-X: Multimodal Models with Unified Multi-granularity Comprehension and Generation. *CoRR*.
- Graves, A.; Fernández, S.; Gomez, F.; and Schmidhuber, J. 2006. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on Machine learning*, 369–376. ACM.
- Gupta, A.; Likhomanenko, T.; Yang, K. D.; Bai, R. H.; Aldeneh, Z.; and Jaitly, N. 2024. Visatronic: A Multimodal Decoder-Only Model for Speech Synthesis. *arXiv:2411.17690*.
- Jin, Y.; Xu, K.; Chen, L.; Liao, C.; Tan, J.; Huang, Q.; Chen, B.; Song, C.; Meng, D.; Zhang, D.; et al. 2024. Unified Language-Vision Pretraining in LLM with Dynamic Discrete Visual Tokenization. In *ICLR*.
- Kefalas, T.; Panagakis, Y.; and Pantic, M. 2024. Large-scale unsupervised audio pre-training for video-to-speech synthesis. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.
- Kim, M.; Hong, J.; and Ro, Y. M. 2023. Lip-to-speech synthesis in the wild with multi-task learning. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5. IEEE.
- Lei, S.; Cheng, X.; Lyu, M.; Hu, J.; Tan, J.; Liu, R.; Xiong, L.; Jin, T.; Li, X.; and Zhao, Z. 2024. Uni-Dubbing: Zero-Shot Speech Synthesis from Visual Articulation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 10082–10099.
- Lei, W.; Liu, L.; and Wang, J. 2024. Bridge to non-barrier communication: gloss-prompted fine-grained cued speech gesture generation with diffusion model. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, 6333–6341.
- Leybaert, J.; Aparicio, M.; and Alegria, J. 2010. 19 The Role of Cued Speech in Language Development of Deaf Children. *The Oxford Handbook of Deaf Studies, Language, and Education, Volume 1*, 276.
- Leybaert, J.; and LaSasso, C. J. 2010. Cued speech for enhancing speech perception and first language development of children with cochlear implants. *Trends in amplification*, 14(2): 96–112.
- Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, 19730–19742. PMLR.
- Li, Y.; Zhang, Y.; Wang, C.; Zhong, Z.; Chen, Y.; Chu, R.; Liu, S.; and Jia, J. 2024. Mini-Gemini: Mining the Potential of Multi-modality Vision Language Models. *CoRR*.
- Liu, H.; Chen, Z.; Yuan, Y.; Mei, X.; Liu, X.; Mandic, D.; Wang, W.; and Plumbley, M. D. 2023. AudioLDM: Text-to-Audio Generation with Latent Diffusion Models. In *International Conference on Machine Learning*, 21450–21474. PMLR.
- Liu, L.; and Feng, G. 2019. A pilot study on mandarin chinese cued speech. *American Annals of the Deaf*, 164(4): 496–518.
- Liu, L.; Feng, G.; Beautemps, D.; and Zhang, X.-P. 2020a. Re-synchronization using the hand preceding model for multi-modal fusion in automatic continuous cued speech recognition. *IEEE Transactions on Multimedia*, 23: 292–305.

- Liu, L.; Hueber, T.; Feng, G.; and Beautemps, D. 2018. Visual Recognition of Continuous Cued Speech Using a Tandem CNN-HMM Approach. In *Interspeech*, 2643–2647.
- Liu, L.; Li, J.; Feng, G.; and Zhang, X.-P. S. 2019. Automatic Detection of the Temporal Segmentation of Hand Movements in British English Cued Speech. In *Interspeech*, 2285–2289.
- Liu, L.; and Liu, L. 2023. Cross-modal mutual learning for cued speech recognition. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5. IEEE.
- Liu, L.; Liu, L.; and Li, H. 2024. Computation and parameter efficient multi-modal fusion transformer for cued speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.
- Liu, Y.; Gu, J.; Goyal, N.; Li, X.; Edunov, S.; Ghazvininejad, M.; Lewis, M.; and Zettlemoyer, L. 2020b. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8: 726–742.
- Mi, Z.; Wang, K.-C.; Qian, G.; Ye, H.; Liu, R.; Tulyakov, S.; Aberman, K.; and Xu, D. 2025. I Think, Therefore I Diffuse: Enabling Multimodal In-Context Reasoning in Diffusion Models. *arXiv:2502.10458*.
- Papadimitriou, K.; and Potamianos, G. 2021. A fully convolutional sequence learning approach for cued speech recognition from videos. In *2020 28th European Signal Processing Conference (EUSIPCO)*, 326–330. IEEE.
- Podell, D.; English, Z.; Lacey, K.; Blattmann, A.; Dockhorn, T.; Müller, J.; Penna, J.; and Rombach, R. 2023. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv:2307.01952*.
- Reddy, C. K.; Gopal, V.; and Cutler, R. 2021. DNSMOS: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6493–6497. IEEE.
- Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.
- Sun, Q.; Cui, Y.; Zhang, X.; Zhang, F.; Yu, Q.; Wang, Y.; Rao, Y.; Liu, J.; Huang, T.; and Wang, X. 2024. Generative multimodal models are in-context learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14398–14409.
- Sun, Q.; Yu, Q.; Cui, Y.; Zhang, F.; Zhang, X.; Wang, Y.; Gao, H.; Liu, J.; Huang, T.; and Wang, X. 2023. Emu: Generative pretraining in multimodality. In *The Twelfth International Conference on Learning Representations*.
- Taal, C. H.; Hendriks, R. C.; Heusdens, R.; and Jensen, J. 2010. A short-time objective intelligibility measure for time-frequency weighted noisy speech. In *2010 IEEE international conference on acoustics, speech and signal processing*, 4214–4217. IEEE.
- Wu, C.; Chen, X.; Wu, Z.; Ma, Y.; Liu, X.; Pan, Z.; Liu, W.; Xie, Z.; Yu, X.; Ruan, C.; et al. 2025. Janus: Decoupling visual encoding for unified multimodal understanding and generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 12966–12977.
- Wu, S.; Fei, H.; Qu, L.; Ji, W.; and Chua, T.-S. 2024a. NExT-GPT: Any-to-Any Multimodal LLM. In *International Conference on Machine Learning*, 53366–53397. PMLR.
- Wu, Y.; Zhang, Z.; Chen, J.; Tang, H.; Li, D.; Fang, Y.; Zhu, L.; Xie, E.; Yin, H.; Yi, L.; et al. 2024b. Vila-u: a unified foundation model integrating visual understanding and generation. *arXiv:2409.04429*.
- Yemini, Y.; Shamsian, A.; Bracha, L.; Gannot, S.; and Fetaya, E. 2024. LipVoicer: Generating Speech from Silent Videos Guided by Lip Reading. In *The Twelfth International Conference on Learning Representations*.
- Yu, L.; Shi, B.; Pasunuru, R.; Muller, B.; Golovneva, O.; Wang, T.; Babu, A.; Tang, B.; Karrer, B.; Sheynin, S.; et al. 2023. Scaling autoregressive multi-modal models: Pretraining and instruction tuning. *arXiv:2309.02591*, 2(3).
- Zhang, Y.; Liu, L.; and Liu, L. 2023. Cuing without sharing: A federated cued speech recognition framework via mutual knowledge distillation. In *Proceedings of the 31st ACM International Conference on Multimedia*, 8781–8789.
- Zhou, C.; YU, L.; Babu, A.; Tirumala, K.; Yasunaga, M.; Shamis, L.; Kahn, J.; Ma, X.; Zettlemoyer, L.; and Levy, O. 2024. Transfusion: Predict the Next Token and Diffuse Images with One Multi-Modal Model. In *The Thirteenth International Conference on Learning Representations*.