2025 International Conference on Intelligent Computing
July 26-29, Ningbo, China
https://www.ic-icc.cn/2025/index.php

CDSS: Innovating Cross Differential Attention for Robust Monaural Multi-Speaker Audio-Visual Speech Separation

Yinlong Zhang[1], Jinjiang Liu[1], Jiawei Jin[1], Jiuxin Lin[1], and Zhiyong Wu[1,2(✉)]

[1] Shenzhen International Graduate School, Tsinghua University, Shenzhen, China
[2] The Chinese University of Hong Kong, Hong Kong SAR, China
zywu@se.cuhk.edu.hk

Abstract. Target Speaker Extraction (TSE) based on visual cues has been widely adopted and further extended to Audio-Visual Multi-Speaker Speech Separation (AV-MSS) through either simultaneous multi-speaker processing or recursive approaches. However, in real-world scenarios, obtaining complete visual information for all speakers is often impractical due to data collection constraints. Existing methods mostly use basic self-attention mechanisms to model correlations between separated speech streams to mitigate missing visual cues. Nevertheless, these approaches overlook the critical distinction between speech signals with auxiliary visual information and those without, resulting in performance degradation when modalities are incomplete. To address this, we propose a novel Cross Differential Attention (CDA) mechanism that performs cross-modal differentiation, effectively highlighting the salient disparities be-tween modalities. This design enables the model to adaptively emphasize in-formative, modality-specific features, thereby significantly improving robust-ness and effectiveness in both complete and missing-visual scenarios. Extensive experiments validate our method's superiority, demonstrating state-of-the-art performance on both two-speaker and three-speaker mixture tasks.

Keywords: Speech Separation, Audio-Visual, Cross Differential Attention, Multi-Speaker Scenarios.

1 Introduction

The cocktail party problem[1-3], characterized by overlapping speech from multiple speakers in noisy environments, remains a key challenge in speech separation research. In monaural scenarios, multi-modal techniques particularly those integrating visual cues[4, 5], had been introduced to enhance separation performance. One notable approach is Audio-Visual Target Speaker Extraction (AV-TSE)[6], which uses visual information to extract the speech of a single target speaker. Building upon AV-TSE, AV-MSS[7] extends the capability to scenarios involving multiple speakers. By leveraging multiple visual cues, AV-MSS achieves ordered separations, aligning each separated speech signal with its corresponding visual input. Additionally, AV-MSS incorporates

further processing on initially separated signals, effectively enhancing the quality of separation and mitigating the impact of noise and interference in complex multi-speaker environments.

Fig. 1. Diagram of our overall AV-MSS framework under the missing visual cue condition. This figure depicts the 3-mix scenario with two available visual cues.

The integration of multi-modal inputs and post-processing has firmly established AV-MSS as a powerful solution for addressing the complexities of real-world speech separation scenarios. However, practical applications continue to face significant challenges, particularly when visual information is incomplete or entirely unavailable due to data acquisition constraints. Factors such as a speaker moving off-screen or the camera being turned off in real-world scenarios can result in missing video information[8–11], further complicating the task.

To mitigate this issue, some AV-MSS methods employ self-attention mechanisms across separated speech signals from different speakers [7, 12], focusing on the correlations between them. While these approaches have shown some effectiveness, their reliance on correlations alone has a limited impact. Recent research [13] highlights that the differences between speech signals also carry valuable information, which is often overlooked. This is particularly true in scenarios characterized by an imbalance between audio and visual signals, where leveraging these differences is more appropriate. Current Differential Transformer [14] used Differential Attention (DA) which introduces differential operations into the attention mechanism to enhance modeling diversity. However, they only perform self-differentiation on a single input and assume uniformly distributed noise—so-called "common-mode signals"—during the attention differential process. This assumption does not hold for AV-MSS tasks, where multiple initially separated speech signals are processed in parallel, rendering the single-input paradigm inapplicable.

To alleviate these deficiencies, we propose the Cross Differential Attention (CDA) mechanism, which performs cross-modal differentiation by effectively highlighting the salient disparities between modalities. This allows the model to enhance its performance and robustness under both complete and missing-visual scenarios. Based on this mechanism, we further develop the Cross Differential Attention-based Audio-Visual Speech Separation network (CDSS) to address the challenges of multi-speaker

separation with incomplete visual information. CDSS introduces three novel modules designed to fully leverage the differences between speech signals: the Audio-Visual Cross-Interaction Module (AVC), the Audio-Visual Cross Differential Attention Module (AV-CDA), and the Audio-Visual Recovery Module (AVR).

The AVC module, inspired by previous networks [15], handles the initial audio separation process. The AV-CDA module, incorporating the proposed Cross Differential Attention (CDA) mechanism, serves as the core of the model, explicitly capturing the differences between speech signals with or without visual cue to enhance separation performance. Although our CDA mechanism employs subtraction for differences, similar to prior work [13,14,16], it is important to emphasize that our approach is the first to handle multiple inputs for attention, with a computation pipeline that differs in detail from those of previous studies. Additionally, we have carefully designed two sets of linear modules for different purposes, which is validated through our outlier test. These distinctions are key contributions of our model.

Furthermore, in preliminary experiments involving training on datasets with missing visual information, we observed an unintended behavior of the AV-CDA module: while it effectively enhanced the audio corresponding to regions without visual cues, it occasionally degraded the separation quality in regions where visual information was available. This suggests that the model may implicitly average representations across modalities—leveraging information from video-aligned speech to compensate for missing parts—thereby compromising the fidelity of well-aligned segments. However, this loss can be partially mitigated by incorporating additional visual signals, upon which the separation performance in visually missing regions can be further enhanced. To this end, we propose the Audio-Visual Recovery (AVR) module, which comprises two components: the Video Recovery module for reconstructing video-aligned speech, and the Audio Recovery module for enhancing audio-only segments. Together, these components promote a more balanced and robust performance, particularly in scenarios with incomplete multimodal data.

Following the data selection and preprocessing protocols established in previous work [17], we trained the CDSS model on the VoxCeleb2 [18] dataset. For evaluation, we adopted the same protocol and tested the model on VoxCeleb2, as well as two additional datasets: LRS3-TED [19] and TCD-TIMIT [20]. Our experiments cover both 2-mix and 3-mix scenarios, conducted under both with and without visual cue conditions. The flow for the without visual cue condition is illustrated in Fig 1. In all scenarios, our model achieves state-of-the-art performance, surpassing existing methods.

Ultimately, we highlight the contributions of this work as follows:

- We propose the CDA mechanism, which utilizes the differences between speech signals to address the challenge of missing visual cues, achieving a significant improvement in conditions without visual information. Unlike traditional differential attention mechanisms that perform self-differentiation on a single input during attention computation, our CDA mechanism exploits cross-modal differences between multiple signals. It is specifically designed for the AV-MSS task, and extensive outlier experiments further demonstrate the effectiveness of our approach.

- We introduce the Video Recovery module for recovering the video-aligned portion and the Audio Recovery module for enhancing the portion of the speech signal without visual cue, working together to improve the overall separation quality and further enhance the model's performance.
- We evaluate our model across three datasets (VoxCeleb2, LRS3-TED, and TCD-TIMIT) in both 2-mix and 3-mix settings, achieving state-of-the-art performance in all conditions, including those without visual cues, validating its effectiveness in diverse settings.

2 Related Work

Our investigation specifically targets monaural speech separation, intentionally excluding comparative evaluations with multi-channel methods. While leveraging inter-signal discrepancies to enhance separation performance is a well-established principle, recent studies have primarily focused on applying this concept to general speech separation tasks.

Our main contribution is the introduction of a novel dual-path attention architecture that uniquely exploits the cross-modal discrepancy between audio-visual signals and audio-only signals within the AV-MSS context. This represents the first systematic implementation of such a mechanism in AV-MSS, demonstrating superior performance through dedicated parallel processing modules specifically designed to this setting. Additionally, we conducted supplementary outlier experiments to validate that our modules align with the intended design, further confirming the effectiveness of our approach.

2.1 Audio-only Speech Separation

In social activity scenarios, auditory environments are often characterized by complex mixtures of multiple speakers' voices, along with reverberation and background noise [21,22]. MSS and TSE are widely adopted to address these challenges. MSS, aims to estimate the individual speech sources of each speaker directly from the mixture signal [23–26]. However, it struggles with accurately associating separated speech with the corresponding speakers [27]. Inspired by human auditory perception [28], TSE emulates selective auditory attention by treating the person of interest as the target speaker. In audio-only speech separation, TSE utilizes the enrolled speech [29, 30] of the target speaker as a reference, eliminating the need for prior knowledge of the total number of speakers. By leveraging this reference, the neural network can effectively extract the target speaker's voice from the mixture while mitigating the limitations of blind source separation.

2.2 Audio-Visual Speech Separation

In Audio-Visual Speech Separation (AVSS), both AV-TSE and AV-MSS utilize visual information, such as lip movement [17, 31] and facial features [6], to enhance

separation performance. Building on advancements in audio-only speech separation, several studies have extended these methods to AVSS[15, 32], with some introducing novel architectures[13, 17]. These approaches typically fuse audio and visual features using techniques such as cross-attention[15] or concatenation[13, 17, 32], followed by deeper processing to improve performance. As discussed earlier, AV-MSS expands on AV-TSE by enabling simultaneous multi-speaker processing[17, 32] or recursive strategies[33], often incorporating refinement within a unified framework[12] or a two-stage process[7, 34]. However, in real-world scenarios, challenges such as speakers moving off-screen or cameras being turned off can result in missing visual information [8–11], where existing methods exhibit suboptimal performance. Moreover, research addressing cases with completely absent visual cues remains scarce, despite its critical importance for practical applications. This gap underscores a significant opportunity for future research and innovation.

2.3 Signal Exclusivity

Numerous studies have investigated the concept of exclusivity between signals as a strategy to enhance performance in tasks such as denoising, extraction, and separation across various domains. In Natural Language Processing (NLP), differential transformer [14] architectures utilize differential attention to harness exclusivity in attention signals, effectively reducing attention noise. In computer vision (CV), reverse attention techniques have been successfully employed in applications such as object detection [16] and polyp segmentation [35]. Similarly, in the speech domain, the principle of exclusivity between noise and target speech has been leveraged for tasks including speech enhancement [36], recognition [37], and TSE [13]. However, the application of exclusivity to MSS remains largely underexplored, particularly under conditions where visual cues are absent. This gap highlights a promising direction for future research.

2.4 Differential Attention

The Differential Transformer [14], introduced Differential Attention (DA), significantly enhancing modeling diversity and achieving success in tasks such as long-context modeling, key information retrieval, hallucination mitigation, in-context learning, and reduction of activation outliers. These achievements demonstrate the effectiveness of differential attention. However, our work builds upon and extends this framework, specifically in the context of Audio-Visual Multi-Speaker Speech Separation (AV-MSS). Our primary innovation lies in refining the differential attention mechanism within AV-MSS, enabling more efficient attention design, especially in scenarios with missing visual information, where our method outperforms existing approaches.

The key differences between our approach and DA are as follows. First, while DA handles only a single input at a time, it first applies linear projections to the input signal A, obtaining A_1 and A_2. It then computes output A' only considering the self-difference within a single signal:

$$A' = \mathrm{DA}(A_1 - A_2) \tag{1}$$

In contrast, our method processes multiple speech signals in parallel. For instance, when given two input signals A and B, we calculate the output accounting for the differences between the signals in both directions:

$$A', B' = \mathrm{CDA}(A - B, B - A) \tag{2}$$

This allows our model to better handle the complex relationships between multiple signals, especially in multi-speaker scenarios like AV-MSS, where signals from different sources need to be separated effectively.

Second, we do not rely on predefined assumptions about signal differences. Instead, we use two distinct linear projectors to independently learn the representations of speech signals with and without visual cues, effectively modeling the discrepancies and enhancing separation performance, particularly when visual cues are missing. In contrast, DA assumes uniformly distributed noise ("common-mode signals") and performs self-differentiation based on this assumption, which is not suitable for AV-MSS.

In summary, while DA excels in attention for single signals, our approach adapts and improves this mechanism for AV-MSS, processing multiple separated speech signals simultaneously and handling missing visual cues more effectively, resulting in significant performance improvements.

3 Methodology

Building upon the paradigm of time-domain speech separation models, we propose CDSS, which integrates an audio encoder, a visual encoder, an audio decoder, and a separator. The architecture of the proposed model is illustrated in Figure 2. For clarity, we initially focus on a 3-speaker mixture scenario with two visual cues as input, where the model's output consists of the predicted audio signals for each individual speaker. It is important to note that our model is designed to handle a variety of scenarios, including both complete and incomplete data, as well as 2-mix and 3-mix situations. While each of these scenarios has slight differences in the corresponding formulas, the underlying paradigm remains consistent. The 3-speaker mixture scenario is chosen as an illustrative example because it is the most complex case, and we believe it offers a clearer understanding of our model's process for the reader.

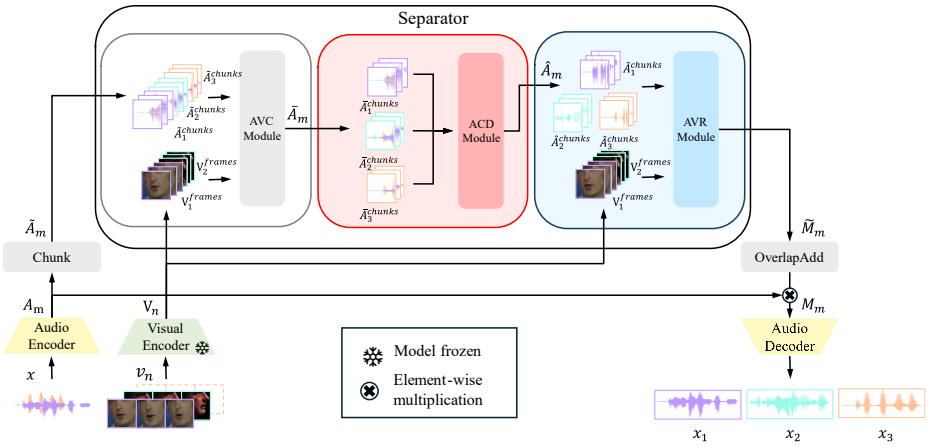

Fig. 2. Diagram of the overall CDSS structure consists of six components: the Visual Encoder, Audio Encoder, Chunk, Separator, OverlapAdd, and Audio Decoder.

3.1 Audio Encoder

The audio encoder extracts the features A_m from the speech mixture sequence $x \in \mathbb{R}^{T_x}$ using the 1D convolution operation with a kernel size L and stride $L/2$:

$$A_m = \text{Conv1D}\left(x, L, \frac{L}{2}\right) \in \mathbb{R}^{D \times T_a} \tag{3}$$

where $T_a = \lceil 2 \times T_x / L \rceil$ with proper zero padding and D is the audio feature dimension.

For the 2D audio feature A_m, Chunk operation divides A_m into chunks of length K with a hop size of $K/2$. These chunks are then concatenated and repeated to form a reshaped 4D audio chunked feature:

$$\tilde{A}_m = \text{Chunk}(A_m) \in \mathbb{R}^{M \times D \times K \times S} \tag{4}$$

where M is the number of the speakers in the mixture and S is the number of chunks, which is specifically designed to match the length of the visual feature.

3.2 Visual Encoder

We build upon prior research in audio-visual speech separation that employs a pre-trained, frozen lip embedding extractor. This extractor consists of a 3D convolution layer, an 18-layer ResNet, and multi-layer TCN networks. Specifically, the Visual Encoder takes in a gray-scale image sequence v_n as input and outputs a visual feature $V_n \in \mathbb{R}^{N \times D \times S}$, where N represents the number of visual cues, D denotes the visual feature dimension, and S indicates the length of the visual feature.

3.3 Separator

As shown in Figure 2, the separator comprises three key modules: the Audio-Visual Cross-Interaction Module (AVC), the Audio-Visual Cross Differential Attention Module (AV-CDA), and the Audio-Visual Recovery Module (AVR). For clarity, unless otherwise specified, in the following sections, m, n, d, k, and s represent positive integers, constrained by $m \leq M$, $n \leq N$, $d \leq D$, $k \leq K$, and $s \leq S$, where M, N, D, K, and S are previously defined.

Audio-Visual Cross-Interaction Module (AVC). The AVC builds on a previous model that integrates audio and visual features through three transformer blocks: IntraTransformer, CrossTransformer, and InterTransformer. The IntraTransformer and InterTransformer use self-attention, while the CrossTransformer applies cross-attention.

Given the audio inputs $\tilde{A}$, the IntraTransformer is first applied to extract frequency features from the audio:

$$\tilde{A}'_m[m,:,:,s] = \text{IntraTransformer}\big(\tilde{A}_m[m,:,:,s]\big) \tag{5}$$

Next, the CrossTransformer fuses the audio and visual features at the same temporal granularity. The visual features V_n are fused with the audio features, and the result is concatenated with the audio features lacking visual information:

$$\begin{aligned} &\tilde{A}'_{m1}, \tilde{A}'_{m2} = \tilde{A}'_m[:N], \tilde{A}'_m[N:] \\ &\tilde{A}''_{m1} = \text{CrossTransformer}(V_n[n,:,:], \tilde{A}'_{m1}[n,:,k,:]) \\ &\tilde{A}''_m = \text{Concatenate}(\tilde{A}''_{m1}, \tilde{A}'_{m2}) \end{aligned} \tag{6}$$

Finally, the InterTransformer extracts temporal features from the fused representation:

$$\bar{A}_m = \text{InterTransformer}\big(\tilde{A}''_m[m,:,k,:]\big) \tag{7}$$

where $\bar{A}_m \in \mathbb{R}^{M \times D \times K \times S}$.

Audio-Visual Cross Differential Attention Module (AV-CDA). The AV-CDA module employs a novel cross differential attention (CDA) mechanism, integrated within transformer blocks, to enhance the input without visual cue. As illustrated in Figure 3, for mixtures involving three speakers, the first two speakers are differentiated from the third. Worth mentioning is that this approach is maintained even when visual cues are not missing. The video components are processed by the first pair of linear modules, while the audio-only components are processed by the second pair. Here, Q_i, K_i, and V_i are computed for each speaker from the input.

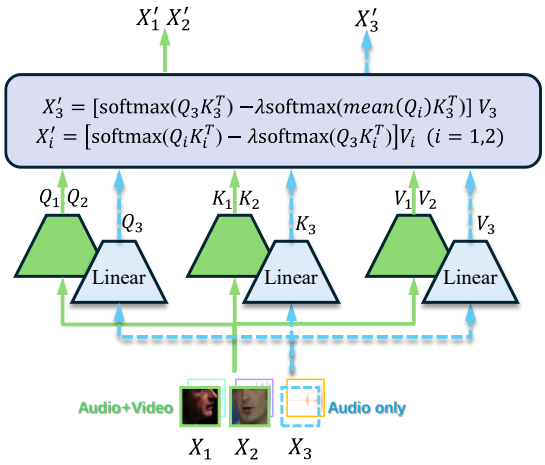

Fig. 3. Diagram illustrating our proposed Cross Differential Attention (CDA) mechanism.

To ensure clarity, we first delve into the internal mechanism of the proposed CDA. Specifically, we detail the attention computation process, and subsequently define the input and output of the overall AV-CDA module. In CDA, both branches independently compute their self-attention matrices, and then perform cross-subtraction by removing the attention weights of the opposite branch. To enhance flexibility, a learnable parameter λ is introduced, which differs from [13]. When multiple speakers are subtracted, as depicted in the figure, the mean of all speakers' attention matrices is used for calculation, unlike [13, 14], which only deal with two inputs. Given three inputs X_1, X_2, X_3 with shape $\mathbb{R}^{t\times d}$ (where t is the time length and d is the feature dimension), the computations proceed as follows, setting $i = 1,2$:

$$\begin{aligned} Q_i &= X_iW_1^Q, K_i = X_iW_1^K, V_i = X_iW_1^V \;\; (i = 1,2) \\ Q_3 &= X_3W_2^Q, K_3 = X_3W_2^K, V_3 = X_3W_2^V \end{aligned} \tag{8}$$

The cross differential mechanism is then applied, where $\sigma(\cdot)$ denotes the softmax operation:

$$\begin{aligned} X_i' &= (\sigma(\frac{Q_iK_i^T}{\sqrt{d}}) - \lambda\sigma(\frac{Q_3K_i^T}{\sqrt{d}}))V_i \;\; (i = 1,2) \\ X_3' &= (\sigma(\frac{Q_3K_3^T}{\sqrt{d}}) - \lambda\sigma(\frac{\text{mean}(Q_1, Q_2)K_3^T}{\sqrt{d}}))V_3 \end{aligned} \tag{9}$$

Based on the above process, the final output of the AV-CDA module is defined as:

$$\text{AV-CDA}(X_1, X_2, X_3) = X_1', X_2', X_3' \tag{10}$$

This CDA mechanism is applied across both the temporal and frequency dimensions. To clarify, we continue with the example shown in Figure 3. The input is first split as follows:

$$\bar{A}_{m1}, \bar{A}_{m2}, \bar{A}_{m3} = \bar{A}_m \tag{11}$$

Here, $\bar{A}_{mi} \in \mathbb{R}^{D \times K \times S}$, where $i \in 1,2,3$. The cross differential mechanism is then applied sequentially over the temporal and frequency dimensions. First the cross differential mechanism is applied across the temporal dimension for each slice s:

$$\bar{A}'_m = \text{AV-CDA}(\bar{A}_{m1}[:,:,s], \bar{A}_{m2}[:,:,s], \bar{A}_{m3}[:,:,s]) \tag{12}$$

The outputs are then updated as:

$$\bar{A}'_{m1}, \bar{A}'_{m2}, \bar{A}'_{m3} = \bar{A}'_m \tag{13}$$

Similarly, the mechanism is applied across the frequency dimension for each time slice k:

$$\hat{A}_m = \text{AV-CDA}(\bar{A}'_{m1}[:,k,:], \bar{A}'_{m2}[:,k,:], \bar{A}'_{m3}[:,k,:]) \tag{14}$$

The final output, $\hat{A}_m$, is obtained with a shape of $\mathbb{R}^{M \times D \times K \times S}$.

Audio-Visual Recovery Module (AVR). The audio-visual recovery module is designed to enhance the quality of separation. As mentioned earlier, we observed that when training the model exclusively in scenarios without visual cue, the AV-CDA module caused the part without visual cue to increase while the part with visual cue decreased. To address this, we first employ cross-attention to recover the diminished video part using the video input. This is followed by self-attention to further enhance the part with visual cue, an approach validated by our ablation study.

We introduce both the video recovery and audio recovery modules, each implemented as a transformer block. The process begins with video recovery, followed by concatenation with the part without visual cue:

$$\begin{aligned} &\hat{A}_{m1}, \hat{A}_{m2} = \hat{A}_m[:N], \hat{A}_m[N:] \\ &\hat{A}'_{m1} = \text{VideoRecovry}(V_n[n,:,:], \hat{A}_{m1}[n,:,k,:]) \\ &\hat{A}'_m = \text{Concatenate}(\hat{A}'_{m1}, \hat{A}_{m2}) \end{aligned} \tag{15}$$

Next, we apply self-attention across all speakers in the time dimension:

$$\widetilde{M}_m = \text{AudioRecovery}\big(\hat{A}'_m[:,d,k,:]\big) \tag{16}$$

The final output, $\widetilde{M}_m$, has a shape of $\mathbb{R}^{M \times D \times K \times S}$.

3.4 Audio Decoder

First, the OverlapAdd operation is applied to obtain masks, which are then element-wise multiplied with the raw mixture to isolate the target speech components. Subsequently, the audio decoder employs a transposed convolution layer to reconstruct the target speech:

$$M = \text{OverlapAdd}(\widetilde{M}_m) \in [0,1]^{M \times D \times T_a} \tag{17}$$

The separated speech signals, denoted as $\hat{s}$ are computed as follows:

$$\hat{s} = \text{TransposedConv1D}\left(A_m \otimes \widetilde{M}_m\right) \in \mathbb{R}^{M \times T_a} \tag{18}$$

Here, $\otimes$ represents element-wise multiplication.

4 Experiments

4.1 Datasets

We use three datasets in our experiments: VoxCeleb2, LRS3-TED, and TCD-TIMIT. VoxCeleb2 is used for both training and testing, while LRS3-TED and TCD-TIMIT are used exclusively for testing. To ensure fair evaluation, we retrained all previous models to ensure a fair comparison and the speaker identities in the training and test sets of VoxCeleb2 do not overlap. All utterances are between 4 and 6 seconds in duration.

For audio, we follow the methodology outlined in MuSE[17] to generate audio mixtures. Specifically, we select 48,000 utterances from 800 speakers in the VoxCeleb2 training set and 36,237 utterances from 118 speakers in the test set. Unlike traditional approaches TSE, we focus on multi-speaker separation and do not introduce additional disturbances. The model is trained on two-speaker (2-mix) and three-speaker (3-mix) scenarios, generating 20,000, 5,000, and 3,000 samples for 2-mix, and 10,000, 2,500, and 1,500 samples for 3-mix. These samples are combined during training to ensure balanced performance across different mixture conditions. To simulate missing visual cues during training, we introduce a 20% probability of omitting the entire visual data for one speaker.

Since the mixture audio is generated randomly, we ensure that the video corresponding to the last audio source in the mixture is designated as missing. For testing on the LRS3 and TCD-TIMIT datasets, we follow prior work [7] by selecting 1,200 videos from each dataset's original test set, comprising 600 mixtures for 2-mix and 400 for 3-mix scenarios.

For the visual modality, all video frames are resampled to 25 FPS. Videos from VoxCeleb2 and LRS3-TED, originally at a resolution of 224×224, can be processed directly using MuSE. For TCD-TIMIT, which has an original resolution of 1920×1080, we manually cropped the lip region and resized it to 224×224 to ensure compatibility with MuSE.

4.2 Implementation Setup

Following [15], the model is trained using the Adam optimizer [38] with an initial learning rate of 1.5×10^{-4}. The learning rate is halved if there is no decrease in the validation loss over the course of 3 epochs. Training is stopped if the validation loss does not improve over 5 consecutive epochs.

Speech quality and intelligibility are evaluated using the perceptual evaluation of speech quality (PESQ)[39] metric, while the scale-invariant signal-to-distortion ratio (SI-SDR)[40] is used as the loss function, defined as:

$$\mathcal{L} = \mathcal{L}_{\text{SI-SDR}}(\hat{s}, s) = -10\log_{10}\left(\frac{||\frac{\langle \hat{s}, s\rangle s}{||s||^2}||^2}{||\hat{s} - \frac{\langle \hat{s}, s\rangle s}{||s||^2}||^2}\right) \tag{19}$$

4.3 Experiments Results

We evaluate the model under two scenarios. In the first scenario, "no missing", we examine its performance with all visual cues, demonstrating that the model's performance does not degrade, but instead improves. In the second scenario, "missing one", we assess the model's robustness to the loss of one visual cue, highlighting its ability to maintain strong performance even when the visual information of one speaker is missing.

For baseline comparisons, we adhere to established audio-visual benchmarks in prior works, given that our contribution resides not in pioneering multimodal integration for speech separation but in enhancing separation efficacy within existing architectural paradigms. Consequently, the experimental evaluation is specifically confined to contemporary audio-visual separation systems.

Table 1. The testing results for the "no missing" scenario are presented in the table.

Network	Params	Voxceleb2				LRS3				TCD-TIMIT			
		SI-SDR		PESQ		SI-SDR		PESQ		SI-SDR		PESQ	
		2-mix	3-mix	2-mix	3-mix	2-mix	3-mix	2-mix	3-mix	2-mix	3-mix	2-mix	3-mix
AV-ConvTasNet[32]	10.5M	10.31	5.35	1.99	1.43	11.64	4.25	2.05	1.37	13.07	4.89	2.42	1.53
MuSE[17]	14.3M	10.95	5.63	2.11	1.43	11.59	4.26	2.07	1.38	12.91	6.76	2.29	1.69
AV-SepFormer[15]	28M	14.05	9.81	2.39	1.84	16.01	8.97	2.55	1.82	18.77	11.71	2.87	2.09
RAVSS[12]	32M	14.36	**10.88**	2.46	**1.99**	15.89	10.28	2.63	1.97	18.05	13.42	2.99	2.43
Ours	32M	**14.40**	10.74	**2.47**	1.96	**16.34**	**11.15**	**2.65**	**2.02**	**19.49**	**14.62**	**3.10**	**2.52**

Performance with Visual Cues. The experimental results presented in Table 1 demonstrate that our model outperforms others in the 2-mix scenario on the VoxCeleb2 dataset, with a slight drop in performance in the 3-mix case. While this may not be surprising, it is noteworthy that our model maintains similar performance in the 3-mix scenario with visual cues on VoxCeleb2 and demonstrates significantly better performance across all other conditions. Importantly, our model achieves this while utilizing a similar number of parameters as AV-SepFormer, yet outperforms it by a large margin. Furthermore, while our model achieves comparable performance to RAVSS in a narrow, specific task, it does so with significantly fewer parameters, highlighting its efficiency and superior overall performance. Specifically, our model demonstrates remarkable generalization ability, significantly outperforming existing methods in cross-domain evaluations. It achieves the best results on both the LRS3 and TCD-TIMIT

datasets, while RAVSS experiences a considerable drop in performance compared to previous work.

To explain this, VoxCeleb2 and LRS3 are known as "wild" datasets due to their real-world recording conditions, in contrast to the "studio" datasets like TCD-TIMIT, which are collected in controlled environments. Despite the challenges posed by VoxCeleb2's less predictable conditions, our model demonstrates strong robustness in these high-quality studio settings, showcasing its excellent generalization across diverse data conditions. In contrast, other methods fail to exhibit similar performance, particularly when tested on cross-domain data.

Table 2. Testing results for the "missing one" scenario.

Dataset	Net-Work	SI-SDR						PESQ					
		2-v	2-m	Δ	3-v	3-m	Δ	2-v	2-m	Δ	3-v	3-m	Δ
VoxCeleb2	RAVSS	14.36	13.61	0.75	**10.88**	9.24	1.64	2.47	2.39	0.08	**1.99**	1.81	0.18
	CDSS	**14.4**	**14.37**	**0.03**	10.74	9.89	0.85	2.47	**2.43**	**0.04**	1.96	**1.82**	**0.14**
LRS3	RAVSS	15.89	15.23	0.66	10.28	6.93	**3.35**	2.63	2.47	0.16	1.97	1.7	0.27
	CDSS	**16.34**	**15.69**	**0.65**	**11.15**	**7.29**	3.86	**2.65**	**2.52**	**0.13**	**2.02**	**1.84**	**0.18**
TCD-TIMIT	RAVSS	18.05	16.66	1.39	13.42	10.46	2.96	2.99	2.78	0.21	2.43	2.22	0.21
	CDSS	**19.49**	**18.62**	**0.87**	**14.62**	**12.67**	**1.95**	**3.1**	**2.94**	**0.16**	**2.52**	**2.34**	**0.18**

Robustness to Missing One Visual Cue This comparison highlights the robustness of our model in scenarios without visual cue. Previous models have employed minimal methods for this task, resulting in poor performance. For instance, AV-SepFormer achieves a negative SI-SDR in this scenario. Therefore, we primarily compare our model with the previous best-performing model, RAVSS. Unlike its originally reported performance in the paper, we reimplemented the model, identified some bugs in its code, and fixed them.

As shown in Table 2, our model consistently outperforms RAVSS across all three datasets in both the 2-mix and 3-mix scenarios under the "m" condition. Additionally, we achieve the smallest drop between the "v" and "m" conditions in nearly all cases. Notably, in cases involving two speakers with one missing visual cue, our task bears similarity to TSE, but with a critical distinction: while TSE focuses on extracting the speech of a single target speaker, our approach extracts speech for all speakers. Despite this added complexity, our model surpasses the TSE model AV-SepFormer, while RAVSS performs worse in comparison.These experiments demonstrate the resilience of our model in handling missing visual cues, showcasing its ability to maintain high performance even under challenging conditions.

Ablation Study. Table 3 demonstrates the effectiveness of each module used in our study.

- Row a: This represents the baseline model with only the AVC module, which captures intra- and inter-chunk correlations of T-domain speech features and fuses audio and visual features via cross-attention. We refined the structure by reducing redundancy, decreasing the number of layers, and limiting cross-modal interactions.

- Row b: Adding the AV-CDA module significantly improves performance, particularly in the "missing one" scenario, with a 13.58 dB gain. For the "no missing" case, the improvement is 0.44 dB, addressing the limitations of previous models in leveraging speech content disparity effectively.
- Row c and Row d: The Video Recovery and Audio Recovery modules are evaluated individually and jointly. Compared with Row b, while the Audio Recovery module (Row c) offers marginal gains, the Video Recovery module (Row d) provides notable improvements, especially in video-related tasks.
- Row e: Although the Audio Recovery module may seem to contribute little at first, as suggested by the comparison between Row b and Row c, the comparison between Row d and Row e reveals a 0.4 dB improvement under the "missing one" condition. This indicates that the Audio Recovery module alone is insufficient to support components without visual cue. Only when Audio Recovery follows Video Recovery can the framework effectively leverage the available information, thereby justifying the chosen module sequence.

Table 3. The testing results for the ablation study in VoxCeleb2 dataset.

No.	AV-CDA	VR	AR	SI-SDR				PESQ			
				2-v	2-m	3-v	3-m	2-v	2-m	3-v	3-m
a	×	×	×	13.53	-0.02	9.28	-2.99	2.36	1.14	1.8	1.07
b		×	×	13.97	13.56	9.43	8.43	2.39	2.33	1.82	1.66
c		×		14.1	13.59	9.7	8.4	2.42	2.37	1.85	1.67
d			×	14.39	13.99	10.66	9.66	2.46	2.42	1.95	1.85
e				14.4	14.37	10.74	9.89	2.47	2.43	1.96	1.82

Outlier Test. To systematically validate the distinct learning characteristics of the dual linear modules in our CDA mechanism, we conducted an outlier detection experiment using the cross-domain LRS3 dataset. This investigation specifically focuses on the video-specific linear module and audio-only-specific linear module, aiming to empirically demonstrate their differential information acquisition capabilities.

Building upon the 3-mix "missing one" paradigm, we implemented a modified input configuration that strategically alters the data positioning scheme compared to the baseline CDA mechanism architecture (Figure 3). As shown in Table 4, the baseline performance without module swaps is shown in row a. When substituting the first two video elements while retaining their original video-specific processing modules (row b), the performance remains stable, confirming the video-specific linear module's acquisition capabilities. Furthermore, we swap the last two elements, and the cross-module substitution - applying the audio-only-specific module to video elements and vice versa (row c) - results in performance degradation.

To enhance experimental comprehensiveness, we extended this evaluation to the 2-mix scenario. As shown in the results, rows b and c for the 2-mix case are identical, both showing a drop compared to row a. This equivalency arises because all substitutions in 2-mix inevitably involve cross-module interactions, unlike the 3-mix case where partial substitutions preserve modality consistency. We also observe that the performance degradation in 2-mix is smaller than 3-mix. The result suggests that the two

sets of linear modules in the CDA mechanism leverage more distinct information in scenarios with a higher number of mixes, as the increased presence of video parts amplifies the differentiation.

Table 4. The testing results for the outlier test

No.	SI-SDR				PESQ			
	2-v	2-m	3-v	3-m	2-v	2-m	3-v	3-m
a	15.8	15.71	10.06	8.81	2.59	2.55	1.95	1.83
b	15.66	15.57	10.08	8.83	2.58	2.55	1.96	1.83
c	15.66	15.57	9.47	8.25	2.58	2.55	1.87	1.78

These experimental findings conclusively validate our architectural design: the dual linear modules develop complementary specialization - the video module excels in visual pattern recognition while the audio-only module focuses on acoustic feature extraction, achieving effective cross-modal disentanglement through modality-adaptive processing.

5 Conclusions

In this paper, we proposed CDSS, a novel model for audio-visual speech separation capable of addressing both with and without visual cue scenarios. The core innovation lies in our introduction of the cross differential attention (CDA) mechanism, which is the first to explicitly exploit the cross-modal discrepancy between audio with and without visual cues to enhance separation performance in AV-MSS tasks. Specifically, CDSS employs distinct linear projections for audio segments with available visual information and those without, enabling the model to compute cross-modal differences that guide more effective separation.

To further reinforce this design, we introduced the Audio-Visual Recovery (AVR) module, which separately enhances the reconstruction performance for video-aligned and audio-only speech. Comprehensive experiments including controlled outlier scenarios demonstrate that CDSS not only achieves state-of-the-art performance on both 2-mix and 3-mix tasks under varying visual availability, but also validates the theoretical motivations behind our design.

While CDSS is currently optimized for up to three concurrent speakers, future work may explore extending this approach to more complex mixtures and adapting the system to real-world challenges such as visual occlusion, reverberation, and dynamic background noise.

References

1. Cherry, E.C.: Some experiments on the recognition of speech, with one and with two ears. The Journal of the acoustical society of America. 25, 975–979 (1953).

2. Bronkhorst, A.W.: The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions. Acta acustica united with acustica. 86, 117–128 (2000).
3. Haykin, S., Chen, Z.: The cocktail party problem. Neural computation. 17, 1875–1902 (2005).
4. Cutler, R., Davis, L.: Look who's talking: Speaker detection using video and audio correlation. In: 2000 IEEE international conference on multimedia and expo. ICME2000. Proceedings. Latest advances in the fast changing world of multimedia (cat. No. 00TH8532). pp. 1589–1592. IEEE (2000).
5. Golumbic, E.Z., Cogan, G.B., Schroeder, C.E., Poeppel, D.: Visual input enhances selective speech envelope tracking in auditory cortex at a "cocktail party." Journal of Neuroscience. 33, 1417–1426 (2013).
6. Gao, R., Grauman, K.: Visualvoice: Audio-visual speech separation with cross-modal consistency. In: 2021 IEEE/CVF conference on computer vision and pattern recognition (CVPR). pp. 15490–15500. IEEE (2021).
7. Cheng, H., Liu, Z., Wu, W., Wang, L.: Filter-recovery network for multi-speaker audio-visual speech separation. In: The eleventh international conference on learning representations (2023).
8. Afouras, T., Chung, J.S., Zisserman, A.: My lips are concealed: Audio-visual speech enhancement through obstructions. arXiv preprint arXiv:1907.04975. (2019).
9. Hong, J., Kim, M., Choi, J., Ro, Y.M.: Watch or listen: Robust audio-visual speech recognition with visual corruption modeling and reliability scoring. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 18783–18794 (2023).
10. Dai, Y., Chen, H., Du, J., Wang, R., Chen, S., Wang, H., Lee, C.-H.: A study of dropout-induced modality bias on robustness to missing video frames for audio-visual speech recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 27445–27455 (2024).
11. Wang, J., Pan, Z., Zhang, M., Tan, R.T., Li, H.: Restoring speaking lips from occlusion for audio-visual speech recognition. In: Proceedings of the AAAI conference on artificial intelligence. pp. 19144–19152 (2024).
12. Pan, T., Liu, J., Wang, B., Tang, J., Wu, G.: RAVSS: Robust audio-visual speech separation in multi-speaker scenarios with missing visual cues. In: Proceedings of the 32nd ACM international conference on multimedia. pp. 4748–4756 (2024).
13. Tao, R., Qian, X., Jiang, Y., Li, J., Wang, J., Li, H.: Audio-visual target speaker extraction with reverse selective auditory attention. arXiv preprint arXiv:2404.18501. (2024).
14. Ye, T., Dong, L., Xia, Y., Sun, Y., Zhu, Y., Huang, G., Wei, F.: Differential transformer. In: Proceedings of ICLR 2025 (2025).
15. Lin, J., Cai, X., Dinkel, H., Chen, J., Yan, Z., Wang, Y., Zhang, J., Wu, Z., Wang, Y., Meng, H.: Av-sepformer: Cross-attention sepformer for audio-visual target speaker extraction. In: ICASSP 2023-2023 IEEE international conference on acoustics, speech and signal processing (ICASSP). pp. 1–5. IEEE (2023).
16. Chen, S., Tan, X., Wang, B., Hu, X.: Reverse attention for salient object detection. In: Proceedings of the european conference on computer vision (ECCV). pp. 234–250 (2018).
17. Pan, Z., Tao, R., Xu, C., Li, H.: Muse: Multi-modal target speaker extraction with visual cues. In: ICASSP 2021-2021 IEEE international conference on acoustics, speech and signal processing (ICASSP). pp. 6678–6682. IEEE (2021).
18. Chung, J.S., Nagrani, A., Zisserman, A.: Voxceleb2: Deep speaker recognition. arXiv preprint arXiv:1806.05622. (2018).

19. Afouras, T., Chung, J.S., Zisserman, A.: LRS3-TED: A large-scale dataset for visual speech recognition. arXiv preprint arXiv:1809.00496. (2018).
20. Harte, N., Gillen, E.: TCD-TIMIT: An audio-visual corpus of continuous speech. IEEE Transactions on Multimedia. 17, 603–615 (2015).
21. Wang, Q., Muckenhirn, H., Wilson, K., Sridhar, P., Wu, Z., Hershey, J., Saurous, R.A., Weiss, R.J., Jia, Y., Moreno, I.L.: Voicefilter: Targeted voice separation by speaker-conditioned spectrogram masking. arXiv preprint arXiv:1810.04826. (2018).
22. Delcroix, M., Zmolikova, K., Kinoshita, K., Ogawa, A., Nakatani, T.: Single channel target speaker extraction and recognition with speaker beam. In: 2018 IEEE international conference on acoustics, speech and signal processing (ICASSP). pp. 5554–5558. IEEE (2018).
23. Luo, Y., Mesgarani, N.: Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation. IEEE/ACM transactions on audio, speech, and language processing. 27, 1256–1266 (2019).
24. Luo, Y., Chen, Z., Yoshioka, T.: Dual-path rnn: Efficient long sequence modeling for time-domain single-channel speech separation. In: ICASSP 2020-2020 IEEE international conference on acoustics, speech and signal processing (ICASSP). pp. 46–50. IEEE (2020).
25. Subakan, C., Ravanelli, M., Cornell, S., Bronzi, M., Zhong, J.: Attention is all you need in speech separation. In: ICASSP 2021-2021 IEEE international conference on acoustics, speech and signal processing (ICASSP). pp. 21–25. IEEE (2021).
26. Wang, Z.-Q., Cornell, S., Choi, S., Lee, Y., Kim, B.-Y., Watanabe, S.: TF-GridNet: Making time-frequency domain models great again for monaural speaker separation. In: ICASSP 2023-2023 IEEE international conference on acoustics, speech and signal processing (ICASSP). pp. 1–5. IEEE (2023).
27. Yu, D., Kolbæk, M., Tan, Z.-H., Jensen, J.: Permutation invariant training of deep models for speaker-independent multi-talker speech separation. In: 2017 IEEE international conference on acoustics, speech and signal processing (ICASSP). pp. 241–245. IEEE (2017).
28. Xu, C., Rao, W., Chng, E.S., Li, H.: Time-domain speaker extraction network. In: 2019 IEEE automatic speech recognition and understanding workshop (ASRU). pp. 327–334. IEEE (2019).
29. Ge, M., Xu, C., Wang, L., Chng, E.S., Dang, J., Li, H.: L-spex: Localized target speaker extraction. In: ICASSP 2022-2022 IEEE international conference on acoustics, speech and signal processing (ICASSP). pp. 7287–7291. IEEE (2022).
30. Mu, Z., Yang, X., Sun, S., Yang, Q.: Self-supervised disentangled representation learning for robust target speech extraction. In: Proceedings of the AAAI conference on artificial intelligence. pp. 18815–18823 (2024).
31. Pan, Z., Ge, M., Li, H.: USEV: Universal speaker extraction with visual cue. IEEE/ACM Transactions on Audio, Speech, and Language Processing. 30, 3032–3045 (2022).
32. Wu, J., Xu, Y., Zhang, S.-X., Chen, L.-W., Yu, M., Xie, L., Yu, D.: Time domain audio visual speech separation. In: 2019 IEEE automatic speech recognition and understanding workshop (ASRU). pp. 667–673. IEEE (2019).
33. Xu, X., Dai, B., Lin, D.: Recursive visual sound separation using minus-plus net. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 882–891 (2019).
34. Yao, Z., Pei, W., Chen, F., Lu, G., Zhang, D.: Stepwise-refining speech separation network via fine-grained encoding in high-order latent domain. IEEE/ACM Transactions on Audio, Speech, and Language Processing. 30, 378–393 (2022).
35. Fan, D.-P., Ji, G.-P., Zhou, T., Chen, G., Fu, H., Shen, J., Shao, L.: Pranet: Parallel reverse attention network for polyp segmentation. In: International conference on medical image computing and computer-assisted intervention. pp. 263–273. Springer (2020).

36. Zheng, C., Peng, X., Zhang, Y., Srinivasan, S., Lu, Y.: Interactive speech and noise modeling for speech enhancement. In: Proceedings of the AAAI conference on artificial intelligence. pp. 14549–14557 (2021).
37. Hu, Y., Hou, N., Chen, C., Chng, E.S.: Interactive feature fusion for end-to-end noise-robust speech recognition. In: ICASSP 2022-2022 IEEE international conference on acoustics, speech and signal processing (ICASSP). pp. 6292–6296. IEEE (2022).
38. Kingma, D.P.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980. (2014).
39. Rix, A.W., Beerends, J.G., Hollier, M.P., Hekstra, A.P.: Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs. In: 2001 IEEE international conference on acoustics, speech, and signal processing. Proceedings (cat. No. 01CH37221). pp. 749–752. IEEE (2001).
40. Le Roux, J., Wisdom, S., Erdogan, H., Hershey, J.R.: SDR–half-baked or well done? In: ICASSP 2019-2019 IEEE international conference on acoustics, speech and signal processing (ICASSP). pp. 626–630. IEEE (2019).