A Study of Dropout-Induced Modality Bias on Robustness to Missing Video Frames for Audio-Visual Speech Recognition

Yusheng Dai[†], Hang Chen[†], Jun Du[†*], Ruoyu Wang[†], Shihao Chen[†], Haotian Wang[†], Chin-Hui Lee[‡]
[†] University of Science and Technology of China, Hefei, China
[‡]Georgia Institute of Technology, Atlanta, America
jundu@ustc.edu.cn

Abstract

Advanced Audio-Visual Speech Recognition (AVSR) systems have been observed to be sensitive to missing video frames, performing even worse than single-modality models. While applying the common dropout techniques to the video modality enhances robustness to missing frames, it simultaneously results in a performance loss when dealing with complete data input. In this study, we delve into this contrasting phenomenon through the lens of modality bias and uncover that an excessive modality bias towards the audio modality induced by dropout constitutes the fundamental cause. Next, we present the Modality Bias Hypothesis (MBH) to systematically describe the relationship between the modality bias and the robustness against missing modality in multimodal systems. Building on these findings, we propose a novel Multimodal Distribution Approximation with Knowledge Distillation (MDA-KD) framework to reduce over-reliance on the audio modality, maintaining performance and robustness simultaneously. Finally, to address an entirely missing modality, we adopt adapters to dynamically switch decision strategies. The effectiveness of our proposed approach is evaluated through comprehensive experiments on the MISP2021 and MISP2022 datasets. Our code is available at https://github.com/dalision/ModalBiasAVSR.

1. Introduction

Audio-Visual Speech Recognition (AVSR) is a multimodal application inspired by human speech perception. It outperforms single-modality models by incorporating noise-invariant complementary information from visual cues, especially in noisy environments. Driven by increasingly large open-source datasets and models [1–4], AVSR has achieved significant advancements across various benchmarks with a simple end-to-end design [5, 6].

Recent research on AVSR focuses on more challenging real-life scenarios. Techniques such as reinforcement learning [7] and carefully designed fusion architecture [8–10] are used to accommodate varying noise levels and overlapping speech. Self-supervised learning [11] and automatic labeling techniques [12] are applied facing insufficient audio-visual pairs. Meanwhile, various synchronization modules have been developed for audio-visual alignment.[13–15]. However, restricted to the open-source datasets [1, 2, 16], most studies often assume that each video is recorded in relatively high quality, without blurring, corruption, or loss. Moreover, there is growing evidence to suggest that current advanced AVSR systems are highly susceptible to perturbations in video modality [17, 18], resulting in significant performance degradation even perform worse than single-modality models [19, 20].

Missing video modality is a crucial and common problem for AVSR applied in real-life scenarios [1, 17, 19, 20]. It arises from various causes, including losses induced by network latency or hardware limitations, as well as errors in lip movement tracking due to occlusion and side-face. Most researchers utilize dropout techniques [1] on video training data to improve robustness against missing modalities [19–23]. It has been demonstrated to effectively mitigate the out-of-distribution (OOD) issue and alleviate performance degradation without additional inference consumption or complex modules. However, it leads to new challenges on real-life scenarios with low-quality input. In our early experiments on MISP datasets [24, 25], a contradictory phenomenon could be observed in Figure 1: *while applying the dropout strategy to video training data enhance the robustness against missing video modality, it also leads to performance degradation when dealing with complete data input.*

*Corresponding author. This work was supported by the National Natural Science Foundation of China under Grant No. 62171427.

[1]Distinguished from the classic dropout that randomly deactivates nodes during neural network training, dropout in this paper specifically refers to a data augmentation technique that partially or entirely replaces original video frames with padding frames.

On the other hand, all AVSR systems consistently lag behind unimodal ASR when facing completely missing video.

We attempt to analyze the reasons behind the above-mentioned phenomenon from the perspective of modality bias. Existing multimodal applications can be categorized into two types: (1) modality-balanced systems, in which each modality contributes relatively equally to the model decision, such as Multimodal Emotion Recognition (MER) [26] and Hate Speech Detection (HSD) [27]; (2) modality-biased systems that over-relies on certain modality that contains more task-related information. AVSR is a typical modality-biased system dominated by audio. Therefore, an intuitive insight suggests that although dropout on the video modality could address the OOD problem between the training and inference stages, it may exacerbate the modality bias on audio, subsequently demonstrating robustness towards missing video input.

In this paper, we first verify this intuitive hypothesis in Section 2 by quantitatively analyzing the differences between AVSR and unimodal automatic speech recognition (ASR). The results uncover that the modality bias essentially represents a shift from a multimodal to a unimodal distribution on audio modality in latent representation space. Next in Section 3, we extend our findings to more general mulitmodal applications and propose the Modality Bias Hypothesis (MBH) to systematically describe the relationship between modality bias and robustness to missing modality. In Sections 4 and 5, we are committed to achieving two objectives: improving the robustness of AVSR without degradation with complete input, and ensuring that AVSR consistently outperforms ASR when faced with severe or complete video missing. To this end, we present Multimodal Distribution Approximation with Knowledge Distillation (MDA-KD), in which the robust student model leverages hidden knowledge extracted by a relatively unbiased teacher model to prevent the distribution of task-relevant representations from transfering into a unimodal distribution. The method is observed to enhance missing robustness through the learning of complementary information from the other modality and utilizing context information from adjacent frames. For video severely or entirely missing situations, adapters are adopted to the modality-specific branch to dynamically switch decision bias dominated by modality-specific representations. The key contributions can be summarized as follows:

• We investigate dropout-induced modality bias and uncover that it fundamentally manifests as a shift from a multimodal to a unimodal distribution of audio modality in the hidden representation subspace as detailed in Section 2.

• We propose using the Modality Bias Hypothesis (MBH) to systematically describe the decision-making process influenced by modal bias in a multimodal system, as well as the relationship between modal bias and modality missing robustness as detailed in Section 3.

• We propose Multimodal Distribution Approximation with Knowledge Distillation (MDA-KD) to enhance robustness against missing video and avoid performance degradation with complete input. For entirely missing modalities, adapters are adopted to dynamically switch decision bias to the specific modality as detailed in Section 5.

• We achieve top AVSR performances on MISP2021 and MISP2022 datasets while maintaining robustness against missing video frames as detailed in Section 7.

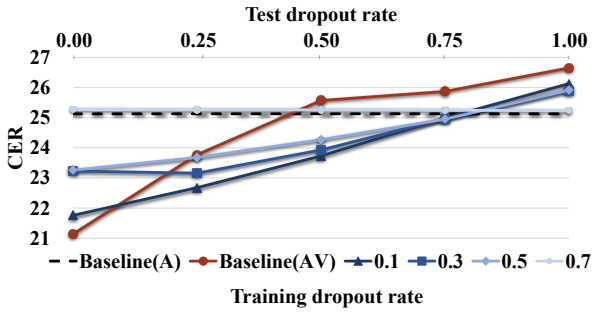


Figure 1. CER (in %) degradation curves of AVSR trained with different dropout rates on video frames. Compared with the baseline AVSR without dropout (in red), other AVSR systems perform better with missing input but worse with complete data input. As the training dropout rate increases, the CER curve of AVSR gradually converges to that of ASR (dotted line).

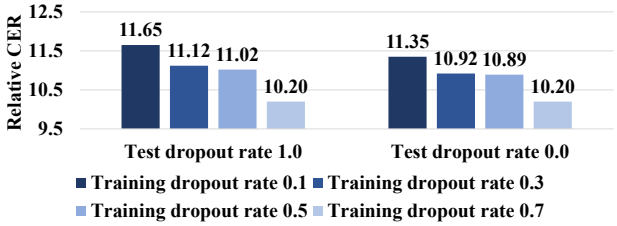


Figure 2. Two groups of similarity analysis between ASR and AVSR transcriptions. In both groups, an increase in the similarity of recognition transcriptions is observed as the training dropout rate increases. The similarity is measured by relative CER (in %), where the ASR transcription replaces the ground truth.

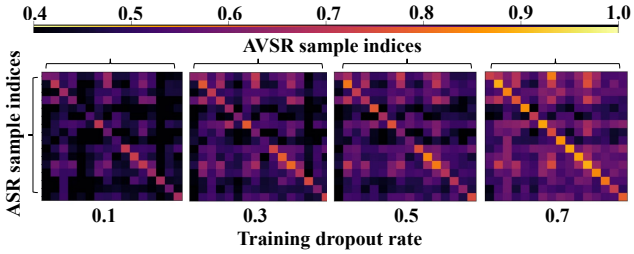


Figure 3. Similarity matrices of intermediate representations between ASR and different AVSR settings. As training dropout rates increase, the diagonal lines become brighter, indicating closer proximity between the multimodal and the unimodal distributions of the latent decisive subspace in AVSR.

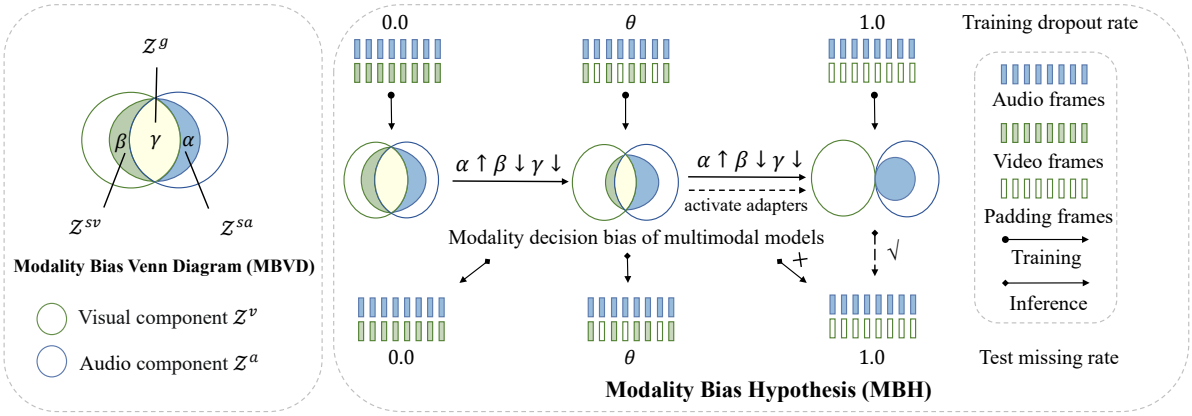


Figure 4. An illustration of the Modality Bias Hypothesis (MBH). In the left subplot, the task-relevant component (shaded part) of the latent representations consists of Z^{sa}, Z^{sv} and Z^g, representing audio-specific, visual-specific decision features and modality-general decisive features respectively. The corresponding proportions are denoted by α, β, and γ. The right subplot shows a dynamic process of decisive bias with an increasing training dropout rate. Dropout leads to a consistent modality bias on audio, regardless of the extent of the missing.

2. Dropout-Induced Modality Bias

We investigate the contradictory phenomenon 1 by examining the character error rate (CER) across five Mandarin AVSR systems varying training dropout rates (from 0.0 to 0.7) and testing video missing rates (from 0.0 to 1.0). As shown in Figure 1, two trends are observed: (1) in terms of absolute CER, the model trained with a higher dropout rate deteriorate more on no-missing complete multimodal data and slightly missing video frames, but it performs better on severely and entirely missing video frames; and (2) in term relative performance, the CER degradation curve of the AVSR model trained with a higher dropout rate tends to converge to the unimodal ASR recognition curve. We further ensure whether the similarity of performance degradation curves directly corresponds to the recognition transcription similarity of ASR and AVSR in Figure 2. As we expected, an increase in training dropout rate leads to higher transcription similarity between AVSR and ASR across different test settings.

To understand this, we investigate the discrepancy in decisive patterns of ASR and each AVSR. We aim to quantify the divergence between latent decision distributions of these models by measuring the distance of intermediate representation samples. Through random sampling of complete audio-visual data batches, we generate intermediate layer representations using the encoder of ASR or AVSR trained at different dropout rates. Figure 3 illustrates cosine distance-based similarity matrices for the intermediate representations between ASR and different AVSR configurations. The diagonal elements in each subplot represent the similarity between intermediate representations from the same inputs. Notably, with an increase training dropout rate, these diagonal lines brighten, signifying a rise in intermediate representation similarity. This suggests closer proximity of the AVSR multimodal distribution in the latent decisive subspace to the unimodal distribution of ASR.

Through the aforementioned three experiments, we have discovered that increasing the training dropout rate on video data leads to increased similarity between AVSR and ASR in the performance degradation curves, recognition results, and intermediate representation subspace distribution. The findings reveal the significant impact of dropout in introducing effectively perturbs the distribution of multimodal training data. It leads to a shift from multimodal joint distribution to unimodal distribution, resulting in a decision bias towards audio during the decision-making process, as reflected in the output similarity of ASR. We refer to this phenomenon induced by dropout as dropout-induced modality bias. Although dropout-induced bias enhances the robustness of missing video data to some extent, we emphasize that it contradicts the primary design of AVSR as a robust application in noisy environments with supplementary visual cues. The introduction of artificial noise (padding frames) in video data induces the model to converge toward trivial solutions, leading to an excessive dependence on the audio modality. This over-reliance, in turn, leads to a degradation in performance when presented with complete multimodal input in a noisy environment.

3. Modality Bias Hypothesis (MBH)

In this section, we propose the Modality Bias Hypothesis (MBH) based on the Modality Bias Venn diagram (MBVD) to systematically describe the relationship between modality bias and robustness to missing modality.

Modality Bias Venn Diagram As shown in Figure 4 on the left, the MBVD depicts the components of the latent decisive feature of multimodal systems in the form of a Venn Diagram. It is a variant of the Modality Venn Diagram (MVD) employed in multimodal knowledge distillation [28]. Without loss of generality, we take AVSR as an

example and define $\mathcal{X}^a$, $\mathcal{X}^v$, and $\mathcal{Y}$ as the original feature space of audio, video and label space, respectively. The decisive feature z, commonly a form of intermediate layer representation, consists of two modality components z^a (blue circles) and z^v (green circle). We denote $I(\cdot)$ as mutual information and $I(\cdot|\cdot)$ as conditional mutual information. The task-relevant decisive feature z^u ($I(z, y)$) is depicted by the shaded region and can be further divided into three components. z^g ($I(z^a, z^v, y)$) represents modality-general decisive features, while z^{sa} ($I(z^u, z^a|z^g)$) and z^{sv} ($I(z^u, z^v|z^g)$) represent modality-specific decisive features. We denote their proportions in z^u as α, β, and γ, respectively. These features collectively contribute to determining the final task output $\hat{y}$. For AVSR, a higher α represents a greater decision bias of the model on the audio modality, focusing more on speech than lip movements. A larger γ indicates a model's inclination towards modality synergy by maximizing the mutual information between modalities for decision-making, as in some modality-balanced models [26, 27]. Furthermore, z^u is generated by the original features x^a, x^v as $g\left(x^a, x^v; \phi\right)$, where $g(\phi)$ can be seen as a neural network-based transfer such as an encoder with parameters ϕ. Therefore, the decision process of the multimodal system can be decomposed into two steps, following the Bayesian process: the MBVD hidden decisive feature generation step and the decision step:

$$P\left(y \mid x^a, x^v\right) = P\left(y \mid z^\mu\right) P\left(z^\mu \mid x^a, x^v\right) \qquad (1)$$

Modality Bias Hypothesis Based on MBVD, we give a systematic description of the relationship between modality bias and robustness to missing modality in the view of MBH. As shown in Figure 4 on the right, by applying dropout with different rates $k_i \in [0, 1]$ on video training data, the original video feature space $\mathcal{X}^v$ can be split into a series of subsets $\{\mathcal{X}^v_{k_1}, \mathcal{X}^v_{k_2}, ..., \mathcal{X}^v_{k_n}\}$. The samples from space $\mathcal{X}^a \times \mathcal{X}^v_{k_i}$ are denoted as dyads $\left(x^a, x^v_{k_i}\right)$. Compared to the model trained on complete multimodal datas $(x^a, x^v_{0.0})$, the model trained on data pairs (x^a, x^v_θ) with a video dropout rate $\theta_{train} \in (0.0, 1.0)$ exhibits a greater decision bias on audio modality with larger α, smaller β, and γ. As θ approaches 1.0, the task-relevant decisive feature z_u becomes steadily dominated by the audio-specific decisive feature z_a, resulting in a transformation from a bimodal distribution in the latent representation subspace to a unimodal one. The decision pattern of the multimodal model shifts from $p(y|z_u)$ to $p(y|z_a)$.

During the inference stage, these multimodal models display different modality biases. For the model trained on complete multimodal data or dropout on audio with a larger γ, they tend to search general information shared among modalities. This hypothesis effectively explains the observed experimental phenomena in previous studies. For modality-biased models, such as Multimodal Sentiment Analysis (MSA) [22] dominated by text, Multimodal Speech Enhancement (MSE) [29] dominated by audio, as well as AVSR dominated by audio [21, 23, 30], it has been observed that applying dropout on the primary modality helps alleviate modality bias and brings about slight improvements when dealing with complete input. On the other hand, the AVSR model with larger α and smaller γ values tends to focus more on speech and neglect complementary information from lip movements. When dealing with partially or completely missing video data, the model with larger α shows its robustness, which aligns well with the aforementioned experimental observations.

4. Multimodal Distribution Approximation with Knowledge Distillation (MDA-KD)

For the robustness training of modality-bias systems, it is crucial to avoid dropout-induced modality bias on the primary modality. Dropout indeed alleviates the OOD problem to some extent but encourages multimodal models to pursue trivial solutions at the same time. Ideal robust multimodal models are expected to achieve two goals: (1) learn to extract mutual information across modalities rather than relying on a certain modality when facing complete paired input, and (2) learn to complement information from the other modality and utilize context information from adjacent frames. To prevent excessive modality bias caused by dropout, we propose a novel Multimodal Distribution Approximation with Knowledge Distillation (MDA-KD) framework to constrain the distribution of the multimodal feature space during the robustness training phase.

Unlike traditional knowledge distillation methods, firstly, the teacher model is trained on the complete multimodal data pairs, while the student model is trained on missing video data. The teacher model is relatively unbiased with a higher proportion of modality-general decisive features z^g in the MBVD space. During the training process of the student model, the teacher model serves as an anchor point, preventing the student model from shifting towards a unimodal distribution on the audio modality. Note that the difference between teacher and student models in our method is modality bias varies, rather than size, architecture as in common KD methods [31–34]. Additionally, distillation occurs at the hidden layer rather than the logistic outputs, aiming to minimize the distances between decision distribution samples of the teacher and student models and further constrain the intermediate representation subspace distribution of the student model. In practice, we take the knowledge from the intermediate representation of the cross-modal encoder layers.

Here, we adopt the symbol definitions from Section 3 and provide a formal description of MDA-KD. For a naturally modal-biased multimodal system, the data samples from original feature space $\mathcal{X}^a \times \mathcal{X}^v_{k_i} \times \mathcal{Y}$ can be de-

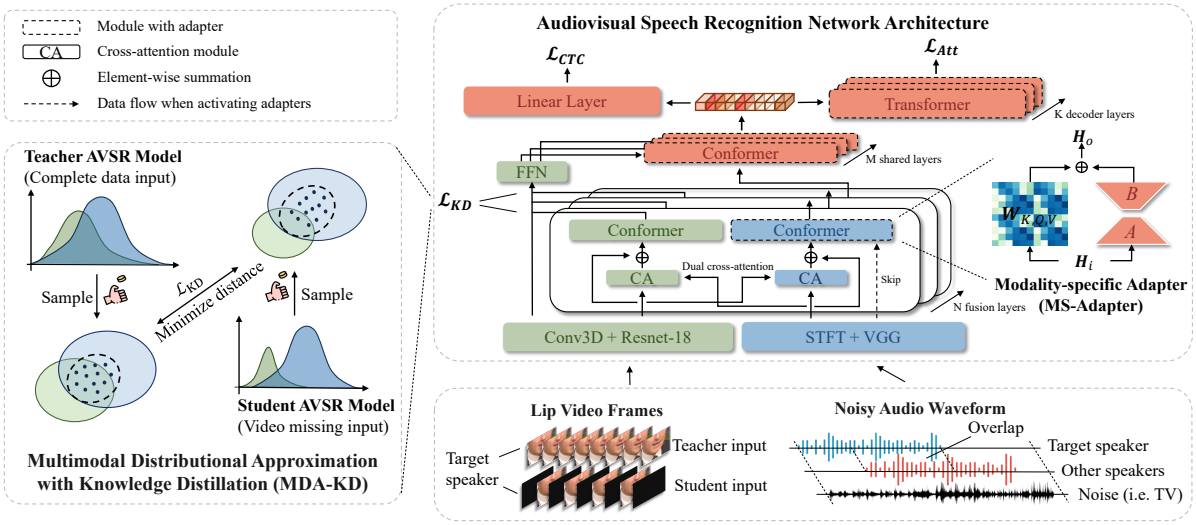


Figure 5. Overall framework of the proposed AVSR system. We address challenging real-world scenarios involving missing video frames and noisy speech with an overlap rate exceeding 40% during both the training and testing stages. In MDA-KD, latent knowledge is sampled from the latent distribution of the teacher model with complete data input. This latent konwledege serves as an anchor point to prevent dropout-induced modality bias during the robustness training of the student network. For entirely missing video input, the MS-Adapter is activated to enable a dynamic decision switch.

noted as triples $\left(x^a, x^v_{k_i}, y\right)$. For simplicity, we denote $x^v_{0.0}$ as x^v. The teacher model $Te(\phi)$ is first trained on complete multimodal data (x^a, x^v, y) model with parameters ϕ, and the model's decision process can be formulated as $P_{te}\left(y \mid x^a, x^v\right)$ in a Bayesian decision problem. We assume that the teacher model is a neural network $g\left(\phi\right)$ and it is trained by minimizing the following loss function, a form of multitask learning.

$$Te(\phi) = \min_{\phi} \mathcal{L}_{\text{MLT}}\left(g\left(x^a, x^v; \phi\right), y\right), \tag{2}$$

$$\begin{aligned}\mathcal{L}_{\text{MLT}}(x^a, x^v; \phi) = \lambda \log P_{\text{CTC}}(y \mid x^a, x^v) \\ + (1-\lambda) \log P_{\text{Att}}(y_i \mid x^a, x^v),\end{aligned} \tag{3}$$

where the tunable parameter $\lambda \in [0, 1]$ is used to balance the sequence-level Connectionist Temporal Classification (CTC) loss and the frame-wise Cross Entropy (CE) loss, which serve as the standard end-to-end ASR training objectives. During the training of the student model, the dropout strategy is applied to the secondary modality v, while the teacher model is frozen with complete multimodal data as input. It is important to note that the student and teacher models have the same network architecture. From the perspective of MBVD, the whole decision process of the multimodal model can be divided into hidden feature generation step and decision step.

$$P_{st}\left(y \mid x^a, x^v_{k_i}\right) = P_{st}\left(y \mid z^\mu\right) P_{st}\left(z^\mu \mid x^a, x^v_{k_i}\right), \tag{4}$$

$$P_{te}\left(y \mid x^a, x^v\right) = P_{te}\left(y \mid z^\mu\right) P_{te}\left(z^\mu \mid x^a, x^v\right), \tag{5}$$

where $z^\mu \in \mathbb{R}^{d\mu}$ represents the combined representation of modality-specific decisive features $z^{sa} \in \mathbb{R}^{da}$, $z^{sv} \in \mathbb{R}^{dv}$, and modality-general decisive features $z^g \in \mathbb{R}^{dg}$. The tuple (z^{sa}, z^{sv}, z^g) represents a sample drawn from the MBVD hidden features space, denoted as $\mathcal{Z}^{sa} \times \mathcal{Z}^{sv} \times \mathcal{Z}^g$.

Initialized on the parameter of the teacher model, we introduce an additional loss term to constrain the dynamic process of the student model's MBVD feature distribution in robust training. The distance between batch samples from the student and the teacher model is used to approximate the difference of distribution, which serves as a form of frame-level knowledge distillation.

$$\begin{aligned}\mathcal{L}_{\text{KD}}(x^a, x^v, x^v_k; \phi_{te}, \phi_{st}) &= \text{KL}\left(S_{te}, S_{st}\right), \\ S_{te} &= \sigma_T\big(\text{Sample}(P_{te}(z^\mu \mid x^a, x^v))\big), \\ S_{st} &= \sigma_T\big(\text{Sample}(P_{st}(z^\mu \mid x^a, x^v_{k_i}))\big),\end{aligned} \tag{6}$$

where $\sigma_T(x)$ denotes the SoftMax function with temperature T and $Sample$ represents the sample function. This distribution approximation serves two main purposes. Firstly, during training, when the student network encounters a missing modality feature $x^v_{k_i}$, the convergence of the student's decisive feature $z^u = g(x^a, x^v_{k_i}; \phi_{st})$ towards the teacher's decisive feature $z^u = g(x^a, x^v; \phi_{te})$ encourages the utilization of contextual information from $x^v_{k_i}$. Additionally, with the dual cross-attention design, the process complements the information extracted from x^a, effectively addressing the condition of missing frames and promoting out-of-distribution generality. On the other hand, the KD loss is used to minimize the distance between the distributions of the teacher and student models, preventing the student model from converging to trivial solutions. Subsequently, we train the student model jointly with a weighted sum of the standard training loss and distillation loss:

$$\mathcal{L}_{\text{MLT}}(x^a, x^v, x_k^v; \phi_{te}, \phi_{st}) = \beta \mathcal{L}_{\text{KD}}(x^a, x^v, x_k^v; \phi_{te}, \phi_{st}) + (1-\beta)\mathcal{L}_{\text{MLT}}(x^a, x_k^v; \phi_{st}). \quad (7)$$

5. Modality-Specific Adapter (MS-Adapter)

As illustrated in Figure 4 on the right, when facing severely or entirely missing video data, we consider it unreliable to continue employing a synergistic decision-making strategy like MDA-KD with relatively high values of γ and β. Padding frames lack sufficient contextual information and may introduce noise. Therefore, in such scenarios, a dynamic switch in decision strategy from $P(y|z^u)$ to $P(y|z^a)$ is necessary as a complement to MDA-KD. In view of the success of adapters applied in foundation model fine-tuning [35–38], we attempt to extend it to address the modality missing issue in multimodal models. For clarity, we refer to this extension as Modality-Specific Adapter (MS-Adapter). Specifically, LORA [39] is adopted to self-attention layers in the audio branch, marked with a dashed box in Figure 5. These adapters perform residual-style feature blending with the original pre-trained features. The residual weight could be represented as low-rank matrices $\Delta W \in \mathbb{R}^{d\times d}$, and it could be decomposed into a pair of fan-in and fan-out linear layers with weights $A \in \mathbb{R}^{r\times d}$ and $B \in \mathbb{R}^{d\times r}$ ($r \ll d$). The reparametrization operation can be formulated below.

$$H_o = H_i(W_0 + \Delta W) = H_i(W_0 + BA) \quad (8)$$

By activating the MS-Adapter, we can dynamically switch the decision-making pattern by activating the adapters. We highlight two advantages of the MS-Adapter. First, a substantial amount of unpaired unimodal training data and data augmentation techniques could be used in the training process of the adapters. Second, the adapter training process provides an opportunity to modify the computation pathway. As illustrated in Figure 5 with dashed arrows, in both training and inference stage with audio-only input, the computation flow of the video branch will be directly cut off, and the modality fusion cross-attention module will be skipped to reduce computational costs.

6. Experiment Settings

Dataset We conduct our experiments on MISP2021 [24] and MISP2022 [40]. These two open-source datasets present a large-scale audio-visual corpus recorded in real-life home TV scenarios with multiple groups of speakers chatting simultaneously. Multiple microphone arrays and cameras are used to collect far/middle/near-field audio and far/middle-field video. Compared to the carefully recorded videos in LRS2 [1] and LRS3 [2] from BBC interviews and TED talks, MISP datasets offer static shooting perspectives with diverse resolutions, including naturally blurred and obstructed frames. The videos are accompanied by various background noises and high speech overlap rates (42% in training set and 49% in test set). Compared oracle segment-level AVSR task in MISP201, MISP2022 presents a more challenging task of session-level AVSR without oracle speaker diarization results. To avoid limitations associated with noise simulation, all experiments are evaluated exclusively on far-field data, which aligns well with common in-car, office meeting, or smart home scenarios.

Implementation Detail We strictly adhere to the approaches outlined in [18] for model training and network architectures. We initialize the AVSR model with two pre-trained unimodal models and fine-tune it in an end-to-end manner. As shown in Figure 5, the AVSR model is a dual-branch network where $N = 3$, $M = 9$ and $K = 6$. For the loss function in Equation 3, we set λ to 0.7 and CTC loss consists of the same weighted intermediate CTC [41] losses in 3, 6, 9, 12 layers. In Equation 4, we use 0.1 for β. We follow [18] to establish two baselines A0 and AV0 trained on complete modality data with dropout techniques. AV0 is fine-tuned based on A0 and a pre-trained ResNet-18 encoder with a 3D-CNN head. [3]

Dropout Settings Similar to [19], we evaluate the robustness to missing video modality with various dropout methods and rates: Segment Dropout, Utterance Dropout, and Interval Dropout. Testing involves dropout rates from 0.0 to 1.0 in 0.25 intervals. Results from the three dropout methods are averaged at each rate to obtain overall dropout results. When conducting ablation studies, segments with naturally missing video frames (17%) are excluded from the test set, ensuring a consistent and controlled video missing rate. In our method, during training, A certain proportion of sample is assigned a random dropout method from the above three methods and an extra one from [21] with an optimized dropout rate. In both training and testing stages, we pad the missing video frame pixels with zeros instead of using interpolation or repetition methods. We conduct a hyper-parameter search over the training dropout rate and found that 0.5 is optimal for our method (when D_prod is 0.5). This rate implies that half of the video frames in a selected sample are padded with zeros. [3]

7. Experiments and Result Analysis

7.1. Overall Comparison of Experiment Settings

In Table 1, we conduct key parameter analysis and abolation study of the proposed methods on the MISP2022 dataset with oracle speaker diarization results. We first explore the impact of dropout probability in training videos. In contrast to AV1, AV2 introduces half of the complete data pairs. As a result, it mitigates dropout-induced modality

[3] More details can be found in Appendix.

Model	Training setttings					Test dropout rate				
	Dropout	D_{prob}	Init.	MDA-KD	MS-Adapter	0.00	0.25	0.50	0.75	1.00
A0	✗	0.0	Random	✗	✗	25.13	25.13	25.13	25.13	25.13
AV0	✗	0.0	A0	✗	✗	21.14	23.77	25.57	25.87	26.65
AV1	✔	1.0	A0	✗	✗	23.26	23.68	24.27	24.95	25.91
AV2	✔	0.5	A0	✗	✗	21.72	22.56	23.37	24.46	25.64
AV3	✔	0.5	AV0	✗	✗	21.53	22.47	23.65	24.55	25.90
AV4	✔	0.5	AV0	✔	✗	21.38	22.18	23.20	24.40	25.70
AV5	✔	0.5	AV0	✔	✗	21.11	21.77	22.78	24.02	25.45
AV6	✔	0.5	AV0	✔	✔	**21.11**	**21.77**	**22.78**	**24.02**	**24.94**

Table 1. An overall comparison in CER (%) of different system configurations. Different from the dropout rate, D_{prob} represents the proportion of data with missing frames in the training set. Init. refers to the network initialization method.

Insert part	Rank	DA	Params(MB)	CER(%)
Encoder	32	✗	4.50	25.35
Encoder	32	✔	4.50	25.08
En&Decoder	32	✔	9.00	25.20
Encoder	64	✔	9.00	25.08
En&Decoder	64	✔	18.00	25.05
Encoder	128	✔	18.00	25.01
En&Decoder	128	✔	36.00	**24.94**

Table 2. Performance analysis of MS-Adapter. DA means data augmentation, including speed perturbation and utterance concat.

bias to some extent, since a higher proportion of complete data tend to encourages the model to learn general information across modalities. This finding aligns with previous research [19], highlighting the superiority of utterance dropout over random frame dropout (the former means a larger D_{prob}). Next, AV3 is trained based on AV0, which means the subsequent optimized processing starts from a relatively stable convergence state with complete input. In the robust training stage, the balanced state tends to be disrupted when trained on incomplete modality pairs, searching for a new optimization coverage range. However, when trained on complete data pairs, the scenario is reversed. Thus, while AV3 outperforms AV2 with low test missing rates, it lags behind when facing severe video absence, illustrating a tug-of-war dynamic without clear guidance.

Next, we validate the effectiveness of MDA-KD. Compared with AV3, AV5 demonstrates superior performance for both complete and missing video modality inputs. AV4 successfully achieves our goal of enhancing robustness without any performance degradation on complete input (21.11% vs. 21.14%). This implies that the teacher model AV0 provides an explicitly optimized target in robustness training. It effectively constrains the distribution shift to the audio modality, preventing excessive modality bias caused by dropout. Furthermore, in AV4, we restrict the flow of audio data into the video branch within the dual cross-attention module. Consequently, a performance drop is observed across all test suites, highlighting the effectiveness of MDA-KD in leveraging the dual cross-attention module to extract modality-general information from audio for complementing missing information. Subsequently, we integrate MS-Adapters into the audio branch in AV6 based on AV5. Consequently, the performance with audio-only input improves to a 24.94% CER, surpassing A0 for the first time (24.94% vs. 25.13%). These results show the effectiveness of MS-Adapters by dynamically switching to the decision patterns on audio modality with audio-only input.

Method	Test dropout rate				
	0.00	0.25	0.50	0.75	1.00
Cascade Utt [19]	22.54	23.89	25.23	26.05	28.15
AV Dropout Utt [21]	22.00	23.37	25.35	26.21	26.78
Dropout Utt [20]	22.08	23.21	24.56	25.08	25.46
Ours	**21.11**	**21.77**	**22.78**	**24.02**	**24.94**

Table 3. A CER(%) comparison with other dropout methods.

7.2. Validation of MS-Adapter

We further explore three key factors in MS-Adapter adaptation: data augmentation, insert part and rank dimension. In Table 2, we observe a decrease in CER from 25.45% (AV4) to 25.35%, and it further improves to 25.08% with data augmentation doubling audio training data. These results suggest that the adapter adaptation effectively enhances the robustness of AVSR with completely missing video, requiring only an additional 4.50MB in parameters. It provides an opportunity to apply data augmentation that is effective for unimodal model training and to use extra unpaired data. Next, increasing the ranks and the quantity of adapters results in further performance gains at the expense of a larger parameter. The best performance, achieving 24.94%, is shown in the bottom row and attained with the adapter inserted in both encoder and decoder blocks.

7.3. Comparisons with Other Dropout Techniques

As shown in Table 3, we compare our proposed framework with three widely used dropout techniques [19–21]. Cascade Utt employs a separable cascade structure, where an AV model is superimposed on an audio-only model. Inputs are then routed through either the audio-only path or

Benchmark	System	Training Data		Backbone	Obj. Function	CER / cpCER(%)
		A	V			
MISP2021	SJTU [42]	300 hours	LRW-1000	Conformer	ED + SE	34.02
	NIO [43]	3300 hours	LRW-1000 [4]	Transformer	ED	25.07
	USTC [18]	500 hours	w/o extra data	Conformer	ED	24.58
	Ours	1000 hours	w/o extra data	Conformer	ED + InterCTC	**21.53**
MISP2022	NIO [44]	3300 hours	LRW-1000	Conformer	ED	29.58
	XMU [45]	2100 hours	LRW-1000	Conformer	ED + InterCTC	31.88
	NPU [46]	1300 hours	w/o extra data	E-Branchformer	ED + InterCTC	29.13
	Ours	1000 hours	w/o extra data	Conformer	ED + InterCTC	**28.06**

Table 4. A Comparison of the state-of-the-art systems. InterCTC refers to Intermediate CTC loss [41], the ED loss is formulated in Equation (3) and SE represents the mean square error loss. We use evaluate the performance using the concatenated minimum-permutation character error rate (cpCER) [47] metric for the session-level AVSR task.

the AV path with a probability of p_1. AV Dropout Utt randomly drops either the entire video or the entire audio segments with a probability of p_2. Dropout Utt exclusively drops the video segments with a probability of p_3. We adopt the optimal dropout settings from [19], where $p_1 = 0.25$, $p_2 = 0.25$, and $p_3 = 0.5$. For Cascade Utt, we follow [19] to build the network and maintain comparable parameters numbers. As a result, our proposed methods outperforms the other three techniques in all test suites and does not cause performance degradation.

7.4. Comparisons with State-of-the-art Systems

Finally, we compare our system with the state-of-the-art systems on the MISP2021 and MISP2022 challenges[18, 42–45, 48] as shown in Table 4. With Recognizer Output Voting Error Reduction (ROVER) [49], we rescore the output transcripts of A0, AV0, and A6 mentioned in Table 1. In the MISP2021 utterance-level AVSR challenge with oracle speaker diarization results, our system outperforms the previous SOTA system by achieving an absolute CER reduction of 3.05% from 24.58% to 21.53%. Our top-performing system, AV6, attains a CER of 22.13%. Moving to the MISP2022 session-level AVSR challenge, we build our diarization system closely adhering to [50]. We secure a ROVER cpCER score of 28.06% and obtain the best system score with a cpCER of 28.55%. When oracle segmentations are utilized, our system achieves a ROVER CER score of 21.80% and the best model score of 21.53% in CER.

8. Related Works

Modality Missing in Multimodal Learning The prevalent issue of missing modalities in multimodal applications has prompted research that specifically targets severe modality absences. Generative models [51, 52] and meta-learning predict missing modalities using available or few-shot paired samples. Balanced models utilize joint multimodal representations [53–55]. Models addressing modality bias employ data augmentation methods like modality dropout [19, 22] to tackle out-of-distribution challenges. For AVSR, we prioritize efficiency and opt for dropout due to its plug-and-play nature and lightweight implementation. More discussion could be found in Appendix.

Video Modality Robustness in AVSR To enhance performance on low-resolution videos, visual extractors are commonly pre-trained on relatively high-quality videos with isolated words [5] or acoustic pseudo-labeling classification tasks [18]. Addressing situations involving corruption, Hong et al. [17] have designed an explicit scoring module to identify reliable streams and effectively manage input scenarios. Regarding the issue of missing video frames, most researchers have applied dropout techniques to enhance missing robustness [19–23]. In classical dropout methods, frame level dropout is utilized in [23] and utterance-level dropout is applied in AV-Hubert [21]. As a recent work focusing on this issue, Chang et al. [19] unify test suites of missing videos. However, the proposed binary evaluation metric overly emphasizes relative robustness trends, neglecting absolute performance. Compared to the methods mentioned earlier, we explore the problem of missing video frames from the perspective of modality bias. Leveraging classical techniques and simple designs, our approach achieves both performance and robustness without introducing additional inference time. It adapts to various scenarios of frame absence through a unified model.

9. Conclusion

In this work, we discover and analyze the essence of dropout-induced modality bias. Based on these findings, we proposed MBH to provide a systematic description of the relationship between modality bias and missing robustness in multimodal systems. Consequently, we propose a new multimodal distribution approximation with knowledge distillation approach to deal with missing video frames for AVSR. Furthermore, we apply adapters to handle videos with both severe and complete missing rates. For future work, we intend to validate our findings in this study across a wide range of multimodal applications beyond AVSR.

References

[1] Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Zisserman. Lip reading sentences in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6447–6456, 2017. 1, 6

[2] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. LRS3-TED: a large-scale dataset for visual speech recognition. *arXiv preprint arXiv:1809.00496*, 2018. 1, 6

[3] Hang Chen, Jun Du, Yusheng Dai, Chin Hui Lee, Sabato Marco Siniscalchi, Shinji Watanabe, Odette Scharenborg, Jingdong Chen, Bao Cai Yin, and Jia Pan. Audio-visual speech recognition in misp2021 challenge: Dataset release and deep analysis. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, volume 2022, pages 1766–1770, 2022. 3

[4] Shuang Yang, Yuanhang Zhang, Dalu Feng, Mingmin Yang, Chenhao Wang, Jingyun Xiao, Keyu Long, Shiguang Shan, and Xilin Chen. Lrw-1000: A naturally-distributed large-scale benchmark for lip reading in the wild. In *2019 14th IEEE international conference on automatic face & gesture recognition (FG 2019)*, pages 1–8. IEEE, 2019. 1, 8

[5] Pingchuan Ma, Stavros Petridis, and Maja Pantic. End-to-end audio-visual speech recognition with conformers. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7613–7617. IEEE, 2021. 1, 8

[6] Xichen Pan, Peiyu Chen, Yichen Gong, Helong Zhou, Xinbing Wang, and Zhouhan Lin. Leveraging unimodal self-supervised learning for multimodal audio-visual speech recognition, 2022. 1

[7] Chen Chen, Yuchen Hu, Qiang Zhang, Heqing Zou, Beier Zhu, and Eng Siong Chng. Leveraging modality-specific representations for audio-visual speech recognition via reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 12607–12615, 2023. 1

[8] Bo Xu, Cheng Lu, Yandong Guo, and Jacob Wang. Discriminative multi-modality speech recognition. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pages 14433–14442, 2020. 1

[9] Jianwei Yu, Shi-Xiong Zhang, Jian Wu, Shahram Ghorbani, Bo Wu, Shiyin Kang, Shansong Liu, Xunying Liu, Helen Meng, and Dong Yu. Audio-visual recognition of overlapped speech for the lrs2 dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6984–6988. IEEE, 2020.

[10] Joanna Hong, Minsu Kim, Daehun Yoo, and Yong Man Ro. Visual context-driven audio feature enhancement for robust end-to-end audio-visual speech recognition. *arXiv preprint arXiv:2207.06020*, 2022. 1

[11] Alexandros Haliassos, Pingchuan Ma, Rodrigo Mira, Stavros Petridis, and Maja Pantic. Jointly learning visual and auditory speech representations from raw data. *arXiv preprint arXiv:2212.06246*, 2022. 1

[12] Pingchuan Ma, Alexandros Haliassos, Adriana Fernandez-Lopez, Honglie Chen, Stavros Petridis, and Maja Pantic. Auto-AVSR: Audio-visual speech recognition with automatic labels. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023. 1

[13] George Sterpu, Christian Saam, and Naomi Harte. Attention-based audio-visual fusion for robust automatic speech recognition. In *Proceedings of the 20th ACM International conference on Multimodal Interaction*, pages 111–115, 2018. 1

[14] George Sterpu, Christian Saam, and Naomi Harte. How to teach DNNs to pay attention to the visual modality in speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:1052–1064, 2020.

[15] Yuchen Hu, Ruizhe Li, Chen Chen, Heqing Zou, Qiushi Zhu, and Eng Siong Chng. Cross-Modal Global Interaction and Local Alignment for Audio-Visual Speech Recognition. *arXiv preprint arXiv:2305.09212*, 2023. 1

[16] Triantafyllos Afouras, Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Zisserman. Deep audio-visual speech recognition. *IEEE transactions on pattern analysis and machine intelligence*, 44(12):8717–8727, 2018. 1

[17] Joanna Hong, Minsu Kim, Jeongsoo Choi, and Yong Man Ro. 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*, pages 18783–18794, 2023. 1, 8

[18] Yusheng Dai, Hang Chen, Jun Du, Xiaofei Ding, Ning Ding, Feijun Jiang, and Chin-Hui Lee. Improving Audio-Visual Speech Recognition by Lip-Subword Correlation Based Visual Pre-training and Cross-Modal Fusion Encoder. In *2023 IEEE International Conference on Multimedia and Expo (ICME)*, pages 2627–2632. IEEE, 2023. 1, 6, 8, 3

[19] Oscar Chang, Otavio de Pinho Forin Braga, Hank Liao, Dmitriy Dima Serdyuk, and Olivier Siohan. On robustness to missing video for audiovisual speech recognition. *Transactions on Machine Learning Research (TMLR)*, 2022. 1, 6, 7, 8, 3, 4

[20] Takaki Makino, Hank Liao, Yannis Assael, Brendan Shillingford, Basilio Garcia, Otavio Braga, and Olivier Siohan. Recurrent neural network transducer for audio-visual speech recognition. In *2019 IEEE automatic speech recognition and understanding workshop (ASRU)*, pages 905–912. IEEE, 2019. 1, 7, 4

[21] Bowen Shi, Wei-Ning Hsu, Kushal Lakhotia, and Abdelrahman Mohamed. Learning audio-visual speech representation by masked multimodal cluster prediction, 2022. 4, 6, 7, 8

[22] Devamanyu Hazarika, Yingting Li, Bo Cheng, Shuai Zhao, Roger Zimmermann, and Soujanya Poria. Analyzing modality robustness in multimodal sentiment analysis, 2022. 4, 8

[23] Shiliang Zhang, Ming Lei, Bin Ma, and Lei Xie. Robust audio-visual speech recognition using bimodal DFSMN with multi-condition training and dropout regularization. In *ICASSP 2019-2019 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 6570–6574. IEEE, 2019. 1, 4, 8

[24] Hang Chen, Hengshun Zhou, Jun Du, Chin-Hui Lee, Jingdong Chen, Shinji Watanabe, Sabato Marco Siniscalchi,

Odette Scharenborg, Di-Yuan Liu, Bao-Cai Yin, Jia Pan, Jian-Qing Gao, and Cong Liu. The First Multimodal Information Based Speech Processing (Misp) Challenge: Data, Tasks, Baselines And Results. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 9266–9270, 2022. 1, 6

[25] Jun Du, Chin-Hui Lee, Jingdong Chen, Shinji Watanabe, Sabato Marco Siniscalchi, and Odette Scharenborg. Multimodal Information Based Speech Processing (MISP) Challenge 2022. https://mispchallenge.github.io/mispchallenge2022/, 2022. Accessed: 2023-06-26. 1

[26] Jinming Zhao, Ruichen Li, and Qin Jin. Missing modality imagination network for emotion recognition with uncertain missing modalities. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2608–2618, 2021. 2, 4

[27] Mengmeng Ma, Jian Ren, Long Zhao, Davide Testuggine, and Xi Peng. Are multimodal transformers robust to missing modality? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18177–18186, 2022. 2, 4

[28] Zihui Xue, Zhengqi Gao, Sucheng Ren, and Hang Zhao. The modality focusing hypothesis: Towards understanding cross-modal knowledge distillation, 2022. 3, 1

[29] Hang Chen, Jun Du, Yu Hu, Li-Rong Dai, Bao-Cai Yin, and Chin-Hui Lee. Correlating subword articulation with lip shapes for embedding aware audio-visual speech enhancement. *Neural Networks*, 143:171–182, 2021. 4

[30] Pan Zhou, Wenwen Yang, Wei Chen, Yanfeng Wang, and Jia Jia. Modality attention for end-to-end audio-visual speech recognition. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6565–6569. IEEE, 2019. 4

[31] Xianing Chen, Qiong Cao, Yujie Zhong, Jing Zhang, Shenghua Gao, and Dacheng Tao. Dearkd: data-efficient early knowledge distillation for vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12052–12062, 2022. 4

[32] Francisco Rivera Valverde, Juana Valeria Hurtado, and Abhinav Valada. There is more than meets the eye: Self-supervised multi-object detection and tracking with sound by distilling multimodal knowledge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11612–11621, 2021.

[33] Zihui Xue, Sucheng Ren, Zhengqi Gao, and Hang Zhao. Multimodal knowledge expansion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 854–863, 2021.

[34] Baoyun Peng, Xiao Jin, Jiaheng Liu, Dongsheng Li, Yichao Wu, Yu Liu, Shunfeng Zhou, and Zhaoning Zhang. Correlation congruence for knowledge distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5007–5016, 2019. 4

[35] Renrui Zhang, Rongyao Fang, Wei Zhang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-adapter: Training-free clip-adapter for better vision-language modeling. *arXiv preprint arXiv:2111.03930*, 2021. 6

[36] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International Conference on Machine Learning*, pages 2790–2799. PMLR, 2019.

[37] Sylvestre-Alvise Rebuffi, Hakan Bilen, and Andrea Vedaldi. Learning multiple visual domains with residual adapters. *Advances in neural information processing systems*, 30, 2017.

[38] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters. *International Journal of Computer Vision*, pages 1–15, 2023. 6

[39] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021. 6

[40] Zhe Wang, Shilong Wu, Hang Chen, Mao-Kui He, Jun Du, Chin-Hui Lee, Jingdong Chen, Shinji Watanabe, Sabato Siniscalchi, Odette Scharenborg, et al. The multimodal information based speech processing (misp) 2022 challenge: Audio-visual diarization and recognition. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023. 6

[41] Jaesong Lee and Shinji Watanabe. Intermediate loss regularization for ctc-based speech recognition. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6224–6228. IEEE, 2021. 6, 8

[42] Wei Wang, Xun Gong, Yifei Wu, Zhikai Zhou, Chenda Li, Wangyou Zhang, Bing Han, and Yanmin Qian. The sjtu system for multimodal information based speech processing challenge 2021. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 9261–9265. IEEE, 2022. 8

[43] Gaopeng Xu, Song Yang, Wei Li, et al. Channel-Wise AV-Fusion Attention for Multi-Channel Audio-Visual Speech Recognition. In *Proc. ICASSP 2022*, pages 9251–9255. IEEE, 2022. 8

[44] Sang Wang Gaopeng Xu, Xianliang Wang et al. The NIO system for audio-visual diarization and recognition in MISP challenge 2022. https://mispchallenge.github.io/mispchallenge2022/papers/task2/Track2_NIO.pdf, 2022. 8

[45] Tao Li, Haodong Zhou, Jie Wang, Qingyang Hong, and Lin Li. The XMU System for Audio-Visual Diarization and Recognition in MISP Challenge 2022. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–2. IEEE, 2023. 8

[46] He Wang, Pengcheng Guo, Pan Zhou, and Lei Xie. Mlca-avsr: Multi-layer cross attention fusion based audio-visual speech recognition. *arXiv preprint arXiv:2401.03424*, 2024. 8

[47] Shinji Watanabe, Michael Mandel, Jon Barker, Emmanuel Vincent, Ashish Arora, Xuankai Chang, Sanjeev Khudan-

pur, Vimal Manohar, Daniel Povey, Desh Raj, et al. Chime-6 challenge: Tackling multispeaker speech recognition for unsegmented recordings. *arXiv preprint arXiv:2004.09249*, 2020. 8

[48] Pengcheng Guo, He Wang, Bingshen Mu, Ao Zhang, and Peikun Chen. The NPU-ASLP System for Audio-Visual Speech Recognition in MISP 2022 Challenge. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–2. IEEE, 2023. 8

[49] Jonathan G Fiscus. A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (ROVER). In *Proc. asrU 1997*, pages 347–354. IEEE, 1997. 8

[50] Ming Cheng, Haoxu Wang, Ziteng Wang, Qiang Fu, and Ming Li. The whu-alibaba audio-visual speaker diarization system for the misp 2022 challenge. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–2. IEEE, 2023. 8

[51] Qiuling Suo, Weida Zhong, Fenglong Ma, Ye Yuan, Jing Gao, and Aidong Zhang. Metric Learning on Healthcare Data with Incomplete Modalities. In *IJCAI*, volume 3534, page 3540, 2019. 8, 4

[52] Lei Cai, Zhengyang Wang, Hongyang Gao, Dinggang Shen, and Shuiwang Ji. Deep adversarial learning for multi-modality missing data completion. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 1158–1166, 2018. 8, 4

[53] Zilong Wang, Zhaohong Wan, and Xiaojun Wan. Transmodality: An end2end fusion method with transformer for multimodal sentiment analysis. In *Proceedings of The Web Conference 2020*, pages 2514–2520, 2020. 8, 4

[54] Hai Pham, Paul Pu Liang, Thomas Manzini, Louis-Philippe Morency, and Barnabás Póczos. Found in translation: Learning robust joint representations by cyclic translations between modalities. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6892–6899, 2019. 4

[55] Jiale Li, Hang Dai, Hao Han, and Yong Ding. Mseg3d: Multi-modal 3d semantic segmentation for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21694–21704, 2023. 8

[56] A Varga, HJM Steeneken, et al. Noisex-92: A database and an experiment to study the effect of additive noise on speech recognition systems. *Speech Commun*, 12(3):247–253, 1993. 1

[57] Lukas Drude, Jahn Heymann, Christoph Boeddeker, and Reinhold Haeb-Umbach. Nara-wpe: A python package for weighted prediction error dereverberation in numpy and tensorflow for online and offline processing. In *Speech Communication; 13th ITG-Symposium*, pages 1–5. VDE, 2018. 1

[58] Christoph Boeddecker, Jens Heitkaemper, Joerg Schmalenstroeer, et al. Front-end processing for the CHiME-5 dinner party scenario. In *Proc. CHiME 2018*, pages 35–40, 2018. 1, 3

[59] Desh Raj, Daniel Povey, and Sanjeev Khudanpur. GPU-accelerated guided source separation for meeting transcription, 2022. 3

[60] Mengmeng Ma, Jian Ren, Long Zhao, Sergey Tulyakov, Cathy Wu, and Xi Peng. Smil: Multimodal learning with severely missing modality. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 2302–2310, 2021. 4

[61] Itai Gat, Idan Schwartz, Alexander Schwing, and Tamir Hazan. Removing bias in multi-modal classifiers: Regularization by maximizing functional entropies. *Advances in Neural Information Processing Systems*, 33:3197–3208, 2020. 4

[62] Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2901–2910, 2017. 4

[63] Yangyang Guo, Liqiang Nie, Harry Cheng, Zhiyong Cheng, Mohan Kankanhalli, and Alberto Del Bimbo. On modality bias recognition and reduction. *ACM Transactions on Multimedia Computing, Communications and Applications*, 19(3):1–22, 2023. 4

A Study of Dropout-Induced Modality Bias on Robustness to Missing Video Frames for Audio-Visual Speech Recognition

Supplementary Material

10. Additional Experiments

Analysis of Latent Space Distribution Samples We further analyze the latent space distribution samples of the proposed robust AVSR model and achieve following two conclusions: (1) In Figure 7, we observe that MDA-KD effectively avoids dropout-induced bias and make sure the model to employ a collaborative decision strategy, even with video frames missing input. (2) In Figure 8, we demonstrate that the model decision-making pattern indeed dynamically switches to an audio-dominant one by activating the MS-Adapter when facing complete video missing input.

Analysis of Zero-shot Noise Robustness We further evaluate the system performance with zero-shot noise. Specifically, we simulate the noise speech with unseen Babble noise from NOISEX [56] and the near-field audio captured by a head-worn microphone at 0 dB, -2.5 dB, and -5.0 dB SNR levels. In Table 3, we reuse the symbols from Table 3. The results demonstrate that the proposed modality-unbiased model, AV6, outperforms both the modality-biased model AV1 and the unimodal model A0 in both Near Field and Far Field settings with in-set noise. More importantly, we highlight the advantage of zero-shot noise robustness of the proposed method across all SNR levels, aligning with the target of AVSR as a robust system for real-world applications.

Analysis of Computational Consumption We analyze the computational efficiency in FLOPS with audio-only input to demonstrate the effectiveness of reducing computation by activating the MS-Adapter and interrupting the data flow in the video branch. Upon activating the MS-Adapter, data solely flows through the audio branch, requiring only 3.89 GFLOPS with 94.21 M parameters for computation. This contrasts favorably with conventional methods that necessitate padding video tensor inputs, consuming 12.64 GFLOPS with totaling 144.78 M parameters.

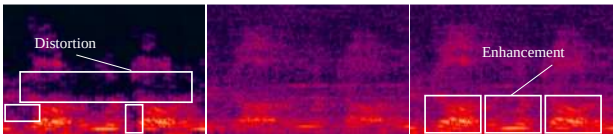
Experiment Details on Different Test Dropout Methods In Figure 9, we provide more comprehensive experimental results and present performance degradation curves across all three test suites (Segment Dropout, Utterance Dropout, and Interval Dropout) to facilitate further research.

11. Distinctions between MBVD and MVD

There are three key distinctions between the Modality Bias Venn Diagram (MBVD) and the Modality Venn Diagram (MVD) [28]. Firstly, MBVD focuses on the latent space to describe the decision pattern of a multimodal model, while MVD space is essentially another form of the original feature space. Secondly, for the generation order, MBVD maps from the original feature space $\mathcal{X}$ to the decisive feature space $\mathcal{Z}$, while MVD follows the opposite direction. Lastly, MBVD is employed to describe modality bias in decision-making processes, whereas MVD is utilized for knowledge distillation.

Models	Near Field	Far Field	Zero-shot Babble Noise		
			0dB	-2.5dB	-5dB
A0	18.10	25.13	33.52	62.17	75.76
AV1	17.71	23.26	29.40	51.63	63.80
AV6	**16.86**	**21.11**	**26.67**	**44.97**	**55.65**

Table 5. CER comparison of zero-shot noise roubustness.



CPU-GSS Enhanced Original GPU-GSS Enhanced

Figure 6. Spectral analysis of GSS-enhanced signals

12. Limitations of the work

Modality dropout presents two facets. On one hand, it could address the out-of-distribution (OOD) issue by missing modalities. On the other hand, if applied on supplementary modalities, it can induce dropout-induced modality bias in modality-biased systems. For our further exploration, we realize the manifestation of these characteristics is related to input quality. In this work, we focus on real-world TV room scenarios with relatively low-resolution video and noisy speech. Under such conditions, dropout-induced modality bias is observed prominently. While for high-quality datasets, such as LRS2 and LRS3, dropout serves more as a form of data augmentation, and the dropout-induced modality bias are mitigated by high-quality input. Nevertheless, in all conditions, the proposed MDA-KD and MS-Adapter consistently lead to relative improvements to original dropout method.

13. Implement Details

Data Processing Details We apply conventional signal processing algorithms, such as Weighted Prediction Error (WPE) [57] and Guided Source Separation (GSS) [58], to

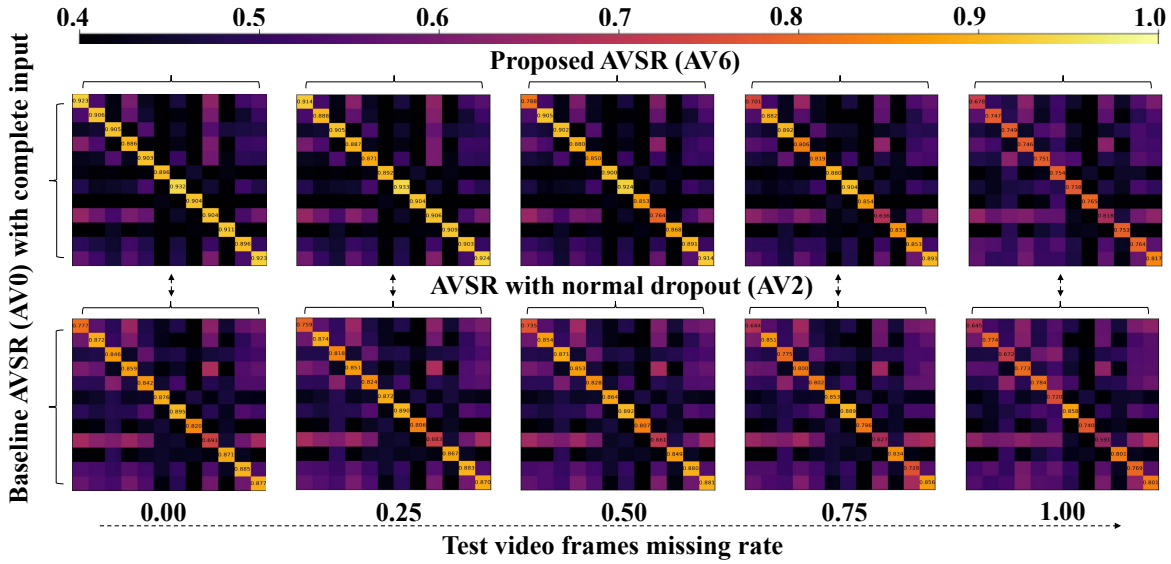


Figure 7. We investigate the decision discrepancies between the proposed robust AVSR (AV6) and the AVSR trained using the normal dropout technique (AV2) across different test video frames missing rates. Similar to Figure 3, we quantify the divergence by calculating the cosine distance similarity of latent decision distribution samples from both models with missing video frames input and those of AV0 with complete data input. The latter samples represent an ideal collaborative decision strategy. Each diagonal element in the cosine distance-based similarity matrix represents the similarity between intermediate representations with the same sample index but may have different missing rates. As a result, two prominent phenomena emerge. (1) In vertical comparison between AV6 and AV2, the sample similarities of AV6 consistently surpass those of AV2 along the diagonal line, indicating a closer approximation to the ideal collaborative decision distribution in latent space. These results suggest that MDA-KD enables AV6 to adopt a decision strategy similar to AV0, whether facing complete input or missing video frames, effectively utilizing content information and modality general information audio modality. (2) In horizontal comparison, in the first row, the diagonal elements in each subplot consistently darken as the missing rate increases, and the last subplot darkens sharply with the shift of decisive bias on audio modality upon activating the MS-Adapter. This trend is less pronounced in the second row, as AV2 exhibits an excessive modality bias on audio modality, deviating from the collaborative decision strategy.

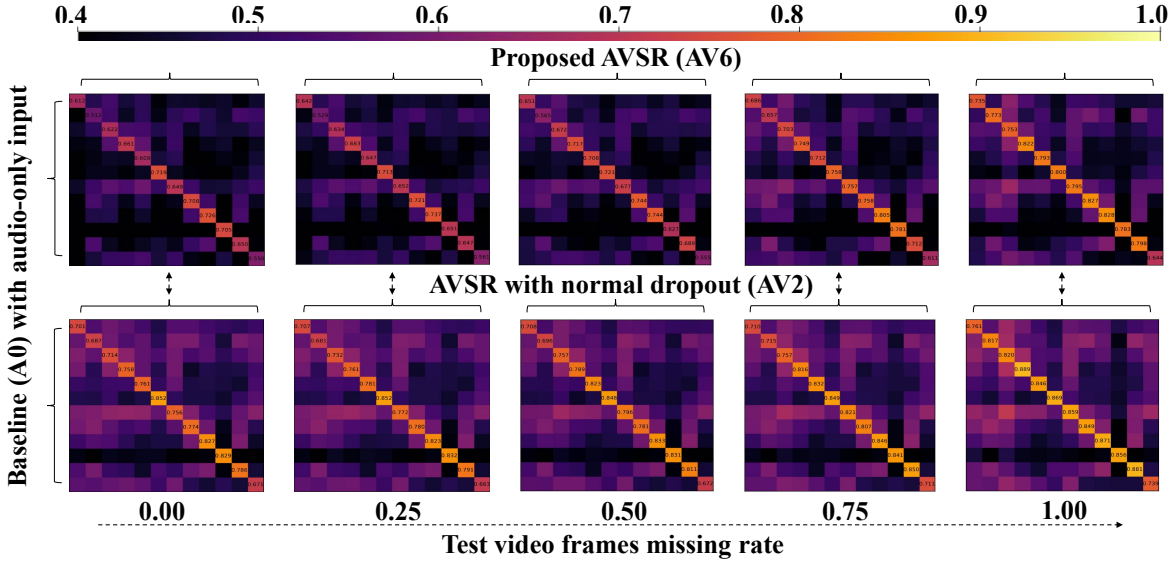


Figure 8. We compare the decision discrepancies between AV6 and AV2 with A0, revealing two distinct phenomena. (1) In the first row, the diagonal line of the last subplot sharply brightens compared to the former four subplots, indicating the effectiveness of the MS-Adapter in dynamically switching the decisive pattern towards the audio-dominant one. (2) In comparison to the first row, the diagonal line of the second row remains consistently bright across various missing video frame rate inputs. This further confirms that AV2 is a modality-biased model that consistently relies on the audio modality.

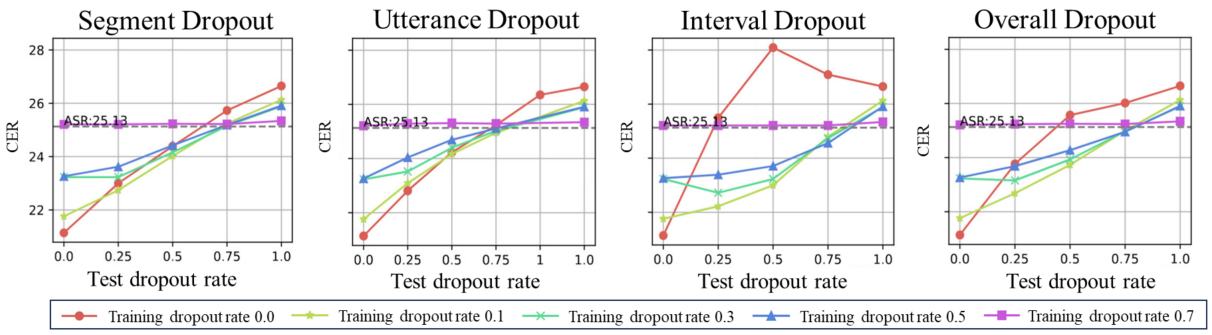


Figure 9. Performance degradation curves of AVSR systems with different training dropout rate test in different test dropout methods.

multichannel far/middle-field audio for dereverberation and source signal separation in both the training and test sets. Specifically, we utilize a GPU-accelerated version of GSS [59]. As shown in Figure 6, it effectively enhances the spectral speech components for the target speaker while minimizing speech distortion compared to the CPU version [58]. Then We apply a short fourier transform and mel filter to obtain 80-dimensional Fbank frames in the frequency domain, with a 0.25s window length and a 0.01s frame shift, using a 16k sample rate. For video, following [3], we acquire grayscale lip ROI with 88×88 pixels before inputting it into the network. In ASR training, all enhanced far/near/middle-field audio is used, employing various data augmentation techniques, such as adding noise, Room Impulse Response (RIR) convolution, speed perturbation, and concatenating nearby segments to create a 10-fold training set. The technique of concatenating nearby segments effectively generates a longer segment, providing additional content information. This technique can used in both training and decoding phrases. For VSR, we pre-train the visual frontend on far/middle-filed video following [18] by correlating lip shapes with syllabic HMM states (3168 Senone units). In AVSR training, the audio and visual branches are initialized with pre-trained ASR and VSR representations. We create an 8-fold training set, incorporating two effective data augmentation techniques: (1) matching synchronous audio and video segments recorded in different fields and (2) concatenating nearby segments in both video and audio.

Training Implementation Details All conformers in our network use the same set of hyperparameters ($n_{\text{head}} = 8$, $d_{\text{model}} = 512$, $d_{\text{ffn}} = 2048$, $CNN_{kernel} = 5$). The decoder consists of six transformer blocks ($n_{head} = 8$, $d_{\text{model}} = 512$, $d_{\text{ffn}} = 2048$). For unimodal model training, we strictly adhere to [18]. In robustness training for this work, all models are optimized using Adam with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and a learning rate of 0.0012. For MDA-KD implementation in Section 7, we utilize the intermediate representation samples from the output of ResNet-18 and the first layer of Conformer in the video branch in practice. For further exploration, we successfully validate that the output of the multimodal encoder exhibits similar effectiveness in achieve both missing robustness and accuracy with complete input. The learning rate undergoes a linear warm-up during the first 3000 steps and subsequently decreases proportionally to the inverse square root of the step number. We train for 12 epochs with a training batch size of 128, utilizing 4 NVIDIA Tesla A100 48GB GPUs. For MS-Adapter adaptation, we train 5 epochs with a batch size of 144 and a learning rate of 0.0002. During decoding, the beam size is set to 10 in beam search. Additionally, a 6-layer transformer-based language model trained on the transcription of the training set is employed in decoding, with a weight of 0.2, although it brings negligible performance improvement.

Dropout Setting Details Segment Dropout, Utterance Dropout, and Interval Dropout are employed to simulate missing video modality in different scenarios. Segment dropout occurs when contiguous segments of video frames are dropped, which often occurs when the lips are covered or when the person is in a side-face pose. Utterance Dropout refers to dropping the entire video, which represents situations where the camera is turned off. Interval Dropout means dropping ($dropout\ rate < 0.5$) or preserving ($dropout\ rate > 0.5$) video frames at a fixed interval, indicating missing due to network latency or hardware computation bottleneck. Unlike previous work [19], we have simplified the test suites by removing frame-level random dropout to ensure experimental reproducibility. Furthermore, the starting position for segment dropout is randomly determined. Considering our study on modality bias and robustness, the focus lies more on the dropout rate than the dropout method.

14. More Discussions on Related Works

Missing Modality in Multimodal Learning The missing modalities problem is common in multimodal applications, whether in the training or testing stage, and has attracted a lot of research interest. For modality-balanced models like Multimodal Emotion Recognition (MER) and multimodal sensor fusion in autonomous driving, the mainstream approach is to learn joint multimodal representations to capture intra- or inter-modal features cross modalities [53, 54]. For modality-biased models, data augmentation methods such as modality dropout effectively address out-of-distribution issues [19–21]. In cases of severe modality absence, generative models [51, 52] and meta-learning based methods [60] are used to directly predict the missing modalities based on available modalities or a few-shot paired samples. For AVSR, we prioritize efficiency and opt for dropout due to its plug-and-play nature and lightweight implementation.

Modality Bias in Multimodal Learning The modality bias is observed in many multimodal applications, since there is a direct correlation between a specific modality and the target task, leading to one modality dominating the decision-making process [61]. In the VQA, several de-bias methods have been proposed. New datasets following the answer distribution balancing rule have been constructed to address the language prior problem [62]. Guo et al. [63] develop plug-and-play loss function methods that can adaptively learn the feature space for each label. Gat et al. [61] have proposed a method based on the log-Sobolev inequality. Although many studies have been conducted on removing bias, there is a lack of conception or mathematical models to describe model bias and limited research on the impact of bias on the modality missing problem.

Dropout-Induced Modality Bias on Mulitmodal Tasks For AVSR, this excessive modality bias towards audio is a double-edged sword, as it brings robustness to missing video data while degrading the performance of a multimodal model on complete multimodal data. It causes the model to tend towards trivial solutions and ignore optimal ones. As a result, the model neglects visual cues, making it sensitive to perturbations in speech. This contradicts the intention of AVSR as a multimodal robust speech recognition application in noisy environments.

For other multimodal applications, Hazarika et al. investigate the robustness of Multimodal Sentiment Analysis (MSA), which is a multimodal classifier with text, visual, and audio as input [22]. By applying dropout on the training text, the robustness against missing text can be achieved without compromising the original performance. These findings seem to be inconsistent with the degradation results observed in AVSR. While the truth is that the common MSA system exhibits a severe modality bias dominated by text, and it is sensitive to perturbations in text but robust to other modalities. Applying dropout on text helps to mitigate over-reliance and encourages the model to leverage supplementary information across modalities. A similar phenomenon has been observed in AVSR when applying dropout on the audio modality [20, 21]. Interestingly, in our research on video robustness in AVSR, video is a supplementary modality within the system rather than the dominant one. As a result, we emphasize that it is important to first determine whether the system has a dominant or supplementary modality when studying the robustness of a specific modality within a multimodal bias system.