

Not Only Vision: Evolve Visual Speech Recognition via Peripheral Information

Zhaoxin Yuan^{1,2}, Shuang Yang^{1,2}, Shiguang Shan^{1,2} and Xilin Chen^{1,2}

¹ State Key Laboratory of AI Safety, Institute of Computing Technology,
Chinese Academy of Sciences, Beijing 100190, China

² University of Chinese Academy of Sciences, Beijing 100049, China

yuanzhaoxin20@mails.ucas.ac.cn, {shuang.yang, sgshan, xlchen}@ict.ac.cn

Abstract

Is visual information alone sufficient for visual speech recognition (VSR) in challenging real-world scenarios? Humans do not rely solely on visual information for lip-reading but also incorporate additional cues, such as speech-related context and prior knowledge about the task. However, existing methods have largely overlooked such external information in automatic VSR systems. To systematically explore the role of such information for VSR, we introduce the concept of Peripheral Information. We categorize it into three types based on the relevance to the spoken content: (1) Contextual Guidance (e.g., topic or description of speech), (2) Task Expertise (e.g., human prior experience in lip-reading), and (3) Linguistic Perturbation (irrelevant signals processed alongside meaningful information). Considering the disparity that peripheral information provides additional clues with varying significance while visual input serves as the most direct source for VSR, we propose a framework that introduces a hierarchical processing strategy to handle different modalities. With visual-specific adaptation and a dynamic routing mechanism for multi-modal information, our approach reduces the impact of modality conflicts effectively and enables selective utilization of peripheral information with varying relevance. Leveraging readily available peripheral information, our model achieves a WER of 22.03% on LRS3. Further experiments on AVSpeech demonstrate its generalization in real-world scenarios.

1. Introduction

Visual Speech Recognition (VSR) has gained wide attention due to its important applications, including assisting speech interpretation in noisy environments[42], and offering a communication method for patients with speech disabilities [28]. However, it remains challenging in real-world scenarios. Issues such as low image resolution, poor lighting conditions, and dynamic head poses [12] introduce substantial noise and ambiguity into the visual signal, which

collectively complicate accurate interpretation. Variations in speaking styles and individual differences in lip movements further exacerbate the difficulty of the task.

To address these challenges, studies have made significant efforts on optimizing the processing and learning of visual information. Some approaches attempt to expand the visual input region from lip to face to capture richer visual cues [1, 19]. More recent works leverage audio-visual data and self-supervised learning method to extract more discriminative visual features [14, 15, 39, 49, 52]. These efforts have led to more robust and general representations of visual signals, resulting in substantial gains in VSR. With these advances, VSR systems have become more capable than ever. However, in real-world scenarios, visual information is often compromised due to occlusions, low-resolution inputs, or the visual ambiguity of homophonous phonemes, which makes recognition challenging. A natural question arises: could there be other complementary information that further enhance VSR, or, *is optimizing visual information processing the only path toward robust VSR?*

The human lip-reading process offers a potential alternative perspective. When humans perform lip-reading, they do not rely solely on visual signals but instead subconsciously integrate information beyond visual as assistance. For example, background information about the speech, such as the setting or topic of discussion and some knowledge of the speaker, or familiarity with common linguistic patterns, constrains the space of possible interpretations. Furthermore, humans leverage their accumulated experience with lip-reading itself, recognizing that certain lip movements can correspond to multiple words with distinct meanings. These human abilities are well acknowledged yet remains unexplored in automatic VSR systems.

In this work, we present the first attempt to integrate information beyond visual dynamics into the automatic VSR process. While prior studies in Automatic Speech Recognition (ASR) have successfully leveraged additional context to enhance recognition accuracy [4, 6, 17, 18, 20, 29, 36, 41], the role of such information in VSR has been largely

overlooked. This may be due to the greater complexity of learning from visual modality, making integration less straightforward. Moreover, conventional definitions of context primarily focus on information directly related to the spoken content, such as preceding words or linguistic dependencies. While such information is undoubtedly helpful, it can be difficult to obtain in certain scenarios, limiting the applicability of context-aware models. Additionally, this restricted interpretation of context hinders further exploration of external cues that could aid recognition.

We expand the concept of context to peripheral information, with the goal of capturing a broader range of external cues that remained unexplored. Besides conventional content-based dependencies, peripheral information also encompasses task-relevant knowledge and even factors that may seem unrelated. We categorize peripheral information into three levels based on its relevance to the spoken content: (1) Contextual Guidance, including background like the speech topic or speaker information; (2) Task Expertise, referring to human cognitive abilities, such as recognizing common word pairings and grammatical patterns to resolve ambiguities; and (3) Linguistic Perturbation, covering irrelevant elements like noisy or unrelated text that may disrupt recognition. Integrating peripheral information into VSR is challenging due to its disparity with visual input, as well as its varying relevance to target content. To address this, we propose a framework that hierarchically processes the multi-level inputs and dynamically routes multimodal information, enabling effective use of peripheral information in recognition.

The contributions of this work can be summarized as:

- We present the first attempt to extend VSR beyond visual signals. Through a systematic investigation, we demonstrate that different types of peripheral information can positively contribute to recognition, emulating human-like lip-reading capabilities.
- We propose a new framework that integrates visual information with peripheral information to jointly infer the spoken content. By a hierarchical processing approach for multimodal inputs, it enables the effective utilization of information at different levels.
- We evaluate our approach on the widely used LRS3 dataset, achieving a word error rate (WER) of 22.03%, outperforming previous methods trained on similar amounts of lip-reading data. Furthermore, we demonstrate its generalization to the more complex AVSpeech dataset, highlighting the robustness of our approach in real-world scenarios.

2. Related Work

2.1. Visual Speech Recognition

Visual Speech Recognition (VSR) has evolved significantly driven by advancements in computer vision and machine

learning techniques. Early VSR approaches relied heavily on hand-crafted visual features [13, 23, 32]. These methods extracted features such as lip contours or motion trajectories using techniques like Scale Invariant Feature Transform (SIFT) [30] and Active Appearance Models (AAM) [21]. However, these hand-crafted features were often limited in their ability to capture complex spatiotemporal dynamics of lip movements. The advent of deep learning marked a paradigm shift in VSR. Starting with CNN-based architectures for spatial pattern extraction and RNN variants for temporal modeling [5, 9], the field has progressed to Transformer-based models that capture long-range spatiotemporal dependencies in lip movements [1, 25, 39]. These innovations have led to substantial performance gains. For instance, the word recognition accuracy on the LRW dataset [9] has improved from 61.1% [9] (2017) to 95.0% [3] (2024).

Despite recent advancements, VSR still remains a challenging task. The small size of lip regions makes them highly susceptible to practical issues such as occlusion, low resolution, extreme pose variations, and poor lighting, which degrade the quality of visual features. Additionally, visemes (visually similar pronunciations) are difficult to distinguish solely from lip movements, particularly for homophones. Human lip-readers naturally try to overcome these challenges by integrating broader information beyond lip movements (e.g. priors about speaking scenario and topics, and accumulated experiences of VSR task). In contrast, current machine lip-reading systems remain constrained to visual-only inputs. This work aims to explore the potential of the peripheral information beyond visual dynamics for VSR, mirroring human lip-reading strategies.

2.2. Context-Aware Speech Recognition

Integrating contextual information into automatic speech recognition (ASR) has been explored to enhance the recognition of rare or out-of-vocabulary words, such as named entities, technical terms, and numerical expressions. Existing approaches can be broadly categorized into explicit contextual biasing and broader context utilization. Several studies [4, 17, 18, 36] have introduced explicit biasing modules in ASR models toward recognizing infrequent or unseen words. They incorporate contextual information such as domain-specific keywords or personalized word lists to adapt to user-specific vocabulary and improve accuracy without modifying the underlying model architecture.

Recent works [6, 20, 29] have explored leveraging large language models (LLMs) to enhance ASR robustness by incorporating contextual information, such as previous utterances and talk descriptions. By utilizing the advanced capabilities of LLMs, these methods enable context-aware ASR, significantly refining accuracy.

While prior ASR studies primarily focus on leveraging

content-related contextual information, our work extends this notion by introducing the concept of peripheral information, which encompasses a more diverse set of elements. This includes not only conventional context cues related to spoken content but also experiential knowledge from human lip-reading processes and potential disturbance factors. In VSR, visual representations are complex due to factors such as occlusions, lighting variations, and the ambiguous nature of lip movements, making the incorporation of additional signals less straightforward. Careful design is required to effectively integrate peripheral information. By systematically exploring the contributions of these multi-level information sources, we aim to provide a broader and more comprehensive perspective to improve VSR.

3. Method

In this section, we present a more detailed explanation of peripheral information as well as our approach to integrating it into Visual Speech Recognition.

3.1. Peripheral Information

As introduced in the previous sections, we categorize peripheral information into three types based on their relevance to the specific speech content.

Contextual Guidance refers to background information that is closely related to the speech content, helping to refine recognition by narrowing the set of possible hypotheses. For instance, a TED talk titled “Niels Diffrient: Rethinking the Way We Sit Down” suggests a focus on design and human body measurements, shaping expectations for relevant vocabulary. Such guidance includes the title of the speech, a brief content overview, the speaker’s name, and a short description of their background. It is not limited to these mentioned elements and may also include other forms of contextual guidance depending on circumstances.

Task Expertise refers to the cognitive strategies humans employ in VSR, leveraging both phonetic knowledge and linguistic priors. Familiarity with the correspondence between lip movements and phonemes helps constrain possible interpretations, especially when similar lip patterns correspond to multiple sounds. Additionally, prior exposure to frequent word sequences and syntactic structures aids in resolving ambiguities.

Linguistic Perturbation denotes the presence of disturbance or misleading information in real-world scenarios. Correctly identifying and utilizing relevant information while filtering out perturbative or irrelevant signals is crucial for a robust VSR system. Similar to noise injection for image or audio in training process for CV or acoustic tasks, linguistic perturbation adds controlled randomness at semantic level, forcing the model to distinguish useful and misleading signals for VSR.

3.2. Model Architecture

Integrating peripheral information into VSR poses two key challenges: First, there exists an inherent disparity between visual and textual modalities, stemming from their distinct representational characteristics. Visual features encode fine-grained spatiotemporal patterns of lip movements, preserving continuous motion trajectories and subtle articulatory dynamics. In contrast, textual information is discrete and symbolic, capturing high-level semantic meaning. The fundamental difference makes the alignment and fusion less straightforward. Second, peripheral information encompasses cues with varying degrees of relevance to the spoken content, some provide reinforce recognition, while others introduce perturbations.

To address the challenges, we propose a novel framework for robustly integrating peripheral information for VSR. It is illustrated in Figure 1 and consists of two main stages: multimodal encoding and decoding. In the first stage, a visual encoder extracts fine-grained spatiotemporal features from lip movements, while peripheral information is structured and integrated with visual features to form a multimodal input. In the second stage, a large language model enhanced with a novel Synergy LoRA decodes the multimodal input and generates the transcripts.

3.2.1. Stage 1: Multimodal Encoding

Visual Encoder. To extract speech-related representations from raw video inputs, we utilize a visual encoder that processes sequences of facial movements centered on the lips. The visual encoder captures semantic and phonetic attributes of speech by modeling the spatiotemporal dynamics of lip motion. The resulting features are then transformed into the LLM’s representation space, enabling processing and understanding by it.

We use a pretrained visual encoder [39] that follows a commonly used Transformer[45]-based architecture in VSR. Let it be denoted as \mathcal{E}_v . Given an input raw video sequence $I_v = \{i_1, i_2, \dots, i_T\}$, where $i_t \in \mathbb{R}^{d_x}$ represents the video frame at time step t , the visual encoder produces a sequence of visual features:

$$F_v = \mathcal{E}_v(I_v), \quad (1)$$

where $F_v = \{f_1, f_2, \dots, f_T\}$, and $f_t \in \mathbb{R}^{d_f}$ represents the extracted feature vector.

Modality Adapter. The modality adapter transforms visual features extracted by the encoder into the LLM’s embedding space. However, visual features typically have high temporal resolution (e.g., 25fps), while speech generally flows only 2–3 words per second. To retain as much temporal information as possible while align visual features with textual information temporally, we apply a $2\times$ average downsampling, followed by a projection layer that maps the features into the LLM’s representation space. The visual

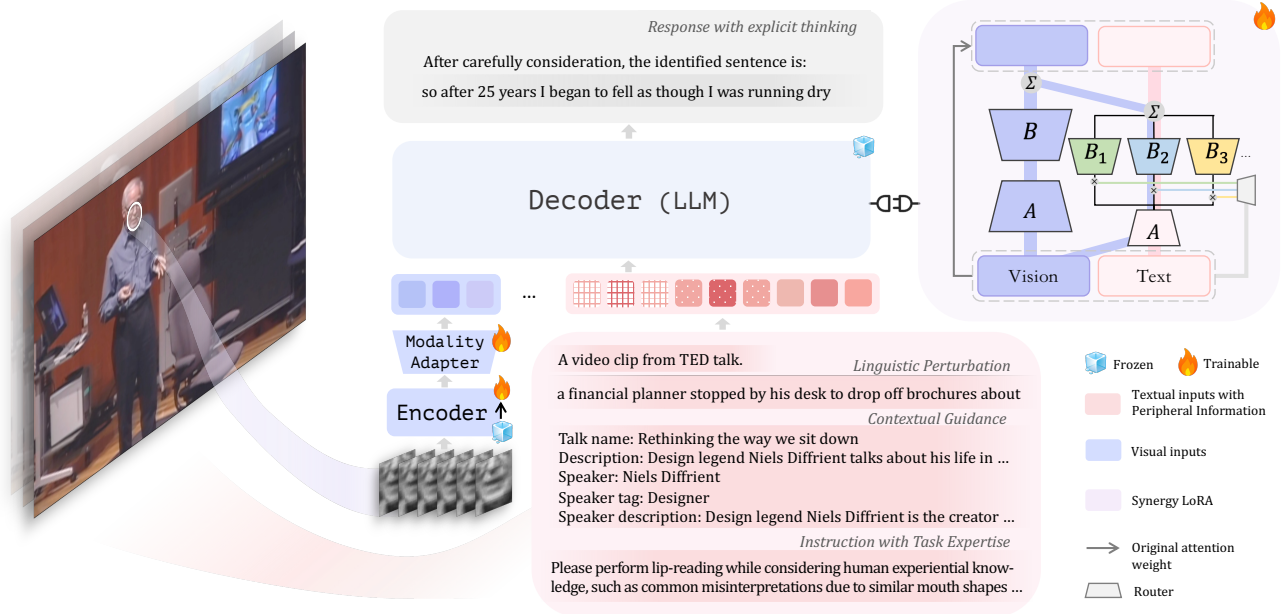


Figure 1. **Overview of our proposed framework.** Visual features are extracted by a pretrained encoder and then mapped to the LLM embedding space through a learnable projector. A structured input and the output constraint mechanism enable effective utilization of peripheral information at different levels. Synergy LoRA is introduced within the attention mechanism of the LLM for adaptation.

features are downsampled and projected into the LLM’s embedding space as

$$E_v = W_p \cdot \text{downsample}(F_v), \quad (2)$$

where $E_v \in \mathbb{R}^{\lfloor T/2 \rfloor \times d_e}$ represents the final visual embeddings, $W_p \in \mathbb{R}^{d_e \times d_f}$ is a learnable projection matrix, and $\text{downsample}(F_v) \in \mathbb{R}^{\lfloor T/2 \rfloor \times d_f}$ denotes the downsampled visual features.

Peripheral Information Integration. We integrate the three distinct types of peripheral information into VSR in a textual format. Additionally, we introduce the directive “transcribe the video to text” to explicitly instruct the model to perform video transcription as part of the VSR task. Figure 1 illustrates how these components are implemented in practice. In the following section, we detail the specific implementation of peripheral information.

Contextual Guidance provides background knowledge related to the speech content, such as the title, speaker biography, and topic descriptions. These elements are consolidated into a structured textual prompt like “Speech title: X Description: Y Speaker: Z.”

Task Expertise pertains to the cognitive mechanisms humans employ in VSR, utilizing both phonological awareness and linguistic priors. To introduce this cognitive capability into our framework, we propose a two-step integration mechanism: (1) *Task Awareness*: We refine the instruction prompt from generic directives (e.g., “Transcribe the speech to text.”) to task-specific guidance that explicitly highlights potential challenges, such as “Please perform lip-reading while considering human experiential knowledge,

such as common misinterpretations ...”. Such explicit instruction guides the model to capture the difficulties of the task, allowing the model to establish an informed connection between visual input and linguistic content. (2) *Output Constraint*: Before generating the final transcript, the model is required to output an explicit reasoning statement, simulating the human cognitive process of self-monitoring. For example, instead of directly producing the transcription, the model first generates a phrase “After carefully consideration, the identified sentence is”, thereby making the reasoning process explicit. Although we provide specific formats for prompts and constraints, our method does not depend on fixed wording. We demonstrate that this structured two-step mechanism enhances robustness.

Real-world scenarios often contain extraneous or misleading information, which can hinder reliable model performance. We introduce controlled *linguistic perturbation* designed to help the model distinguish relevant content from irrelevant noise. Linguistic perturbation sources from both randomly generated tokens and extracted fragments from unrelated natural text. They are inserted either between sentences in contextual guidance, when available, or immediately before the instruction such as “Transcribe the speech” or its modified version incorporating task-specific expertise. This approach encourages the model to focus on meaningful information while effectively filtering out irrelevant content.

Together, these three types of information constitute the concept of peripheral information proposed in this work. They span a spectrum from directly related to entirely unrelated to the spoken content. The text containing peripheral

information is processed by the tokenizer and embedding layer of the LLM to obtain the text embedding E_t of length L . It is subsequently combined with the visual embedding E_v along the temporal axis to construct the multimodal representation:

$$S = [E_v; E_t], \quad (3)$$

where $S \in \mathbb{R}^{T' \times d_e}$, and $T' = \lfloor T/2 \rfloor + L$ denotes the combined length of the visual and text embeddings along the temporal dimension.

3.2.2. Stage 2: Decoding

The decoding stage leverages a LLM to process multimodal inputs and infer the spoken content. We introduce a novel adaptation module, Synergy LoRA, to enhance the model’s ability to accommodate diverse inputs.

LLM. The LLM is responsible for decoding from the multimodal input sequence S to the target output Y . This generation process for Y can be expressed as:

$$P(Y|S) = \prod_{m=1}^M P(y_m|S, y_{<m}), \quad (4)$$

where $y_{<m} = \{y_1, y_2, \dots, y_{m-1}\}$ represents the previously generated tokens. The LLM models this conditional distribution by processing the multimodal input sequence, leveraging its reasoning capabilities to predict each word y_m in an autoregressive manner.

During training, it learns to generate the correct transcript Y given the input S . The optimization objective is to maximize the likelihood of the target output with ground-truth transcript, formulated as:

$$\mathcal{L} = - \sum_{m=1}^M \log P(y_m|S, y_{<m}). \quad (5)$$

By minimizing this negative log-likelihood loss, the model improves its ability to generate accurate transcriptions in an autoregressive manner.

During inference, the model generates the output iteratively. Given an initial multimodal input S , it starts by predicting y_1 and subsequently conditions each next token y_m on the previously generated tokens $y_{<m}$. The process continues until a stop token defined in the LLM is met.

Synergy LoRA. To adapt the LLM to VSR task, we fine-tune it using Low-Rank Adaptation (LoRA) [16]. However, it poses a notable challenge in bridging the gap between two input modalities. Visual features differ significantly from the high-level semantic representations that LLMs are designed to process. When adapting the LLM with LoRA, a larger rank is required. Furthermore, as shown in Figure 2, including peripheral information can reduce the performance of the best configuration that does not use it. This suggests that the model encounters difficulties or conflicts

when adapting to two modalities with distinct characteristics. To address this, we introduce Synergy LoRA, a coordinated adaptation mechanism that prioritizes the processing of visual features while dynamically incorporating peripheral information.

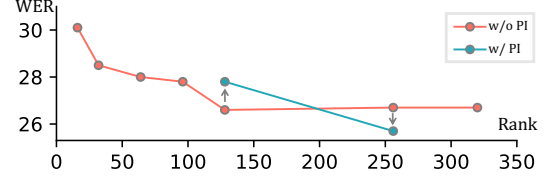


Figure 2. **WER under different LoRA rank.** w/ PI indicates incorporating peripheral information and w/o indicates without it.

Specifically, we introduce a specialized module to optimize visual information considering its critical role in VSR, minimizing the impact of modality disparities. Beyond this, we design a multimodal Mixture-of-Experts[38] (MoE) module for adaptation across the entire multimodal input sequence. It enables a *synergistic* adaptation to the distinct characteristics of each information level, from directly relevant visual information to auxiliary context cues and irrelevant perturbations. Inspired by recent findings that LoRA adaptation matrices (A matrices) exhibit similarities across tasks [43, 51], we design Synergy LoRA such that all experts share the same A matrix, while the B matrices are dynamically routed to process multimodal inputs. Notably, Synergy LoRA is applied to the query and value of attention weights for all layers in the LLM.

Let the attention weights for visual inputs be denoted as W_v . After adaptation, the updated weights are given by:

$$W'_v = W_v + \Delta W_v, \quad (6)$$

where ΔW_v represents the update introduced by the visual-specific module. Denote attention weights for the complete multimodal inputs as W_c , the MoE-based module updates them as:

$$W'_c = W_c + \sum_{i=1}^K g_i \cdot \Delta W_i, \quad (7)$$

where K is the number of experts, g_i is the gating score for expert i , and ΔW_i represents the update introduced by expert i . The gating scores g_i are computed using a linear layer followed by a softmax function:

$$g_i = \text{softmax}(W_g h)_i, \quad (8)$$

where h is the hidden representation of the input sequence, and W_g is learnable parameters of the router.

In general, the attention weights for visual inputs are derived from both the visual-specific module and the MoE-based module. After adaptation, the attention weights for the visual inputs are $W'_v + \sum_{i=1}^K g_i \cdot \Delta W_{v,i}$, where $\Delta W_{v,i}$ represents the additional adjustment introduced by expert

i. For textual inputs, the attention weights are adjusted solely by the MoE-based module. Let the original attention weights for textual inputs be W_t , it is then updated as $W_t + \sum_{i=1}^K g_i \cdot \Delta W_{t,i}$. This design ensures that visual inputs are first processed independently to capture fine-grained details, and then further refined with high-level semantic cues from textual peripheral information.

4. Experiment

4.1. Datasets

LRS3-TED[2]. LRS3-TED is a popular lipreading dataset, comprised of speech samples from TED and TEDx videos, which covers a wide range of challenging conditions, including thousands of speakers, large pose variations, diverse lighting conditions, different resolution, accent and so on. It covers diverse topics and is widely adopted for evaluating visual speech recognition methods for being the largest publicly available one of its kind.

AVSpeech[11]. AVSpeech is another challenging large-scale audio-visual speech dataset with thousands hours of samples extracted from YouTube, representing a broader range of real-world speaking scenarios. The data spans a wide variety of scenarios, speakers, face poses, accents and so on. Its diverse and complex characteristics enable the evaluation of methods in handling real-world, challenging scenarios. As the original data lacks transcriptions, we employ Whisper[35] for automatic annotation and take the English portion of it for evaluation.

Peripheral Information. Contextual Guidance in peripheral information are collected from readily available sources in our experiments. For LRS3, we used a pre-collected dataset from Kaggle¹ as well as additional data collected from YouTube links associated with the videos. For AVSpeech, video titles and descriptions were obtained directly from their corresponding YouTube links. Linguistic perturbation, includes two types of token-level perturbations: completely random tokens and randomly sampled fragments from natural text. The natural text was sourced from news corpus². More details are in Appendix 6.

4.2. Evaluation and Implementation Details

Word error rate (WER) is adopted as the evaluation metric for VSR, which is defined as $WER = (S + D + I)/N$, where S, D, I, N represent the number of words substituted, deleted, inserted, and referenced. We adopt the pre-trained AV-HuBERT[39] as the visual encoder and Llama3.1-8B-Instruct[10] as the LLM by default, unless otherwise specified. Due to page limitations, more implementation details are provided in Appendix 7, along with a further discussion of the visual encoder in Appendix 8.2.

¹<https://www.kaggle.com/datasets/thegupta/ted-talk/data>

²https://huggingface.co/datasets/fancyzhx/ag_news

Table 1. **Results on LRS3.** Comparing with prior works, ours with peripheral information outperforms when utilizing comparable amounts of unlabelled (Unlab.) and labelled (Lab.) video data. * denotes self-training on the unlabelled data. † denotes inclusion of extra audio or language model. For consistency, results cited for comparison are rounded to one decimal place, as reported in most previous studies.

Method	Unlab. hours	Lab. hours	WER (%) ↓
<i>Fully Supervised Models</i>			
Zhang et al. [48]	-	863	60.1
Ma et al. [25]	-	595	43.3
Prajwal et al. [33]	-	698	40.6
Ma et al. [26]	-	1,459	31.5
Ma et al. [27]	-	3,448	19.1
<i>Self-supervised Pre-training & Supervised Fine-tuning</i>			
Ma et al. [24]	1,759	433	38.8
Shi et al. [39]	1,759	433	28.6/26.9*
Zhu et al. [52]	1,759	433	28.4†
Haliassos et al. [14]	1,759	433	27.8/24.4*
Yeo et al. [47]	1,759	433	27.6†
Yeo et al. [46]	1,759	433	25.4†
Cappellazzo et al. [7]	1,759	433	25.3†
Prajwal et al. [34]	1,759	433	24.3†
Haliassos et al. [15]	1,759	433	22.3*
Ours	1,759	433	22.03†
<i>Trained using non-publicly available datasets</i>			
Afouras et al. [1]	-	1,519	58.9
Shillingford et al. [40]	-	3,886	55.1
Serdyuk et al. [37]	-	90,000	17.0
Liu et al. [22]	3,652	3,068	16.9
Chang et al. [8]	-	100,000	12.8

4.3. Main Results

Table 1 presents a comparison of our method with state-of-the-art models on the widely used lip-reading dataset LRS3. Benefiting from peripheral information, our approach achieves a WER of 22.03%, outperforming existing methods trained on similar amounts of lip-reading data. Notably, our model surpasses even those approaches [14, 15, 39] that uses self-training on much more data.

Our model consistently outperforms prior works that utilize LLMs [7, 46] (25.40% and 25.30%) both with (22.03%) and without (24.92%) peripheral information. These results highlight the potential of leveraging peripheral information for VSR and effectiveness of our method.

4.4. Detailed Analysis and Discussion

4.4.1. Synergy LoRA

We compare Synergy LoRA with different designs both with and without peripheral information, demonstrating the soundness of its design as well as its advantage in handling hierarchical information inputs. Results are shown in Table 2 with indices corresponding to Figure 3.

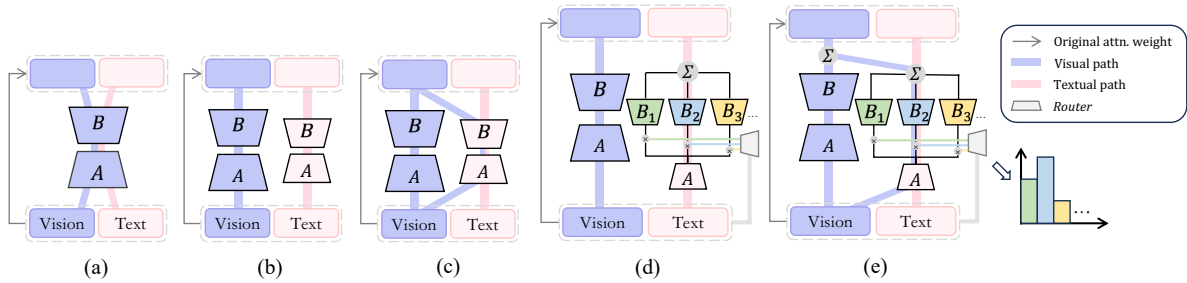


Figure 3. **Different Adaptation Designs.** Shorthand notations are used to describe the relationships between vision and text: & stands for shared parameters, | for separate, and MoE for Mixture of Experts. (a) vanilla LoRA, V&T: same adaptation parameters for both modalities. (b) Distinct Paths, V | T: separate paths for vision and text. (c) Vision-Specific Path, V | V&T: vision has an independent path, while text path shares parameters with vision. (d) MoE for Text, V | MoE(T): separate paths, with text featuring an MoE adaptation. (e) **Synergy LoRA**, V | MoE(V&T): independent adaptation for vision, while text MoE shares parameters with vision.

Table 2. **LoRA configuration and Results.** We show results on LRS3 for various LoRA configurations, with indices corresponding to Figure 3. The WER values are reported for two scenarios: with and without peripheral information. Numbers in parentheses indicate reduction in WER compared to the w/o column.

Design	Rank	WER (%)	
		w/o	w/
(a) V&T	128	25.25	23.92 (−1.33)
(b) V T	96 32	24.92	23.64 (−1.28)
(c) V V&T	96 32	25.04	22.78 (−2.26)
(d) V MoE(T)	96 8 × 4	25.17	22.44 (−2.73)
(e) V MoE(V&T)	96 8 × 4	25.03	22.03 (−3.00)

Without peripheral information. When no peripheral information is provided, separated adaptation (b-e) for visual and textual modalities perform better than using a single LoRA module (a), even though they have identical total rank sizes. This indicates that the separation design for two modalities in Synergy LoRA allows each to specialize in learning intra-modal adaptations, effectively mitigating the challenges posed by modality differences.

With peripheral information. When full peripheral information is introduced, we find that sharing certain parameters between visual and textual adaptations during the adaptation process yields significantly greater relative improvements compared to scenarios without peripheral information. This can be observed by comparing the relative gains between (b) and (c), as well as (d) and (e). Joint optimization enables the dominant visual modality to guide the textual modality, extracting meaningful information to enhance recognition. The MoE-based design of Synergy LoRA demonstrates clear advantages in handling peripheral information, as it can automatically route signals to the most relevant experts, achieving hierarchical utilization of diverse information sources.

Expert visualization. To further understand the MoE-based design in Synergy LoRA, we visualize the weights assigned to different experts for a randomly sampled instance from LRS3 test set shown in Figure 4. More ex-



Figure 4. **Visualization of expert weight.** We show the weight of experts in the last layer of LLM for a random sample from LRS3.

amples are provided in Appendix 11. These visualizations reveal distinct patterns in how the experts specialize in handling specific types of textual information. Experts 4 exhibit specialization with a high weight for structural elements of text (e.g., “and”, “to”, colons, periods). General descriptive phrases and instruction tend to activate Expert 2 more frequently. Detailed and content-specific descriptions are predominantly handled by Experts 1 and 3. This specialization suggests that these experts focus on fine-grained information, enabling the model to capture nuanced details that are essential for accurate recognition.

4.4.2. Peripheral Information

Here, we present the contributions from three different types of peripheral information.

Contextual Guidance. We show the contribution of different types of contextual guidance to recognition, considering both individual and progressively accumulated scenarios. The results are presented in Table 3. Among all the contextual guidance, descriptive information, including description of speech and speaker, typically exhibits a higher degree of correlation with the target content and accounts for the largest performance improvement. The results demonstrate that contextual guidance effectively narrowing down the potential vocabulary of the speech content, thereby reducing uncertainty and errors.

Task Expertise. It originates from human experience. Such information has often been overlooked, yet it represents a crucial capability that humans rely on when perform-

Table 3. **Comparison of Contextual Guidance.** WER results are reported for each type individually and accumulatively, without task expertise and linguistic perturbation.

Contextual Guidance Type	WER (%)	
	Individual	Accumulated
<i>Without</i>	25.03	
Speech Title	24.28	
Speech Description	24.06	23.46
Speaker Name	24.33	23.30
Speaker Tags	24.57	23.19
Speaker Description	24.22	22.99

ing recognition tasks. We present two implementations of task expertise(A and B), each with unique prompt and constraint. A provides a more coarse-grained specification, B offers a finer-grained level of detail. Detailed prompts are provided in Appendix 10.

Table 4 demonstrates the contribution of this experiential information under various settings, along with corresponding ablations. The effectiveness of it remains robust in the absence of contextual guidance (see the third and fifth row). With the contextual guidance of speech title and description, task expertise shows consistent improvement, increasing from 23.46% in Table 3 to 23.18% and 23.09% under Types A and B, respectively. This enhancement stems from both the prompting and the output constraints. When either component is removed, the model’s performance improvement diminishes. The decline is more pronounced when output constraints are removed, indicating that explicit reasoning processes significantly aid the model in recognizing task-specific characteristics in a manner akin to human cognition. Task expertise can effectively enhance the model’s deeper understanding of the task, thereby reducing errors.

Table 4. **Ablation for Task Expertise.** Complete task expertise includes the full two-step mechanism, while “-task-aware prompt” and “-output constraint” indicate the exclusion of each respective component. Contextual guidance* only includes speech title and description in this table.

Configuration	WER (%)	
	Type A	Type B
<i>without peripheral information</i>	25.03	
<i>with contextual guidance*</i>	23.46	
Complete task expertise	24.62	24.44
+ contextual guidance*	23.18	23.09
- task-aware prompt	24.83	24.93
- output constraint	24.91	25.02

Linguistic Perturbation. We compare the effects of linguistic perturbations under different settings, with results presented in Table 5. When incorporating the title and description of the speech as contextual guidance, training with perturbation information enables the model to learn how to filter out disturbances, thereby resulting a lower WER (from 23.46% to 23.12%). In the absence of other peripheral

information, linguistic perturbation still leads to improvements. Furthermore, the results show that both sources of perturbations enhance the model’s robustness: completely random generated tokens and extracted fragments from natural text. Moreover, our Synergy LoRA proves to be more effective at learning to handle noisy contextual information compared with vanilla LoRA, further demonstrating the advantages of this hierarchical design. Further discussion on linguistic perturbation can be found in Appendix 9.2.

Table 5. **Ablation for Linguistic Perturbation.** Contextual guidance* here only includes speech title and description.

Linguistic Perturbation	Contextual Guidance*	LoRA	WER (%)
-	-	<i>Synergy LoRA</i>	25.03
✓ Random	-	Synergy LoRA	24.62
✓ Natural text clip	-	Synergy LoRA	24.65
✓ Random	✓	Synergy LoRA	23.12
✓ Random	✓	vanilla LoRA	24.05

Results on AVSpeech. We further validate our method on the AVSpeech dataset, which contains large-scale real-world data with diverse and challenging scenarios. Results are shown in Table 6. Our method of introducing different peripheral information as well as Synergy LoRA demonstrates consistent improvement on this challenging dataset, reflecting strong generalization capabilities.

Table 6. **Results on AVSpeech.** We compare the WER under different configurations. Desc. is video description. Results in parentheses correspond to the use of vanilla LoRA for comparison with Synergy LoRA.

Guidance		Expertise	Perturbation	WER (%)
Title	Desc.			
-	-	-	-	46.24 (48.83)
✓	-	-	-	42.51
✓	✓	-	-	41.35
-	-	✓	-	44.51
-	-	-	✓	45.62
✓	✓	✓	✓	40.23 (44.58)

5. Conclusion

In this work, we introduced the concept of peripheral information for VSR, extending beyond the conventional reliance on visual cues. To effectively incorporate these auxiliary signals, we proposed a hierarchical processing framework that enables the complementary fusion of multimodal information. Our model demonstrates strong performance on the widely used LRS3 dataset, while additional experiments on AVSpeech further validate its generalization ability in complex scenarios. This study offers a broader perspective on the role of information beyond visual dynamics in lip-reading. For future work, exploring richer sources of peripheral information and seamlessly integrating them into VSR systems would be an intriguing direction. We hope this study inspires further research both within and beyond the domain of visual speech recognition.

Acknowledgements

This work is partially supported by National Natural Science Foundation of China (No. U24A20332, 62461160331, 62276247). The authors would like to thank Yuheng Fan, Zimo Fan, Tianyue Wang, Bingquan Xia and Yuanhang Zhang for their help in the completion of this work.

References

- [1] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. Deep lip reading: a comparison of models and an online application. *arXiv preprint arXiv:1806.06053*, 2018. 1, 2, 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. 6, 1
- [3] Young Jin Ahn, Jungwoo Park, Sangha Park, Jonghyun Choi, and Kee-Eung Kim. Syncvsr: Data-efficient visual speech recognition with end-to-end crossmodal audio token synchronization. *arXiv preprint arXiv:2406.12233*, 2024. 2
- [4] Petar S Aleksic, Mohammadreza Ghodsi, Assaf Hurwitz Michaely, Cyril Allauzen, Keith B Hall, Brian Roark, David Rybach, and Pedro J Moreno. Bringing contextual information to google speech recognition. In *Interspeech*, pages 468–472, 2015. 1, 2
- [5] Yannis M Assael, Brendan Shillingford, Shimon Whiteson, and Nando De Freitas. Lipnet: End-to-end sentence-level lipreading. *arXiv preprint arXiv:1611.01599*, 2016. 2
- [6] Ye Bai, Jingping Chen, Jitong Chen, Wei Chen, Zhuo Chen, Chuang Ding, Linhao Dong, Qianqian Dong, Yujiao Du, Kepan Gao, et al. Seed-asr: Understanding diverse speech and contexts with llm-based speech recognition. *arXiv preprint arXiv:2407.04675*, 2024. 1, 2
- [7] Umberto Cappellazzo, Minsu Kim, Honglie Chen, Pingchuan Ma, Stavros Petridis, Daniele Falavigna, Alessio Brutti, and Maja Pantic. Large language models are strong audio-visual speech recognition learners. *arXiv preprint arXiv:2409.12319*, 2024. 6, 3
- [8] Oscar Chang, Hank Liao, Dmitriy Serdyuk, Ankit Shahy, and Olivier Siohan. Conformer is all you need for visual speech recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2024. 6
- [9] Joon Son Chung and Andrew Zisserman. Lip reading in the wild. In *Asian Conference on Computer Vision*, 2017. 2
- [10] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. 6, 2, 3
- [11] Ariel Ephrat, Inbar Mosseri, Oran Lang, Tali Dekel, Kevin Wilson, Avinatan Hassidim, William T Freeman, and Michael Rubinstein. Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation. *arXiv preprint arXiv:1804.03619*, 2018. 6, 1
- [12] Souheil Fenghour, Daqing Chen, Kun Guo, Bo Li, and Perry Xiao. Deep learning-based automated lip-reading: A survey. *IEEE Access*, 9:121184–121205, 2021. 1
- [13] Alan J Goldschen, Oscar N Garcia, and Eric D Petajan. Continuous automatic speech recognition by lipreading. In *Motion-Based recognition*. Springer, 1997. 2
- [14] 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, 6, 2, 3
- [15] Alexandros Haliassos, Rodrigo Mira, Honglie Chen, Zoe Landgraf, Stavros Petridis, and Maja Pantic. Unified speech recognition: A single model for auditory, visual, and audio-visual inputs. *arXiv preprint arXiv:2411.02256*, 2024. 1, 6
- [16] Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. LoRA: Low-Rank Adaptation of large language models. In *Proc. ICLR*, 2022. 5
- [17] Christian Huber, Juan Hussain, Sebastian Stüker, and Alexander Waibel. Instant one-shot word-learning for context-specific neural sequence-to-sequence speech recognition. In *IEEE Automatic Speech Recognition and Understanding Workshop*, 2021. 1, 2
- [18] Mahaveer Jain, Gil Keren, Jay Mahadeokar, Geoffrey Zweig, Florian Metze, and Yatharth Saraf. Contextual rnn-t for open domain asr. *arXiv preprint arXiv:2006.03411*, 2020. 1, 2
- [19] Alexandros Koumparoulis, Gerasimos Potamianos, Youssef Mroueh, and Steven J Rennie. Exploring roi size in deep learning based lipreading. In *AVSP*, 2017. 1
- [20] Egor Lakomkin, Chunyang Wu, Yassir Fathullah, Ozlem Kalinli, Michael L Seltzer, and Christian Fuegen. End-to-end speech recognition contextualization with large language models. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2024. 1, 2, 4
- [21] Yuxuan Lan, Barry-John Theobald, Richard Harvey, Eng-Jon Ong, and Richard Bowden. Improving visual features for lip-reading. In *Auditory-visual speech processing 2010*, 2010. 2
- [22] Xubo Liu, Egor Lakomkin, Konstantinos Vougioukas, Pingchuan Ma, Honglie Chen, Ruiming Xie, Morrie Doulaty, Niko Moritz, Jachym Kolar, Stavros Petridis, et al. Synthsvr: Scaling up visual speech recognition with synthetic supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 6
- [23] Juergen Luetin and Neil A Thacker. Speechreading using probabilistic models. *Computer vision and image understanding*, 1997. 2
- [24] Pingchuan Ma, Rodrigo Mira, Stavros Petridis, Björn W Schuller, and Maja Pantic. Lira: Learning visual speech representations from audio through self-supervision. *arXiv preprint arXiv:2106.09171*, 2021. 6
- [25] Pingchuan Ma, Stavros Petridis, and Maja Pantic. End-to-end audio-visual speech recognition with conformers. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2021. 2, 6
- [26] Pingchuan Ma, Stavros Petridis, and Maja Pantic. Visual speech recognition for multiple languages in the wild. *Nature Machine Intelligence*, 2022. 6

- [27] Pingchuan Ma, Alexandros Haliassos, Adriana Fernandez-Lopez, Honglie Chen, Stavros Petridis, and Maja Pantic. Auto-avs: Audio-visual speech recognition with automatic labels. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2023. 6, 2, 3
- [28] Harry McGurk and John MacDonald. Hearing lips and seeing voices. *Nature*, 1976. 1
- [29] Kento Nozawa, Takashi Masuko, and Toru Taniguchi. Enhancing large language model-based speech recognition by contextualization for rare and ambiguous words. *arXiv preprint arXiv:2408.08027*, 2024. 1, 2
- [30] Samuel Pachoud, Shaogang Gong, and Andrea Cavallaro. Macro-cuboid based probabilistic matching for lip-reading digits. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2008. 2
- [31] Ankita Pasad, Bowen Shi, and Karen Livescu. Comparative layer-wise analysis of self-supervised speech models. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023. 3
- [32] Eric Petajan, Bradford Bischoff, David Bodoff, and N Michael Brooke. An improved automatic lipreading system to enhance speech recognition. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, 1988. 2
- [33] KR Prajwal, Triantafyllos Afouras, and Andrew Zisserman. Sub-word level lip reading with visual attention. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, 2022. 6
- [34] KR Prajwal, Triantafyllos Afouras, and Andrew Zisserman. Speech recognition models are strong lip-readers. In *InterSpeech 2024*, pages 2425–2429, 2024. 6
- [35] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, 2023. 6
- [36] Kanthashree Mysore Sathyendra, Thejaswi Muniyappa, Feng-Ju Chang, Jing Liu, Jinru Su, Grant P Strimel, Athanasios Mouchtaris, and Siegfried Kunzmann. Contextual adapters for personalized speech recognition in neural transducers. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2022. 1, 2
- [37] Dmitriy Serdyuk, Otavio Braga, and Olivier Siohan. Transformer-based video front-ends for audio-visual speech recognition for single and multi-person video. *arXiv preprint arXiv:2201.10439*, 2022. 6
- [38] Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*, 2017. 5
- [39] Bowen Shi, Wei-Ning Hsu, Kushal Lakhotia, and Abdelrahman Mohamed. Learning audio-visual speech representation by masked multimodal cluster prediction. *arXiv preprint arXiv:2201.02184*, 2022. 1, 2, 3, 6
- [40] Brendan Shillingford, Yannis Assael, Matthew W Hoffman, Thomas Paine, Cian Hughes, Utsav Prabhu, Hank Liao, Hasim Sak, Kanishka Rao, Lorraine Bennett, et al. Large-scale visual speech recognition. *arXiv preprint arXiv:1807.05162*, 2018. 6
- [41] Jiwon Suh, Injae Na, and Woohwan Jung. Improving domain-specific asr with llm-generated contextual descriptions. In *Conference of the International Speech Communication Association*, pages 1255–1259. International Speech Communication Association, 2024. 1
- [42] William H Sumbly and Irwin Pollack. Visual contribution to speech intelligibility in noise. *The journal of the acoustical society of america*, 1954. 1
- [43] Chunlin Tian, Zhan Shi, Zhijiang Guo, Li Li, and Cheng-Zhong Xu. Hydralora: An asymmetric lora architecture for efficient fine-tuning. *Advances in Neural Information Processing Systems*, 37:9565–9584, 2025. 5
- [44] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023. 2, 3
- [45] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017. 3
- [46] Jeong Hun Yeo, Seunghee Han, Minsu Kim, and Yong Man Ro. Where visual speech meets language: Vsp-llm framework for efficient and context-aware visual speech processing. *arXiv preprint arXiv:2402.15151*, 2024. 6, 3
- [47] Jeong Hun Yeo, Minsu Kim, Jeongsoo Choi, Dae Hoe Kim, and Yong Man Ro. Akvs: Audio knowledge empowered visual speech recognition by compressing audio knowledge of a pretrained model. *IEEE Transactions on Multimedia*, 2024. 6
- [48] Xingxuan Zhang, Feng Cheng, and Shilin Wang. Spatio-temporal fusion based convolutional sequence learning for lip reading. In *Proceedings of the IEEE/CVF International conference on Computer Vision*, 2019. 6
- [49] Yuanhang Zhang, Shuang Yang, Shiguang Shan, and Xilin Chen. Es3: Evolving self-supervised learning of robust audio-visual speech representations. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 1
- [50] Ziqiang Zhang, Long Zhou, Junyi Ao, Shujie Liu, Lirong Dai, Jinyu Li, and Furu Wei. Speechut: Bridging speech and text with hidden-unit for encoder-decoder based speech-text pre-training. *arXiv preprint arXiv:2210.03730*, 2022. 3
- [51] Jiacheng Zhu, Kristjan Greenewald, Kimia Nadjahi, Haitz Saez De Ocariz Borde, Rickard Br uel Gabriellsson, Leshem Choshen, Marzyeh Ghassemi, Mikhail Yurochkin, and Justin Solomon. Asymmetry in low-rank adapters of foundation models. *arXiv preprint arXiv:2402.16842*, 2024. 5
- [52] Qiushi Zhu, Long Zhou, Ziqiang Zhang, Shujie Liu, Binxing Jiao, Jie Zhang, Lirong Dai, Daxin Jiang, Jinyu Li, and Furu Wei. Vatlm: Visual-audio-text pre-training with unified masked prediction for speech representation learning. *IEEE Transactions on Multimedia*, 2023. 1, 6