Not Only Vision: Evolve Visual Speech Recognition via Peripheral Information

Supplementary Material

6. Datasets

LRS3. LRS3[2] stands as the largest publicly-available transcribed dataset for audio-visual speech recognition, containing over 430 hours of TED talk presentations. Figure 5 illustrates some frames in the dataset. With a rich vocabulary exceeding 50,000 words and more than 5,500 speakers across presentations. These videos exhibit substantial variations in recording conditions: varying image resolutions, diverse lighting conditions, dynamic head poses, and a wide range of speaking styles and accents (e.g. native or non-native English speakers). Such diversity in visual and linguistic variations makes LRS3 a challenging benchmark for evaluating VSR systems. The test set comprises approximately one hour of footage with speakers entirely distinct from the training set, ensuring rigorous speaker-independent evaluation.

Figure 5. Example frames in LRS3.

AVSpeech. The AVSpeech dataset[11] captures real-world speech scenarios by mining YouTube videos, resulting in a massive collection of 4,700+ hours of segments from more than 290k videos. Some frames of the dataset are shown in Figure 6. The facial regions in these videos exhibit substantial variations: from consumer-grade webcam footage to professional broadcast quality, under natural daylight or artificial illumination, and with spontaneous head movements during speech. Natural occlusions frequently occur, such as hair covering parts of the face, hands gesturing near the mouth, or overlay text and graphics in broadcast videos. These facial variations, combined with the dataset's diverse speaker demographics, pose significant challenges for VSR systems. In our experiments, our evaluation focuses on the English portion of the dataset, which amounts to approximately 1,322 hours of speech content. Additionally, as the original test set is relatively large at 150 hours, we randomly sampled an approximately one-hour subset for test, which maintains consistency with the LRS3 test set scale.

Figure 6. Example frames in AVSpeech.

Contextual Guidance. For LRS3, contextual guidance is collected from both Kaggle and metadata from YouTube links of the videos. For AVSpeech, nearly all samples (approximately 97%) are paired with video titles and descriptions, which are directly extracted from their YouTube links. The final statistics of contextual guidance and examples are shown in Table 7.

To ensure consistency and usability, all collected information went through a simple filtering process:

- Descriptions were truncated to three sentences or 100 words, whichever came first.
- Website URLs embedded in the descriptions were replaced with a generic placeholder, `url`, to produce concise and clean text.

This filtering step aimed to retain only the most relevant and readable portions of the content-related anchors.

Table 7. **Overview of contextual guidance of LRS3 and AVSpeech.** Availability indicates proportion of samples.

Type	**Availability**	**Example**
LRS3		
Scenario	100%	A speech from TED talk
Speaker Name	99%	Niels Diffrient
Speaker Tags	31%	Designer
Speaker Description	31%	Design legend Niels Diffrient is the creator of the ...
Speech Title	99%	Rethinking the way we sit down
Speech Description	99%	Design legend Niels Diffrient talks about his life in ...
AVSpeech		
Title	99%	Malware in the industrial world
Description	99%	Dewan Chowdhury, MalCrawler A talk at Kaspersky ...

7. Implementation Details

For each video sample, the input consists of a sequence of lip-centered regions of interest (ROIs), each cropped to 96 × 96 pixels from individual frames. This sequence is processed by the visual encoder to extract the corresponding visual features. We use the AV-HuBERT[39] Large encoder, which outputs feature representations with a dimension of $d_f = 1024$, while the large language model LLaMA-3-8B-Instruct[10] has an embedding dimension of $d_e = 4096$. For downsampling by averaging, if the feature sequence length is not perfectly divisible by the downsampling factor, the last few feature frames are discarded until divisibility is ensured. Within the LLM, the synergy LoRA include a visual-specific module with a rank of 96 and an alpha of 192, as well as a Mixture-of-Experts (MoE)-like module applied to the complete input sequence. The MoE module consists of $K = 4$ expert modules, each with a rank of 8 and an alpha of 16. A dropout rate of 0.05 is applied to all modules during training to enhance generalization. For inguistic perturbations, random tokens or text fragments are inserted into the sentence with a probability of 0.2, where the inserted tokens have a length ranging from 8 to 16 tokens.

All models are trained for 8 epochs using the AdamW optimizer and a reciprocal learning rate scheduler. The projector and synergy LoRA are trainable throughout duiring the whole training process. The visual encoder is unlocked for training during the last 4 epochs. And during last 4 epochs, we saved checkpoints every 0.5 epoch, which were then averaged to produce the final model weights. During inference, beam search with a width of 8 was employed. All training and evaluations were conducted on 8 NVIDIA A100 40GB GPUs.

8. Additional Experiments

We conducted a series of additional experiments to investigate the impact of different components in our framework. As the early-stage exploration, experiments in section 8.1, 8.2 and 9.1 differ slightly from the main experimental setup and were performed using LLaMA-2-7B[44]. Specifically, we employed the standard LoRA for fine-tuning the LLM, with a rank of 256 and an alpha of 512. The model was trained for 16 epochs, and the encoder was kept frozen unless otherwise stated. Despite these differences, it does not affect the generality of the conclusions.

8.1. Necessity of the Visual Encoder

The visual encoder plays a crucial role in VSR systems by processing and extracting meaningful visual features from raw input videos. In our primary approach, peripheral information and visual features are integrated as parallel inputs to the LLM, enabling rich interactions between these two modalities during decoding. To investigate whether visual information is essential for effectively leveraging peripheral information, we implement a variant that operates exclusively in the text domain and compare it with our proposed approach that embeds visual features.

Results are presented in Table 8. Specifically, we train a baseline model (No.1 in the table), which consists of a frozen AV-HuBERT encoder and a trainable 6-layer Transformer decoder. The generated Top-5 predictions are gathered for our experiments. These predictions are then refined using a LLM fine-tuned with or without contextual guidance (No.2 in the table). This refinement process operates solely in the text domain, where the LLM optimizes the predictions based on contextual guidance. The prompt templates used for this text-only refinement approach are detailed in Table 17.

Table 8. **Comparison of text-only refinement versus visual feature embedded approach.** Contextual guidance mentioned in the table includes speech title and description. Gray numbers indicate relative decreases compared to baseline. Lower WER is better.

No.	Configuration	Model	WER (%)
1	Best Hypothesis	Baseline	29.4
2	Top-5 Hypotheses	LLM w/ LoRA	29.2 (-0.2)
	Top-5 Hypotheses + Contextual Guidance		28.8 (-0.6)
3	Visual Feature	Visual Feature Embedded	26.6 (-2.8)
	Visual Feature + Contextual Guidance	LLM w/ LoRA	24.7 (-4.7)

We observed two points: (i) Incorporating visual features alone shows more effectiveness than text-only refinement with Top-5 hypotheses. This gap suggests that visual features preserve crucial speech characteristics that may be lost in text hypotheses, providing a more reliable foundation for further improvement with contextual guidance. (ii) The impact of contextual guidance varies substantially between the two approaches. With visual features, contextual guidance brings an additional 1.9% WER reduction, whereas in the text-only setting, it only contributes a 0.4% improvement. This difference indicates that the semantic information from contextual guidance can be more effectively utilized when combined with detailed speech representations, as the model can better resolve ambiguities by aligning contextual peripheral information with visual features. To summarize, preserving speech information through visual features is crucial for effective recognition. Text-only approach, despite leveraging the same powerful language model and contextual guidance, cannot fully compensate for the absence of original speech signals.

8.2. Visual Encoder Selection

Having demonstrated the importance of visual features in our framework, we further investigate how different visual encoders affect the model's performance by comparing three representative architectures from different training paradigms: AV-HuBERT [39], RAVEn [14], and Auto-AVSR [27]. Each leverages a unique training approach, where AV-HuBERT utilizes cluster-based self-supervised learning (SSL), RAVEn employs the common cross-modal masked prediction for SSL, and Auto-AVSR adopts end-to-end (E2E) training for audio-visual speech recognition. We adopt RAVEn Large[3] in the experiments for fair comparison, which has a similar model size to AV-HuBERT Large[4] used in our submission. For Auto-AVSR, we only use its VSR encoder trained on 1,759 hours of VSR data[5] in an E2E framework, as it achieves comparable performance to AV-HuBERT Large and RAVEn Large on VSR tasks.

Table 9. **Results under different visual encoders.**

Visual Encoder	WER (%)
AV-HuBERT[39]	26.6
RAVEn[14]	83.7
Auto-AVSR[27]	—

Under the same training configuration (frozen encoder, LoRA-tuned LLM, without peripheral information) as AV-HuBERT, both the other two visual encoders exhibited significant limitations, as shown in Table 9: RAVEn showed inferior performance, while the Auto-AVSR encoder failed to converge in training. RAVEn's inadequate performance in our achitecture aligns with the findings reported by Cappellazzo et al. [7], where it fails to match the performance of a frozen AV-HuBERT encoder even after fine-tuning. The superior performance of AV-HuBERT could be attributed to its cluster-based self-supervised leaning paradigm, which enables learned representations to better align with phonetic and linguistic information[31, 46, 50]. For Auto-AVSR, we hypothesize that its convergence failure might be attributed to its end-to-end supervised training nature, where the encoder's features are specifically optimized for its corresponding decoder and may converge to local optima or regions that are less adaptable to other decoders such as LLMs when trained on limited labeled data (433 hours in our case).

8.3. Effect of Different Large Language Models

Table 10. **Results fo different LLMs on LRS3.**

LLM	LLaMA-2-7B	LLaMA-3.1-8B	LLaMA-3.1-8B-Instruct
WER (%)	23.60	23.24	23.18

To investigate how different language models affect the overall performance, we experimented three model, LLaMA-2-7B[44], LLaMA-3.1 and LLaMA-3.1-Instruct[10]. The results are reported under identical training conditions and peripheral information settings, shown in Table 10. Improvements from LLaMA-2 to LLaMA-3 series suggest that advanced understanding and reasoning capabilities of LLM contribute to more effective utilization of peripheral information, leading to improved accuracy.

8.4. Effect of Down-sampling Strategies

Table 11. **WER under different down-sampling rates and methods.**

Rate	Method	WER (%)
1×	-	25.34
2×	Concatenation	25.30
2×	Average Pooling	25.25
3×	Average Pooling	25.84

Due to the significant temporal disparity between visual features (25 frames per second) and corresponding speech text (around 2-3 words per second), we perform down-sampling for the visual features to make the transition from visual modality to LLM's pre-learned textual space more easily. Experiments were conducted with original LoRA for LLaMA-3.1-8B-Instruct, and peripheral information is not integrated. As shown in Table 11, a 2× average pooling provides the optimal balance between performance and efficiency. It slightly outperforms the model without down-sampling as well as concatenation, while higher down-sampling rates leads to performance degradation. Although concatenation preserves more raw information, the model may struggle to effectively align and utilize it, leading to a marginal decline in performance.

9. Peripheral Information

9.1. Where different peripheral information work

Where Contextual Guidance works. To understand how contextual information enhances recognition, we performed an analysis on two types of overlap between contextual guidance (CG) and ground truth (GT) transcripts: full word overlap and content word overlap (excluding stopwords like "the", "is"). As shown in Table 12, incorporating contextual guidance reduces the error rate from 26.6% to 24.7%, while the overall non-stopword overlap rate between CG and GT is 8.1%. For overlapping stopwords, the accuracy improves from 72% to 81%. However, this improvement only contributes 0.7% to the total 1.9% error rate reduction. The remaining comes from better recognition of common words and sentence structure.

This result indicates that contextual guidance's impact extends beyond direct word matching. Firstly, it helps the model generate more grammatically coherent sequences. Moreover, it improves recognition accuracy for words not present in the peripheral information. The model uses contextual guidance to build a semantic framework to better infer and predict rather than simply matching keywords, lead-

[3] https://github.com/ahaliassos/raven

[4] https://github.com/facebookresearch/av_hubert

[5] https://github.com/mpc001/auto_avsr

Table 12. **Analysis of Contextual Guidance (CG)'s impact.**

Evaluation Metrics	Configuration	Value (%)
Overall Recognition Performance		
WER	w/o CG	26.6
	w/ CG	24.7
Word Overlap Analysis		
CG-GT overlap rate	All words	26.9
	w/o stopwords	8.1
Word accuracy in CG-GT overlap regions		
Accuracy	w/o CG	72.0
	w/ CG	81.0

ing to more accurate and coherent transcriptions overall.

Where Task Expertise and Linguistic Perturbation Works. In our evaluation, we categorized words into stopwords and non-stopwords, calculating accuracy rates after sequence alignment using edit distance. The alignment process identifies substitution, deletion, and insertion errors by matching each word in the hypothesis with the reference transcript, ensuring a fair comparison across different word types. To isolate the impact of each method, we conducted controlled experiments comparing models with and without task expertise and linguistic perturbation while keeping all other conditions identical. In this experiment, we used the prompt *Transcribe the speech and then correct possible errors* to guide the model's behavior. The generation constraint was set as *Transcript after correction*. Our accuracy calculation only considers substitution errors (incorrect predictions) by dividing the number of correct predictions by the total word count per category. Results are shown in Table 13.

Table 13. **Impact of task expertise and linguistic perturbation on different word categories.**

Method	Stopword Accuracy	Non-stopword Accuracy	WER (%)
VSR-adapted LLM	81.2	72.6	26.6
+ Task Expertise	+1.5	+1.1	26.2
+ Linguistic Perturbation	-0.2	-0.2	26.1

We observed that task expertise improved the accuracy for stopwords and non-stopwords by 1.5% and 1.1% respectively, indicating that human-sourced experiential information enhances word-level prediction accuracy. This balanced improvement across word types suggests that task expertise help on both functional errors (in stopwords) and semantic errors (in content words). The improvement over stopwords benefits from the linguistic rules presented by LLM, and the improvement over content words may be benefited from the common sense and the general priors learned by LLM which helps to improve the predictions.

While linguistic perturbation shows a slight decrease of about 0.2% in accuracy for both categories when considering only substitution errors, its improvement in overall WER suggests that linguistic perturbation excels at maintaining the prediction completeness. This finding indicates that linguistic perturbation helps the model better understand natural language flow and reduce errors such as word omissions or redundant insertions, while making the model more robust to irrelevant information in the input.

Different versions of speech descriptions in contextual guidance. Three versions of speech description for LRS3 are used in our experiments as shown in Table 14, including ***Raw***, ***Filtered*** and ***Summarized***. The ***Raw*** version refers to the original unprocessed speech description obtained from the source website. Since these descriptions often contain irrelevant promotional content at the end and embedded hyperlinks throughout the text, we obtained a ***Filtered*** version using simple rules as described in Appendix 6. To extract concise and relevant information from original long and semantically noisy speech descriptions for TED talks, we designed a structured prompt shown in Table 17. This process generates our ***Summarized*** version, where the assistant may help remove semantically irrelevant content and maintain consistency across summarie.

Table 14. **Effect of Linguistic Perturbation under different speech description settings.** Character indicates the key properties of each description type, while WER shows performance *with / without* linguistic perturbation.

Speech Description	Character	WER (%)
Raw	Long & Noisy	24.9 / 26.3
Filtered	Short & Clean	24.7 / 24.9
Summarized	Short & Formulaic	25.5 / 25.1

9.2. Further Discussion on Linguistic Perturbation

Prior studies in ASR [20] have leveraged title and description information associated with speech to enhance recognition performance. They introduce context perturbations during decoding to examine the model's sensitivity to noise, showing that the absence of context or its substitution with random words from the training data leads to slight performance degradation. Our work differs from it in three key aspects:

First, rather than focusing solely on contextual guidance in VSR, we broaden the perspective to consider both task expertise and linguistic perturbations. We show that, under an appropriate methodological framework, these factors can positively contribute to recognition performance rather than acting as detrimental noise.

Second, our approach that introduce linguistic perturbation during training explicitly enhances the model's ability to process noise. In contrast to methods that only introduce perturbations during decoding, we incorporate them at the training stage while keeping contextual information intact. During decoding, we remove these perturbations to preserve clean input. Our method allows the model to generalize better to diverse inputs.

Finally, we introduce a novel adaptation module, Synergy LoRA, that enables the effective integration of multi-level information. This mechanism facilitates a more efficient utilization of various linguistic cues, further improving robustness and generalization across different input scenarios.

9.3. Cases of Peripheral Information Benefits

Table 15 illustrates how different types of peripheral information enhance recognition accuracy. Cases highlighted in green represent better predictions using given peripheral information, while those in red indicate cases where the WER becomes worse. In Case 1 and 2, the speech title and description provide semantic clues (underlined "disaster" and "men") that help recover key phrases, while the model without them fails completely. Case 3 demonstrates how scene information helps recognize common speech phrases, correctly outputting the closing statement. Case 4 shows how our model successfully disambiguates phonetically similar content, correctly recognizing "computers". These cases highlight how our peripheral information integration approach guides the model toward semantically coherent recognition, outperforming the context-free model.

We present more success and failure examples in Table 16. In the majority of successful cases in our experiments, we observed that contextual cues are either directly or indirectly connected with the spoken content to be recognized, thereby providing substantial improvements. For instance, in Case 1, the presence of "Biologist" and "tail" in the contextual information likely facilitated the prediction of "animal" and "end". In Case 2, although the introduction of contextual information did not lead to completely accurate sentence prediction, the correct prediction of the key word "wash" significantly improved the semantic alignment with the ground truth.

Analysis of failure cases reveals that errors often occur when the target utterance has limited relevance to the general topic or provided contextual information, where contextual guidance may potentially mislead the recognition process. As illustrated in Case 6, the presence of the word "algorithms" in the speech title and description introduced incorrect predictions, where "super real" was mistakenly recognized as "algorithm". Future work should explore mechanisms that enable the model to selectively utilize contextual information, distinguishing between relevant and potentially misleading context.

10. Examples of Prompt Used in Our Work

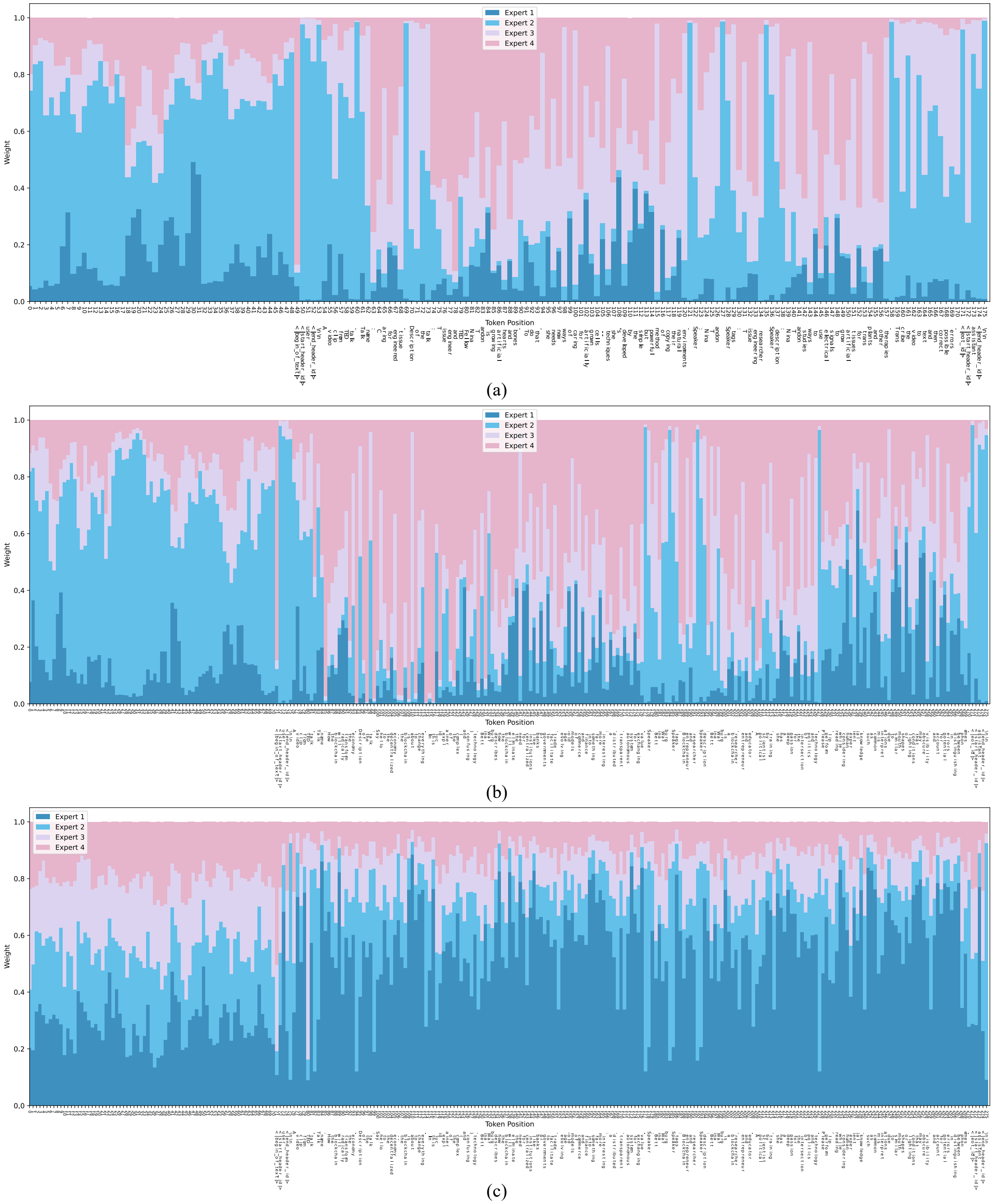
Prompts for Peripheral Information Integration. Table 17 illustrates the representative prompts under different settings in our experiments. For example, when incorporating scenario contextual guidance and task expertise on LRS3 dataset, we use "A speech from TED talk. Transcribe the speech and then correct possible errors." as the input prompt. Similar patterns can be applied to generate other prompts with different types of peripheral information.

11. More Visualizations

As shown in Figure 7, we conducted more visualizations of the expert load in Synergy LoRA. Samples are randomly selected from the LRS3 dataset.

Notably, at the 32nd attention layer, we observed a consistent activation pattern. Different components of the input, such as structural elements, instructional information, and fine-grained descriptions, tend to activate specific experts with significantly higher weights. This suggests that the MoE module effectively captures distinct types of information, assigning specialized experts accordingly.

However, although the first attention layer also exhibits some activation patterns, they are not as pronounced as those in higher layers. A possible explanation is that lower layers in LLMs generally focus on capturing basic lexical or syntactic structures, whereas deeper layers tend to model more abstract and contextual information.



(a)

(b)

(c)

Figure 7. **More visualizations of expert weight.** **(a)** Sample 1, layer 32 of LLM; **(b)** Sample 2, layer 32; **(c)** Sample 2, layer 1. Specific text of each token is shown below the figure.

Table 15. **Success Cases on the LRS3 test set.**

Case 1 **Scene:** TED Talk *Instruction: A speech from TED talk. Transcribe the speech.*
GT: thank you for your time
w/: thank you for your time
w/o: second verse time
Case 2 **Speech Title:** These robots come to the rescue after a disaster
GT: that was a huge problem at the haiti earthquake
w/: that was a huge problem at the hurricane earthquake
w/o: that was a huge problem in the ann
Case 3 **Speech Title:** The single biggest health threat women face **Description:** Surprising, but true: More women now die of heart disease than men, yet cardiovascular research ...
GT: she had to impersonate a man
w/: she had to do impersonate a man
w/o: she had to do an emergency back
Case 4 **Task Expertise Type A** *Instruction: Transcribe the speech and correct possible errors.* *Response: Transcript after correction: {transcript}*
GT: you learn more about how computers work
w/: you'll learn more about how computers work
w/o: to learn more about how you build this work
Case 5 **Task Expertise Type B** *Instruction: Please perform lip-reading while considering human experiential knowledge, such as common misinterpretations due to similar mouth shapes or lighting conditions that may obscure visibility, and account for potential errors in distinguishing between phonemes.* *Constrain: After carefully consideration, the identified sentence is: {transcript}*
GT: your job needs to be challenging
w/: your job needs to be challenging
w/o: your job is to be challenging
GT: and then something falls off the wall
w/: and then something falls off the wall
w/o: and then something falls off the walk

Table 16. **More Success and Failure Cases on the LRS3 test set.**

Case 1 **Scene:** TED Talk **Speech Title:** Learning from the gecko's tail **Description:** Biologist Robert Full studies the amazing gecko, with its supersticky feet and tenacious climbing skill. But high-speed footage reveals that the gecko's tail harbors perhaps the most surprising talents of all.
GT: look at the end to see the animal
w/: look at the end of the animal
w/o: look at the editors of the camera
Case 2 **Scene:** TED Talk **Speech Title:** The simple power of hand-washing **Description:** Myriam Sidibe is a warrior in the fight against childhood disease. Her weapon of choice? A bar of soap.
GT: did you learn to wash your hands at home
w/: did you learn to wash your hands the right way
w/o: did you learn to raise your hand to ask for help
Case3 **Scene:** TED Talk **Speech Title:** A global food crisis may be less than a decade away **Description:** Sara Menker quit a career in commodities trading to figure out how the global value chain of agriculture works. Her discoveries have led to some startling predictions: "We could have a tipping point in global food and agriculture if surging demand surpasses the agricultural system's structural capacity to produce food," she says.
GT: china is constrained in terms of how much more land it actually has available for agriculture and it has massive
w/: china is constrained in terms of how much more land it actually has available for agriculture and it has massive
w/o: chinese constrains determine how much more land it actually has available for agriculture and it has massive
Case 4 **Scene:** TED Talk **Speech Title:** 3 ways to fix a broken news industry **Description:** Something is very wrong with the news industry. Trust in the media has hit an all-time low; we're inundated with sensationalist stories, and consistent, high-quality reporting is scarce, says journalist Lara Setrakian.
GT: i do believe we can fix what's broken
w/: i do believe that we can fix what's broken
w/o: i do believe we can fix what's broken
Case 5 **Scene:** TED Talk **Speech Title:** The power of the informal economy **Description:** Robert Neuwirth spent four years among the chaotic stalls of street markets, talking to pushcart hawkers and gray marketers, to study the remarkable S̈ystem D,ẗhe world's unlicensed economic network. Responsible for some 1.8 billion jobs, it's an economy of underappreciated power and scope.
GT: there's nothing underground about it
w/: there's nothing on the ground about it
w/o: there's nothing ungrounded about it
Case 6 **Scene:** TED Talk **Speech Title:** How algorithms shape our world **Description:** Kevin Slavin argues that we're living in a world designed for – and increasingly controlled by – algorithms. In this riveting talk from TEDGlobal, he shows how these complex computer programs determine: espionage tactics, stock prices, movie scripts, and architecture.
GT: it's super real and it's happening around you
w/: it's an algorithm you feel that it's happening around you
w/o: it's in its humid real and it's happening around you

Table 17. **Examples of prompts and output constraint (when incorporating task expertise) in our experiments.** Gray text includes the dataset, experiment setting or purpose.

LRS3: Instruciton, without any peripheral information
Transcribe the video to text.
LRS3: Contextual Guidance (full) and instruction
A video clip from TED talk.\nSpeech Title: How the blockchain will radically transform the economy.\nDescription of the talk:\nSay hello to the decentralized economy – the blockchain is about to change everything. In this lucid explainer of the complex (and confusing) technology, Bettina Warburg describes how the blockchain will eliminate the need for centralized institutions like banks or governments to facilitate trade, evolving age-old models of commerce and finance into something far more interesting: a distributed, transparent, autonomous system for exchanging value.\nSpeaker: Bettina Warburg\nSpeaker tags: Blockchain entrepreneur and researcher\nSpeaker description: Bettina Warburg is a blockchain researcher, entrepreneur and educator. A political scientist by training, she has a deep passion for the intersection of politics and technology. \nTranscribe the video to text.
LRS3: Contextual Guidance (Speech Title and description) and instruction
A video clip from TED talk.\nSpeech Title: How the blockchain will radically transform the economy.\nDescription of the talk:\nSay hello to the decentralized economy – the blockchain is about to change everything. In this lucid explainer of the complex (and confusing) technology, Bettina Warburg describes how the blockchain will eliminate the need for centralized institutions like banks or governments to facilitate trade, evolving age-old models of commerce and finance into something far more interesting: a distributed, transparent, autonomous system for exchanging value. \nTranscribe the video to text.
LRS3: Task Expertise Type A
Instruction: Transcribe the speech and then correct possible errors. *Constrain:* Transcript after correction:{transcript}
LRS3: Task Expertise Type B
Instruction: Please perform lip-reading while considering human experiential knowledge, such as common misinterpretations due to similar mouth shapes or lighting conditions that may obscure visibility, and account for potential errors in distinguishing between phonemes. *Constrain:* After carefully consideration, the identified sentence is: {transcript}
LRS3: Text-only refinement w/o Contextual Guidance
Instruction: You are familiar with visual speech recognition (VSR) and transcript re-scoring. You have a few transcripts generated by a VSR model. Your task is to generate the most likely transcript from them. IMPORTANT: Format your response as #transcript#. No other text should be included.\n \n ### Input: {nbest_list}\n \n ### Response:
LRS3: Text-only refinement w/ Contextual Guidance
Instruction: You are familiar with visual speech recognition (VSR) and transcript re-scoring. Your task is to work with a speech from TED talk titled "{name}". The description of this talk is "{description}". Several candidate transcripts have been generated by a VSR model for this speech. Your task is to generate the most likely transcript from them. IMPORTANT: Format your response as #transcript#. \n \n ### Input: \n {nbest_list} \n \n ### Response:
LRS3: Prompt for a ***Summarized*** version of peripheral information
This is a description of a ted talk, please summarize it concisely and comprehensively in a few sentences, within {num} words. Begin with "A TED talk ..." Description: "{description}"