Speech-Aware Long Context Pruning and Integration for Contextualized Automatic Speech Recognition

Yiming Rong[12*], Yixin Zhang[12*], Ziyi Wang[12*], Deyang Jiang[12], Yunlong Zhao[12], Haoran Wu[1†], Shiyu Zhou[1†], Bo Xu[12†]

[1]The Key Laboratory of Cognition and Decision Intelligence for Complex Systems, Institute of Automation, Chinese Academy of Sciences, China
[2]School of Artificial Intelligence, University of Chinese Academy of Sciences, China
{rongyiming2022, zhangyixin2024, wuhaoran2018, shiyu.zhou, xubo}@ia.ac.cn

Abstract

Automatic speech recognition (ASR) systems have achieved remarkable performance in common conditions but often struggle to leverage long-context information in contextualized scenarios that require domain-specific knowledge, such as conference presentations. This challenge arises primarily due to constrained model context windows and the sparsity of relevant information within extensive contextual noise. To solve this, we propose the SAP^2 method, a novel framework that dynamically prunes and integrates relevant contextual keywords in two stages. Specifically, each stage leverages our proposed Speech-Driven Attention-based Pooling mechanism, enabling efficient compression of context embeddings while preserving speech-salient information. Experimental results demonstrate state-of-the-art performance of SAP^2 on the SlideSpeech and LibriSpeech datasets, achieving word error rates (WER) of 7.71% and 1.12%, respectively. On SlideSpeech, our method notably reduces biased keyword error rates (B-WER) by 41.1% compared to non-contextual baselines. SAP^2 also exhibits robust scalability, consistently maintaining performance under extensive contextual input conditions on both datasets.

Code — https://github.com/jymh/SAP2-ASR

1 Introduction

Automatic speech recognition (ASR) systems have achieved remarkable performance in common scenarios, approaching human-level accuracy in routine speech recognition tasks (Radford et al. 2023; Gao et al. 2022; Xu et al. 2025). However, current challenges no longer center around baseline transcription accuracy but instead focus on effectively recognizing speech in complex, context-rich scenarios (Chang et al. 2021; Ni et al. 2024; Wang et al. 2024c; Sudo et al. 2024). For example, in conference videos, speakers commonly utilize visual aids, particularly slides containing large amounts of text (Wang et al. 2024b,a). Such OCR-derived contexts usually include important entities and domain-specific terminology. In Figure 1, despite that keywords like "glaucoma" constitute only a small portion of the speech content, they are crucial for comprehending domain-specific information. However, processing extensive textual inputs introduces considerable computational overhead and the risk of incorporating irrelevant information when the textual contexts span hundreds of thousands of words.

Existing contextualized ASR methods mainly follow two paradigms: (1) deep biasing E2E models that apply attention mechanisms to influence the decoding of ASR model (Pundak et al. 2018; Chang et al. 2021; Han et al. 2021, 2022; Huang et al. 2023; Sudo et al. 2024; Yu et al. 2024), and (2) SpeechLLM models that treat contextual keywords as prompts (Lakomkin et al. 2024; Yang et al. 2024b; Bai et al. 2024). Deep biasing models face performance gaps compared with large-scale pre-trained models (Lakomkin et al. 2024) and exhibit architectural rigidity. SpeechLLM models are more flexible and robust when incorporating contextual keywords. However, when dealing with long context information, simply treating contextual keywords as prompts poses challenges to effective long-context processing and robust noise filtering.

Neuroscience study on speech recognition (Sohoglu et al. 2012) indicates that human perceptual clarity improves significantly when speech aligns with prior contextual knowledge, demonstrating the human brain's active integration of prior knowledge during speech recognition. Our motivation is to simulate this active integration process, prune and compress noisy contextual information to knowledge more aligned with speech contents. We propose SAP^2 (**S**peech-**A**ware Context **P**runing with **S**peech-Driven **A**ttention-based **P**ooling), a novel framework that actively filters long-form contextual keywords and integrates them into ASR systems without sacrificing efficiency. Our contributions are as follows:

- **Effective Long-Context Processing**: We introduce a novel Speech-Aware Context Pruning framework, instructing a SpeechLLM to dynamically prune OCR-derived textual contexts and retain only keywords directly relevant to the speech content.
- **Cross-Modal Context Compression**: We propose a Speech-Driven Attention-based Pooling mechanism that

*These authors contributed equally.
†Corresponding author.

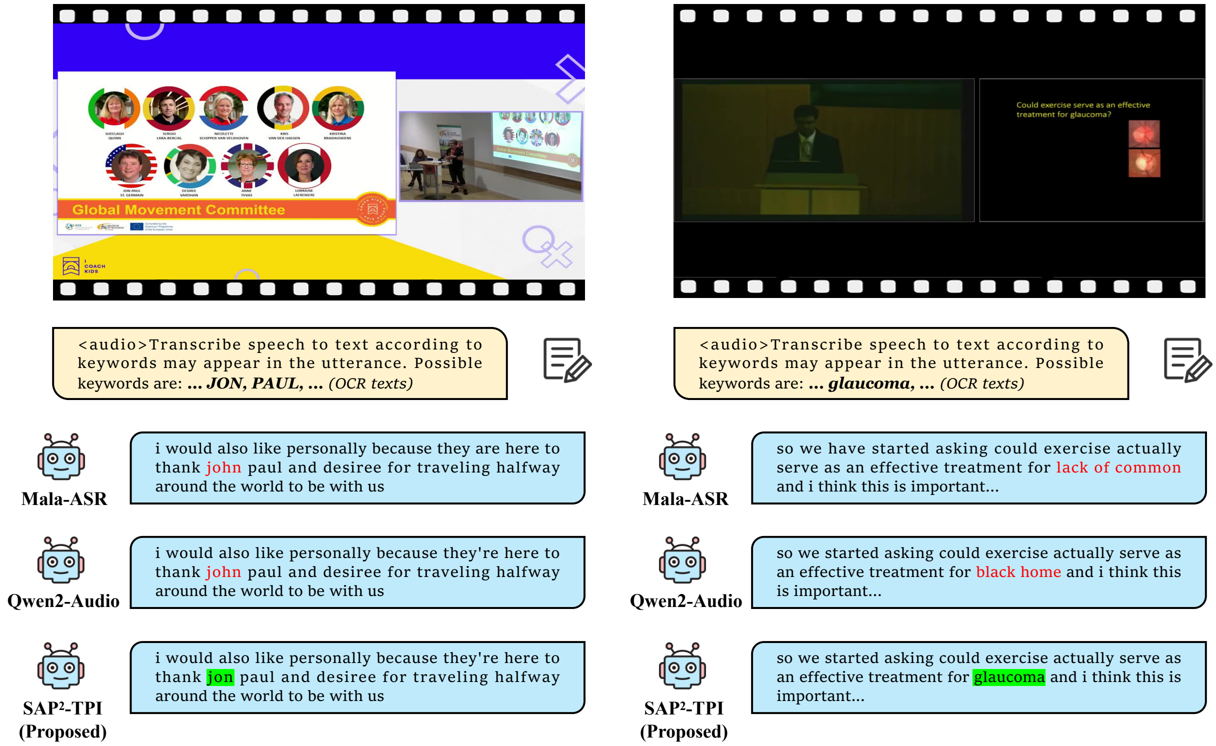

Figure 1: Comparisons of the previous SOTA MaLa-ASR, the baseline Qwen2-Audio-PC, and our SAP^2-TPI on the SlideSpeech test set. For each sample, red texts indicate recognition errors in proper nouns, while green-highlighted texts showcase corrections made by SAP^2-TPI. These two examples represent typical cases of proper noun recognition: the left demonstrates SAP^2-TPI's accuracy in recognizing rare personal names, and the right highlights its capability in identifying medical terminology.

leverages cross-modal alignment between speech and text. This pooling strategy compresses extensive textual inputs into concise, speech-relevant context embeddings.

- **State-of-the-Art Performance and Robustness** Experimental results on SlideSpeech and LibriSpeech validate the superiority of our proposed SAP^2 method, achieving a WER of 7.71% and 1.12%, respectively. Particularly, our approach achieves a notable 41.1% relative improvement in biased keyword recognition over non-contextual baselines on SlideSpeech. Furthermore, SAP^2 demonstrates robust scalability, maintaining consistent performance under extensive contextual input conditions on both datasets.

2 Method

2.1 Problem Formulation

Given a sequence of acoustic features $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T]$ of length T extracted by an audio encoder, the objective of speech recognition is to maximize the probability of output text label $\boldsymbol{Y} = [\boldsymbol{y}_1, \ldots, \boldsymbol{y}_L]$ of length L. In contextualized ASR scenarios, additional context information $\boldsymbol{Z} = [\boldsymbol{z}_1, \ldots, \boldsymbol{z}_C]$ is provided in the dataset $\mathcal{D} = \left\{\boldsymbol{X}^{(i)}, \boldsymbol{Y}^{(i)}, \boldsymbol{Z}^{(i)}\right\}_{i=1}^{N}$. The probability of an auto-regressive ASR model decoding the text label $\boldsymbol{Y}$ is formulated as:

$$p_\theta(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{Z}) = \prod_{j=1}^{L} p_\theta(\boldsymbol{y}_j|\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{Y}_{<j}), \tag{1}$$

where θ is the parameter of the ASR model. For many deep biasing ASR models, $\boldsymbol{Z}$ is fed into a biasing decoder through cross-attention layers. In large pre-trained ASR models like Whisper or SpeechLLM models like Qwen2-Audio, $\boldsymbol{Z}$ is concatenated with acoustic features $\boldsymbol{X}$. We unify the presentation, ignore the implementation details and treat $\boldsymbol{Z}$ as a conditional variable. Thus, the overall training objective is:

$$\max_\theta \mathbb{E}_{(\boldsymbol{X},\boldsymbol{Y},\boldsymbol{Z})\sim\mathcal{D}}[p_\theta(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{Z})]. \tag{2}$$

2.2 Speech-Aware Context Pruning

For contextual keywords $\boldsymbol{Z} = [\boldsymbol{z}_1, \ldots, \boldsymbol{z}_C]$, when C is a very large number, typically ten times larger than L, there might be few keywords truly relevant to the speech. This is a common case in conference recording videos, where slides related to a 10-second utterance contain thousands of words of background information. Suppose only c core keywords $\tilde{\boldsymbol{Z}} = [\tilde{\boldsymbol{z}}_1, \ldots, \tilde{\boldsymbol{z}}_c]$ will benefit the recognition of $\boldsymbol{X}$, which means

$$p_\theta(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{Z}) = \int p_\theta(\boldsymbol{Y}|\boldsymbol{X}, \tilde{\boldsymbol{Z}}) p(\tilde{\boldsymbol{Z}}|\boldsymbol{X}, \boldsymbol{Z}) d\tilde{\boldsymbol{Z}}. \tag{3}$$

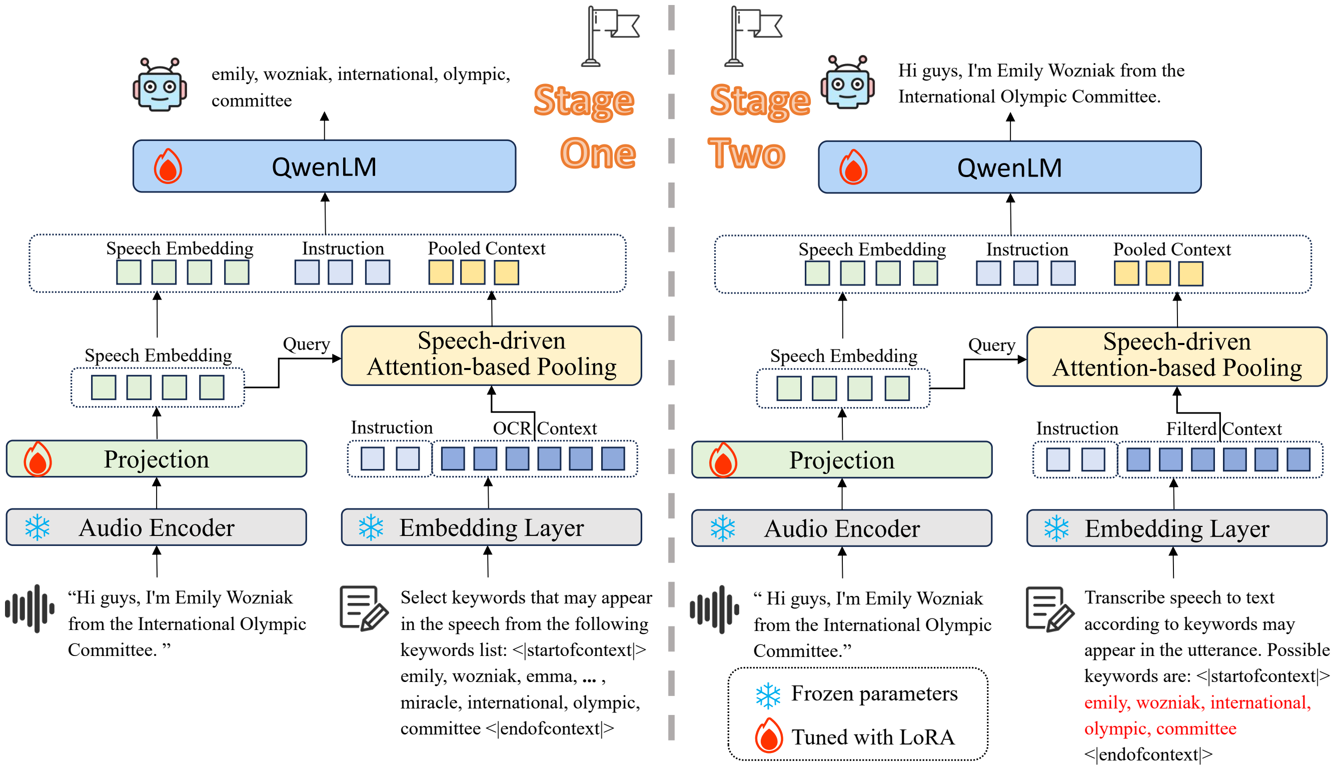

Figure 2: The overall architecture of proposed SAP^2-TPI framework: In stage one, long contextual keywords are pruned based on speech to reduce irrelevant information. Pruned contextual keywords are inputs of stage two, used for contextualized speech recognition. Speech-driven attention-based pooling is utilized in both stages to compress context embeddings.

In practical settings, when text label $\boldsymbol{Y}$ and context information $\boldsymbol{Z}$ are given, the truly beneficial core keywords $\tilde{\boldsymbol{Z}} = \boldsymbol{Z} \cap \boldsymbol{Y}$ is the keywords that appear in the label, the original dataset can be augmented as $\mathcal{D} = \left\{\left(\boldsymbol{X}^{(i)}, \boldsymbol{Y}^{(i)}, \boldsymbol{Z}^{(i)}, \tilde{\boldsymbol{Z}}^{(i)}\right)\right\}_{i=1}^{N}$. Thus, we can suppose that $\tilde{\boldsymbol{Z}}$ is a deterministic variable, $p_\theta(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{Z})$ is derived as:

$$p_\theta(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{Z}) = p_\theta(\boldsymbol{Y}|\boldsymbol{X}, \tilde{\boldsymbol{Z}}) p(\tilde{\boldsymbol{Z}}|\boldsymbol{X}, \boldsymbol{Z}).$$

Suppose that the context pruning model $p(\tilde{\boldsymbol{Z}}|\boldsymbol{X}, \boldsymbol{Z})$ is parameterized by ψ, the overall objective of the system is:

$$\max_{\theta,\psi} \mathbb{E}_{(\boldsymbol{X},\boldsymbol{Y},\boldsymbol{Z},\tilde{\boldsymbol{Z}})\sim\mathcal{D}} \left[p_\theta(\boldsymbol{Y}|\boldsymbol{X}, \tilde{\boldsymbol{Z}}) p_\psi(\tilde{\boldsymbol{Z}}|\boldsymbol{X}, \boldsymbol{Z})\right]. \quad (4)$$

This derivation decomposes the original task into two stages. Prior work like Jayanthi et al. (2023) explores this two-stage prune-and-integrate pipeline. However, the pruning model p_ψ is mainly based on similarity-based algorithms, which makes it hard to generalize to unseeing textual contents and speech. In Figure 2, we utilize a SpeechLLM model as the context pruning model. It accepts the input of long contextual keywords and speech, and output core keywords relevant to the speech using speech-driven attention-based pooling. The core keywords generated in step one is concatenated with the speech to perform contextualized ASR. Regarding this **Two-stage Pruning-Integration (TPI)** manager, we denote the pipeline as SAP^2-TPI. Unless specified otherwise, SAP^2 refers to the two-phase trained variant (SAP^2-TPI). During the training time, ASR model p_θ and speech-driven context pruning model p_ψ can be trained simultaneously, which means the training objective can be written as

$$\begin{aligned} &\max_{\theta} \mathbb{E}_{(\boldsymbol{X},\boldsymbol{Y},\tilde{\boldsymbol{Z}})\sim\mathcal{D}} \left[p_\theta(\boldsymbol{Y} \mid \boldsymbol{X}, \tilde{\boldsymbol{Z}})\right], \\ \text{s.t.} \quad &\max_{\psi} \mathbb{E}_{(\boldsymbol{X},\boldsymbol{Z},\tilde{\boldsymbol{Z}})\sim\mathcal{D}} \left[p_\psi(\tilde{\boldsymbol{Z}} \mid \boldsymbol{X}, \boldsymbol{Z})\right]. \end{aligned} \quad (5)$$

During the inference stage, $p_\theta(\boldsymbol{Y}|\boldsymbol{X}, \tilde{\boldsymbol{Z}})$ relies on $\tilde{\boldsymbol{Z}}$ produced by the context pruning model, so the speech recognition task is conducted after contextual keywords are pruned.

2.3 Speech-driven Attention-based Pooling

In SpeechLLM models, when the length of contextual keywords C is too large or even exceeds the maximum context window of the LLM, the input inevitably suffers from content truncations or critical semantic loss. To solve this, we propose a Speech-driven Attention-based Pooling method that reduces the context length while preserving speech-relevant information. The idea is to weight contextual keywords with speech attention scores and conduct pooling on the weighted text features. We calculate the attention score between speech embedding $\boldsymbol{h}^x$ and the contextual keywords embedding $\boldsymbol{h}^z$ to determine the weight of tokens. Then we perform window-wise pooling to contract the embedding of keywords. Please refer to the Appendix for deductions and detailed implementations.

This speech-driven attention-based pooling mechanism can be used in both the SpeechLLM ASR model and the context pruning model, combining the cross-attention module in deep biasing ASR models and simplicity of SpeechLLM ASR models.

3 Experiments

In this section, we conduct experiments to investigate the effectiveness of our framework.

3.1 Experimental Setup

Dataset We conduct experiments on the **SlideSpeech** (Wang et al. 2024b) dataset and the **LibriSpeech** (Panayotov et al. 2015) dataset. The contextual keywords of SlideSpeech are slide-derived OCR texts, and the contextual keywords of LibriSpeech is dynamically constructed following the approach in Le et al. (2021).

To investigate the impact of extended contextual information on speech recognition accuracy and to evaluate the effectiveness of our proposed long-context pruning and integration framework, we augment SlideSpeech OCR contexts by extracting semantic keywords from five consecutive slides. Specifically, using the Jaccard index to group slides into coherent clusters, combining multiple consecutive slides together.

To assess how augmented contexts would influence ASR performance, we analyze two metrics:

- ***Keyword Coverage Rate*** (The ratio of core keywords in ASR texts to text length) increases from 4.01% (single slide) to 6.51% (five slides), confirming more contextual keywords that may aid ASR are included.
- ***Information Rate*** (The ratio of core keywords to total keywords) decreases from 1.82% (single slide) to 0.64% (five slides), reflecting much higher noise levels in extended contexts. This demonstrates the critical need for robust noise-resistant methods to extract salient information from longer OCR contexts.

Evaluation Metrics We follow the evaluation metrics of previous works of contextualized ASR (Huang et al. 2023; Yu et al. 2024; Yang et al. 2024b; Wang et al. 2024b). We evaluate the results using WER, biased word error rate (B-WER), unbiased word error rate (U-WER), and the Recall of words in both the biasing list and transcript. U-WER is calculated for words not in the contextual keywords list, which reflects the foundational ASR capabilities of the model. B-WER is computed for words in the contextual keywords list, which indicates the model's capability of accurately integrating core keywords.

3.2 Implementation Details

We fine-tune Qwen2-Audio-7B-Instruct (hereafter referred to as Qwen2-Audio) for context pruning and speech recognition tasks. During fine-tuning, we employ LoRA adapters to update both the multimodal projector and LLM backbone

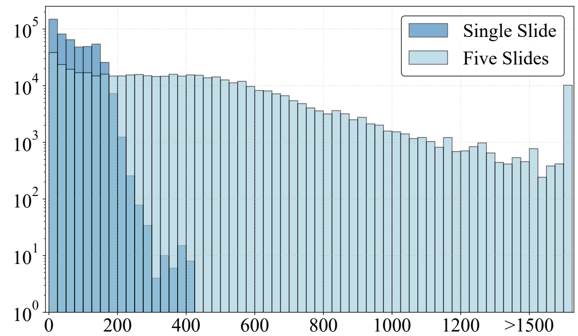

Figure 3: Distribution of tokenized keyword counts. (x-axis: number of tokens, linear scale; y-axis: instance count, log scale). Single-slide contexts exhibit mean/median token lengths of 64.57/53, while five-slide contexts reach 402.81/332.

while keeping the speech encoder frozen. Please refer to the Appendix for details of hyper-parameters.

Speech-driven attention-based pooling mechanism is applied in both stages. In each stage, we only compress contextual keywords, excluding instruction prompts. Without specific clarification, pooling window size is set to 2. Please refer to the Appendix for detailed instruction prompts.

3.3 Experiment Results on SlideSpeech

We first validate our method on SlideSpeech which contains natural long context OCR suitable for contextualized ASR scenarios. Table 1 demonstrates the performance of the SAP^2-TPI model. Under both 1-slide and 5-slide context settings, SAP^2-TPI **achieves state-of-the-art** results on the two subsets of SlideSpeech. WER performance achieves a relative reduction of 30.19% and 10.45% over the previous SOTA MaLa-ASR on the S95 and L95 subsets, respectively. From the table, we can draw the follow conclusions:

- The SlideSpeech dataset features complex conference presentations regarding professional regions. The performance gap between uncontextualized and contextualized methods is obvious. Our SAP^2-TPI using 5-slide context obtains a SOTA WER of 7.71%
- SAP^2-TPI demonstrates strong keywords integration abilities. B-WER of SAP^2-TPI trained on L95 dataset decreases 41.1% compared with Qwen2-Audio fine-tuned without contextual keywords. Notably, SAP^2-TPI simultaneously maintains superior contextual keyword recall 95.59%, demonstrating its effectiveness in both recognition accuracy and contextual relevance. B-WER/Recall performances of SAP^2-TPI achieve an absolute improvement of 2.49%/2.48% and 0.35%/0.28% compared with MaLa-ASR on S95 and L95 subsets, respectively.
- In the 5-slide scenarios compared to the 1-slide scenario, SAP^2-TPI demonstrates stronger performance on the larger subset L95 while also improving overall WER and B-WER on the smaller subset S95. This reflects SAP^2-TPI's robust capabilities in processing long-context in-

Train	Model	Contextual Keywords	Test			
			WER	U-WER	B-WER	Recall ↑
S95 (161h)	SlidesSpeech (Wang et al. 2024b)	✗	21.22	20.83	26.60	73.51
	MaLa-ASR (Yang et al. 2024b)	✗	11.80	11.71	13.52	86.71
	Qwen2-Audio (Chu et al. 2024)	✗	10.79	10.85	7.92	90.33
	CPP (Huang et al. 2023)	1slide	20.95	20.73	24.05	76.10
	LCB-net (Yu et al. 2024)	1slide	19.21	18.89	23.70	76.48
	MaLa-ASR (Yang et al. 2024b)	1slide	11.26	11.52	7.67	92.50
	SAP^2-TPI (Proposed)	1slide	8.04	**8.13**	5.29	94.85
	SAP^2-TPI (Proposed)	5slides	**7.86**	8.17	**5.18**	**94.98**
L95 (473h)	SlidesSpeech (Wang et al. 2024b)	✗	12.89	12.90	12.70	87.43
	MaLa-ASR (Yang et al. 2024b)	✗	8.61	8.72	7.34	92.84
	Qwen2-Audio (Chu et al. 2024)	✗	10.76	10.91	7.71	91.42
	CPP (Huang et al. 2023)	1slide	12.38	12.60	9.32	90.86
	LCB-net (Yu et al. 2024)	1slide	12.02	12.24	9.03	91.12
	MaLa-ASR (Yang et al. 2024b)	1slide	8.46	8.73	4.89	95.31
	SAP^2-TPI (Proposed)	1slide	7,92	8.16	4.71	95.47
	SAP^2-TPI (Proposed)	5slides	**7.71**	**8.08**	**4.54**	**95.59**

Table 1: Performance comparison of our SAP^2-TPI and other models evaluated on SlideSpeech test dataset, trained on S95/L95.

formation. We will further discuss robustness and generalization to longer contexts of SAP^2-TPI in Section 3.7.

3.4 Experiment Results on LibriSpeech

To compare our SAP^2-TPI method with more current contextualized ASR models, we conduct experiments on the widely used LibriSpeech dataset. We analyze WER(U-WER/B-WER) with biasing list lengths of 100, 500, and 1000 on test-clean and test-other subsets. The conclusions drawn from Table 2 align with those in Section 3.3:

- SAP^2-TPI also achieves a **SOTA** performance on LibriSpeech across all biasing list settings, reaching an lowest WER of 1.12%. SAP^2-TPI achieves absolute reductions of 0.48% and 0.74% on WER compared with previous SOTA Gong et al. (2024) on test-clean (N=100) and test-clean (N=500), with even large absolute reductions of 0.8% and 1.18% on test-other (N=100) and test-other(N=100).
- SAP^2-TPI exhibits strong robustness of processing longer context, increasing only 0.1% WER on test-clean subset when 10x more biasing keywords are provided. B-WER on test-clean and test-other subsets also faintly increases when integrating more contextual keywords.

3.5 Ablation Studies

In this section, we aim to clarify the effectiveness of two-stage speech-aware context pruning and the proposed speech-driven attention-based pooling mechanism.

To investigate whether it's useful to prune long context in two stages, we compare our SAP^2-TPI with directly concatenating keywords with speech features and jointly training pruning and recognition model.

Prompt Concatenation (PC) Following Yang et al. (2024b), we directly concatenate contextual keywords with speech features. We design different instructions according to two situations: contextual keywords are provided, or no keywords are available. Please refer to the Appendix for detailed prompts.

Joint Pruning-Integration (JPI) Leveraging Qwen2-Audio's multi-task learning capability, we explore training a single model for context pruning and speech recognition. We train the model to sequentially output pruned keywords and recognition results. Please refer to the Appendix for reasons and deductions.

Furthermore, to clarify the effectiveness of speech-driven attention-based pooling, we compare the ASR performance and time consumption of our proposed SAP^{2}methods to Qwen2-Audio with different stage-wise styles.

Ablation Study on Speech-Aware Context Pruning In Table 3, among the three proposed methods, SAP^2-TPI demonstrates consistent superiority in biased keyword prediction accuracy, achieving the lowest B-WER across all configurations. Notably, SAP^2-PC attains the optimal U-WER of 7.94% on the L95 dataset, while its contextual precision lags behind SAP^2-TPI by 2.03-3.15% in B-WER metrics. SAP^2-JPI exhibits overall WER degradation and reduced contextual sensitivity. Table 4 demonstrates that SAP^2-JPI achieves a lower F1-score of predicting speech-relevant contextual keywords, likely due to interference between its dual objectives of keyword selection and speech recognition. This performance gap highlights the critical importance of effective context pruning, where suboptimal keyword selection directly impacts recognition quality.

We can also conclude from Table 3 how the proposed three methods perform when contextual keywords are augmented. While SAP^2-PC and SAP^2-JPI show marginal improvements or even alter for the worse with extended context windows, SAP^2-TPI achieves constant WER reductions of

Model	N=100		N=500		N=1000	
	test-clean	test-other	test-clean	test-other	test-clean	test-other
DB-RNNT+DB-LM (Le et al. 2021)	1.98 (1.5/5.7)	5.86 (4.9/14.1)	2.09 (1.6/6.2)	6.09 (5.1/15.1)	2.14 (1.6/6.7)	6.35 (5.1/17.2)
Attention-based DB + BPB beam search (Sudo et al. 2024)	2.75 (2.3/6.0)	5.60 (4.9/12.0)	3.21 (2.7/7.0)	6.28 (5.5/13.5)	3.47 (3.0/7.7)	7.34 (6.4/15.8)
Biasing fusion + SpeechLLM (Gong et al. 2024)	1.6 (1.3/5.5)	3.8 (2.6/13.5)	1.9 (1.4/6.0)	3.9 (2.7/14.2)	–	
CTC-Assisted LLM-Based ASR (Yang et al. 2024a)	1.27 (1.00/3.67)	2.72 (2.16/8.02)	1.33 (1.03/3.92)	3.04 (2.40/9.04)	1.33 (1.00/4.16)	2.99 (2.31/9.33)
SAP^2-TPI (Proposed)	1.12 (0.89/**3.03**)	2.60 (2.18/**6.55**)	1.16 (0.93/**3.09**)	2.72 (2.23/**7.29**)	1.22 (0.96/**3.50**)	2.94 (2.33/**8.62**)

Table 2: Performance (WER(U-WER/B-WER)) (%) comparison of our SAP^2-TPI method and different methods on the LibriSpeech test dataset with different biasing list length (N=100, 500, 1000).

Train	Model	Contextual Keywords	Test			
			WER	U-WER	B-WER	Recall ↑
S95 (161h)	SAP^2-PC	1slide	**7.85**	**7.96**	6.36	93.77
	SAP^2-JPI	1slide	8.12	8.25	6.39	93.71
	SAP^2-TPI	1slide	8.04	8.13	**5.29**	**94.85**
	SAP^2-PC	5slides	7.93	8.13	6.24	93.88
	SAP^2-JPI	5slides	7.92	**7.70**	9.62	90.79
	SAP^2-TPI	5slides	**7.86**	8.17	**5.18**	**94.98**
L95 (473h)	SAP^2-PC	1slide	7.93	**7.94**	7.74	92.26
	SAP^2-JPI	1slide	8.07	8.08	7.85	92.31
	SAP^2-TPI	1slide	**7.92**	8.16	**4.71**	**95.47**
	SAP^2-PC	5slides	7.86	8.04	6.26	93.88
	SAP^2-JPI	5slides	7.98	8.05	7.34	92.80
	SAP^2-TPI	5slides	**7.71**	8.08	**4.54**	**95.59**

Table 3: Performance comparison of SAP^2-PC, SAP^2-JPI and SAP^2-TPI evaluated on SlideSpeech test dataset, trained on S95/L95 datasets.

Model	Contextual Keywords	F1-score
SAP^2-JPI	1slide	20.45
SAP^2-TPI	1slide	93.85
SAP^2-JPI	5slides	21.19
SAP^2-TPI	5slides	94.48

Table 4: F1-score of filtered contextual keywords of SAP^2-JPI and SAP^2-TPI on SlideSpeech test dataset, trained on L95.

2.2% (S95) and 2.7% (L95) under extended context conditions. This indicates that SAP^2-TPI possesses superior contextual robustness, capable of effectively integrating long-form contextual information without performance saturation.

Ablation Study on Speech-driven Attention-based Pooling To clarify the effectiveness of Speech-driven attention-based pooling, we compare Qwen2-Audio models fine-tuned with the pooling strategy with those without. The experiments are conducted on SlideSpeech L95 dataset with 1-slide and 5-slide contextual keywords, evaluated on test set only. Table 5 illustrates that performances of PC and TPI improve when applied with Speech-driven attention-based pooling, while JPI deteriorates. The performance degradation of JPI may be because the pooling of text embeddings increases the ambiguity of context representations and intensifies the complexity of dual objectives. What's more, when larger amounts of contextual keywords are provided, the improvement brought by Speech-driven attention-based pooling becomes more obvious.

What's more, we analyze the training and inference time of Qwen2-Audio-PC, Qwen2-Audio-TPI and SAP^2-TPI. We train three models on SlideSpeech L95 5-slide dataset using 7 40G A100 GPUs and infer on 15-slide dataset using 1 40G A100 GPU in Table 6. The training and inference time of SAP^2-TPI is close to those of Qwen2-Audio-PC, indicating that two-stage pipeline doesn't introduce much complexity. For Qwen2-Audio-TPI and SAP^2-TPI, the decrease in training time and inference time of Stage II is not obvious, as the number of contextual keywords is mostly less than 10 after context pruning. However, training time for context pruning significantly decreases by 24.97%, inference time also decreases by 20.20%. Our ablation study indicates that Speech-driven attention-based pooling plays an important role when there are large amounts of contextual keywords.

3.6 Analysis on Pooling Window Size

To analyze model performance as the compression rate increases, we train SAP^2-PC, SAP^2-JPI and SAP^2-TPI with pooling window sizes of 2, 4 and 8 on L95 (5 slides) dataset. Figure 4 demonstrates that WER slightly increases when window size increases. However, SAP^2-TPI constantly outperforms the other two methods, while SAP^2-PC is most robust to pooling window size changes.

Train	Contextual Keywords	Model	Test			
			WER	U-WER	B-WER	Recall ↑
L95 (473h)	1slide	*Qwen2-Audio*				
		+PC	8.22	8.35	6.34	93.82
		+JPI	7.97	8.13	5.81	94.39
		+TPI	**7.92**	8.19	**4.12**	**96.00**
		SAP²				
		+PC	7.93	**7.94**	7.74	92.26
		+JPI	8.07	8.05	7.34	92.80
		+TPI	**7.92**	8.16	4.71	95.47
	5slides	*Qwen2-Audio*				
		+PC	7.94	8.14	6.18	93.94
		+JPI	7.99	8.19	5.26	94.94
		+TPI	7.90	8.16	**4.35**	**95.81**
		SAP²				
		+PC	7.86	**8.04**	6.26	93.88
		+JPI	7.98	8.05	7.34	92.80
		+TPI	**7.71**	8.08	4.54	95.59

Table 5: Ablation study on effectiveness of Speech-driven attention-based pooling.

Method	Training time (s)	Inference time (s) (15 slides)
Qwen2-Audio-PC	60134.5	6054
Qwen2-Audio-TPI (Stage I)	51003.8	1960
Qwen2-Audio-TPI (Stage II)	30120.3	5047
SAP^2-TPI (Stage I)	38267.9	1564
SAP^2-TPI (Stage II)	29608.8	5076

Table 6: Comparison of TPI methods' training time of each stage with or without Speech-driven attention-based pooling. Qwen2-Audio-TPI and SAP^2-TPI are trained on L95 dataset (5 slides) with 7 40G A100 GPUs. ASR results are inferred on SlideSpeech 15-slide test set using 1 40G A100 GPU.

3.7 Analysis of Inference Generalization Across Variable Context Lengths

In Section 3.3, we analyzed the robustness of SAP^2-TPI when trained and tested with varying numbers of slides. Since retraining the model for each context length is impractical, we specifically investigate whether SAP^2-TPI exhibits stronger robustness than Qwen2-Audio-TPI on the challenging SlideSpeech dataset.

Both models (trained on L95 5-slide dataset) are evaluated on test sets with 1, 3, 5, 7, 9, 15, and 25 slides (Please refer to the Appendix for statistics of test sets). As shown in Figure 5, SAP^2-TPI consistently outperforms the baseline when contextual slides exceed 3. Increasing slides from 5 to 15 expands the mean/median token lengths from 402.81/332 to 1024.88/964. The Keyword Coverage Rate rises from 6.51% to 8.50%, while the Information Rate drops from 0.64% to 0.33%. Despite these changes, WER slightly improves from 7.71% to 7.67%, demonstrating robust adaptation to longer contexts.

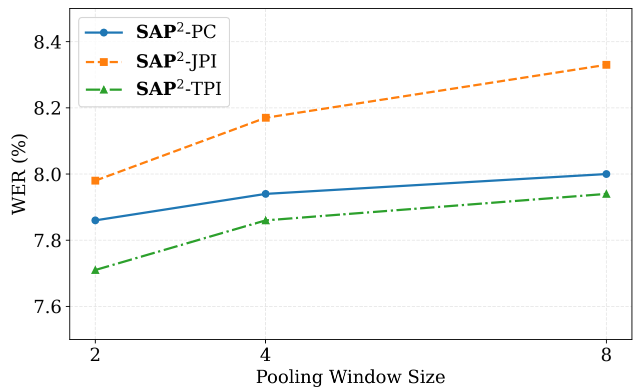

Figure 4: Analysis of the impact of pooling window sizes on WER. Experiments are conducted on L95 dataset (5 slides).

These results underscore SAP^2-TPI's superior generalization capability, with performance gains becoming increasingly significant as context length grows.

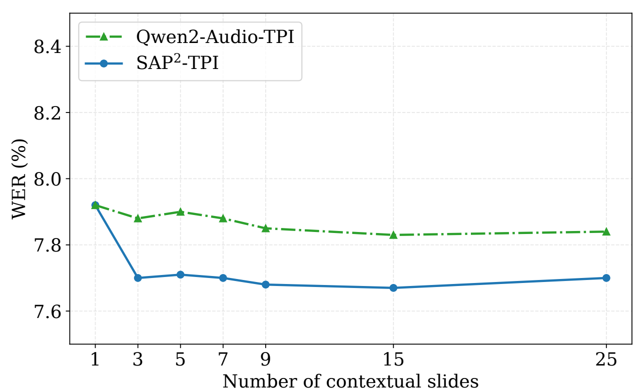

Figure 5: WER across context lengths (1, 3, 5, 7, 9, 15, 25 slides) for SAP^2-TPI and Qwen2-Audio-TPI, both trained on 5-slide segments (L95 dataset)

4 Conclusion

In this paper, we tend to address two limitations of contextualized ASR systems: (1) constrained context window capacity for document-level long-form processing, and (2) the performance degradation caused by noisy long context information. We propose SAP^2 that synergizes dynamic keyword pruning with speech-adaptive attention pooling for contextual embedding compression. Extensive experiments on SlideSpeech and LibriSpeech verify SAP^2's superiority and stable performance under various context settings. Because of the limitations of computational resources and data, we only use contexts splitted as single words, future directions will explore semantically meaningful phrasal contexts and multimodal fusion with visual embeddings. We also hope this study motivates future investigations into context-aware speech recognition paradigms.

Acknowledgements

This work is supported by the National Key R&D Program of China (No.2022ZD0116405).

References

Bai, Y.; Chen, J.; Chen, J.; Chen, W.; Chen, Z.; Ding, C.; Dong, L.; Dong, Q.; Du, Y.; Gao, K.; et al. 2024. Seed-asr: Understanding diverse speech and contexts with llm-based speech recognition. *arXiv preprint arXiv:2407.04675*.

Chang, F.-J.; Liu, J.; Radfar, M.; Mouchtaris, A.; Omologo, M.; Rastrow, A.; and Kunzmann, S. 2021. Context-aware transformer transducer for speech recognition. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 503–510. IEEE.

Chu, Y.; Xu, J.; Yang, Q.; Wei, H.; Wei, X.; Guo, Z.; Leng, Y.; Lv, Y.; He, J.; Lin, J.; et al. 2024. Qwen2-audio technical report. *arXiv preprint arXiv:2407.10759*.

Gao, Z.; Zhang, S.; McLoughlin, I.; and Yan, Z. 2022. Paraformer: Fast and accurate parallel transformer for non-autoregressive end-to-end speech recognition. *arXiv preprint arXiv:2206.08317*.

Gong, X.; Lv, A.; Wang, Z.; and Qian, Y. 2024. Contextual Biasing Speech Recognition in Speech-enhanced Large Language Model. In *Interspeech 2024*, 257–261.

Han, M.; Dong, L.; Liang, Z.; Cai, M.; Zhou, S.; Ma, Z.; and Xu, B. 2022. Improving end-to-end contextual speech recognition with fine-grained contextual knowledge selection. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 8532–8536. IEEE.

Han, M.; Dong, L.; Zhou, S.; and Xu, B. 2021. Cif-based collaborative decoding for end-to-end contextual speech recognition. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6528–6532. IEEE.

Huang, K.; Zhang, A.; Yang, Z.; Guo, P.; Mu, B.; Xu, T.; and Xie, L. 2023. Contextualized end-to-end speech recognition with contextual phrase prediction network. *arXiv preprint arXiv:2305.12493*.

Jayanthi, S. M.; Kulshreshtha, D.; Dingliwal, S.; Ronanki, S.; and Bodapati, S. 2023. Retrieve and Copy: Scaling ASR Personalization to Large Catalogs. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*, 631–639.

Lakomkin, E.; Wu, C.; Fathullah, Y.; Kalinli, O.; Seltzer, M. L.; and Fuegen, C. 2024. End-to-end speech recognition contextualization with large language models. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 12406–12410. IEEE.

Le, D.; Jain, M.; Keren, G.; Kim, S.; Shi, Y.; Mahadeokar, J.; Chan, J.; Shangguan, Y.; Fuegen, C.; Kalinli, O.; Saraf, Y.; and Seltzer, M. L. 2021. Contextualized Streaming End-to-End Speech Recognition with Trie-Based Deep Biasing and Shallow Fusion. *ArXiv*, abs/2104.02194.

Ni, Z.; Han, M.; Chen, F.; Meng, L.; Shi, J.; Lv, P.; and Xu, B. 2024. ViLaS: Exploring the Effects of Vision and Language Context in Automatic Speech Recognition. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 11366–11370. IEEE.

Panayotov, V.; Chen, G.; Povey, D.; and Khudanpur, S. 2015. Librispeech: An ASR corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5206–5210.

Pundak, G.; Sainath, T. N.; Prabhavalkar, R.; Kannan, A.; and Zhao, D. 2018. Deep context: end-to-end contextual speech recognition. In *2018 IEEE spoken language technology workshop (SLT)*, 418–425. IEEE.

Radford, A.; Kim, J. W.; Xu, T.; Brockman, G.; McLeavey, C.; and Sutskever, I. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, 28492–28518. PMLR.

Sohoglu, E.; Peelle, J. E.; Carlyon, R. P.; and Davis, M. H. 2012. Predictive top-down integration of prior knowledge during speech perception. *Journal of Neuroscience*, 32(25): 8443–8453.

Sudo, Y.; Shakeel, M.; Fukumoto, Y.; Peng, Y.; and Watanabe, S. 2024. Contextualized Automatic Speech Recognition With Attention-Based Bias Phrase Boosted Beam Search. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 10896–10900. IEEE.

Wang, H.; Kurita, S.; Shimizu, S.; and Kawahara, D. 2024a. SlideAVSR: A Dataset of Paper Explanation Videos for Audio-Visual Speech Recognition. *arXiv preprint arXiv:2401.09759*.

Wang, H.; Yu, F.; Shi, X.; Wang, Y.; Zhang, S.; and Li, M. 2024b. SlideSpeech: A Large Scale Slide-Enriched Audio-Visual Corpus. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 11076–11080. IEEE.

Wang, Z.; Rong, Y.; Jiang, D.; Wu, H.; Zhou, S.; and Xu, B. 2024c. CIEASR: Contextual Image-Enhanced Automatic Speech Recognition for Improved Homophone Discrimination. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 915–924.

Xu, K.-T.; Xie, F.-L.; Tang, X.; and Hu, Y. 2025. FireRedASR: Open-Source Industrial-Grade Mandarin Speech Recognition Models from Encoder-Decoder to LLM Integration. *arXiv preprint arXiv:2501.14350*.

Yang, G.; Ma, Z.; Gao, Z.; Zhang, S.; and Chen, X. 2024a. CTC-Assisted LLM-Based Contextual ASR. *Proc. SLT*.

Yang, G.; Ma, Z.; Yu, F.; Gao, Z.; Zhang, S.; and Chen, X. 2024b. MaLa-ASR: Multimedia-Assisted LLM-Based ASR. *arXiv preprint arXiv:2406.05839*.

Yu, F.; Wang, H.; Shi, X.; and Zhang, S. 2024. LCB-Net: Long-Context Biasing for Audio-Visual Speech Recognition. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 10621–10625. IEEE.