VISH-ASR: Viterbi-guided Selection of Hotwords for Speech-LLM Contextual Automatic Speech Recognition

Anonymous ACL submission

Abstract

Prompt-based Speech LLM ASR can incorporate contextual information by injecting hotwords into the prompt. Yet biasing performance is highly sensitive to candidate quality. Thus we propose VISH (Viterbi-guided Selection of Hotwords), a lightweight, training-free, plug-and-play pre-selection module that verifies and filters hotwords before prompt injection. VISH requires no modification to the Speech LLM and performs streaming, synchronous multi-word scoring on CTC posteriors, efficiently evaluating thousands of candidates via prefix sharing. Across multiple Speech-LLM ASR frameworks, hotword list sizes, in-domain and out-of-domain biasing sets, and noisy conditions, VISH-ASR consistently outperforms prior contextual biasing baselines. Ablations further show robustness to thresholding, and a bidirectional segment-level variant provides additional gains. Overall, VISH offers an efficient, model-agnostic solution for reliable hotword integration in Speech-LLM ASR.

1 Introduction

End-to-end automatic speech recognition (ASR) systems perform well on general-domain corpora (Li et al., 2022; Ji et al., 2024), but remain vulnerable to hotwords: long-tail yet semantically critical terms, such as contact names, locations, and product names, that are sparsely represented in training data. Since such terms vary across users and tasks, they are naturally provided as external lexical context (Shi et al., 2024; Hori et al., 2018; Chen et al., 2019a). Contextual ASR has therefore been explored through weighted finite-state transducers (WFSTs) (Bakhturina et al., 2022), neural contextual injection (Pundak et al., 2018; Jain et al., 2020; Variani et al., 2020; Zhao et al., 2019). Recently, Speech Large Language Models (LLMs) that combine speech encoders with pretrained language models (Song et al., 2023; Asano et al., 2025; Peng et al., 2026). However, robust hotword recognition with Speech LLMs remains challenging.

Speech-LLM-based contextual ASR generally follows two paradigms. The first modifies the model architecture or relies on fine-tuning to learn hotword-aware representations, achieving strong performance but requiring substantial data and computation and offering limited flexibility for dynamic contexts (Fu et al., 2025; Ren et al., 2026; Jing et al., 2025; Kong et al., 2025). The second adopts a training-free, plug-and-play strategy, directly injecting hotwords into the prompt without modifying the Speech LLM (Yang et al., 2023; Chen et al., 2024; Gong et al., 2025). While enabling flexible personalization and deployment, prompt-based injection is sensitive to candidate-list quality: large lists weaken conditioning and introduce interference that degrades both hotword and background-word recognition (Gong et al., 2025). Therefore, prompt-based systems commonly first retrieve a compact, high-confidence hotword subset for prompt construction (Yang et al., 2023).

However, most prompt-based contextual ASR pipelines construct prompts through text-only pre-selection (Yang et al., 2024; Asano et al., 2025), such as ranking candidates from 1-best or N-best hypotheses or matching decoded text against hotword entries (Chen et al., 2019b; Yang et al., 2024). Because these methods rely on discrete recognition outputs rather than verifying candidate-specific phonetic evidence in the audio, they are sensitive to ASR errors, acoustic variability, and noise. Under acoustic or domain shift, this lack of acoustic grounding can lead to false selections and missed hotwords, limiting robust deployment.

Acoustic-based alternatives remain insufficient

for prompt-based contextual ASR. Cross-modal retrieval methods (Gong et al., 2025; Hou et al., 2025; Li et al., 2023) rank hotwords using utterance-level audio-text embeddings, but provide only coarse semantic evidence rather than candidate-level phonetic verification. Search-graph-based approaches, such as on-the-fly WFST composition (Hori et al., 2007) and lattice rescoring (Wang et al., 2023), leverage acoustic scores but require internal decoding artifacts unavailable in black-box Speech LLMs. Therefore, existing acoustic approaches either lack fine-grained verification or cannot serve as portable plug-and-play pre-selectors.

To address these limitations, we propose Viterbi-guided Selection of Hotwords (VISH), a lightweight, training-free, and plug-and-play acoustic verifier for prompt-based contextual ASR. Its single-word variant, S-VISH, uses constrained Viterbi alignment on CTC frame posteriors to score whether a candidate hotword is acoustically supported. For large candidate inventories, M-VISH jointly scores candidates through prefix-tree computation sharing, enabling efficient pre-selection. We further introduce Bidirectional M-VISH (Bi-MVISH) to improve robustness by combining forward and backward acoustic evidence. Requiring neither modification nor fine-tuning of the underlying Speech LLM, VISH supplies acoustically verified hotwords for prompt construction.

The key contributions of this paper are as follows:

1. We analyze the relationship between selection recall and hotword recognition performance, and propose Viterbi-guided Selection of Hotwords (VISH), a training-free, plug-and-play pre-selector for prompt-based contextual ASR. Based on constrained CTC Viterbi alignment, VISH assigns frame-level acoustic confidence scores to bias words for reliable filtering.

2. VISH consistently improves hotword recognition while preserving general-word recognition across multiple baselines in both in-domain and out-of-domain scenarios, enabling effective contextual biasing without degrading overall transcription quality.

3. VISH remains robust under challenging acoustic conditions, consistently outperforming baselines in noisy environments by leveraging acoustic evidence for hotword selection rather than relying solely on text-based matching.

4. We further introduce bidirectional multi-word VISH (Bi-MVISH), a segment-level streaming variant that leverages limited future context to improve hotword selection and further enhance the robustness of VISH.

2 Methodology

2.1 Speech LLM-based Contextual ASR

Speech-LLM ASR systems typically adopt an encoder–adapter–LLM architecture, where speech representations are mapped into the LLM embedding space and decoded with optional textual prompts. In prompt-based contextual ASR, candidate hotwords are appended to the prompt for flexible personalization. However, large or inaccurate candidate lists can introduce interference and degrade recognition (Gong et al., 2025). Therefore, a pre-selection module is needed to filter a compact set of high-confidence hotwords before prompt construction.

2.2 A Plug-and-Play Pre-Selector for Speech-LLMs

VISH (Viterbi-guided Selection of Hotwords) operates as an external, training-free acoustic pre-selector before prompt construction. As illustrated in Figure 1, it takes a candidate bias list and frame-level CTC posteriors as input. The posteriors can be obtained from an accessible CTC encoder within the Speech LLM or from a frozen external CTC model applied to the same audio. VISH scores the candidate hotwords and selects a compact subset for prompt construction, requiring neither fine-tuning nor access to the Speech-LLM decoder.

2.3 Viterbi-guided Selection of Hotword

2.3.1 Single-Word VISH

Single-word Viterbi-guided Selection of Hotwords (S-VISH) estimates acoustic support for each candidate hotword using CTC Viterbi scoring (Xi et al., 2025b, 2024, 2025c,a). Given an utterance $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_T)$, a CTC acoustic model outputs frame-level posteriors

$$p_t(k) = P(\pi_t = k \mid \mathbf{x}), \qquad k \in \Sigma \cup \{\varnothing\}, \quad (1)$$

where Σ is the character vocabulary, $\varnothing$ is the blank symbol, and $\boldsymbol{\pi}$ is a CTC path. For a candidate word w represented by $\mathbf{y} = (y_1, \ldots, y_L)$, S-VISH computes

$$V(\mathbf{y} \mid \mathbf{x}) = \max_{\boldsymbol{\pi}:\, B(\boldsymbol{\pi})=\mathbf{y}} \sum_{t=1}^{T} \log p_t(\pi_t), \qquad (2)$$

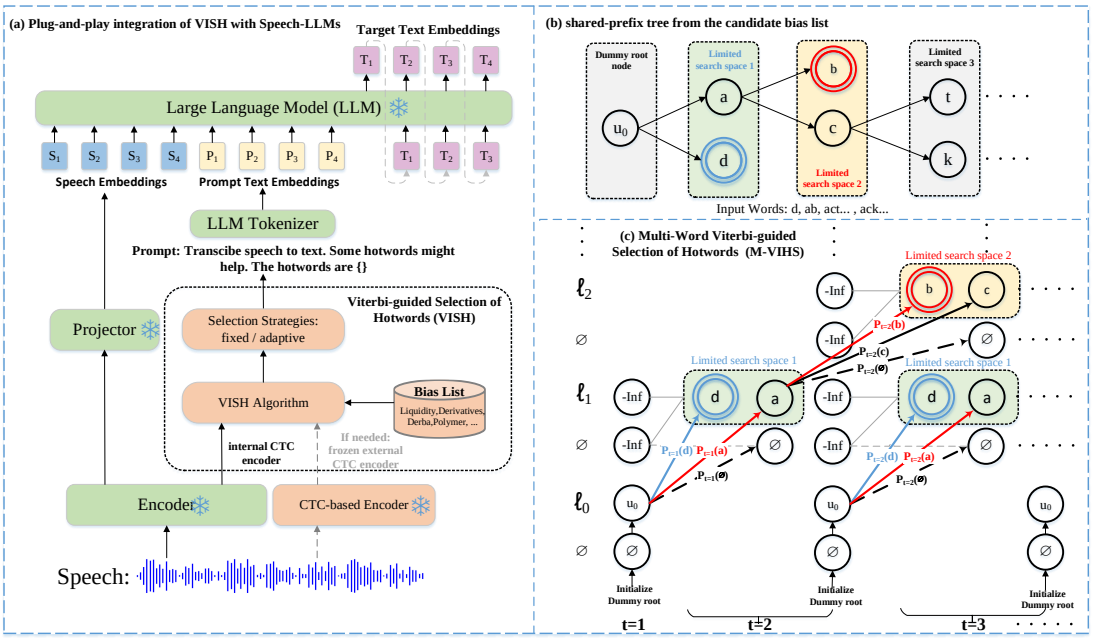

Figure 1: Illustration of VISH integration and search procedure. (a) Plug-and-play integration of VISH with Speech-LLMs. (b) Construction of the shared-prefix tree from the candidate bias list. (c) M-VISH inference procedure: At each time step t, perform Viterbi alignment state transitions within a limited search space and record the scores of the terminal nodes. $p_t(k)$ means a CTC acoustic model outputs frame-level posteriors, mentioned in Equation 1. ℓ denotes the deepth in the prefix tree, defined in Equation 7.

where $B(\cdot)$ is the standard CTC collapsing operator. This score is obtained by Viterbi dynamic programming on the extended target sequence

$$\mathbf{y}' = (\varnothing, y_1, \varnothing, \ldots, \varnothing, y_L, \varnothing), \qquad S = 2L + 1. \quad (3)$$

Let $\alpha_t(s)$ denote the best log-score ending at state s of $\mathbf{y}'$ at frame t. For $t \geq 2$,

$$\alpha_t(s) = \log p_t(y'_s) + \max\{\alpha_{t-1}(s),\ \alpha_{t-1}(s-1),\ \alpha_{t-1}(s-2)\}, \quad (4)$$

where the $s-2$ transition is allowed only if $y'_{s-2} \neq \varnothing$; out-of-range states are set to $-\infty$. The best score for w is

$$m_w = \max\{\alpha_T(S-1),\ \alpha_T(S)\}. \quad (5)$$

Let N_w be the number of non-blank frames in the corresponding best path. The length-normalized confidence score is

$$score_w = \frac{m_w}{N_w}. \quad (6)$$

2.3.2 Multi-Word VISH

Independently scoring each candidate with S-VISH incurs a cost that grows linearly with the bias-list size. To efficiently handle large candidate sets, M-VISH organizes the bias list $W = \{w^{(1)}, \ldots, w^{(N)}\}$ into a prefix tree, where each node u represents a character prefix of depth $\ell(u)$. Computation is shared among candidates with common prefixes. For each node u, the corresponding CTC extended target sequence is defined as

$$\mathbf{y}'_u = (\varnothing, y_1, \varnothing, y_2, \ldots, y_{\ell(u)}, \varnothing), \qquad S_u = 2\ell(u) + 1. \quad (7)$$

For each active node u, M-VISH maintains a Viterbi score vector $\alpha_t^u(s)$ on $\mathbf{y}'_u$ and a length accumulator $\gamma_t^u(s)$ for the corresponding best path. The scores are updated using Equation (4). The acoustic support of node u at frame t is defined as

$$\pi_t(u) = \max_{1 \leq s \leq S_u} \alpha_t^u(s). \quad (8)$$

At each frame, M-VISH retains only the top-K nodes ranked by $\pi_t(u)$ ($K = 40$ in our experiments) and expands their children by extending the parent DP states. For a word-ending node corresponding to w, the best observed $\pi_t(u)$ is recorded as m_w, together with the alignment length N_w from $\gamma_t^u(s)$. The final length-normalized confidence score is computed using Equation (6), as detailed in Algorithm 1.

M-VISH incorporates two decoding designs. First, a dummy root node prevents spurious ac-

Algorithm 1: Multi-Word Viterbi-guided Selection of Hotwords (M-VISH)

Input: CTC posteriors $\{p_t\}_{t=1}^{T}$; bias list W; path budget K.
Output: Length-normalized scores $\{s_w \mid w \in W\}$.

Build a trie over W with a dummy root node (no character); each edge corresponds to a character.
Initialize per-word best log-scores $m_w \leftarrow -\infty$ and alignment lengths $N_w \leftarrow 0$ for all $w \in W$.
Initialize active node list P with the children of the root.
foreach $u \in P$ **do**
  Initialize DP states $\alpha_1^u(\cdot)$ and frame counts $\gamma_1^u(\cdot)$ according to standard CTC Viterbi initialization.
for $t \leftarrow 1$ **to** T **do**
  PSD skip frames.
  foreach $u \in P$ **do**
    Update $\alpha_t^u(s)$ and $\gamma_t^u(s)$ for all $s \in \{1, \ldots, S_u\}$ using the Viterbi recurrence on $\mathbf{y}'_u$ (Eq. (4) adapted per node).
    Compute $\pi_t(u) \leftarrow \max_{1 \le s \le S_u} \alpha_t^u(s)$ and let $s_u^* = \arg\max_s \alpha_t^u(s)$.
  Prune P to the top-K nodes ranked by $\pi_t(u)$.
  foreach $u \in P$ *that is a word-ending node corresponding to* w **do**
    if $\pi_t(u) > m_w$ **then**
      $m_w \leftarrow \pi_t(u)$
      $N_w \leftarrow \gamma_t^u(s_u^*)$
  foreach $u \in P$ *with children* **do**
    For each child v of u, initialize or extend its DP states $\alpha_t^v(\cdot)$ and $\gamma_t^v(\cdot)$ from the final states of u.
  Initialize and add active node into list P with the children of the root.
foreach $w \in W$ **do**
  $score_w \leftarrow m_w / N_w$
return $\{score_w\}$

tivations near the trie root under aggressive pruning, with DP states initialized only from its children. Second, Peak Suppression Decoding (PSD) skips frames whose CTC blank posterior exceeds 0.997 and excludes them from path expansion, reducing unnecessary computation and suppressing low-confidence prefix propagation.

Overall, M-VISH reuses Viterbi computation across shared prefixes and provides efficient, training-free acoustic pre-selection. While S-VISH requires $O(T \sum_{w \in W} L_w) = O(TN\bar{L})$ time to score a bias list of N words with average length $\bar{L}$, M-VISH reduces the complexity to $O(TK\bar{L})$ by restricting the number of active prefixes to a fixed budget K per frame. Since $K \ll N$ in practice, this reduction yields substantial computational savings while preserving selection accuracy.

2.3.3 Bidirectional M-VISH

Bidirectional Multi-Word VISH (Bi-MVISH) improves selection robustness by applying M-VISH to both the original and time-reversed CTC posterior sequences of each audio segment. The backward pass can exploit stronger acoustic cues near word endings, mitigating score underestimation when a hotword begins with weak or ambiguous frames. For each candidate, Bi-MVISH takes the maximum of the forward and backward length-normalized scores as its final confidence score. Although requiring future frames within each segment, Bi-MVISH remains compatible with segment-level streaming inference and introduces no additional training.

3 Experimental Setups

3.1 Dataset

- **LibriSpeech** (Panayotov et al., 2015) is a 960-hour English audiobook ASR corpus. We use the standard *test-clean* and *test-other* subsets for in-domain evaluation.
- **SPGISpeech** (O'Neill et al., 2021) is an earnings-call ASR corpus containing over 5,000 hours of speech with rich financial terminology. We use it to evaluate out-of-domain hotword biasing.
- **WHAM!** (Wichern et al., 2019) provides real-world ambient noise. We construct *test-clean-noisy* and *test-other-noisy* by mixing LibriSpeech test utterances with randomly sampled WHAM! noise at SNRs uniformly distributed from 0 to 20 dB.

Following (Le et al., 2021), we treat rare words as hotwords by excluding the 5,000 most frequent words from the training or reference text. Rare words occurring in a test utterance are ground-truth hotwords, while distractors are sampled from the remaining rare-word vocabulary. For LibriSpeech and SPGISpeech, we construct candidate lists at scales $N \in \{100, 500, 1000, 2000\}$ by combining ground-truth hotwords with sampled distractors.

3.2 Evaluation

3.2.1 Evaluation Metrics

- We evaluate hotword pre-selection using recall:

$$\text{Recall} = \frac{|W_{\text{gt}} \cap W_{\text{pred}}|}{|W_{\text{gt}}|}, \tag{9}$$

where W_{gt} is the set of ground-truth hotwords in the utterance and W_{pred} is the selected set for prompting.

- Following contextual ASR practice (Le et al., 2021), we report WER, biased word error rate (B-WER), and unbiased word error rate (U-WER). B-WER and U-WER are computed on words inside and outside the bias list, respectively.

3.2.2 Inference Protocol for Plug-and-Play Biasing

As described in Section 2.2, the plug-and-play VISH module operates independently of the Speech-LLM architecture and relies solely on frame-level CTC character posteriors. Specifically, we adopt two different Speech-LLM architectures, which are Prompt-SLAM-ASR (Yang et al., 2024) and Qwen3-ASR (Shi et al., 2026). For both models, the fine-tuned WavLM-Large encoder from Prompt-SLAM-ASR (Yang et al., 2024) is utilized as the CTC encoder of the pre-selector. After filtering, selected hotwords are injected into both systems using an identical prompt template: "Transcribe speech to text. Some hotwords might help. The hotwords are {}".

3.2.3 Selection Strategies

The final bias list W_{pred} is constructed from VISH scores using one of two strategies:

- **Adaptive Selection** selects up to five highest-scoring candidates satisfying $s_w > \tau$, where τ is determined on the development set by maximizing recall. This strategy is used when a matched development set is available.
- **Fixed Selection** selects the top-K candidates, with $K = 10$ in all experiments. This strategy avoids threshold tuning and is suitable when no matched development set is available.

4 Results and Analysis

4.1 Correlation Between Hotword Recall and B-WER

Figure 2 shows a clear negative correlation between hotword pre-selection recall and final B-WER. Across bias-list sizes and selection strategies, higher recall consistently corresponds to lower B-WER, suggesting that accurate hotword retrieval is closely associated with improved biased-word recognition. Since B-WER directly measures recognition accuracy on bias words and reflects the downstream effect of pre-selection, we use it as the primary metric in subsequent experiments.

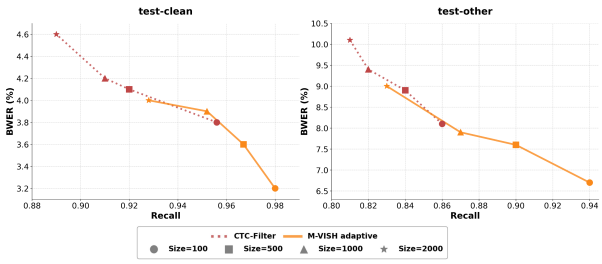

Figure 2: Recall and B-WER correlation of M-VISH adaptive and the CTC-Filter (Yang et al., 2024) pre-selection methods on LibriSpeech test-clean/other using Prompt-SLAM-ASR as the Speech-LLM. Threshold τ in Adaptive selection strategy is set to 0.74.

4.2 VISH Performance of In-domain Testsets

We first evaluate VISH on the in-domain LibriSpeech *test-clean* and *test-other* sets. As shown in Table 1, directly injecting the full bias list provides limited benefit at $N = 100$, but sharply increases WER and B-WER at $N = 500$ and fails to decode at larger scales. This confirms that pre-selection is necessary to control prompt interference as the candidate list grows.

Results in Table 1 show that VISH consistently reduces B-WER across all bias-list scales while maintaining overall transcription accuracy. Among the evaluated pre-selection methods, Bi-MVISH with Adaptive Selection achieves the strongest performance. On *test-other*, it reduces B-WER from 8.0% to 6.2% at $N = 100$, a 22.5% relative improvement over the strongest baseline. The advantage persists at larger scales, with B-WER reduced to 7.5% at $N = 1000$ and 8.6% at $N = 2000$, while overall WER also decreases. Similar trends on *test-clean* confirm that VISH improves hotword recognition without degrading general transcription quality.

4.3 Semantic Generalization: Out-of-Domain Biasing

Figure 3 evaluates VISH under domain shift on SPGISpeech, whose financial terminology differs substantially from LibriSpeech. VISH consistently improves biased-word recognition while maintaining overall transcription accuracy. Compared with the text-only CTC-Filter baseline, M-VISH and Bi-MVISH under Adaptive Selection reduce B-WER from 19.1% to 14.8% and 14.1% at $N = 100$, corresponding to relative improvements of 22.5% and

Table 1: Performance comparison on LibriSpeech **test-clean/test-other** using Prompt-SLAM-ASR as the Speech-LLM. “GT-Bias” injects ground-truth hotwords into the prompt (oracle upper bound). “w/o biasing terms” injects an empty bias field. The results are independent of the bias list size N, so only reported once. “w/o filtering strategies” injects the full bias list without pre-selection; “✗” denotes decoding failure due to the over-length context. “WFST Fusion Boosting” is a standalone WFST-based decoding system without an LLM, while “WFST-Filter” applies WFST as a first-stage filter within the Prompt-SLAM-ASR pipeline. “fixed” and “adaptive” denote Fixed Top-K ($K = 10$) and Adaptive Selection ($\tau = 0.74$), respectively. The best results are highlighted in **bold**, and the second-best results of B-WER are highlighted in underline.

Biasing Strategies	**GT-Bias**		**N=100**		**N=500**		**N=1000**		**N=2000**	
	WER	**B-WER**	**WER**	**B-WER**	**WER**	**B-WER**	**WER**	**B-WER**	**WER**	**B-WER**
WFST Fusion Boosting (Yao et al., 2021)	1.5/3.3	3.2/8.6	2.3/4.8	3.9/10.4	5.1/8.8	6.7/15.6	8.2/13.1	9.4/20.6	14.4/20.3	15.4/26.8
Prompt-SLAM-ASR (Yang et al., 2024)										
w/o biasing terms (base)	2.0 / 4.2				9.3 / 20.0					
w/o filtering strategies	1.2/2.4	2.8/5.8	1.8/3.7	5.8/11.2	29.8/28.8	31.9/35.4	✗	✗	✗	✗
+*WFST-Filter*	1.2/2.4	2.8/5.8	1.4/3.0	3.8/10.1	1.5/3.1	4.4/11.3	1.4/3.2	4.5/12.7	1.6/3.4	5.2/13.8
+*CTC-Filter (Yang et al., 2024)*	1.1/2.7	2.8/6.0	1.3/2.7	3.7/8.0	1.3/3.0	3.9/9.0	1.3/3.0	4.2/9.3	1.4/3.2	4.4/10.0
+*Bias Retrieval (Gong et al., 2025)*	1.1/2.7	2.8/6.0	1.3/2.7	3.8/8.3	1.3/3.0	4.0/8.7	1.3/3.1	4.1/9.0	1.4/3.1	4.2/9.3
+*S-VISH fixed*	1.2/2.4	2.8/5.8	1.3/2.7	3.5/7.0	1.3/3.0	4.0/8.4	1.4/3.0	4.2/8.6	1.4/3.2	4.4/9.2
+*S-VISH adaptive*			1.2/2.6	3.3/7.3	1.3/2.9	4.1/8.4	1.3/2.9	4.1/8.5	1.4/3.1	4.3/9.0
+*M-VISH fixed*			1.3/2.9	3.5/6.7	1.4/3.0	3.7/7.8	1.4/3.0	4.1/7.9	1.4/3.2	4.3/9.0
+*M-VISH adaptive*			1.2/2.6	3.2/6.7	1.3/2.8	3.6/7.6	1.3/2.9	3.9/7.9	1.4/3.1	4.0/9.0
+*Bi-MVISH adaptive*			**1.2/2.6**	**3.2/6.2**	**1.3/2.7**	**3.4/6.9**	**1.3/2.8**	**3.7/7.5**	**1.4/3.0**	**3.9/8.6**

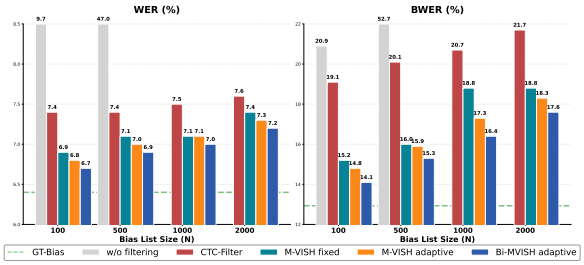

Figure 3: Performance comparison on the out-of-domain SPGISpeech test set using Prompt-SLAM-ASR as the Speech-LLM. “GT-Bias" injects ground-truth hotwords into the prompt (oracle upper bound). “fixed" and “adaptive" refer to Fixed Top-K (K=10) and Adaptive Selection (τ=0.72), respectively. When the bias-list is too large, w/o filtering method is unable to decode successfully.

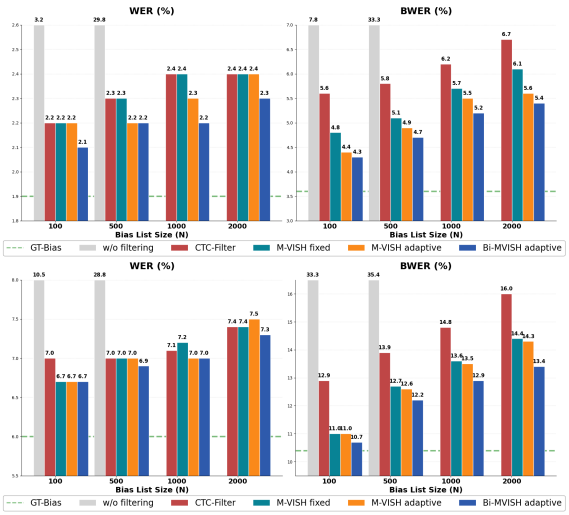

Figure 4: Performance comparison on the challenging acoustic conditions librispeech test-clean-noisy (above part) and test-other-noisy (below part) using Prompt-SLAM-ASR as the Speech-LLM. “fixed" and “adaptive" refer to Fixed Top-K (K=10) and Adaptive Selection (τ=0.72), respectively. When the bias-list is too large, w/o filtering method is unable to decode successfully.

26.2%, while lowering WER from 7.4% to 6.8% and 6.7%. At $N = 2000$, Bi-MVISH still reduces B-WER from 21.7% to 17.6% and WER from 7.6% to 7.2%. These results confirm that acoustically grounded pre-selection remains effective under domain shift and large candidate inventories.

4.4 Acoustic Robustness: Noisy Conditions

Figure 4 evaluates VISH under additive noise using LibriSpeech mixed with WHAM!. Across both noisy test sets and all bias-list scales, VISH consistently outperforms the CTC-Filter baseline in B-WER. For example, at $N = 1000$ on *test-other-noisy*, M-VISH and Bi-MVISH under Adaptive Selection reduce B-WER from 14.8% to 13.5% and 12.9%, respectively. Bi-MVISH provides larger gains than forward-only M-VISH, indicating the benefit of bidirectional scoring under noise. In particular, noise may weaken acoustic evidence at

word onsets and cause forward alignment to underestimate hotword confidence; backward alignment can complement these weak cues using stronger evidence near word endings. Section 4.6.3 provides further analysis.

4.5 Plug-and-Play Generalization to Qwen3-ASR

Table 2 evaluates VISH with Qwen3-ASR 0.6B and 1.7B. Since Qwen3-ASR has no integrated CTC encoder, we use the same frozen external CTC encoder to provide frame-level posteriors for both CTC-Filter and VISH. Without fine-tuning Qwen3-ASR, VISH consistently reduces B-WER while maintaining or improving overall WER across candidate-list scales. On *test-other* at $N = 2000$, Bi-MVISH reduces B-WER from 13.2% to 10.2% for the 0.6B model and from 11.0% to 9.1% for the 1.7B model, corresponding to relative improvements of 22.7% and 17.3%, respectively. These results demonstrate that VISH generalizes as a plug-and-play pre-selector across Speech-LLM architectures.

4.6 Ablation Study

4.6.1 Robustness of selecting threshold τ

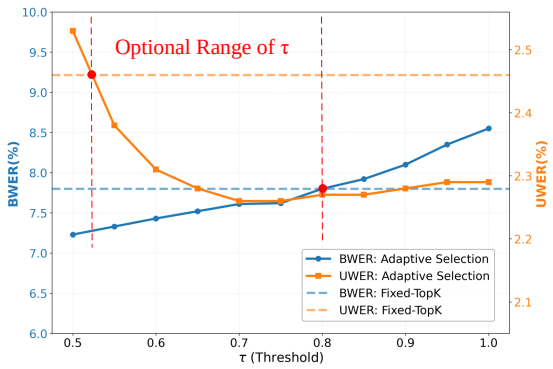

Figure 5: B-WER and U-WER under different τ values of test-other dataset, at the number of biasing terms $N = 500$. The optional range of τ (bounded by the red dashed lines) corresponds to the region where the adaptive selection method achieves both lower B-WER and lower U-WER than the fixed top-$K(K = 10)$ strategy.

We evaluate the sensitivity of Adaptive Selection to τ at $N = 500$. Figure 5 reports B-WER and U-WER on *test-other* under different thresholds, where the optional range denotes values for which Adaptive Selection outperforms Fixed Selection in both metrics.

Figure 6 compares the optional ranges across *test-clean*, *test-other*, and SPGISpeech. The ranges are relatively wide and substantially overlap across datasets, indicating that Adaptive Selection is not highly sensitive to τ. This overlap further suggests that a threshold tuned on one development condition can remain effective when a matched development set is unavailable.

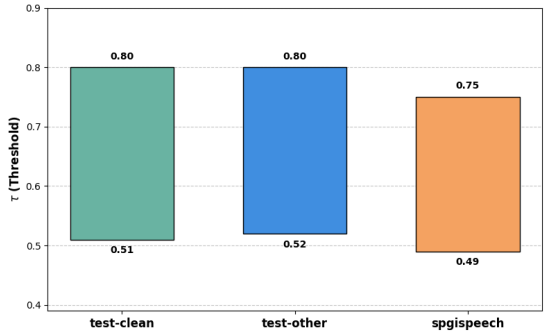

Figure 6: The resulting optional τ ranges estimated for different testing datasets, at the number of biasing terms $N = 500$.

4.6.2 Comparison of Selection Strategies under Fixed Top-K

We compare the recall of M-VISH and CTC-Filter at $N = 1000$ using the same Fixed Top-K selection rule, with $K \in \{3, 6, 9, 12\}$. As shown in Table 3, M-VISH consistently achieves higher recall across all datasets and values of K. This result indicates that VISH provides more accurate hotword ranking, contributing to its downstream recognition gains.

Table 3: Hotword recall (%) under different fixed top-$K(K = 10)$ selection strategies across different datasets, with a bias candidate list size of $N = 1000$.

Dataset	Selection Strategy	Recall with Fixed Top-K (%)			
		3	6	9	12
test-clean	CTC-Filter	61.1	86.1	91.0	93.4
	M-VISH	**70.1**	**89.5**	**95.2**	**98.4**
test-other	CTC-Filter	67.8	78.6	82.5	87.5
	M-VISH	**74.6**	**82.1**	**86.8**	**94.3**
spgispeech	CTC-Filter	42.2	79.5	80.2	80.5
	M-VISH	**75.6**	**82.4**	**92.4**	**94.0**

4.6.3 Bi-MVISH for boosted hotword verification

Although VISH relies on frame-level CTC evidence, forward Viterbi alignment may underestimate hotwords with weak or ambiguous onsets

Table 2: Performance comparison on LibriSpeech test-clean/other using Qwen3-ASR as the Speech-LLM. “GT-Bias” injects ground-truth hotwords into the prompt (oracle upper bound). “w/o filtering strategies” injects the full bias list without pre-selection. “fixed” and “adaptive” denote Fixed Top-K ($K = 10$) and Adaptive Selection ($\tau = 0.74$). The best results are highlighted in **bold**, and the second-best results of B-WER are highlighted in underline.

Biasing Strategies	GT-Bias		N=100		N=500		N=1000		N=2000	
	WER	B-WER	WER	B-WER	WER	B-WER	WER	B-WER	WER	B-WER
Qwen3-ASR 0.6B										
w/o filtering strategies	1.5/3.2	4.5/8.2	2.0/4.4	8.1/16.6	2.2/4.6	9.3/19.1	2.2/4.7	9.6/19.4	2.3/4.7	9.7/19.7
+*CTC-Filter (Yang et al., 2024)*			1.8/3.7	6.4/12.6	1.8/3.8	6.6/12.9	1.8/3.8	6.7/13.3	1.8/3.8	6.8/13.2
+*M-VISH fixed*			1.7/3.4	5.2/9.2	1.6/3.4	5.4/9.9	1.7/3.5	5.6/10.4	1.8/3.5	5.9/10.7
+*M-VISH adaptive*			1.6/3.3	5.1/9.2	1.6/3.4	5.1/9.8	1.7/3.5	5.4/10.2	1.7/3.5	5.6/10.6
+*Bi-MVISH adaptive*			1.5/3.3	**5.0/8.9**	1.5/3.3	**5.0/9.3**	1.6/3.4	**5.3/9.8**	1.6/3.5	**5.4/10.2**
Qwen3-ASR 1.7B										
w/o filtering strategies	1.3/2.5	4.0/8.0	1.5/3.1	5.3/12.4	1.6/3.4	6.5/14.6	1.7/3.5	6.8/15.4	1.7/3.6	7.1/15.9
+*CTC-Filter (Yang et al., 2024)*			1.4/2.8	5.2/10.1	1.4/2.8	5.2/10.4	1.4/2.8	5.2/10.7	1.5/2.9	5.4/11.0
+*M-VISH fixed*			1.3/2.6	4.3/8.3	1.3/2.7	4.3/8.7	1.4/2.6	4.8/9.0	1.4/2.8	4.8/9.3
+*M-VISH adaptive*			1.3/2.6	4.2/8.5	1.3/2.7	4.4/9.0	1.4/2.7	4.6/9.2	1.4/2.7	4.7/9.3
+*Bi-MVISH adaptive*			1.3/2.5	**4.0/8.0**	1.3/2.6	**4.1/8.3**	1.4/2.6	**4.3/8.8**	1.4/2.7	**4.5/9.1**

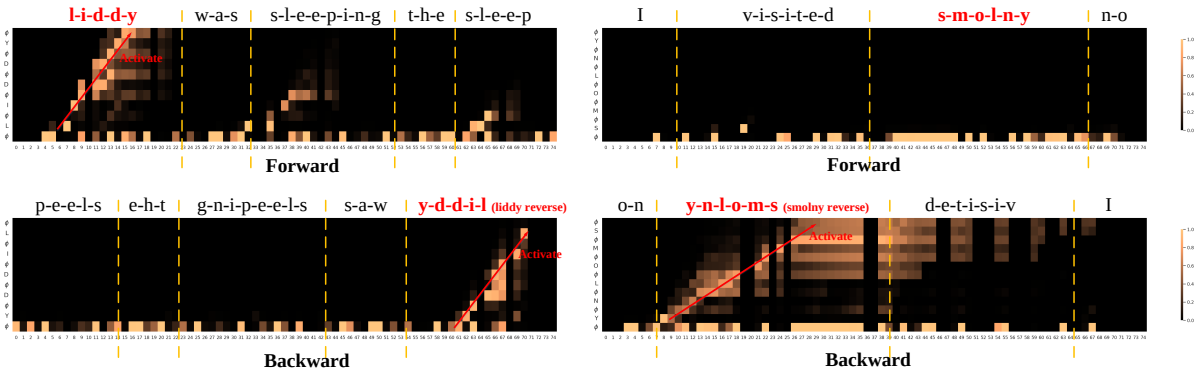

Figure 7: Bi-MVISH alignment for “liddy” (left) and “smolny” (right). Top: forward pass; bottom: backward pass. While both directions detect “liddy”, only backward alignment activates for “smolny” due to its weak onset. This highlights how reverse-time decoding improves robustness by leveraging strong endings. Vertical yellow dotted line denotes the word boundaries.

caused by coarticulation, noise, or channel distortion. Figure 7 illustrates this issue: both forward and backward passes strongly activate for “liddy”, whereas for “smolny”, the forward pass fails near the weak onset while the backward pass detects the word using clearer evidence near its ending. These examples show how Bi-MVISH complements forward alignment and improves hotword verification under challenging acoustic conditions.

5 Conclusion

We proposed VISH, a lightweight, training-free, and plug-and-play pre-selection module for prompt-based contextual Speech-LLM ASR. By leveraging frame-level CTC evidence to verify candidate hotwords, the proposed approach provides acoustically grounded hotword selection without requiring any modification or fine-tuning of the underlying Speech LLM. We execute sevetal extensive experiments including multiple Speech-LLM ASR frameworks, bias-list sizes up to 2000, in-domain and out-of-domain datasets, and noisy conditions. These experiments demonstrate that VISH consistently improves biased-word recognition while preserving general-word accuracy. The proposed approach also shows strong robustness to threshold selection and maintains stable performance across different evaluation settings. These results highlight the effectiveness and practical applicability of acoustically grounded hotword verification for scalable contextual ASR.

Limitations

VISH relies on frame-level CTC posteriors from an accessible or external acoustic model, and its effectiveness may depend on the quality and domain coverage of this model. Our experiments focus on English contextual ASR with predefined candidate hotword lists; performance on other languages, accents, domains, and dynamically changing user vocabularies remains to be evaluated. In addition, inaccurate or privacy-sensitive hotword lists may introduce inappropriate contextual biasing in deployment. Future work will investigate broader multilingual and real-world settings, as well as safeguards for user-provided contextual terms.

Ethical Considerations

This work uses existing English ASR datasets and pretrained speech models for research evaluation only. We do not collect new human-subject data or recruit human annotators. In real-world contextual ASR applications, user-provided hotword lists may contain personal or sensitive information, such as names, contacts, locations, or domain-specific entities. Moreover, inaccurate or malicious candidate lists may introduce inappropriate bias into transcriptions. Therefore, deployments of VISH-ASR should protect contextual vocabulary data, validate user-provided hotwords, and avoid treating biased ASR outputs as ground truth in high-stakes settings such as medical, legal, or financial transcription.

References

Yuya Asano, Sabit Hassan, Paras Sharma, Anthony B Sicilia, Katherine Atwell, Diane Litman, and Malihe Alikhani. 2025. Contextual asr error handling with llms augmentation for goal-oriented conversational ai. In *Proceedings of the 31st International Conference on Computational Linguistics: Industry Track*, pages 374–386.

Evelina Bakhturina, Yang Zhang, and Boris Ginsburg. 2022. Shallow fusion of weighted finite-state transducer and language model for text normalization. In *Proc. ISCA Interspeech*, pages 491–495.

Zhehuai Chen, He Huang, Andrei Andrusenko, Oleksii Hrinchuk, Krishna C Puvvada, Jason Li, Subhankar Ghosh, Jagadeesh Balam, and Boris Ginsburg. 2024. Salm: Speech-augmented language model with in-context learning for speech recognition and translation. In *Proc. IEEE ICASSP*, pages 13521–13525. IEEE.

Zhehuai Chen, Mahaveer Jain, Yongqiang Wang, Michael L Seltzer, and Christian Fuegen. 2019a. End-to-end contextual speech recognition using class language models and a token passing decoder. In *Proc. IEEE ICASSP*, pages 6186–6190. IEEE.

Zhehuai Chen, Mahaveer Jain, Yongqiang Wang, Michael L Seltzer, and Christian Fuegen. 2019b. Joint grapheme and phoneme embeddings for contextual end-to-end asr. In *Interspeech*, pages 3490–3494.

Li Fu, Yu Xin, Sunlu Zeng, Lu Fan, Youzheng Wu, and Xiaodong He. 2025. PAC: Pronunciation-aware contextualized large language model-based automatic speech recognition. *arXiv preprint arXiv:2509.12647*.

Xun Gong, Anqi Lv, Zhiming Wang, Huijia Zhu, and Yanmin Qian. 2025. Br-asr: Efficient and scalable bias retrieval framework for contextual biasing asr in speech llm. *arXiv preprint arXiv:2505.19179*.

Takaaki Hori, Jaejin Cho, and Shinji Watanabe. 2018. End-to-end speech recognition with word-based rnn language models. In *Proc. IEEE SLT*, pages 389–396. IEEE.

Takaaki Hori, Chiori Hori, Yasuhiro Minami, and Atsushi Nakamura. 2007. Efficient wfst-based one-pass decoding with on-the-fly hypothesis rescoring in extremely large vocabulary continuous speech recognition. *IEEE Transactions on audio, speech, and language processing*, 15(4):1352–1365.

Haoxiang Hou, Xun Gong, Wangyou Zhang, Wei Wang, and Yanmin Qian. 2025. Ranking and selection of bias words for contextual bias speech recognition. In *Proc. Interspeech 2025*, pages 5183–5187.

Mahaveer Jain, Gil Keren, Jay Mahadeokar, Geoffrey Zweig, Florian Metze, and Yatharth Saraf. 2020. Contextual RNN-T for open domain ASR. In *Proc. ISCA Interspeech*, pages 11–15.

Shengpeng Ji, Yifu Chen, Minghui Fang, Jialong Zuo, Jingyu Lu, Hanting Wang, Ziyue Jiang, Long Zhou, Shujie Liu, Xize Cheng, and 1 others. 2024. Wavchat: A survey of spoken dialogue models. *arXiv preprint arXiv:2411.13577*.

Ruihao Jing, Cheng Gong, Yu Jiang, Boyu Zhu, Shansong Liu, Chi Zhang, Xiao-Lei Zhang, and Xuelong Li. 2025. Rare word recognition and translation without fine-tuning via task vector in speech models. *arXiv preprint arXiv:2512.21894*.

YuXiang Kong, JunFeng Hou, Jian Tang, Bingqing Zhu, Jicheng Zhang, and Shaofei Xue. 2025. Contextual biasing for llm-based asr with hotword retrieval and reinforcement learning. *arXiv preprint arXiv:2512.21828*.

Duc Le, Mahaveer Jain, Gil Keren, Suyoun Kim, Yangyang Shi, Jay Mahadeokar, Julian Chan, Yuan Shangguan, Christian Fuegen, Ozlem Kalinli, and 1 others. 2021. Contextualized streaming end-to-end speech recognition with trie-based deep biasing and shallow fusion. *arXiv preprint arXiv:2104.02194*.

Jinyu Li and 1 others. 2022. Recent advances in end-to-end automatic speech recognition. *APSIPA Transactions on Signal and Information Processing*, 11(1).

Yuang Li, Yu Wu, Jinyu Li, and Shujie Liu. 2023. Prompting large language models for zero-shot domain adaptation in speech recognition. In *Proc. IEEE ASRU*, pages 1–8. IEEE.

Patrick K O'Neill, Vitaly Lavrukhin, Somshubra Majumdar, Vahid Noroozi, Yuekai Zhang, Oleksii Kuchaiev, Jagadeesh Balam, Yuliya Dovzhenko, Keenan Freyberg, Michael D Shulman, and 1 others. 2021. Spgispeech: 5,000 hours of transcribed financial audio for fully formatted end-to-end speech recognition. *arXiv preprint arXiv:2104.02014*.

Vassil Panayotov and 1 others. 2015. Librispeech: an asr corpus based on public domain audio books. In *Proc. IEEE ICASSP*, pages 5206–5210.

Jing Peng, Yucheng Wang, Bohan Li, Yiwei Guo, Hankun Wang, YanGui Fang, Yu Xi, Haoyu Li, Xu Li, Ke Zhang, Shuai Wang, and Kai Yu. 2026. A survey on speech large language models for understanding. *IEEE Journal of Selected Topics in Signal Processing*, 20(1):2–31.

Golan Pundak, Tara N Sainath, Rohit Prabhavalkar, Anjuli Kannan, and Ding Zhao. 2018. Deep context: end-to-end contextual speech recognition. In *Proc. IEEE SLT*, pages 418–425. IEEE.

Bo Ren, Ruchao Fan, Yelong Shen, Weizhu Chen, and Jinyu Li. 2026. Rlbr: Reinforcement learning with biasing rewards for contextual speech large language models. *arXiv preprint arXiv:2601.13409*.

Xian Shi, Xiong Wang, Zhifang Guo, Yongqi Wang, Pei Zhang, Xinyu Zhang, Zishan Guo, Hongkun Hao, Yu Xi, Baosong Yang, and 1 others. 2026. Qwen3-asr technical report. *arXiv preprint arXiv:2601.21337*.

Xian Shi, Yexin Yang, Zerui Li, Yanni Chen, Zhifu Gao, and Shiliang Zhang. 2024. Seaco-paraformer: A non-autoregressive asr system with flexible and effective hotword customization ability. In *Proc. IEEE ICASSP*, pages 10346–10350. IEEE.

Gan Song, Zelin Wu, Golan Pundak, Angad Chandorkar, Kandarp Joshi, Xavier Velez, Diamantino Caseiro, Ben Haynor, Weiran Wang, Nikhil Siddhartha, and 1 others. 2023. Contextual spelling correction with large language models. In *Proc. IEEE ASRU*, pages 1–8. IEEE.

Ehsan Variani, David Rybach, Cyril Allauzen, and Michael Riley. 2020. Hybrid autoregressive transducer (hat). In *Proc. IEEE ICASSP*, pages 6139–6143. IEEE.

Weiran Wang, Zelin Wu, Diamantino Caseiro, Tsendsuren Munkhdalai, Khe Chai Sim, Pat Rondon, Golan Pundak, Gan Song, Rohit Prabhavalkar, Zhong Meng, and 1 others. 2023. Contextual biasing with the knuth-morris-pratt matching algorithm. *arXiv preprint arXiv:2310.00178*.

Gordon Wichern, Joe Antognini, Michael Flynn, Licheng Richard Zhu, Emmett McQuinn, Dwight Crow, Ethan Manilow, and Jonathan Le Roux. 2019. Wham!: Extending speech separation to noisy environments. *arXiv preprint arXiv:1907.01160*.

Yu Xi, Xiaoyu Gu, Haoyu Li, Jun Song, Bo Zheng, and Kai Yu. 2025a. Masked self-distilled transducer-based keyword spotting with semi-autoregressive decoding. *arXiv preprint arXiv:2505.24820*.

Yu Xi, Hao Li, Baochen Yang, Haoyu Li, Hainan Xu, and Kai Yu. 2024. TDT-KWS: Fast and accurate keyword spotting using token-and-duration transducer. In *Proc. IEEE ICASSP*, pages 11351–11355. IEEE.

Yu Xi, Haoyu Li, Xiaoyu Gu, Yidi Jiang, and Kai Yu. 2025b. MFA-KWS: Effective keyword spotting with multi-head frame-asynchronous decoding. *IEEE Transactions on Audio, Speech and Language Processing*.

Yu Xi, Haoyu Li, Xiaoyu Gu, Hao Li, Yidi Jiang, and Kai Yu. 2025c. Streaming keyword spotting boosted by cross-layer discrimination consistency. In *Proc. IEEE ICASSP*, pages 1–5. IEEE.

Guanrou Yang, Ziyang Ma, Zhifu Gao, Shiliang Zhang, and Xie Chen. 2024. Ctc-assisted llm-based contextual asr. In *Proc. IEEE SLT*, pages 126–131. IEEE.

Zhanheng Yang, Sining Sun, Xiong Wang, Yike Zhang, Long Ma, and Lei Xie. 2023. Two stage contextual word filtering for context bias in unified streaming and non-streaming transducer. *arXiv preprint arXiv:2301.06735*.

Zhuoyuan Yao, Di Wu 0061, Xiong Wang, Binbin Zhang, Fan Yu, Chao Yang, Zhendong Peng, Xiaoyu Chen, Lei Xie, and Xin Lei. 2021. Wenet: Production oriented streaming and non-streaming end-to-end speech recognition toolkit. In *interspeech*, volume 2021, pages 4054–4058.

Ding Zhao, Tara N Sainath, David Rybach, Pat Rondon, Deepti Bhatia, Bo Li, and Ruoming Pang. 2019. Shallow-fusion end-to-end contextual biasing. In *Interspeech*, pages 1418–1422.

A Generative AI Use Disclosure

AI tools were utilized solely for editing and polishing the manuscript to improve clarity and grammar. All core scientific content, including experimental design, data collection, and analysis, was conducted independently by the authors without the use of AI.

B Reproducibility

The submitted supplementary material includes an anonymized implementation of the proposed VISH decoding algorithms, including S-VISH, M-VISH,

and Bi-MVISH. We provide this code to facilitate inspection of the constrained CTC Viterbi scoring, shared-prefix search, and bidirectional hotword verification procedures described in this work. Since VISH is training-free and operates with frame-level CTC posteriors, the implementation can be integrated with compatible acoustic encoders and Speech-LLM ASR systems. We plan to publicly release the implementation after further organization.