Journal of Science and Technology – Smart Systems and Devices
Volume x, Issue x, Month Year, ppp-ppp

End-to-End Vietnamese ASR with Acoustic-Aware Punctuation and Capitalization for Edge Devices

Trang Trinh Thi Huyen, Hoa Tran Nhat, Thuan Nguyen Dinh, Hiep Hoang Van, Vinh La The, Hung Pham Ngoc, Ngoc Mai Xuan*

Hanoi University of Science and Technology, Ha Noi, Vietnam

**Corresponding author email: ngoc.maixuan@hust.edu.vn*

Abstract

Vietnamese automatic speech recognition (ASR) for edge devices remains challenging because practical systems must simultaneously satisfy recognition accuracy, low latency, low memory overhead, and transcript readability. These requirements are particularly demanding for Vietnamese, where tonal and prosodic cues are important for sentence structuring, while many lightweight ASR systems still produce raw lowercase text without punctuation. In this work, we present an edge-ready Vietnamese ASR system built on VietASR Iteration-3 and fine-tuned on nearly 1,000 hours of professionally curated Dolly Vietnamese speech data with punctuation and capitalization annotations. Our work makes three main contributions. First, we develop a deployment-oriented fine-tuning pipeline for a compact 68M-parameter VietASR model, targeting practical offline inference on resource-constrained devices. Second, we introduce an end-to-end structured decoding formulation in which punctuation marks and capitalization indicators are incorporated directly into the output vocabulary and predicted jointly with lexical units, eliminating the need for a separate post-processing module. Third, we validate the system on both recognition and deployment aspects through evaluation on standard Vietnamese benchmarks and on a Qualcomm QCS8550 edge platform. Experimental results show that the fine-tuned model achieves 6.41% WER on the VIVOS test set, while also generating well-structured transcripts with strong punctuation performance (micro-F1 of 0.94). On the target device, the deployed system processes a 30-second utterance with about 380.0 ms encoder latency on CPU and 321.0 ms on NPU, while decoder and joiner costs remain negligible. Comparative analysis further indicates more stable Vietnamese output and better formatted transcripts than Whisper-large-v3-turbo and the pretrained VietASR baseline. These results suggest that compact Vietnamese ASR models can jointly address accuracy, formatting, and deployment constraints, making them suitable for real-time, privacy-preserving, on-device speech applications.

Keywords: Capitalization prediction, RNN-T, edge AI, punctuation restoration, viet-ASR.

1. Introduction

Automatic Speech Recognition (ASR) has become a cornerstone technology for human–computer interaction, enabling applications such as voice assistants, automated captioning, and spoken document processing [1]. Recent advances in deep learning, particularly end-to-end (E2E) architectures based on Transformers and Conformers [2, 3] , have significantly improved recognition accuracy across many languages. However, deploying ASR systems in edge and on-device environments remains challenging due to strict constraints on latency, memory, and power consumption [4].

These challenges are especially pronounced for Vietnamese, a tonal and low-resource language in which prosodic cues—such as pauses, pitch variations, and rhythm—play a crucial role in conveying sentence structure and semantic meaning [5]. In addition, sentence boundaries are signalled through pause duration, pitch resets, and energy contours rather than purely syntactic markers—cues [6], making it difficult to recover well-structured text from speech alone.

Beyond word-level transcription accuracy, real-world Vietnamese ASR applications require rich and readable outputs, including proper capitalization, punctuation, and sentence segmentation. Such formatted transcriptions are essential for downstream tasks such as machine translation, meeting summarization, and spoken document understanding. Nevertheless, many existing ASR systems produce raw, lowercase text without punctuation, which limits their practical usability.

Conventional pipelines address this by cascading a speech recogniser with a separate NLP module for punctuation restoration and capitalisation [7–9]. While effective in server-side settings, cascading designs increase latency, memory footprint, and error propagation—drawbacks that are particularly costly on resource-constrained edge hardware. This limitation motivates the need for ASR models that are not only accurate but also compact, efficient, and capable of generating structured text directly.

Large-scale multilingual foundation models, such as OpenAI's Whisper [10] and Google's Universal Speech Model (USM) [11], have demonstrated strong robustness across languages. However, significant gaps remain when applying these models to Vietnamese ASR in resource-constrained environments:

1) Impracticality for Edge Deployment: Models such as Whisper Large-v3-turbo contain over 0.8 billion parameters and typically require GPUs or specialized accelerators.

p-ISSN 3093-3285
e-ISSN 3093-3315
https://doi.org/10.51316/jst.xxx.ssad.xxxx.xx.x.x
Received: Mar 25, 2026; Revised: May 1, 2026;
Accepted: Jun 3, 2026; Online: Jul 3, 2026

2) Instability in Low-Resource Languages: These models may produce malformed characters, inconsistent casing, or hallucinated text [12], especially under constrained inference conditions.

3) Lack of Integrated Text Formatting in Lightweight Settings: Efficient deployments still rely on cascading pipelines, increasing system complexity and latency.

To address efficiency, the VietASR framework [13] introduced a self-supervised learning pipeline tailored for Vietnamese, combining HuBERT-style [14] pretraining on large-scale unlabeled Vietnamese speech data with an efficient Zipformer architecture. While VietASR achieves strong performance in terms of Word Error Rate (WER), it primarily focuses on acoustic modeling and transcription accuracy, and does not explicitly address the generation of formatted, deployment-ready text.

In this work, we extend a publicly available VietASR pretrained model in a downstream fine-tuning setting, leveraging publicly available Vietnamese datasets [**nguyen2025dollyaudio**] annotated with punctuation and capitalization. Our goal is to enable the model to generate structured text directly from speech, eliminating the need for external NLP post-processing and making it more suitable for edge deployment.

Designing such a system introduces several technical challenges:

- Resource Constraints vs. Accuracy: The model must remain compact (typically under 100M parameters) to enable real-time CPU inference, while achieving accuracy competitive with significantly larger models.
- Acoustic–Prosodic Mapping: Punctuation marks such as commas and periods are not explicitly spoken. The model must learn to map acoustic cues—pause duration, pitch resets, and energy variations—directly to punctuation tokens without relying on purely text-based grammar models.

In this paper, we address these challenges by extending VietASR toward edge-ready structured transcription. Our main contributions are summarized as follows:

- We implement a deployment-oriented fine-tuning strategy for VietASR using nearly 1,000 hours of punctuated and capitalized Vietnamese speech data, enabling the model to generate structured text directly from acoustics.
- An end-to-end approach to prosody-aware punctuation and capitalization generation without external NLP post-processing.
- We demonstrate stable, real-time CPU inference with improved output readability, while maintaining strong recognition accuracy. Experimental results on the VIVOS test set show that the model achieves a Word Error Rate (WER) of 6.41%, validating its effectiveness for practical, on-device Vietnamese ASR

2. Related Work

2.1. Self-Supervised Speech Representation Learning

Self-supervised learning (SSL) has become a cornerstone of modern ASR, enabling models to learn robust speech representations from large-scale unlabeled audio. wav2vec 2.0 introduced masked latent prediction to reduce reliance on labeled data while maintaining strong downstream performance [15]. HuBERT further improved training stability by predicting discretized hidden units obtained via clustering [14]. Despite their effectiveness, most SSL-based models operate on raw waveforms and employ large Transformer encoders, resulting in high computational and memory costs that hinder real-time and on-device deployment, particularly for low-resource languages such as Vietnamese. To address these limitations, VietASR adapts the HuBERT paradigm for practical Vietnamese ASR by using log-Mel filterbank (Fbank) features instead of raw waveforms [13]. In addition, VietASR introduces a Supervised Codebook strategy, where a lightweight ASR model trained on limited labeled data generates pseudo-labels for SSL pretraining. This approach better aligns learned representations with downstream ASR objectives and enables efficient training of compact models suitable for edge deployment [13].

2.2. Efficient and Streaming ASR Architectures

End-to-end architectures such as the RNN-Transducer (RNN-T) are widely adopted for streaming ASR due to their low-latency decoding capabilities [16]. Transformer-based models, including the Conformer, further improve recognition accuracy by modeling both global and local temporal dependencies [2]. However, standard Conformers process speech at a fixed frame rate, leading to redundant computation for acoustically stable segments. The Zipformer architecture [17] mitigates this issue through a multi-resolution encoder that dynamically downsamples intermediate representations before upsampling. This design substantially reduces computational cost while preserving temporal information. Zipformer also incorporates BiasNorm and the ScaledAdam optimizer to improve training stability. Prior work has shown that Zipformer models with fewer than 100M parameters, when trained with the VietASR pipeline, can achieve competitive performance on Vietnamese ASR benchmarks [13].

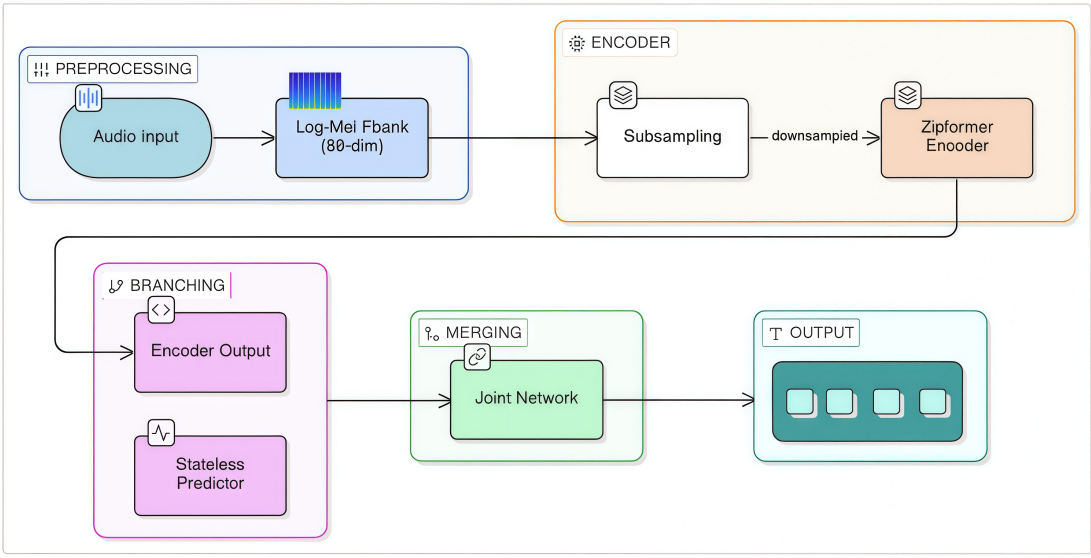

Fig. 1. Overall architecture of the VietASR system. The pipeline consists of four stages: (1) acoustic feature extraction using log-Mel filterbanks, (2) speech representation encoding with a Zipformer encoder, (3) RNN-T decoding with a stateless predictor, and (4) token generation through a joiner network producing BPE tokens with punctuation and capitalization

2.3. Prosody-Aware Punctuation and Capitalization

Punctuation restoration and capitalization are traditionally treated as downstream NLP tasks applied to raw ASR outputs [7–9]. While effective in text-based settings, such approaches rely solely on lexical context and discard important acoustic and prosodic cues, such as pause duration, pitch resets, and energy variations, which naturally signal sentence boundaries in speech. In this work, we adopt an end-to-end formulation in which punctuation marks and capitalization indicators are explicitly included in the ASR output vocabulary and learned directly during acoustic model fine-tuning. By training on punctuated and capitalized transcripts, the model is encouraged to align prosodic patterns with structural tokens, allowing commas, periods, and sentence boundaries to be inferred directly from speech rather than recovered through post-processing. This integrated design removes the need for cascading punctuation restoration modules, reducing both system complexity and inference latency. As a result, punctuation and capitalization are generated jointly with lexical content in a single decoding pass, making the proposed approach particularly suitable for real-time, on-device ASR deployment.

3. Methodology

3.1. Base Model: VietASR Iteration-3

Our system is built upon [13] Iteration-3, whose overall architecture is illustrated in Fig. 1. Which adopts a Zipformer encoder coupled with an RNN-Transducer (RNN-T) decoder. Acoustic features are represented using 80-dimensional log-Mel filterbank (Fbank) features, a widely used and computationally efficient representation for production ASR systems. The resulting model contains approximately 68 million parameters, providing a favorable trade-off between recognition accuracy and deployment feasibility for CPU-based, on-device inference.

The Zipformer architecture employs a multi-resolution, U-Net-like structure that dynamically downsamples intermediate representations and later upsamples them to the original temporal resolution. This design significantly reduces computational cost while preserving essential temporal information, making it well-suited for latency-sensitive and resource-constrained environments [17].

3.2. Tokenization and Output Vocabulary

We employ SentencePiece-based Byte Pair Encoding (BPE) for subword tokenization. To enable structured text generation, the output vocabulary is explicitly extended with tokens representing punctuation marks (., ,, ?, !) and capitalization indicators. Rather than relying on a separate text normalization or punctuation restoration module, capitalization is encoded through dedicated case-aware tokens that are jointly predicted during decoding.

This unified vocabulary design allows the model to learn lexical content, punctuation, and capitalization within a single end-to-end ASR framework. As a result, text formatting decisions are directly conditioned on both acoustic and linguistic context, avoiding error propagation commonly observed in cascading post-processing pipelines.

3.3. Fine-Tuning with Punctuated Speech Data

To adapt VietASR Iteration-3 for deployment-ready transcription, we fine-tune the model on punctuated and capitalized Vietnamese speech, enabling direct generation of structured text. Fine-tuning is conducted on the full model, including the acoustic encoder, predictor, and joiner modules of the RNN-T architecture.

Fine-tuning is performed on Dolly-Audio, a large-scale Vietnamese speech corpus consisting of approximately 1,000 hours of professionally curated recordings. The dataset includes speech from 152 speakers across multiple Vietnamese regions, covering a wide range of accents and speaking styles. All audio is carefully cleaned to remove background noise and music, and utterances are segmented at sentence-level boundaries to preserve natural prosody

Transcriptions are richly annotated with punctuation marks and capitalization, aligned with natural speech pauses and sentence boundaries. The corpus spans diverse domains such as news, education, entertainment, and conversational speech, ensuring linguistic diversity and robustness. Manual sampling indicates an estimated near-zero transcription error rate, making the dataset highly suitable for supervised ASR fine-tuning and structured text modeling

3.4. RNN-Transducer Fine-Tuning Objective

We adopt the Recurrent Neural Network Transducer (RNN-T) loss as the optimization objective during fine-tuning. Given an input acoustic feature sequence

$$\mathbf{x} = (x_1, x_2, \ldots, x_T)$$

and a target token sequence

$$\mathbf{y} = (y_1, y_2, \ldots, y_U),$$

where $\mathbf{y}$ includes lexical subword units, punctuation tokens, and capitalization markers, the RNN-T models the conditional probability:

$$P(\mathbf{y} \mid \mathbf{x}) = \sum_{\pi \in \mathscr{B}^{-1}(\mathbf{y})} P(\pi \mid \mathbf{x}),$$

where π denotes an alignment sequence over the extended vocabulary augmented with a blank symbol, and $\mathscr{B}$ is the standard collapse function that removes blank tokens.

The training loss is defined as the negative log-likelihood:

$$\mathscr{L}_{\text{RNN-T}} = -\log P(\mathbf{y} \mid \mathbf{x}).$$

This formulation allows the model to learn monotonic alignments between speech frames and output tokens, which is critical for streaming and real-time decoding.

3.5. Acoustic-Aware Text Structuring

Unlike conventional ASR pipelines that treat punctuation restoration as a downstream text-only task, our approach incorporates punctuation and capitalization directly into the RNN-T decoding process. By extending the output vocabulary, punctuation marks (., „ ?, !) and capitalization tokens are treated as first-class symbols predicted jointly with lexical units.

During fine-tuning, the model learns to associate acoustic-prosodic cues—such as pause duration, pitch resets, and energy contours—with structural tokens. For example, longer pauses and pitch resets are frequently aligned with period or comma tokens, while capitalization markers are learned through contextual and semantic cues embedded in the acoustic and linguistic representations.

This joint learning strategy enables the ASR system to generate readable, well-formatted Vietnamese text in a single decoding pass, without relying on external punctuation restoration or text normalization modules. As a result, the system remains compact, low-latency, and well-suited for edge and on-device deployment.

4. Experiments

4.1. Implementation Details.

Our implementation is built on the icefall [18] speech recognition framework, which provides optimized training and decoding pipelines for RNN-T–based models. Data preparation and corpus management are handled using lhotse [19], a unified toolkit for large-scale speech data preprocessing and manifest generation. All audio preprocessing steps, including feature extraction, segmentation, and supervision alignment, are performed using lhotse abstractions, enabling scalable and reproducible experimentation across large speech corpora.

4.2. Training and Evaluation Environment.

Model training and evaluation are conducted using the icefall framework with lhotse for data preparation and feature extraction. All experiments are performed on a single NVIDIA A10 GPU (24GB).

We adopt the ScaledAdam optimizer with a tri-stage learning rate schedule (warm-up, constant, decay) for stable convergence. Each training batch contains approximately 1000 seconds of audio. The predictor uses a stateless architecture with hidden size 512 and context size 1, while the joiner also uses a hidden size of 512.

The output vocabulary consists of 2000 SentencePiece BPE units, extended with punctuation and capitalization tokens. Recognition performance is evaluated on the VIVOS test set using WER, together with qualitative assessment of structured output quality.

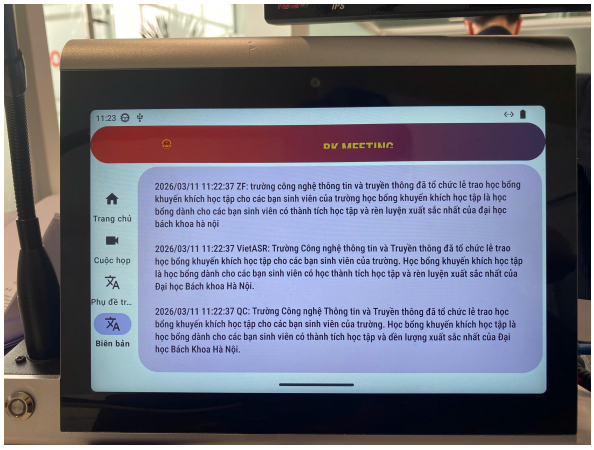

Fig. 2. Our smart meeting-room edge device based on the Qualcomm QCS8550 Edge AI platform

4.3. Edge Deployment and Latency Evaluation.

To assess practical deployment feasibility, latency and on-device inference experiments were conducted on our BKMeeting edge AI device built on the Qualcomm QCS8550 platform, as shown in Fig. 2. The device integrates an octa-core Kryo CPU with one Prime core running at up to 3.36 GHz, four Performance cores up to 2.8 GHz, and three Efficiency cores up to 2.0 GHz. It also supports LPDDR5x memory up to 4200 MHz with a maximum capacity of 16 GB. For AI acceleration, the platform provides a Qualcomm Hexagon NPU with HVX and HMX support, delivering up to 48 TOPS for INT8 inference and 12 TOPS for FP16 inference.

Latency was measured on 30-second audio segments to reflect realistic application scenarios such as voice interaction, dictation, and short meeting-room utterances. All latency results were obtained using single-stream offline decoding without batching. This deployment setup was used specifically to evaluate runtime efficiency and edge inference behavior, whereas model training and recognition benchmarking were carried out on AWS.

5. Results

5.1. Recognition Accuracy

Table 1 presents the Word Error Rate (WER) comparison on the VIVOS official test set [20], Common Voice Vietnamese test split [21], and the VLSP 2020 datasets specifically Set 2 [22] test sets.

The proposed VietASR-Finetuned model achieves a WER of 6.41% on VIVOS and 12.45% on Common Voice. , and 12.73% on VLSP.

Compared with the VietASR-Pretrained baseline, which obtains WERs of 6.32%, 11.46%, and 9.31% on VIVOS, Common Voice, and the weighted average, respectively, the fine-tuned model maintains competitive recognition accuracy while additionally supporting structured outputs with punctuation and capitalization. Moreover, when compared with lightweight PhoWhisper variants, the proposed model achieves substantially better recognition accuracy. Specifically, PhoWhisper-base (74M parameters) obtains WERs of 8.46%, 16.19%, and 43.01%, while PhoWhisper-tiny (39M parameters) obtains WERs of 10.41%, 19.05%, and 49.85%. These results show that our model provides a better accuracy–functionality trade-off for edge-oriented Vietnamese ASR.

Compared with larger models, VietASR-Finetuned demonstrates competitive performance despite its significantly smaller size (68M parameters). While Whisper-large-v3-turbo exhibits higher WER (8.81%, 13.74%, and 20.41% on VIVOS, Common Voice, and VLSP, respectively), the proposed model achieves better accuracy under the same evaluation conditions. Similarly, although Parakeet-CTC attains lower WER on VIVOS, it relies on a much larger model (600M parameters), making it less suitable for edge deployment. We also compare VietASR-Finetuned with lightweight PhoWhisper variants. Although PhoWhisper-small has a larger model size (244M parameters), it obtains WERs of 6.33%, 11.08%, and 32.96% on VIVOS, Common Voice, and VLSP, respectively. In contrast, VietASR-Finetuned achieves better overall robustness, especially on the VLSP evaluation set, while using only 68M parameters. This indicates that the proposed model offers a more favorable trade-off between recognition accuracy, output quality, and edge-deployment feasibility.

These results highlight that deployment-oriented fine-tuning on punctuated and capitalized Vietnamese speech improves output readability while maintaining strong recognition performance.

The gains can be attributed to stronger alignment between acoustic representations and structured linguistic targets, which provide additional supervision signals during training.

Overall, the proposed VietASR-Finetuned model achieves a favorable trade-off between model size, accuracy, and deployment efficiency, making it well-suited for real-time Vietnamese ASR on resource-constrained devices.

5.2. Inference Latency and Throughput

Table 2 reports the mean inference latency, averaged over 100 runs, together with the output characteristics of the evaluated ASR systems for a 30-second audio input. The results show that both VietASR-Pretrained and VietASR-Finetuned achieve substantially lower inference latency than Whisper-large-v3-turbo under the tested deployment settings.

On the QCS8550 platform, the VietASR models require only 321.0–322.0 ms on the NPU and 380.0 ms on the CPU for the encoder, while the decoder

Table 1. Comparison of VietASR with open-source ASR models. The average WER is weighted by the word count of each test set

Model	**Params (M)**	**WER on Evaluation Datasets (%)**		
		VIVOS	**Common Voice**	**VLSP 2020 set 2**
Whisper-large-v3-turbo [10]	809	8.81	13.74	20.41
Parakeet-CTC [23]	600	5.96	8.58	13.60
PhoWhisper [24]				
small	244	6.33	11.08	32.96
base	74	8.46	16.19	43.01
tiny	39	10.41	19.05	49.85
VietASR-Pretrained [13]	68	6.32	11.46	9.31
VietASR-Finetuned	**68**	**6.41**	**12.45**	**12.73**

and joiner introduce only 0.1–0.2 ms of additional latency. In contrast, Whisper-large-v3-turbo has a much heavier decoding process. Although its encoder is executed only once for a 30-second audio input, its autoregressive decoder must be repeatedly invoked to generate the output token sequence. Therefore, the decoder latency reported in Table 2 represents the cost of a decoding step rather than the full accumulated decoding cost. For a typical 30-second utterance with hundreds of generated tokens, this iterative decoding overhead becomes substantial. These results indicate that VietASR avoids the major decoding bottleneck of Whisper-style models and is considerably more efficient for edge deployment.

In terms of output quality, Whisper-large-v3-turbo exhibits unstable transcription behavior in our deployment setting, including malformed characters and inconsistent casing. VietASR-Pretrained provides much faster and more stable inference, but its outputs remain limited to lowercase text without punctuation. After fine-tuning, VietASR-Finetuned preserves nearly the same latency characteristics as the pretrained model while improving transcription usability, producing outputs with capitalization, punctuation, and more stable decoding behavior. This result suggests that the proposed fine-tuning strategy improves output quality without introducing a meaningful runtime penalty.

Even when executed entirely on the QCS8550 CPU, VietASR-Finetuned processes a 30-second audio input with an encoder latency of only 380.0 ms, corresponding to a real-time factor well below 1. More importantly, its decoding overhead is almost negligible because the RNN-T predictor and joiner require only 0.1–0.2 ms. Compared with Whisper-large-v3-turbo and PhoWhisper, whose autoregressive decoders accumulate latency over many generated tokens, the proposed model achieves substantially faster end-to-end inference while also providing more reliable and structured output behavior.

5.3. Output Quality and Decoding Stability

Beyond latency and WER, we qualitatively analyze the readability and stability of generated transcripts. We observe that Whisper-based models, when executed under resource constraints, may produce malformed Vietnamese characters, inconsistent casing, or partially corrupted outputs.

In contrast, the proposed VietASR system consistently generates well-formed Vietnamese text with appropriate capitalization and punctuation. Sentence boundaries align naturally with acoustic pauses, resulting in transcriptions that are immediately usable for downstream applications such as captioning, summarization, or human-facing interfaces.

This stability is attributed to the end-to-end learning of structured tokens, which eliminates the need for external punctuation restoration modules and reduces error propagation across components.

5.4. Punctuation and Capitalization Evaluation

To further assess the quality of structured text generation, we evaluate the model's ability to predict both punctuation and capitalization under two settings: (1) end-to-end evaluation, where predictions are assessed directly on ASR outputs, and (2) conditional evaluation given correct words, where accuracy is measured only on correctly recognized tokens.

This dual evaluation protocol allows us to distinguish between errors caused by speech recognition and those arising from text structuring.

Table 2. Inference latency and output characteristics

Model	Platform	Inference Time			Output Characteristics
		Encoder	Decoder	Joiner	
Whisper-large-v3-turbo	CPU	2054.36 ms	43.08 ms	–	Unstable output; malformed characters; inconsistent casing
	NPU	799.7 ms	13.1 ms	–	
Phowhisper-small	CPU	2244.2 ms	292.8 ms	–	Lowercase only; no punctuation
	NPU	313.3 ms	66.0 ms	–	
Phowhisper-base	CPU	568.4 ms	63.7 ms	–	Lowercase only; no punctuation
	NPU	138.6 ms	10.4 ms	–	
Phowhisper-tiny	CPU	245.0 ms	23.1 ms	–	Lowercase only; no punctuation
	NPU	68.6 ms	4.9 ms	–	
Parakeet-CTC	CPU	1404.5 ms	1.1 ms	–	Unstable output; malformed characters; inconsistent casing
	NPU	396.0 ms	0.6ms	–	
VietASR-Pretrained	CPU	380.0 ms	0.2 ms	0.2 ms	Lowercase only; no punctuation
	NPU	322.0 ms	0.1 ms	0.1 ms	
VietASR-Finetuned	**CPU**	**380.0 ms**	**0.2 ms**	**0.2 ms**	**Capitalization and punctuation; stabilized decoding**
	NPU	**321.0 ms**	**0.1 ms**	**0.1 ms**	

Evaluation Metrics. For punctuation, we report precision, recall, and F1-score for three major classes: comma, period, and question mark. For capitalization, we evaluate four classes: *LOWER*, *UPPER*, *TITLE_CASE*, and *SENTENCE_START*. The evaluation is conducted on a test set containing 66,591 utterances, corresponding to over 1.4 million tokens and more than 125,000 punctuation instances.

End-to-End Performance. Table 3 presents punctuation performance measured directly on ASR outputs. The model achieves a micro-averaged F1-score of 0.94, indicating strong overall performance in generating structured text from raw speech.

Table 3. End-to-end punctuation performance

Punctuation	Precision	Recall	F1
Comma	0.91	0.89	0.90
Period	0.98	0.97	0.97
Question	0.84	0.81	0.82
Micro Avg	0.94	0.93	0.94
Macro Avg	0.91	0.89	0.90

Table 4 reports capitalization performance under the same setting. The model achieves an overall accuracy of 0.99 and a macro-averaged F1-score of 0.93. Performance is near-perfect for the dominant *LOWER* class, while strong results are observed for *SENTENCE_START* and *TITLE_CASE*. The *UPPER* class remains relatively more challenging due to its low frequency.

Table 4. End-to-end capitalization performance

Class	Precision	Recall	F1
LOWER	0.99	1.00	1.00
UPPER	0.88	0.85	0.87
TITLE_CASE	0.93	0.88	0.90
SENTENCE_START	0.96	0.94	0.95
Macro Avg	0.94	0.92	0.93

Performance Given Correct Words. To isolate structuring performance from ASR errors, we further evaluate punctuation only on correctly aligned words. As shown in Table 5, the performance remains consistent, with a micro F1-score of 0.94.

Table 5. Punctuation performance given correct words

Punctuation	Precision	Recall	F1
Comma	0.91	0.89	0.90
Period	0.98	0.97	0.98
Question	0.84	0.82	0.83
Micro Avg	0.95	0.93	0.94
Macro Avg	0.91	0.89	0.90

For capitalization, Table 6 shows that performance improves significantly when conditioned on correct words, reaching near-perfect accuracy and a macro

F1-score of 0.99. In particular, *TITLE_CASE* improves notably, indicating that most capitalization errors originate from ASR mistakes rather than modeling limitations.

Table 6. Capitalization performance given correct words

Class	Precision	Recall	F1
LOWER	1.00	1.00	1.00
UPPER	1.00	1.00	1.00
TITLE_CASE	0.98	0.97	0.97
SENTENCE_START	0.99	0.99	0.99
Macro Avg	0.99	0.99	0.99

Analysis. Several important observations can be drawn:

- **Strong sentence boundary modeling:** High performance on periods (F1 $\approx$ 0.97–0.98) and *SENTENCE_START* (F1 = 0.95) indicates that the model effectively captures sentence boundaries from acoustic-prosodic cues such as pauses and pitch resets.
- **Robust intra-sentence structuring:** Comma prediction achieves an F1-score of 0.90, suggesting that the model learns finer-grained prosodic patterns, although this remains more challenging than sentence boundary detection.
- **Effective proper noun modeling:** The *TITLE_CASE* results (F1 = 0.90 end-to-end, 0.97 conditional) demonstrate that the model can reliably recover named entities such as *Hà Nội* and *Việt Nam*, which are essential for Vietnamese text readability.
- **Minimal impact from ASR errors:** The similarity between end-to-end and conditional punctuation results, together with the improvement in capitalization performance, suggests that punctuation prediction is robust to recognition errors, while capitalization is more sensitive to lexical correctness.

Overall, the results demonstrate that the proposed end-to-end approach can reliably generate well-structured Vietnamese text, jointly modeling punctuation and capitalization within a unified ASR decoding framework without requiring external post-processing modules.

5.5. Hallucination Analysis of Whisper on Vietnamese Data

We analyze the hallucination behavior of Whisper-large-v3-turbo on Vietnamese speech data, where hallucinations refer to transcriptions that have no phonetic or semantic correspondence to the input audio, often occurring under silence or acoustically ambiguous conditions [12].

Table 7. Frequent hallucinated phrases on VLSP

Prediction	Count	Rate	Description
"hãy subscribe cho Ghiền Mì Gõ"	1,972	3.5%	Most frequent bias
"cảm ơn các bạn đã theo dõi và ủng hộ cho kênh lalaschool để không bỏ lỡ những video hấp dẫn"	145	0.26%	YouTube-style outro
"thank you"	82	0.15%	Generic English phrase
Total	2,199	3.9%	–

On the VLSP 2020 dataset, specifically Set 2, which contains 56,428 utterances, we observe a consistent hallucination pattern where the model generates a small set of frequent phrases unrelated to the input. The most common hallucinated outputs are summarized in Table 7.

Such behavior aligns with the concept of *Bag of Hallucinations* (BoH), where hallucinated outputs concentrate on a limited set of recurring phrases. In Vietnamese, these phrases likely originate from online video content seen during pretraining, similar to English patterns such as "thanks for watching".

We further observe that hallucinations are strongly correlated with challenging acoustic conditions, including long silence, low-energy segments, and inputs exceeding the model's segmentation window. In these cases, decoding becomes dominated by the language prior rather than acoustic evidence, producing fluent but incorrect outputs.

Additionally, hallucinations follow a highly skewed distribution, where a small number of phrases account for a large proportion of errors, making them predictable and potentially detectable.

These findings highlight a key limitation of large generative ASR models in deployment settings. Compared to VietASR, which relies on acoustic grounding and structured token prediction, Whisper is more susceptible to hallucination due to its reliance on implicit language modeling. This suggests the need for mitigation strategies such as BoH filtering, confidence scoring, or acoustic alignment verification in practical systems.

6. Conclusion and Future Work

This paper presents an edge-ready Vietnamese ASR system based on a compact 68M-parameter VietASR model, fine-tuned for structured text generation. Through large-scale fine-tuning on nearly 1,000 hours of punctuated and capitalized Vietnamese speech,

the proposed system is able to produce readable, well-formatted transcriptions—including punctuation and capitalization—directly from speech, without relying on external post-processing modules.

Experimental results show that the proposed model achieves 6.41% WER on the VIVOS benchmark, together with a punctuation micro-F1 of 0.94 and near-perfect capitalization performance (macro-F1 of 0.99 given correct words). While maintaining a compact model size suitable for CPU-based execution, the system achieves real-time inference, processing a 30-second utterance in 380.0 ms on CPU and 321.0 ms on NPU. Compared to large-scale multilingual models, the proposed system provides significantly lower latency and more stable decoding behavior than Whisper-large-v3-turbo, avoiding issues such as malformed outputs and hallucinations.

These results confirm that compact ASR models can jointly achieve strong recognition accuracy, structured text generation, and real-time performance on edge devices. By treating punctuation and capitalization as first-class decoding tokens, the system learns to align prosodic cues with sentence structure, resulting in transcriptions that are immediately usable for downstream tasks such as captioning, summarization, and human-facing interfaces.

For future work, we plan to extend the system to streaming ASR with incremental punctuation prediction, improve robustness across regional accents, and explore joint modeling of inverse text normalization and named entity handling within the same framework.

Acknowledgment

This work was financially supported by the Ministry of Education and Training of Vietnam (MOET) under Project No. CT2025.EA.BKA.16. The authors sincerely acknowledge the support provided through this project.

References

[1] D. Jurafsky and J. H. Martin, Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, with Language Models, 3rd ed., 2026.

[2] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, and R. Pang, Conformer: Convolution-augmented Transformer for Speech Recognition, 2020. https://doi.org/10.48550/arXiv.2005.08100

[3] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, Attention Is All You Need, CoRR, vol. abs/1706.03762, 2017 . http://arxiv.org/abs/1706.03762

[4] C. Feng, Y. Lin, S. Zhuo, C. Su, R. K. Ramakrishnan, Z. Yuan, and X. Zhang, Edge-ASR: Towards Low-Bit Quantization of Automatic Speech Recognition Models, 2025. [Online] . Available: https://arxiv.org/abs/2507.07877

[5] L. Do, T. N. Nguyen, T. Pham, V. Do, H. Nguyen, and C. Nguyen, VietSuperSpeech: A Large-Scale Vietnamese Conversational Speech Dataset for ASR Fine-Tuning in Chatbot, Customer Support, and Call Center Applications, 2026. [Online] . Available: https://arxiv.org/abs/2603.01894

[6] A. Jakubiak, P. Stachyra, P. Czubowski, H. Borkowski, S. Latka, R. Izak, K. Jankowski, S. Janicka, and M. Zielinski, Adapting ASR Models for Speech-to-Punctuated-Text Recognition with Utterance Gluing, in Proceedings of the 8th International Conference on Natural Language and Speech Processing (ICNLSP-2025), Southern Denmark University, Odense, Denmark, Aug. 2025, pp. 72–81.

[7] O. Tilk and T. Alumäe, Bidirectional Recurrent Neural Network with Attention Mechanism for Punctuation Restoration. in Interspeech, 2016, pp. 9.

[8] T. Alam, A. Khan, and F. Alam, Punctuation restoration using transformer models for high-and low-resource languages, in Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT 2020), 2020, pp. 132–142.

[9] H. T. T. Uyen, N. A. Tu, and T. D. Huy, Vietnamese capitalization and punctuation recovery models, arXiv preprint arXiv:2207.01312, 2022.

[10] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, Whisper: Robust speech recognition via large-scale weak supervision, arXiv preprint arXiv:2212.01234, 2022.

[11] Y. Zhang, W. Han, J. Qin, Y. Wang, A. Bapna, Z. Chen, N. Chen, B. Li, V. Axelrod, G. Wang, Z. Meng, K. Hu, A. Rosenberg, R. Prabhavalkar, D. S. Park, P. Haghani, J. Riesa, G. Perng, H. Soltau, T. Strohman, B. Ramabhadran, T. N. Sainath, P. Moreno, C.-C. Chiu, J. Schalkwyk, F. Beaufays, and Y. Wu, Google USM: Scaling Automatic Speech Recognition Beyond 100 Languages, 2023. https://doi.org/10.48550/arXiv.2303.01037

[12] M. Barański, J. Jasiński, J. Bartolewska, S. Kacprzak, M. Witkowski, and K. Kowalczyk, Investigation of Whisper ASR Hallucinations Induced by Non-Speech Audio, in ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Apr. 2025, pp. 1–5.

[13] J. Zhuo, Y. Yang, Y. Shao, Y. Xu, D. Yu, K. Yu, and X. Chen, VietASR: Achieving Industry-level Vietnamese ASR with 50-hour labeled data and Large-Scale Speech Pretraining, arXiv preprint arXiv:2505.21527, 2025.

[14] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, Hubert: Self-supervised speech representation learning by masked prediction of hidden units, IEEE/ACM transactions on audio, speech, and language processing, vol. 29, pp. 3451–3460, 2021.

[15] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, wav2vec 2.0: A framework for self-supervised learning of speech representations, Advances in neural information processing systems, vol. 33, pp. 12449–12460, 2020.

[16] A. Graves, Sequence transduction with recurrent neural networks, arXiv preprint arXiv:1211.3711, 2012.

[17] Z. Yao, L. Guo, X. Yang, W. Kang, F. Kuang, Y. Yang, Z. Jin, L. Lin, and D. Povey, Zipformer: A faster and better encoder for automatic speech recognition, arXiv preprint arXiv:2310.11230, 2023.

[18] k2-fsa Project, icefall: A Modern Open-Source Speech Recognition Toolkit, 2022. [Online] . Available: https://github.com/k2-fsa/icefall. Accessed on: Feb. 2, 2026.

[19] P. Żelasko, D. Povey, J. Trmal, and S. Khudanpur, Lhotse: A Speech Data Representation Library for the Modern Deep Learning Ecosystem, 2021. https://doi.org/10.48550/arXiv.2110.12561

[20] H.-T. Luong and H.-Q. Vu, A non-expert Kaldi recipe for Vietnamese speech recognition system, in Proceedings of the Third International Workshop on Worldwide Language Service Infrastructure and Second Workshop on Open Infrastructures and Analysis Frameworks for Human Language Technologies (WLSI/OIAF4HLT2016), 2016, pp. 51–55.

[21] R. Ardila, M. Branson, K. Davis, M. Kohler, J. Meyer, M. Henretty, R. Morais, L. Saunders, F. Tyers, and G. Weber, Common Voice: A Massively-Multilingual Speech Corpus, in Proceedings of the Twelfth Language Resources and Evaluation conference, Marseille, France, May 2020, pp. 4218–4222.

[22] Association for Vietnamese Language and Speech Processing, Automatic Speech Recognition for Vietnamese, 2020. [Online] . Available: https://vlsp.org.vn/vlsp2020/eval/asr

[23] D. Galvez, V. Bataev, H. Xu, and T. Kaldewey, Speed of light exact greedy decoding for rnn-t speech recognition models on gpu, arXiv preprint arXiv:2406.03791, 2024.

[24] T.-T. Le, L. T. Nguyen, and D. Q. Nguyen, PhoWhisper: Automatic Speech Recognition for Vietnamese, in Proceedings of the ICLR 2024 Tiny Papers track, 2024.