# Token-Level Edge–Cloud Speculative Decoding for Real-Time Speech Emotion Captioning

Jiajun Lu
Student ID: 1654675

**Abstract**

This research propose an edge–cloud collaboration framework for Speech Emotion Captioning (SEC): a small on-device language model (SLM) drafts text while a larger cloud language model (LLM) verifies only the hard parts via speculative (draft–verify) decoding. The aim is to preserve near-cloud quality while reducing cloud token usage, latency, and data exposure (audio never leaves the device; only token IDs do). Building on recent token-level edge–cloud SLM and LLM collaboration and cost-aware protocols, the study adapt them to speech by combining local uncertainty measures with content-aware rules to decide when to escalate blocks to the cloud. We will evaluate on public SEC resources and report latency–quality trade-offs versus cloud-only and local-only baselines. The goal is a practical operating point for mobile health and assistive settings where real-time responsiveness and privacy are critical where raw audio remains on device (only token IDs are shared)

## 1 Introduction

Real-time feedback matters in mobile and telemedicine scenarios: delayed responses can limit clinical usefulness for monitoring, coaching, or assistive applications. Edge computing reduces round-trip delays by processing near the data source, and recent 5G/edge studies emphasize low-latency as a cornerstone for healthcare use cases. At the same time, speculative decoding has emerged as a principled way to cut generation time without altering model outputs, by letting a fast drafter propose tokens that a stronger verifier accepts as a longest safe prefix. Token-level edge–cloud collaboration further shows that a small on-device model can generate most tokens while a cloud model corrects only a minority—previously demonstrated for text tasks. We extend these ideas to speech emotion captioning (SEC), where decoding is audio-conditioned and uncertainty can be estimated locally. Recent work in speech (e.g., SpecASR) tailors speculative decoding to ASR and reports sizable latency reductions with no loss in recognition accuracy, suggesting similar benefits may transfer to speech-to-caption pipelines. Our framework keeps audio on device and exchanges only token IDs and accept/patch signals, aiming to minimize network and verification overheads while protecting content.

Building on this, this research study two questions centered on responsiveness: RQ1: Which local uncertainty signals (token-level log-probabilities, entropy/margin, or block-level stability) best predict tokens that require cloud verification in SEC? RQ2: What are the latency, quality and privacy trade-offs relative to local-only and cloud-only baselines?

By focusing on latency and grounding our design in speculative decoding and edge–cloud collaboration, we aim to deliver near-cloud caption quality with faster, more interactive response times appropriate for mobile-health and assistive settings.

## 2 Literature Review

Speculative decoding was introduced in NLP as a draft–verify scheme: a lightweight drafter proposes tokens and a target model accepts the longest safe prefix, preserving the target model's output while cutting decoding steps; this line, originating with Leviathan et al.[1], reports substantial speedups without changing model training or outputs. Building on token-level reasoning, Hybrid SLM–LLM for Edge-Cloud Collaborative Inference [2] shows that an on-device small model can draft while a cloud LLM verifies only difficult tokens, achieving near-LLM quality at a fraction of remote cost—establishing

token-granular collaboration as an alternative to layer partitioning. For multimodal systems, CD-CCA [3] adds a complementary training-time perspective: send only uncertain tokens upstream (UTS), distill knowledge in the cloud (AKD), and return compressed weight updates (DWC), offering a principled way to prioritize what crosses the edge–cloud boundary (demonstrated on text–image). Recently, speculative decoding entered speech via SpecASR [4], which tailors draft length and recycling to LLM-based ASR, yielding 3× speedups but focusing on latency rather than cloud-token cost or privacy. Meanwhile, Speech Emotion Captioning (SEC) reframes emotion understanding as open-ended captioning; SE-Cap [5] establishes an audio-to-caption pipeline, and AlignCap [6] improves robustness via alignment and preference optimization—defining tasks and datasets our work will inherit.

# 3 Research Design

**Task & Dataset**: This study target SEC: given speech, produce a short natural-language caption describing emotions. Core datasets: SE-Cap(EMOSpeech dataset); AlignCap(EMOSEC dataset, based on ESD and IEMOCAP); MER2024 dataset for training/evaluation splits and captions. This study will report BLEU/CIDEr and emotion-coverage/consistency.

**Models**: Edge SLM: Qwen2.5-Omni 3B. Cloud LLM: Qwen2.5-Omni 7B. Same tokenizer enables ID-only communication.

**Protocol (speculative draft-verify)**: The edge SLM decodes in blocks of k tokens. For each block it computes local uncertainty (token NLL/entropy/margin/optional MC-dropout) and content rules (digits, dates, names) to decide whether to call the cloud. If skipped, the block is committed locally. If escalated, the cloud LLM runs teacher-forcing on the block, accepts the longest safe prefix (probability $\geq$ threshold) and replaces the first difficult token (optionally returns a bonus token), then updates its KV cache so only incremental tokens are sent next round. Audio never leaves the device—only token IDs and minimal control signals.

**Latency Measurement**: The study will report end-to-end latency of edge SLM with cloud LLM calls and also compare against cloud-only and local-only baselines.

**Ethics/Privacy**: No raw audio is uploaded; only text tokens are transmitted. We will document remaining leakage risks (e.g., sensitive content in text).

# 4 Timetable

**Phase A (Teaching W4–W5)**: Prepare local-only (3B) and cloud-only (7B) baselines, Implement k-token block drafting on the edge with local "hard-token" gates

**Phase B (Teaching W6-W7)**: Implement Cloud verifier with KV/prefix cache and robust accept-prefix/first-fix rollback.

**Phase C (Teaching W8-W9)**: Main experiments + ablations and error analysis.

**Phase D (Teaching W10-W11)**: Prepare artifacts, demonstrations, model documentation, and oral presentations.

# 5 Expected Contributions

1. A speech-aware edge–cloud speculative decoding protocol that saves remote tokens while preserving SEC quality.
2. A privacy-friendly deployment that never uploads audio.
3. Empirical quality–latency trade-offs on public SEC datasets.
4. Open-sourced scripts for block decoding, gating, and cloud verification. This advances token-level collaboration beyond text into speech captioning.

# References


[1] Yaniv Leviathan, Matan Kalman, and Yossi Matias. "Fast inference from transformers via speculative decoding". In: *International Conference on Machine Learning*. PMLR. 2023, pp. 19274–19286.

[2] Zixu Hao et al. "Hybrid slm and llm for edge-cloud collaborative inference". In: *Proceedings of the Workshop on Edge and Mobile Foundation Models*. 2024, pp. 36–41.

[3] Guanqun Wang et al. "Cloud-device collaborative learning for multimodal large language models". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024, pp. 12646–12655.

[4] Linye Wei et al. "SpecASR: Accelerating LLM-based Automatic Speech Recognition via Speculative Decoding". In: *arXiv preprint arXiv:2507.18181* (2025).

[5] Yaoxun Xu et al. "Secap: Speech emotion captioning with large language model". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 38. 17. 2024, pp. 19323–19331.

[6] Ziqi Liang, Haoxiang Shi, and Hanhui Chen. "Aligncap: Aligning speech emotion captioning to human preferences". In: *arXiv preprint arXiv:2410.19134* (2024).