SEAM: Bridging the Temporal-Semantic Granularity Gap for LLM-based Speech Recognition

Junseok Oh and Ji-Hwan Kim*
Department of Computer Science and Engineering
Sogang University
Seoul, Republic of Korea
{ohjs, kimjihwan}@sogang.ac.kr

Abstract

Speech-LLM integration faces a temporal-semantic granularity gap: speech representations scale with temporal duration while text tokens scale with semantic content. Existing duration-based methods generate embeddings at fixed rates, creating distributional mismatch with LLM pre-training. We propose SEAM (Speech Encoder-Decoder Alignment Module), an encoder-decoder architecture employing variable-rate generation through cross-attention between speech features and text embeddings. SEAM produces embeddings at adaptive rates that closely match natural text distributions while preserving pre-trained knowledge by freezing both speech encoder and LLM. We introduce a multi-stage training strategy and First Token Guidance to improve initial token prediction. SEAM achieves competitive performance on LibriSpeech (2.6%/5.2% WER). More significantly, trained only on LibriSpeech (960h), SEAM achieves 4.7% WER on cross-domain TED-LIUM-v2, demonstrating that integrating LLM's linguistic knowledge enables effective generalization beyond limited speech training data.

1 Introduction

End-to-end automatic speech recognition (ASR) systems have achieved substantial performance improvements by jointly modeling acoustic and linguistic components within unified architectures (Graves and Jaitly, 2014; Deng and Woodland, 2024). Nevertheless, the language modeling capacity in E2E ASR is limited by training on paired speech-text data, which is smaller in scale compared to the text-only corpora that enable LLMs' extensive linguistic knowledge (Brown et al., 2020; Cui et al., 2025). This scale limitation constrains E2E ASR's linguistic coverage and domain robustness, particularly under domain shift and when processing rare words or phonetically similar words.

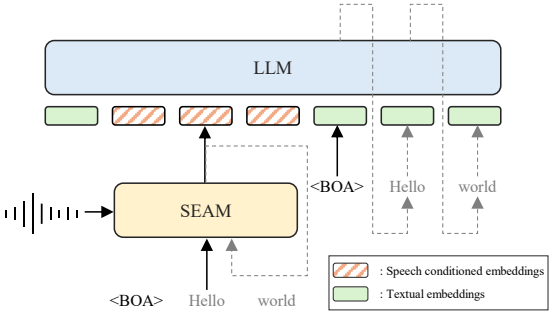

Figure 1: Overview of SEAM approach. SEAM bridges the temporal-semantic granularity gap through variable-rate generation that matches natural text token distributions, enabling LLM processing consistent with pre-training scenarios.

These limitations motivate the integration of LLM capabilities into speech recognition.

However, ASR-LLM integration faces the challenge of bridging the temporal-semantic granularity gap between speech and text representations (Fathullah et al., 2024; Han et al., 2024; Huang et al., 2024). This gap encompasses two main issues: (i) representation space differences between continuous acoustic features and discrete token embeddings, and (ii) temporal-semantic granularity mismatch (Zhang et al., 2025). In vision-language research, Jun et al. (2025) identify a similar challenge in aligning frame-level video features with sentence representations. Speech representations scale with temporal duration (frames per second) while text naturally varies with semantic content density (tokens per semantic unit), creating a misalignment between the temporal resolution at which speech is represented and the semantic granularity at which text is tokenized.

Existing duration-based approaches inadequately address this gap by producing representations that scale linearly with audio duration regardless of semantic content density (Ma et al., 2024; Tang et al., 2023). This creates a distributional mis-

*Corresponding author.

Findings of the Association for Computational Linguistics: EACL 2026, pages 2135–2144
March 24-29, 2026

match: natural text sequences vary in length according to semantic content, while duration-based methods generate embeddings at fixed rates determined by temporal duration. By forcing LLMs to process uniformly-paced duration-based inputs, these methods introduce a structural constraint that diverges from LLM pre-training distributions where sequence length naturally correlates with semantic content.

To address this gap, we propose SEAM (Speech Encoder-Decoder Alignment Module), an encoder-decoder architecture with cross-attention that employs variable-rate generation in the LLM's semantic embedding space. Unlike existing duration-based approaches, SEAM produces embeddings at variable rates that adapt to semantic density rather than audio duration, naturally aligning with LLM pre-training distributions. We use frozen Whisper-large-v2 (Radford et al., 2023) as the speech encoder and Qwen3-4B-Instruct (Yang et al., 2025) as the LLM, with a multi-stage training strategy to preserve pre-trained knowledge while learning cross-modal alignment.

Our main contributions are: (1) SEAM, an encoder-decoder alignment architecture employing variable-rate generation to address sequence length inconsistency in speech-LLM integration, producing embeddings at rates that closely match natural text distributions; (2) a multi-stage training strategy with cross-modal alignment, instruction tuning, and refinement tuning to preserve pre-trained knowledge and address length dependency; (3) First Token Guidance (FTG), a simple yet effective technique that improves initial token prediction by inserting the predicted first assistant token into the prompt; (4) empirical results showing that SEAM effectively leverages LLM's linguistic knowledge for cross-domain generalization, achieving 4.7% WER on TED-LIUM-v2 when trained only on LibriSpeech (960h).

2 Related Work

2.1 ASR-LLM Integration Approaches

Recent work has explored various methods to combine ASR with LLMs. End-to-end approaches include projection-based methods like SLAM-ASR (Ma et al., 2024), which uses simple linear projections trained on LibriSpeech (Panayotov et al., 2015). Other integration approaches have explored connecting speech encoders with LLMs (Yu et al., 2024), integrating pre-trained speech and language models (Hono et al., 2024), and adapting LLMs for speech tasks (Ling et al., 2024; Fathullah et al., 2024).

2.2 Cross-Modal Alignment and Modality Gap

Cross-modal alignment in speech-text integration must address the modality gap, which manifests as both representation space differences and temporal-semantic granularity mismatch (Zhang et al., 2025). This challenge parallels issues in vision-language research: Jun et al. (2025) identify "the level mismatch in aligning frame-level video features with a sentence representation," proposing to bridge this gap by decomposing sentences into semantic units. In speech-text integration, this granularity mismatch manifests as sequence length inconsistency.

Projection-based approaches attempt to address the representation space gap through linear transformations but fail to resolve sequence length inconsistency. SLAM-ASR (Ma et al., 2024) employs linear projection with downsampling, producing embeddings at fixed rates determined by audio duration. This duration-based generation creates a modality gap by forcing LLMs to process uniformly-paced inputs that diverge from their pre-training distribution.

Q-Former architectures use learnable queries but maintain duration-based limitations. SALMONN (Tang et al., 2023) employs a window-level Q-Former adapted from BLIP-2 (Li et al., 2023), segmenting audio into fixed temporal windows. Like projection-based methods, this produces embeddings at fixed rates determined by temporal duration rather than semantic content, creating distributional mismatch with target text.

Recent work has explored more sophisticated approaches. Zhang et al. (2025) propose addressing both representation space and sequence length problems simultaneously, while Fang and Feng (2023) introduce cross-modal regularization techniques. However, the application of attention-based encoder-decoder architectures to cross-modal ASR-LLM integration remains relatively underexplored.

2.3 Encoder-Decoder Architectures and Parameter-Efficient Fine-Tuning

Encoder-decoder architectures with cross-attention have proven effective for sequence-to-sequence learning in ASR. Listen, Attend and Spell (LAS) (Chan et al., 2016) introduced attention-based

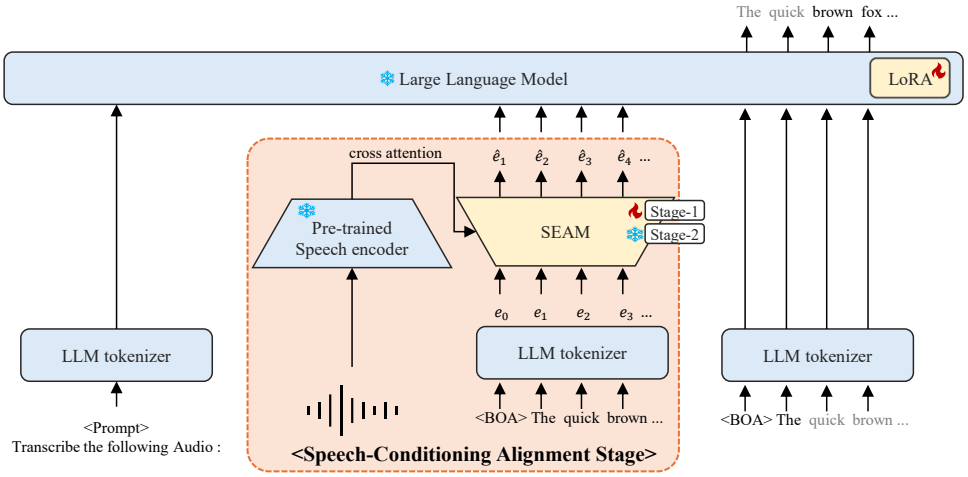

Figure 2: SEAM architecture overview. Dashed box: trainable components in Stage 1 (alignment module). During training, text and audio are both provided; during inference, only audio and the ⟨BOA⟩ token (Beginning of Audio) are used. In Stage 2, the alignment module is frozen and LoRA parameters are trained. In Stage 3, refinement tuning is performed with masking and deletion applied to Stage-1 outputs. Pre-trained components (Whisper-large-v2 encoder and Qwen3 LLM) remain frozen throughout. ⟨EOA⟩ denotes End of Audio.

encoder-decoder (AED) models that use cross-attention to align variable-length speech inputs with text outputs. Building on this foundation, SEAM adapts the AED architecture for cross-modal alignment to address the temporal-semantic granularity gap. Similar to how vision-language methods adjust text granularity to match frame-level features (Jun et al., 2025), SEAM uses cross-attention between speech features (from encoder) and text embeddings (as decoder queries) to learn variable-rate mappings in a shared embedding space that conform to token-level semantic granularity.

Parameter-efficient fine-tuning methods enable effective adaptation of large pre-trained models while preserving their knowledge. LoRA (Hu et al., 2021) introduces trainable low-rank matrices to linear layers, enabling efficient fine-tuning with minimal parameters. SEAM employs LoRA in Stage 2 to adapt the frozen LLM for speech-conditioned generation, applying low-rank adaptation to query, key, value, and feed-forward layers while keeping the alignment module learned in Stage 1 frozen. This stratified approach prevents interference between cross-modal alignment learning and language model adaptation.

3 Speech Encoder-Decoder Alignment Module

3.1 Overview

SEAM employs an encoder-decoder architecture with variable-rate generation in the LLM's semantic embedding space (Figure 2). The key design principle is to decouple sequence length from temporal duration, adjusting representation granularity to match token-level semantics. SEAM consists of: (i) a frozen Whisper encoder providing speech representations, (ii) a trainable alignment module with cross-attention learning variable-rate mappings, and (iii) a frozen LLM performing speech-conditioned generation. This modular design preserves pre-trained knowledge while enabling efficient cross-modal alignment learning.

3.2 Training Strategy

SEAM employs a multi-stage training approach that progressively refines the model through three distinct stages, separating competing optimization objectives while preserving pre-trained knowledge.

Stage 1 (Speech-conditioning alignment): We train the alignment module on speech recognition data to learn cross-modal correspondence between frozen Whisper encoder outputs and frozen Qwen3 LLM token embeddings. This stage focuses on learning to map speech representations into the LLM's semantic embedding space through variable-rate generation, addressing both representation space differences and sequence length inconsistencies. By training to predict embeddings that match the LLM's token embedding space, the alignment module learns semantic-level correspondences rather than simple acoustic-to-text mappings.

Stage 2 (Instruction tuning): We apply

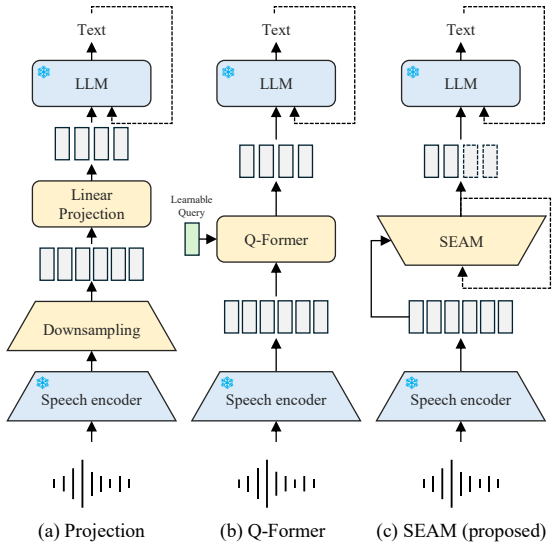

Figure 3: Comparison of speech-LLM integration approaches: (a) Projection-based methods use linear transformations with fixed downsampling; (b) Q-Former approaches use fixed learnable queries creating duration-based representations; (c) SEAM (proposed) uses encoder-decoder architecture enabling variable-rate generation (dashed lines indicate auto-regressive generation).

parameter-efficient fine-tuning using LoRA to adapt the LLM for speech-conditioned text generation with instruction-following capabilities. Task instructions are formatted using Qwen3's chat template to establish the conversational context for ASR tasks. LoRA adaptation enables efficient fine-tuning while preserving the pre-trained LLM's linguistic knowledge.

Stage 3 (Refinement tuning): To address the length-dependent alignment patterns that emerge from teacher-forcing training in Stage 2, we introduce refinement tuning that regularizes the model's reliance on strict length correspondence. We apply random masking and deletion operations to the aligned embeddings generated by the Stage-1 module, forcing the model to handle variable-length prefix sequences. This stage uses transcriptions generated by Stage-1's nearest-neighbor auto-regressive decoding as training data, enabling the model to learn robust corrections of Stage-1 outputs.

3.3 Framework Overview

Given speech features $\mathbf{X} \in \mathbb{R}^{T \times F}$ and text sequence $\mathbf{Y} = [y_1, \ldots, y_N]$, SEAM learns to map speech to aligned embeddings. The Whisper encoder produces speech representations $\mathbf{S} \in \mathbb{R}^{T_s \times d_{\text{speech}}}$, and frozen Qwen3 tokenizer embeddings provide target text embeddings $\mathbf{E} \in \mathbb{R}^{N \times d_{\text{llm}}}$.

During training, the alignment module uses both modalities to learn cross-modal correspondence:

$$\hat{\mathbf{E}} = f_\theta(\mathbf{S}, \mathbf{E}), \tag{1}$$

where f_θ consists of trainable projection layers and a trainable decoder with cross-attention. The alignment module learns to predict embeddings $\hat{\mathbf{E}} \in \mathbb{R}^{N \times d_{\text{llm}}}$ that closely match the target embeddings $\mathbf{E}$ in the LLM's semantic token embedding space. During inference, auto-regressive generation is employed as detailed in Section 3.6.

3.4 Architecture Design

SEAM employs an encoder-decoder architecture where the frozen Whisper-large-v2 encoder provides speech representations and a trainable decoder with cross-attention learns to align these representations with the frozen Qwen3 LLM's semantic token embedding space.

Speech Encoder. We utilize the pre-trained Whisper-large-v2 encoder in a frozen state, which processes 80-dimensional log-mel spectrograms through convolutional layers followed by Transformer layers. The encoder outputs speech representations $\mathbf{S} \in \mathbb{R}^{T_s \times d_{\text{speech}}}$ where $T_s = \lfloor T/2 \rfloor$ due to the 2× downsampling in the convolutional layers and $d_{\text{speech}} = 1280$.

Alignment Module. The alignment module consists of a Transformer decoder following Whisper-large-v2's architecture (32 layers, 20 heads, dimension 1280) and an output projection layer. The decoder is randomly initialized and trained from scratch in Stage 1. Since the Whisper encoder output dimension matches the decoder dimension ($d_{\text{speech}} = d_{\text{dec}} = 1280$), the decoder directly receives speech representations $\mathbf{S} \in \mathbb{R}^{T_s \times 1280}$. The decoder outputs are projected to the LLM dimension through $\mathbf{W}_{\text{out}} \in \mathbb{R}^{1280 \times d_{\text{llm}}}$ where $d_{\text{llm}} = 2560$ for Qwen3-4B.

The cross-attention mechanism enables variable-rate alignment and addresses the temporal-semantic granularity gap by using queries from text embeddings and keys/values from speech features. Target text embeddings $\mathbf{E} \in \mathbb{R}^{N \times d_{\text{llm}}}$ from the Qwen3 tokenizer are first projected to the decoder dimension: $\mathbf{E}' = \mathbf{E}\mathbf{W}_{\text{E-proj}}$ where $\mathbf{W}_{\text{E-proj}} \in \mathbb{R}^{d_{\text{llm}} \times 1280}$. The cross-attention then computes:

$$\mathbf{Q} = \mathbf{E}'\,\mathbf{W}_Q, \quad \mathbf{K} = \mathbf{S}\,\mathbf{W}_K, \quad \mathbf{V} = \mathbf{S}\,\mathbf{W}_V, \tag{2}$$

where $\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V \in \mathbb{R}^{1280 \times d_h}$ are learned attention parameters with head dimension $d_h = 64$.

This mechanism adaptively aggregates acoustic information from local temporal spans in $\mathbf{S}$ to produce token-aligned representations, which are then projected to the LLM's semantic embedding space via $\mathbf{W}_{\text{out}}$.

Attention Mechanism. The decoder employs attention masks to maintain the auto-regressive property essential for generation tasks. This design choice enables effective integration with LLM generation during inference while preserving the learned alignment patterns.

3.5 Training Objectives

Stage 1 (Speech-conditioning alignment). We train only the alignment module while freezing the Whisper encoder and LLM token embeddings to learn mapping into the LLM's semantic embedding space. The alignment objective combines mean squared error (MSE) and cosine similarity to minimize the distance between predicted embeddings $\hat{\mathbf{E}}$ and target LLM token embeddings $\mathbf{E}$, effectively teaching the module to operate within the semantic space where the LLM represents linguistic concepts:

$$\mathcal{L}_{\text{stage1}} = \lambda \|\hat{\mathbf{E}} - \mathbf{E}\|_2^2 + (1-\lambda)(1 - \cos(\hat{\mathbf{E}}, \mathbf{E})) \quad (3)$$

where $\cos(\hat{\mathbf{E}}, \mathbf{E})$ denotes the cosine similarity between $\hat{\mathbf{E}}$ and $\mathbf{E}$, and $\lambda = 0.5$ equally balances magnitude matching (MSE) and directional alignment (cosine similarity). By learning to predict embeddings in the LLM's semantic space, the alignment module acquires representations that naturally align with the LLM's understanding of linguistic semantics.

Stage 2 (Instruction tuning). We apply parameter-efficient fine-tuning using LoRA to adapt the LLM for speech-conditioned text generation with instruction-following capabilities. The training follows a conversation format where the task instruction and aligned speech representations serve as context, and loss is computed only on the assistant's transcription response:

$$\mathcal{L}_{\text{stage2}} = -\frac{1}{N} \sum_{i=1}^{N} \log P(y_i^{\text{assistant}} | y_{<i}^{\text{assistant}}, \mathcal{T}, \hat{\mathbf{E}}) \quad (4)$$

where $y_i^{\text{assistant}}$ represents tokens in the assistant's transcription response, $\mathcal{T}$ is the task instruction "Transcribe the following audio:", and $\hat{\mathbf{E}}$ are aligned speech representations. The instruction tokens are provided as context but excluded from loss computation, enabling the model to learn speech-conditioned generation while following conversational protocols.

Stage 3 (Refinement tuning). To reduce the model's reliance on exact length correspondence, we apply refinement tuning with perturbed embeddings. Given the Stage-1 generated embeddings $\hat{\mathbf{E}}^{\text{stage1}}$ through nearest-neighbor auto-regressive decoding, we apply random masking with probability p_m and deletion with probability p_d to create perturbed sequences $\tilde{\mathbf{E}}$. The training objective remains teacher-forcing but with perturbed inputs:

$$\mathcal{L}_{\text{stage3}} = -\frac{1}{N} \sum_{i=1}^{N} \log P(y_i^{\text{gt}} | y_{<i}^{\text{gt}}, \mathcal{T}, \tilde{\mathbf{E}}) \quad (5)$$

where y_i^{gt} represents ground-truth transcription tokens and $\tilde{\mathbf{E}}$ are the perturbed embeddings from Stage-1 outputs. This training encourages the model to correct imperfect Stage-1 alignments and handle variable-length mismatches, improving robustness to alignment errors during inference.

3.6 Auto-Regressive Inference

SEAM employs auto-regressive generation for both Stage 1 evaluation and complete model inference. Due to the continuous nature of predicted embeddings $\hat{\mathbf{E}}$, directly feeding them back as input would accumulate estimation errors over generation steps. To address this issue, we use nearest neighbor lookup in the LLM token embedding space to find discrete tokens corresponding to predicted embeddings, then use the embeddings of these discrete tokens for subsequent generation steps.

Stage 1 Auto-Regressive Evaluation. Algorithm 1 details the auto-regressive evaluation process for the Stage 1 alignment module. This approach enables WER measurement while preventing error accumulation from continuous embedding feedback.

Full Model Auto-Regressive Inference. After Stage 2 training, the complete SEAM model uses the same nearest neighbor approach to generate embeddings that serve as speech-conditioned context for the instruction-tuned LLM. Algorithm 2 details this two-stage auto-regressive process.

This multi-stage auto-regressive approach ensures alignment quality, instruction-following capability, and robustness to sequence-level variations while preventing error accumulation from continuous embedding feedback.

Algorithm 1 Stage-1 Auto-Regressive Nearest Neighbor Evaluation

Input: Speech features $\mathbf{X} \in \mathbb{R}^{T \times F}$, LLM vocabulary embeddings $\{\mathbf{E}_v\}_{v \in \mathcal{V}}$
Output: Generated token sequence $\mathbf{y}$
$\mathbf{S} \leftarrow \text{WhisperEnc}(\mathbf{X})$
$\mathbf{y} \leftarrow [\langle \text{BOA} \rangle]$
repeat
  $t \leftarrow |\mathbf{y}|$
  $\mathbf{E}_{\leq t} \leftarrow \text{Embed}(\mathbf{y})$
  $\hat{\mathbf{e}}_{t+1} \leftarrow \text{SEAM}(\mathbf{E}_{\leq t}, \mathbf{S})$
  $y_{t+1} \leftarrow \arg\min_{v \in \mathcal{V}} d(\hat{\mathbf{e}}_{t+1}, \mathbf{E}_v)$
  $\mathbf{y} \leftarrow \mathbf{y} \cup \{y_{t+1}\}$
until $y_{t+1} = \langle \text{EOA} \rangle$ or $|\mathbf{y}| > \text{max\_length}$
return $\mathbf{y}$

Algorithm 2 SEAM Full Model Auto-Regressive Inference (Stage 2)

Input: Speech features $\mathbf{X} \in \mathbb{R}^{T \times F}$
Output: Generated transcription text $\mathbf{y}^{\text{assistant}}$
$\mathbf{S} \leftarrow \text{WhisperEnc}(\mathbf{X})$
$\hat{\mathbf{E}} \leftarrow \text{Stage1AutoRegressive}(\mathbf{S})$
$\mathcal{T} \leftarrow$ "Transcribe the following audio:"
prompt $\leftarrow \text{FormatConversation}(\mathcal{T}, \hat{\mathbf{E}})$
$\mathbf{y}^{\text{assistant}} \leftarrow \text{LLM}(\text{prompt})$
return $\mathbf{y}^{\text{assistant}}$

3.7 First Token Guidance

Despite the multi-stage training approach, we observe that instruction-tuned LLMs exhibit lower accuracy in predicting the first assistant token when conditioned solely on speech-aligned embeddings without prior linguistic context. Unlike text-to-text scenarios where the LLM can leverage contextual information from preceding tokens, the initial token in speech transcription must be generated directly from continuous speech-conditioned embeddings. To address this challenge, we introduce **First Token Guidance (FTG)**, a technique that provides explicit guidance for initial token generation by leveraging Stage-1's nearest-neighbor prediction.

In FTG, we first obtain the predicted first token y_1^{pred} through Stage-1's nearest-neighbor decoding in the LLM vocabulary. We then insert the embedding of this predicted token immediately after the assistant marker, providing explicit guidance for the initial generation step. Formally, the LLM generates the assistant response as:

$$\mathbf{y}^{\text{assistant}} = \text{LLM}([\mathcal{T}, \hat{\mathbf{E}}^{\text{stage1}}, \langle \text{assistant} \rangle, \mathbf{E}(y_1^{\text{pred}})]) \tag{6}$$

where $\mathcal{T}$ is the task instruction, $\hat{\mathbf{E}}^{\text{stage1}}$ are the Stage-1 generated embeddings, $\langle \text{assistant} \rangle$ marks the beginning of assistant generation, and $\mathbf{E}(y_1^{\text{pred}})$ is the embedding of the predicted first token from Stage-1 nearest-neighbor lookup. The LLM then continues generation from this guided starting point, producing $y_2, y_3, \ldots$ conditioned on the provided first token. This modification improves first token accuracy and consequently overall transcription quality.

4 Experiments

4.1 Experimental Setup

We train on LibriSpeech (Panayotov et al., 2015), an English speech corpus containing 960 hours of read audiobook speech. For evaluation, we use LibriSpeech test sets and TED-LIUM-v2 (Rousseau et al., 2012) for cross-domain assessment. TED-LIUM-v2 contains spontaneous TED talk presentations, presenting a different domain from LibriSpeech's read speech. We compare against Whisper (Radford et al., 2023), SLAM-ASR (Ma et al., 2024), and SALMONN (Tang et al., 2023). Audio is resampled to 16kHz with 80-dim log-mel features using 25ms windows and 10ms hop. Speech representations come from the frozen Whisper-large-v2 encoder ($d_{\text{speech}} = 1280$), while text embeddings are obtained from the frozen Qwen3-4B-Instruct (Yang et al., 2025) tokenizer embedding table ($d_{\text{llm}} = 2560$).

4.2 Implementation Details

Table 2 summarizes the model architecture and training configuration. The alignment module follows Whisper-large-v2's decoder architecture (32 layers, 20 heads, dimension 1280) but is randomly initialized and trained from scratch in Stage 1. An output projection layer maps from Whisper's dimension (1280) to Qwen3-4B's embedding dimension (2560). For LoRA configuration in Stage 2, we apply low-rank adaptation with rank $r = 32$, scaling factor $\alpha = 64$, and dropout rate 0.1 to the query, key, value, output, gate, up, and down projection layers. In Stage 3, we apply masking probability $p_m = 0.05$ and deletion probability $p_d = 0.1$ to the Stage-1 generated embeddings. Training uses the AdamW optimizer with learning

Table 1: WER (Word Error Rate) (%) on LibriSpeech and TED-LIUM-v2. Whisper refers to Whisper-large-v2. All WER measurements use Whisper's EnglishTextNormalizer for text normalization. For TED-LIUM-v2, Whisper-large-v2 exhibits excessive hallucination; the value in parentheses represents Whisper-medium performance. FTG denotes First Token Guidance (Section 3.7).

Method	**test-clean**	**test-other**	**TED-LIUM-v2**
Whisper (Radford et al., 2023)	2.7	5.2	12.24(7.7)
SLAM-ASR (Ma et al., 2024)	**1.9**	**3.6**	8.8
SALMONN (Tang et al., 2023)	2.2	5.7	4.8
SEAM w/o LLM (Stage-1)	2.7	5.9	7.2
SEAM (Full, Stage-3, w/o FTG)	3.1	6.0	5.7
SEAM (Full, Stage-3, w/ FTG)	2.6	5.2	**4.7**

rate 5e-4, weight decay 0.01, gradient clipping at norm 1.0, and mixed-precision (bfloat16) across 6 NVIDIA RTX A6000 GPUs.

4.3 Evaluation

We evaluate using word error rate (WER) with two different decoding strategies depending on the evaluation objective:

Stage-1 Evaluation (SEAM w/o LLM). To assess the quality of the learned alignment between speech representations and the LLM's semantic token embedding space, we evaluate the Stage-1 model using autoregressive nearest neighbor lookup. This approach directly measures whether the alignment module produces embeddings that are geometrically close to the target token embeddings in the LLM's semantic embedding space.

The distance metric combines MSE and cosine similarity, matching the Stage-1 training objective (Equation 3). We define the alignment distance function:

$$d(\hat{\mathbf{e}}, \mathbf{E}_v) = \lambda \|\hat{\mathbf{e}} - \mathbf{E}_v\|_2^2 + (1-\lambda)(1-\cos(\hat{\mathbf{e}}, \mathbf{E}_v)) \tag{7}$$

where $\lambda = 0.5$ equally balances magnitude and directional alignment. For each predicted embedding $\hat{\mathbf{e}}_i$, we find the nearest neighbor token in the LLM vocabulary $\mathcal{V}$:

$$y_i = \arg\min_{v \in \mathcal{V}} d(\hat{\mathbf{e}}_i, \mathbf{E}_v) \tag{8}$$

Full Model Evaluation (SEAM). For the complete SEAM model after Stage 3 training, evaluation follows the multi-stage auto-regressive inference described in Section 3.6. We use appropriate masking for variable-length sequences to ensure consistent evaluation.

4.4 Results

Variable-Rate Generation Analysis. To understand the problem with duration-based methods, we analyze embedding rates on the LibriSpeech test set (5,559 utterances). When text transcriptions are tokenized, LibriSpeech averages 2.71 tokens per second. However, duration-based speech representations differ significantly: simple downsampling (k=5 in SLAM-ASR) produces 10.0 embeddings per second regardless of semantic content, creating a 3.3× redundancy. This granularity mismatch forces LLMs to process uniformly-paced inputs that deviate from their pre-training distribution where sequence length naturally correlates with semantic content. SEAM addresses this through variable-rate generation: using Stage-1 nearest neighbor evaluation, SEAM produces embeddings averaging 2.99 tokens/sec, closely matching the natural text rate and decoupling sequence length from temporal duration.

ASR Performance and Cross-Domain Generalization. Table 1 presents WER on LibriSpeech test-clean/test-other (ASR performance) and TED-LIUM-v2 (domain robustness). SEAM w/o LLM (Stage-1), using only the alignment module with nearest neighbor lookup, achieves 2.7% WER on test-clean and 5.9% WER on test-other, showing the effectiveness of variable-rate alignment even without LLM refinement. The full SEAM model with Stage-3 refinement tuning achieves 3.1% and 6.0% WER respectively.

Incorporating First Token Guidance (FTG, Section 3.7) yields performance improvements, achieving 2.6% WER on test-clean and 5.2% WER on test-other. On LibriSpeech, SEAM with FTG achieves performance comparable to the baseline Whisper-large-v2 encoder (2.7%/5.2%) when trained on 960 hours of LibriSpeech data. On cross-domain evaluation, SEAM with FTG achieves 4.7% WER on TED-LIUM-v2, which compares favorably to the baseline Whisper encoder (7.7% for Whisper-medium) and SLAM-ASR (8.8%), and is

Table 2: SEAM model parameters and training configuration.

Component	Params	Stage 1	Stage 2
Whisper Encoder	629M	Frozen	Frozen
Whisper Decoder	839M	Trainable	Frozen
Projection Layers	72M	Trainable	Frozen
Qwen3-4B	4,058M	Frozen*	Frozen
LoRA Adapters (r=32)	62M	–	Trainable
Trainable	–	**911M**	**62M**

*Only embeddings (389M) used in Stage 1.

competitive with SALMONN (4.8%) which was trained on considerably more data (4400h vs 960h). This result suggests that integrating LLM's linguistic knowledge through our multi-stage training approach enables effective domain generalization, even when trained exclusively on LibriSpeech. The improvement from FTG demonstrates that providing explicit guidance for initial token generation facilitates the effective use of the LLM's pre-trained knowledge.

5 Conclusion

We have presented SEAM, a Speech Encoder-Decoder Alignment Module that bridges the temporal-semantic granularity gap in speech-LLM integration through variable-rate generation in a shared embedding space. This granularity gap—the mismatch between temporal resolution of speech representations and semantic density of text tokens—parallels similar challenges in vision-language research where frame-level features must align with sentence-level semantics (Jun et al., 2025). By adjusting representation granularity from frame-level temporal resolution to token-level semantic units through variable-rate generation, SEAM addresses the distributional mismatch that limits fixed-rate duration-based approaches (Ma et al., 2024; Tang et al., 2023), achieving competitive performance on speech recognition tasks.

Our experimental results show several key advantages: (1) variable-rate generation that maintains distributional consistency with LLM pre-training, producing embeddings that closely match natural text distributions (2.99 tokens/sec vs 2.71 tokens/sec for natural text) compared to fixed-rate duration-based methods (Ma et al., 2024; Tang et al., 2023); (2) efficient cross-modal alignment learning through a multi-stage training approach that preserves pre-trained knowledge and addresses length-dependency issues through refinement tuning; (3) First Token Guidance (FTG) that addresses the LLM's difficulty in initial token prediction by inserting the predicted first assistant token into the prompt, yielding performance improvements (2.6%/5.2% WER on LibriSpeech test-clean/test-other, matching baseline Whisper performance); and (4) effective cross-domain generalization that leverages LLM's linguistic knowledge: trained exclusively on LibriSpeech (960h), SEAM achieves 4.7% WER on TED-LIUM-v2, which compares favorably to the baseline Whisper encoder (7.7%) and SLAM-ASR (8.8%), and is competitive with SALMONN (4.8%) which was trained on 4400h of data. This result suggests that integrating LLM's pre-trained linguistic knowledge through our multi-stage training approach enables robust generalization to unseen domains, effectively leveraging extensive language modeling capabilities to compensate for limited speech training data.

Future work will include extending SEAM to various speech-LLM tasks beyond ASR, such as speech translation, speech summarization, and speech question answering. Since SEAM learns generalizable speech-text representations through variable-rate alignment while preserving LLM's pre-trained linguistic knowledge, we expect the model to adapt more easily to diverse speech-language tasks and demonstrate robust performance across different domains with minimal domain-specific training data. The modular design of SEAM also enables exploration of different alignment architectures and training strategies for various speech-language integration scenarios. Furthermore, investigating how First Token Guidance can be extended to multi-turn dialogue and conversational speech scenarios presents an interesting direction for future research.

Limitations

This work has several limitations. First, our evaluation is limited to English speech recognition tasks, and generalization to other languages requires further investigation. Second, the current approach relies on specific pre-trained models (Whisper-large-v2 and Qwen3-4B), and adaptation to other model combinations needs systematic study. Third, the multi-stage auto-regressive inference process introduces computational overhead: the Stage 1 alignment module first performs auto-regressive generation with nearest neighbor lookup to produce speech-conditioned embeddings, followed by the LLM's own auto-regressive generation for the final transcription. This dual auto-regressive process increases inference latency compared to single-stage approaches, limiting real-time applications. Fourth, computational efficiency during inference, particularly for real-time applications, requires further optimization beyond the inherent multi-stage bottleneck. Fifth, there exists a potential concern regarding test-set contamination from the LLM's pre-training corpus: Qwen3 may have been exposed to text transcriptions from public speech datasets during pre-training. However, SEAM's Stage-1 alignment module, which performs nearest-neighbor decoding without LLM generation, achieves competitive WER (2.7%/5.9%), demonstrating that the learned speech-to-embedding alignment is effective independent of potential LLM memorization. Finally, the effectiveness of the multi-stage training approach, particularly the impact of Stage 3 refinement tuning with masking and deletion, deserves more comprehensive analysis across diverse datasets and domains.

Acknowledgments

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (RS-2022-II220621, Development of artificial intelligence technology that provides dialog-based multi-modal explainability).

References

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language Models are Few-Shot Learners. *Advances in Neural Information Processing Systems*, 33:1877–1901.

William Chan, Navdeep Jaitly, Quoc Le, and Oriol Vinyals. 2016. Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. In *Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4960–4964.

Wenqian Cui, Dianzhi Yu, Xiaoqi Jiao, Ziqiao Meng, Guangyan Zhang, Qichao Wang, Steven Y. Guo, and Irwin King. 2025. Recent Advances in Speech Language Models: A Survey. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13943–13970.

Keqi Deng and Philip C. Woodland. 2024. Decoupled Structure for Improved Adaptability of End-to-End Models. *Speech Communication*, 163:103109.

Qingkai Fang and Yang Feng. 2023. Understanding and Bridging the Modality Gap for Speech Translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pages 15864–15881, Toronto, Canada.

Yassir Fathullah, Chunyang Wu, Egor Lakomkin, Ke Li, Junteng Jia, Yuan Shangguan, Jay Mahadeokar, Ozlem Kalinli, Christian Fuegen, and Mike Seltzer. 2024. AudioChatLlama: Towards General-Purpose Speech Abilities for LLMs. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5522–5532.

Alex Graves and Navdeep Jaitly. 2014. Towards End-To-End Speech Recognition with Recurrent Neural Networks. In *Proceedings of the 31st International Conference on Machine Learning*, pages 1764–1772.

HyoJung Han, Kevin Duh, and Marine Carpuat. 2024. SpeechQE: Estimating the Quality of Direct Speech Translation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21852–21867.

Yukiya Hono, Koh Mitsuda, Tianyu Zhao, Kentaro Mitsui, Toshiaki Wakatsuki, and Kei Sawada. 2024. Integrating Pre-Trained Speech and Language Models for End-to-End Speech Recognition. In *Findings of the Association for Computational Linguistics: 2024*, pages 13289–13305.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*.

Rongjie Huang, Mingze Li, Dongchao Yang, Jiatong Shi, Xuankai Chang, Zhenhui Ye, Yuning Wu, Zhiqing Hong, Jiawei Huang, Jinglin Liu, Yi Ren,

Yuexian Zou, Zhou Zhao, and Shinji Watanabe. 2024. AudioGPT: Understanding and Generating Speech, Music, Sound, and Talking Head. In *Proceedings of the 38th AAAI Conference on Artificial Intelligence*, pages 23802–23804.

WooJin Jun, WonJun Moon, Cheol-Ho Cho, MinSeok Jung, and Jae-Pil Heo. 2025. Bridging the Semantic Granularity Gap Between Text and Frame Representations for Partially Relevant Video Retrieval. In *Proceedings of the 39th AAAI Conference on Artificial Intelligence*, pages 4166–4174.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. In *Proceedings of the 40th International Conference on Machine Learning*, pages 19730–19742.

Shaoshi Ling, Yuxuan Hu, Shuangbei Qian, Guoli Ye, Yao Qian, Yifan Gong, Ed Lin, and Michael Zeng. 2024. Adapting Large Language Model with Speech for Fully Formatted End-to-End Speech Recognition. In *Proceedings of the 2024 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 11046–11050.

Ziyang Ma, Guanrou Yang, Yifan Yang, Zhifu Gao, Jiaming Wang, Zhihao Du, Fan Yu, Qian Chen, Siqi Zheng, Shiliang Zhang, and Xie Chen. 2024. An Embarrassingly Simple Approach for LLM with Strong ASR Capacity. *Preprint*, arXiv:2402.08846.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: An ASR Corpus Based on Public Domain Audio Books. In *Proceedings of the 2015 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 5206–5210.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine Mcleavey, and Ilya Sutskever. 2023. Robust Speech Recognition via Large-Scale Weak Supervision. In *Proceedings of the 40th International Conference on Machine Learning*, pages 28492–28518.

Anthony Rousseau, Paul Deléglise, and Yannick Estève. 2012. TED-LIUM: An Automatic Speech Recognition Dedicated Corpus. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation*, pages 125–129.

Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, and Chao Zhang. 2023. SALMONN: Towards Generic Hearing Abilities for Large Language Models. In *The Twelfth International Conference on Learning Representations*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 Technical Report. *Preprint*, arXiv:2505.09388.

Wenyi Yu, Changli Tang, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, and Chao Zhang. 2024. Connecting Speech Encoder and Large Language Model for ASR. In *Proceedings of the 2024 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 12637–12641.

Yuhao Zhang, Zhiheng Liu, Fan Bu, Ruiyu Zhang, Benyou Wang, and Haizhou Li. 2025. Soundwave: Less is More for Speech-Text Alignment in LLMs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, pages 18718–18738.