

RASST: Retrieval-Augmented Simultaneous Speech Translation

Jiaxuan Luo¹, Siqi Ouyang², Jiaxing Xu², Lei Li²,
¹Johns Hopkins University, ²Carnegie Mellon University
jluo50@jhu.edu, siqiouya@andrew.cmu.edu

Abstract

Simultaneous speech translation produces target text incrementally from partial speech input. Recent speech large language models have markedly improved SST quality but still struggle with rare and domain-specific terminology. Retrieval augmentation has helped in automatic speech recognition and neural machine translation, but extending it to SST is non-trivial: retrieval must be fast and accurate under partial speech, and the model must decide whether and when to apply retrieved terms during incremental generation. We propose **Retrieval-Augmented Simultaneous Speech Translation (RASST)**, which addresses both challenges. For accurate cross-modal retrieval under partial input, RASST trains a lightweight speech-text retriever that produces chunkwise terminology hints for the Speech LLM via multi-scale retrieval. To use these hints correctly, we synthesize training data that teaches the Speech LLM to decide whether and when to apply each retrieved term. Experiments on the ACL 60/60 development set and the ESO oncology test set show that RASST improves terminology accuracy over InfiniSST by 12.3–31.22% and BLEU by up to 3 points across all tested languages and latency settings. Furthermore, the retriever incurs negligible computational overhead and remains robust as glossary size increases to 10K entries.

1 Introduction

Simultaneous Speech Translation (SST) aims to generate translations incrementally as partial speech input is received. It has wide applications in international conferences and cross-lingual conversations (Ma et al., 2020b; Ren et al., 2020). Recently, Large Language Models (LLMs) have shown strong potential as backbones for SST systems (Wang et al., 2025; Ouyang et al., 2025b; Fu et al., 2025; Cheng et al., 2025). However, they still struggle to accurately translate domain-specific terminology, including technical jargon, person and

Source: Shinzo Abe served as the Prime Minister of Japan.

Reference: 安倍晋三曾担任日本首相。

Plain SST: 新三.阿贝曾担任日本首相。

RASST: 安倍晋三曾担任日本首相。

Table 1: A case showing plain SST without retrieval fails to translate terminology correctly while SST with retrieval is able to do so. English (blue), correct Chinese terms (green), and incorrect ones (red).

location names, and abbreviations. As shown in Table 1, a plain SST system translates "Shinzo Abe" incorrectly as "新三.阿贝", which is a phonetic transliteration rather than the official translation of the politician's name.

Human simultaneous interpreters often rely on external glossaries to translate specialized terminology in real time (Jiang, 2013), and retrieval augmentation has been explored widely in automatic speech recognition (ASR) and neural machine translation (NMT) to improve terminology handling as the glossary size grows (Khandelwal et al., 2021; Huber et al., 2021; Jayanthi et al., 2023; Li et al., 2024, 2025; Wu et al., 2025; Gong et al., 2025; Kong et al., 2025). However, retrieval-augmented SST poses unique challenges. Beyond requiring a cross-modal retriever that is fast and accurate under partial speech input, SST translations, unlike ASR transcripts which are time-aligned with the speech, lag behind the speech. The model must therefore decide not only whether to use a retrieved term, but also when to apply it during incremental translation.

To address these challenges, we propose **Retrieval-Augmented Simultaneous Speech Translation (RASST)**. RASST trains a lightweight speech-text retriever on variable-duration speech contexts with a contrastive loss, and performs multi-scale inference for accurate retrieval under partial input. We further synthesize training data

that teaches the Speech LLM to select relevant retrieved terms and decide when to introduce them during generation, conditioned on the prior speech context and previously generated translation. Across both the ACL 60/60 development set and the ESO oncology test set, RASST consistently outperforms InfiniSST in every language and latency regime evaluated, yielding gains of 12.3–31.22% in terminology accuracy and up to 3 BLEU points. Its retrieval module adds negligible computational cost and scales robustly to glossaries containing 10K entries. Component-wise ablations further confirm the contribution of each part of RASST.

2 Related Works

SST with LLM Recent work shows that simultaneous speech translation (SST) achieves markedly higher quality when large language models (LLMs) serve as the backbone (Ahmad et al., 2024). Koshkin et al. (2024) adapt an LLM to SST via finetuning on synthetic translation trajectories, and Ouyang et al. (2025b) scale this idea with an interleaving architecture for unbounded speech streams. Fu et al. (2025) explore alternative training strategies and data-synthesis pipelines, while Guo et al. (2025) unify streaming transcription and translation with a truncation mechanism that prunes history based on transcribed inputs. However, these systems still struggle with domain-specific terminology and RASST addresses this by integrating a terminology retriever directly into the SST process.

Retrieval-Augmented NMT Retrieval is an effective remedy for translating rare words, named entities, and domain-specific terminology in NMT. Early approaches retrieve translation pieces from similar training pairs (Zhang et al., 2018), mine bilingual terminology with bilingual embeddings (Huck et al., 2019), or fuse dictionary entries into NMT via a Pointer-Disambiguator-Copier (Zhang et al., 2021). kNN-MT (Khandelwal et al., 2021) augments NMT with nearest-neighbor lookup over a large datastore, and follow-up work improves its robustness (Jiang et al., 2021) and efficiency (Meng et al., 2022; Zhu et al., 2023). Other approaches extract and attend over entity candidates (Zeng et al., 2023) or combine retrieval and demonstration with LLMs (Li et al., 2025). For speech translation, Li et al. (2024) prepend retrieved speech segments as demonstrations, and Wu et al. (2025) use sliding-window retrieval to inject

terms into a speech-LLM prompt. All of these target offline translation, whereas RASST integrates retrieval directly into incremental streaming translation.

Contextual-Biasing ASR Recognizing rare words and domain-specific terminology, commonly known as contextual biasing, is also well studied in ASR. Early work splits into two lines: joint optimization of the ASR model with contextual n-gram embeddings (Pundak et al., 2018; Jain et al., 2020), and shallow fusion of biased words into end-to-end models (Zhao et al., 2019), later extended with prefix-trie-based deep biasing (Le et al., 2021a) and unified via deep shallow fusion and trie-constrained biasing (Le et al., 2021b). Huber et al. (2021) support runtime addition of context words through an external memory accessed at decoding. To scale biasing lists from hundreds to tens of thousands of entries, subsequent work leverages neural associative memories (Munkhdalai et al., 2022, 2023; Wu et al., 2023, 2024) and retrieval augmentation (Li et al., 2021; Yusuf et al., 2023; Jayanthi et al., 2023). More recent efforts build on speech foundation models. Yang et al. (2024) use coarse CTC decoding to filter hotwords into an LLM-based ASR prompt, and Gong et al. (2024) inject bias words into a SpeechLLM via prompt formats and a biasing-fusion network. BR-ASR (Gong et al., 2025) decouples bias retrieval from ASR decoding through contrastive learning and curriculum training, scaling to 200k entries at 20ms retrieval latency, while GLCLAP (Kong et al., 2025) retrieves bias words directly from speech via global-local contrastive language-audio pre-training, enabling plug-in biasing on frozen models like Whisper. Finally, PAC (Fu et al., 2026) combines pronunciation-guided context learning with pronunciation-discriminative reinforcement learning. Unlike ASR, where the transcript is time-aligned with the speech, SST translations lag behind the input, so the model must therefore decide whether and when to apply each retrieved term.

3 Problem Formulation

Let $\mathbf{s} = (s_1, \dots, s_T) \in \mathbb{R}^T$ denote a waveform containing T samples, recorded at a sampling rate of f_s Hz. Let l be the chunk duration in seconds and $L = lf_s$ the corresponding number of samples, assumed to be an integer. At each step

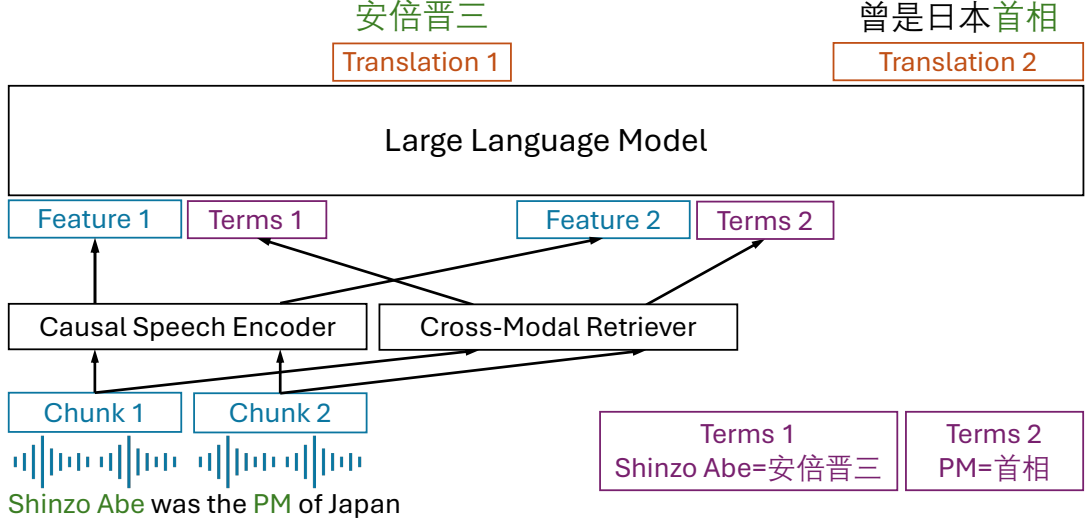


Figure 1: Overview of RASST. RASST interleaves speech chunks, retrieved terms, and translation outputs. For each incoming speech chunk, a causal speech encoder produces its feature representation, and a cross-modal retriever returns the corresponding terms. The retrieved terms are appended to the speech features and passed to the large language model, which decodes the next translation segment.

$i \in \{1, \dots, I\}$, the system receives the fixed-duration speech chunk

$$\mathbf{c}_i = (s_{(i-1)L+1}, \dots, s_{iL}), \quad (1)$$

where $I = T/L$, assuming that T is divisible by L . We denote the speech observed through step i by $\mathbf{c}_{\leq i} = (\mathbf{c}_1, \dots, \mathbf{c}_i)$.

A bilingual glossary is defined as $\mathcal{G} = \{(e_j, \tau_j)\}_{j=1}^M$, where e_j is a source-language term¹ and τ_j is its target-language translation. At each step i , the retriever R_ϕ , parameterized by ϕ , selects a subset of the glossary:

$$G_i = R_\phi(\mathbf{c}_{\leq i}, \mathcal{G}) \subseteq \mathcal{G}. \quad (2)$$

We denote the retrieval history through step i by $G_{\leq i} = (G_1, \dots, G_i)$.

Let $\hat{\mathbf{y}}_i = (\hat{y}_1^i, \dots, \hat{y}_{n_i}^i)$ denote the possibly empty target-language segment emitted at step i by the system, and let $\hat{\mathbf{y}}_{< i} = \hat{\mathbf{y}}_1 \oplus \dots \oplus \hat{\mathbf{y}}_{i-1}$ denote the previously emitted translation prefix. The generation policy is

$$\hat{\mathbf{y}}_i \sim \pi_\theta(\cdot \mid \mathbf{c}_{\leq i}, \hat{\mathbf{y}}_{< i}, G_{\leq i}). \quad (3)$$

If $n_i = 0$, the policy waits for additional speech input. Otherwise, the generated segment is appended to the translation output. Each token emitted at step

¹We use *term* to refer to a specialized word, phrase, or expression associated with a particular profession, scientific discipline, or domain. Examples include technical concepts, method and dataset names, proper names, and acronyms.

i is assigned a delay of il seconds. The complete hypothesis is $\hat{\mathbf{y}} = \hat{\mathbf{y}}_1 \oplus \dots \oplus \hat{\mathbf{y}}_I$. We evaluate translation quality and latency using the hypothesis $\hat{\mathbf{y}}$, reference translation $\mathbf{y}^*$, source transcript $\mathbf{x}^*$, and per-token emission delays.

4 Retrieval-Augmented Simultaneous Speech Translation

In this section, we first present an overview of RASST. We then describe the cross-modal retriever design and the training procedure that enables the Speech LLM to effectively leverage retrieved terms during simultaneous translation.

4.1 RASST Overview

Figure 1 presents an overview of RASST. We adopt the interleaving SST formulation of In-finiSST (Ouyang et al., 2025b), which alternates between incoming speech chunks and incremental translation outputs, and RASST augments this with a retrieval step:

$$\hat{\mathbf{y}}_i \sim \text{SpeechLLM}(\mathbf{c}_1, G_1, \hat{\mathbf{y}}_1, \dots, \mathbf{c}_i, G_i). \quad (4)$$

After each chunk $\mathbf{c}_i$ arrives, the retriever returns candidate terms G_i likely to appear in it, and we append these terms along with their target-language translations to the chunk’s representation. Conditioned on the speech context and the retrieved terms, the LLM then generates the next partial translation $\hat{\mathbf{y}}_i$.

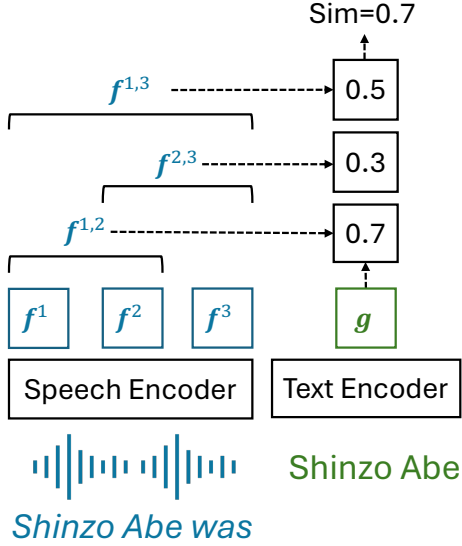


Figure 2: An example of multi-scale inference. The text term is encoded into a single feature vector g , while the speech context is encoded into frame features f^1, f^2, f^3 . We enumerate all windows $(1, 2), (2, 3), (1, 3)$ and compute the corresponding average-pooled features $f^{1,2}, f^{2,3}, f^{1,3}$. The likelihood that the term appears in the speech context is then the maximum cosine similarity between g and any window feature.

4.2 Cross-Modal Retriever

Architecture We adopt a dual-encoder architecture. For text, we use the pre-trained BGE-M3 encoder (Chen et al., 2024) to map each source-language term e_j to a feature vector $g_j \in \mathbb{R}^d$. For speech, we use the Audio Transformer of Qwen3-Omni (Xu et al., 2025) followed by a linear projection to the same dimension d . Since terms are typically short, the speech encoder reads only the current chunk c_i together with a look-back window of duration t_{lookback} , rather than the full history $s_{\leq i}$ as in Equation 2. We denote this input as c'_i and the resulting features as $f_i \in \mathbb{R}^{m \times d}$, with m as the number of frames in c'_i .

Multi-Scale Inference Because terms vary in duration, representing the speech context with a single average-pooled vector loses information. Instead, we pool over windows at multiple scales. As shown in Figure 2, given $f_i = (f_i^1, f_i^2, \dots, f_i^m)$, for each window $1 \leq u < v \leq m$ we compute

$$f_i^{u,v} = \text{AvgPool}(f_i^u, \dots, f_i^v) \in \mathbb{R}^d. \quad (5)$$

The similarity between the speech c'_i and a term e_j is then the maximum cosine similarity over all

windows:

$$\text{Sim}(c'_i, e_j) = \max_{1 \leq u < v \leq m} \text{CosSim}(f_i^{u,v}, g_j), \quad (6)$$

and the retrieved terms are

$$G_i = \text{Top-}K_j(\text{Sim}(c'_i, e_j)) \subseteq \mathcal{G} \quad (7)$$

In practice, we build a FAISS index (Douze et al., 2024) over the term vectors $\{g_j\}_{j=1}^{|\mathcal{G}|}$, then issue a Top- K query for each window feature $f_i^{u,v}$, producing a candidate pool of size up to a multiple of K . We perform a second Top- K selection over this pool by similarity score, keeping the highest score per term when duplicates occur, and finally discard any candidate whose similarity falls below a threshold τ to suppress noise.

Data Synthesis Retriever training requires paired speech and terminology data (s', e_{pos}) . Since terminology-rich speech corpora are scarce, we synthesize training data from two sources: existing ASR datasets and a terminology database.

For ASR data, we obtain word-level timestamps for each transcript using the Montreal Forced Aligner (MFA)², then extract noun phrases as candidate terms using spaCy³ and normalize their surface forms. We use noun phrases because most domain-specific terminology takes this form and they are abundant in standard transcripts. For each term, we carve out a speech context s' of variable duration surrounding it, which gives the retriever flexibility across context lengths at inference time.

For each term in the terminology database, we prompt Gemini-2.5-Flash⁴ to generate several sentences containing the term and synthesize the speech with CosyVoice 3 (Du et al., 2025). The resulting speech is then processed by the same MFA pipeline as the ASR data.

Training Given paired data (s', e_{pos}) , we train the retriever with the InfoNCE loss (van den Oord et al., 2019). In addition to in-batch negatives, we mine hard negatives by using the retriever checkpoint, which is updated every 50 training steps, to retrieve the Top- N most similar terms from the entire training set for each positive e_{pos} . Let E_{neg} be the negative set and $E_{\text{all}} = \{e_{\text{pos}}\} \cup E_{\text{neg}}$. The

²<https://github.com/MontrealCorpusTools/Montreal-Forced-Aligner>

³<https://spacy.io/> and en_core_web_trf model

⁴<https://ai.google.dev/gemini-api/docs/models/gemini-2.5-flash>

retriever loss $\mathcal{L}_{\text{retriever}}$ is then

$$-\log \frac{\exp(\text{Sim}(\mathbf{c}'_i, e_{\text{pos}})/\gamma)}{\sum_{e \in E_{\text{all}}} \exp(\text{Sim}(\mathbf{c}'_i, e)/\gamma)}, \quad (8)$$

where γ is the temperature. Since the timestamp of e_{pos} within s' is known during training through MFA, we bypass the maximum over windows in $\text{Sim}(\cdot)$ and instead set (u, v) to be the smallest window enclosing e_{pos} .

4.3 Speech LLM Training

To teach the Speech LLM to effectively use the retrieved terms G_i during simultaneous translation, we synthesize a corresponding training corpus. Given the speech chunks $c_{1:i}$, we follow InfiniSST’s procedure to produce the translation segments $\hat{y}_{\leq i}$. The remaining piece beyond InfiniSST is the retrieved terms $G_{\leq i}$. A naive approach is to reuse the noun-phrase terms from retriever training and always populate G_i with the ground-truth term. This, however, would leave the model brittle to retriever mistakes at inference time. To reduce this training–inference mismatch, we build a glossary over all noun phrases in the training set and run the cross-modal retriever from Section 4.2 on every speech chunk. Then we add the retrieved distractors in G_i alongside the ground-truth term. The Speech LLM thus learns to translate in the presence of realistic retriever noise.

A training example is shown in Figure 14. We wrap the ground-truth term’s translation in a `<term> . . . </term>` tag, encouraging the Speech LLM to copy ground-truth term translations directly from G_i . We fine-tune with a standard cross-entropy objective applied only to translation tokens $\hat{y}$. Because translations naturally lag behind speech in the training data, the model also receives implicit supervision for *timing*: it observes when a term appears in the speech context and when its translation appears later in the target sequence.

5 Experimental Setups

5.1 Data

Retriever Training We synthesize the retriever training data from GigaSpeech (Chen et al., 2021) and Wikidata⁵. GigaSpeech is a multi-domain English ASR corpus containing 10K hours of high-quality labeled audio. We process it following the procedure in Section 4.2, yielding 4.7M training pairs. The variable speech-context duration is

⁵<https://www.wikidata.org/>

sampled uniformly from 2.88, 3.84, 4.80, 5.76 seconds, corresponding to the inference-time chunk sizes 0.96, 1.92, 2.88, 3.84 plus a $t_{\text{lookback}} = 1.92$ -second look-back. The look-back captures terms that cross chunk boundaries and we set it to 1.92 seconds since most terms are shorter than this, as shown in Figure 7.

For Wikidata, we extract candidate terms from the RDF truthy dump, keeping entities that have both an English `rdfs:label` and at least one `wdt:P31` triple. Full processing details are given in Appendix A. For each retained term, we prompt Gemini-2.5-Flash to generate three natural utterances of 5–15 words containing the term (Figure 15). We then synthesize the utterances with CosyVoice 3⁶, using speaker prompts randomly sampled from GigaSpeech. This results in 2.9M training pairs.

Speech LLM Training We synthesize the Speech LLM training data using the same GigaSpeech ASR dataset. The interleaved En-Zh/De/Ja translation data is synthesized similar to InfiniSST with details in Appendix B. We subsample 12.5K instances for Speech LLM fine-tuning. We build a 140K glossary from GigaSpeech and Wikidata for distractor retrieval, and cap the size of G_i at 20 terms.

Evaluation We evaluate on two datasets. The ACL 60/60 dev set (Salesky et al., 2023) contains five English ACL talks translated into 10 target languages. We evaluate En–Zh/De/Ja directions and feed each unsegmented talk directly to the model. Each talk comes with a technical term list annotated by domain experts. We use these human-curated terminology lists as the retrieval glossary and after source-side deduplication it contains 238 terms.

The ESO test set (Iranzo-Sánchez et al., 2025) comprises five long-form English oncology training videos with human translations into French, Spanish, German, and Slovene. We report En–De direction and use GPT-5.4⁷ to extract and filter domain-specific glossary candidates, including procedures, diagnoses, biomarkers, clinical-trial concepts, and acronyms, and manually verify the retained entries. The final glossary has 212 unique source terms. Appendix I gives the prompts. To test larger retrieval glossary, we expand the ACL

⁶FunAudioLLM/Fun-CosyVoice3-0.5B-2512

⁷<https://developers.openai.com/api/docs/models/gpt-5.4>

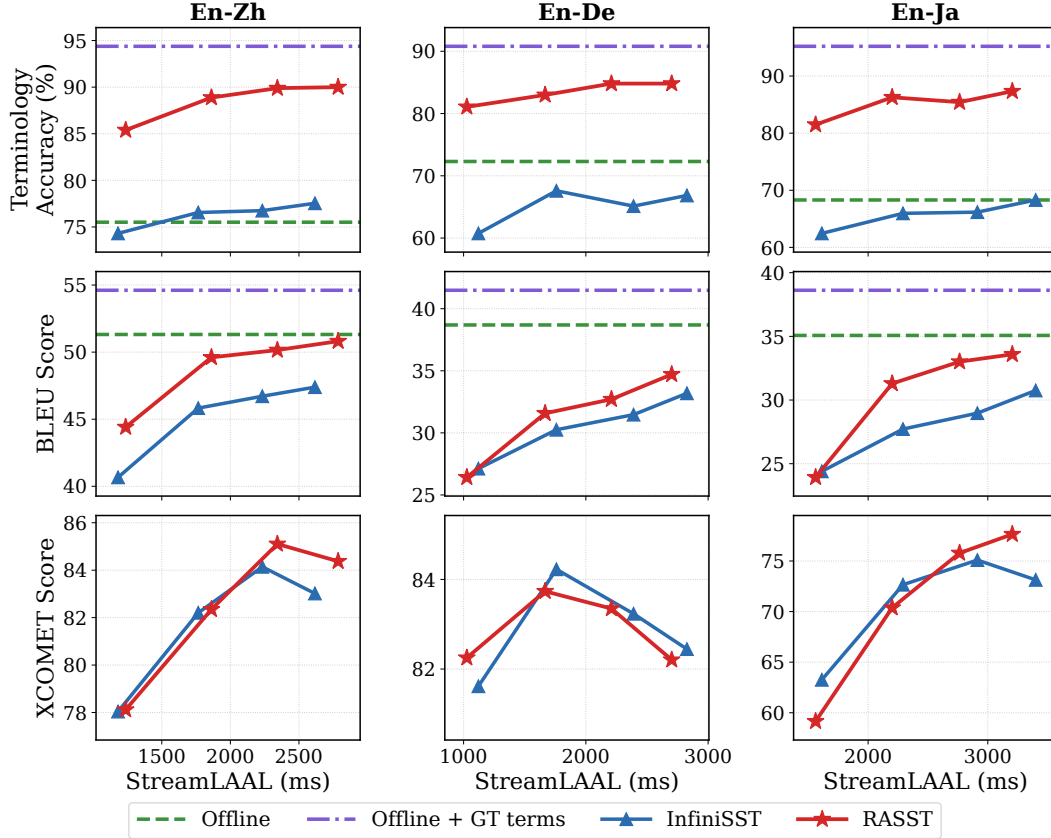


Figure 3: Quality–latency trade-off on ACL 60/60. X-axis is latency measured by StreamLAAL in milliseconds. Y-axis is quality measured by terminology accuracy (exact match), BLEU and XCOMET-XXL. RASST improves terminology accuracy and BLEU in all three directions.

and ESO glossaries to 1K and 10K entries using unseen Wikidata terms from relevant domains.

5.2 Evaluation Metrics

Latency Following IWSLT 2025 practice (Abdumumin et al., 2025), we use StreamLAAL (Papi et al., 2024) as our latency metric. StreamLAAL segments the model hypothesis to align with reference sentences and reports the average latency over the resulting (hypothesis, reference) pairs.

Translation Quality We evaluate aligned sentence triplets (transcript, hypothesis, reference) using SacreBLEU (Papineni et al., 2002; Post, 2018) for surface-form accuracy and XCOMET-XXL (Guerreiro et al., 2024) for semantic accuracy.

Terminology Accuracy We introduce terminology accuracy, which measures how accurately a model translates domain-specific terms. Our primary metric is exact match: for each term occurrence in the reference, we check whether its translation appears verbatim in the aligned hypothesis.

To account for valid orthographic and inflectional variation, we also report a form-aware metric.

This metric accepts differences in case, spacing, and hyphenation across all languages, as well as contiguous German lemma sequences identified by a fixed morphological analyzer. It does not accept synonyms or semantic paraphrases. Appendix F provides the complete evaluation protocol and results.

5.3 Model Configuration

Architecture For the retriever, we use a Qwen3-Omni Audio Transformer⁸ as the speech encoder, followed by a linear projection. For text encoding, we use the dense BGE-M3⁹ encoder and use the contextualized representation at its first [CLS] position as the term vector. For the Speech LLM, we use Qwen3-Omni-30B-A3B-Instruct¹⁰.

Training We train both the cross-modal retriever and the Speech LLM with LoRA (Hu et al., 2022). For retriever training, the number of hard negative

⁸<https://huggingface.co/Atotti/Qwen3-Omni-AudioTransformer>

⁹<https://huggingface.co/BAAI/bge-m3>

¹⁰<https://huggingface.co/Qwen/Qwen3-Omni-30B-A3B-Instruct>

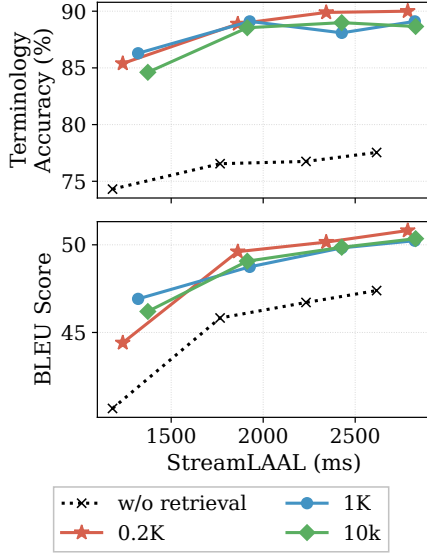


Figure 4: Scaling the glossary from 0.2K to 1K and 10K terms. RASST shows only a slight drop in terminology accuracy and translation quality, remaining well above the no-retrieval baseline.

N is 1024 and the global batch size is 8192. For Speech LLM training, we enable sequence packing with a maximum packed sequence length of 3,072 tokens and use a global batch size of 4 packed sequences. Full hyperparameters and details are provided in Appendix D.

Inference We implement the inference agent in SimulEval (Ma et al., 2020a), varying the chunk size l over 0.96, 1.92, 2.88, 3.84 seconds. The Speech LLM is invoked once per chunk and generates up to 40 new tokens per call. We serve it with vLLM (Kwon et al., 2023) using tensor parallelism of size 2, decode with temperature 0.6, top- $p=0.95$, and top- $k=20$, and strip the assistant’s <term> tags from the hypothesis before scoring.

For retrieval, we build a FAISS inner-product index over ℓ_2 -normalized term embeddings. At each step, the retriever encodes the current speech chunk together with a 1.92-second look-back, and queries the index using only those windows whose right boundary lies within the current chunk. We use Top- $K=10$ and discard candidates with similarity below $\tau=0.78$ ¹¹.

5.4 Baselines

Offline We use Qwen3-Omni-30B-A3B-Instruct to translate pre-segmented speech utterances offline, with the same decoding parameters as

RASST. This serves as an upper bound on translation quality.

Offline + GT Terms We extend the **Offline** baseline by adding the ground-truth terms appearing in each utterance to the prompt in a manner similar to Gong et al. (2024), helping the model handle terminology correctly. This serves as an upper bound on terminology translation accuracy.

InfiniSST InfiniSST (Ouyang et al., 2025b) is a strong SST baseline for simultaneous translation of unbounded speech and achieved top performance in the low-latency track at IWSLT 2025 (Ouyang et al., 2025a; Abdulmumin et al., 2025). We train InfiniSST on the same data and with the same Speech LLM as RASST, but remove all terminology input G_i . This setting can be viewed as RASST without retrieval.

6 Results and Analysis

6.1 Main Results

RASST substantially improves terminology accuracy Figure 3 shows that RASST outperforms InfiniSST in exact-match terminology accuracy across all 12 ACL language–latency settings, with gains of 12.3–21.4 percentage points. Figure 8 presents the English–German results on ESO, where RASST yields larger gains of 26.0–31.2 points across the four latency settings. We attribute the larger gains on ESO to the difficulty of recognizing and translating specialized oncology terms. By contrast, many terms annotated in ACL 60/60 are common words, such as *words*, *model*, and *weights*, which are comparatively easy to translate. Notably, RASST also surpasses the Offline baseline, demonstrating that access to the complete speech context alone is insufficient for reliable terminology translation. Although a gap to Offline + GT Terms remains because retrieval is imperfect, it is typically below five percentage points, indicating that RASST captures most of the benefit of oracle terminology access. The form-aware evaluation in Appendix F confirms that these conclusions are not artifacts of orthographic or inflectional variation.

RASST improves overall translation quality On ACL 60/60, RASST achieves higher BLEU than InfiniSST in 10 of the 12 settings, with a maximum gain of 3 BLEU points. By contrast, the two systems obtain comparable XCOMET scores. One likely explanation is that RASST inserts re-

¹¹We select τ on the dev set to allow at most a 1% drop in recall.

Chunk Size (s)	0.96	1.92	2.88	3.84
Wall-Clock Time (s)	0.036	0.042	0.042	0.043

Table 2: Retriever computation cost across speech chunk sizes. The retriever’s overhead is negligible.

trieved target terms verbatim, which is more favored by a surface-form metric such as BLEU than by a neural semantic metric such as XCOMET. Moreover, although aligned target terms occur in 83.76–86.54% of ACL sentences, they constitute only 10.66–18.06% of the target tokens. Improvements in terminology accuracy therefore affect a limited proportion of the corpus, although expert evaluation frameworks explicitly treat terminology errors as a distinct source of translation quality degradation (Freitag et al., 2021).

RASST is robust to large glossary size Figure 4 shows that RASST remains competitive as the glossary scales to 10K terms. Going from 0.2K to 1K and then 10K terms gradually reduces terminology accuracy by less than 2% and BLEU by less than 1 point, and RASST still clearly outperforms the no-retrieval baseline.

RASST is computationally efficient Table 2 reports the additional computational cost introduced by the retriever. We run evaluation on two NVIDIA RTX A6000 GPUs, serving the Speech LLM with vLLM at tensor parallelism 2 and `gpu_memory_utilization=0.72`, while the retriever encoder fits in the remaining GPU memory. The retriever’s wall-clock time per chunk is at most 3.75% of the chunk size, and becomes smaller as the chunk grows, which makes it negligible in practice.

6.2 Ablation Studies

Retriever hard negatives and threshold We examine the effect of training-time hard negatives and the inference-time threshold τ . In addition to the default retriever trained with $N=1024$ hard negatives, we train two variants with $N=256$ and $N=0$. We evaluate their precision and recall on the GigaSpeech dev set against three glossaries: a 1K glossary built from GigaSpeech noun phrases, and 10K and 100K extensions with Wikidata terms unseen during training. For each (retriever variant, glossary) pair, we sweep τ from 0 to 1 to obtain a precision-recall curve.

Figure 11 shows the results. First, increasing N yields clearly higher recall at the same precision,

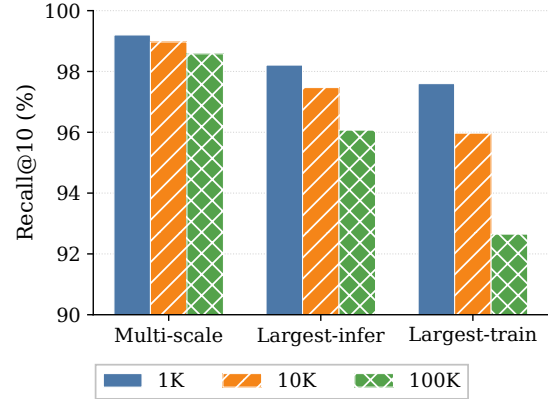


Figure 5: Recall@10 on the GigaSpeech dev set for multi-scale inference in RASST and two variants. Largest-infer uses the same retriever as RASST but queries with only the largest window at inference. Largest-train trains the retriever with the largest window instead of the MFA-localized window. Multi-scale inference combined with MFA-localized training performs best.

confirming the value of hard-negative mining. Second, recall degrades as the glossary grows. Based on these curves, we set $\tau=0.78$ at inference: it costs at most a 1% drop in recall while roughly reducing the retrieval noise by half. Compared with $\tau = 0$ in Figure 12, $\tau = 0.78$ achieves higher terminology accuracy.

Multi-Scale Inference We examine the contribution of multi-scale inference. As described in Section 4.2, RASST’s retriever queries with speech windows at multiple scales and keeps the terms with the highest similarities. We compare against two variants. The first reuses our retriever but queries with only the largest window at inference time, but this introduces a training–inference mismatch, since the retriever was trained with MFA-aligned windows around each term. The second variant removes this mismatch by training the retriever directly with the largest window, without MFA-based window localization. Results are shown in Figure 5. Multi-scale inference combined with MFA-localized training performs best: it loses less than 1% in recall when scaling from 1K to 100K glossary, compared to drops of over 2% and 5% for the two variants. The end-to-end En–Zh evaluation at latency multiplier 2 in Table 3 confirms that this retrieval gain reaches the translation output: multi-scale inference achieves 90.00% terminology accuracy, versus 84.72% for largest-window inference and 73.93% when the retriever

ACL 60/60 En-Ja

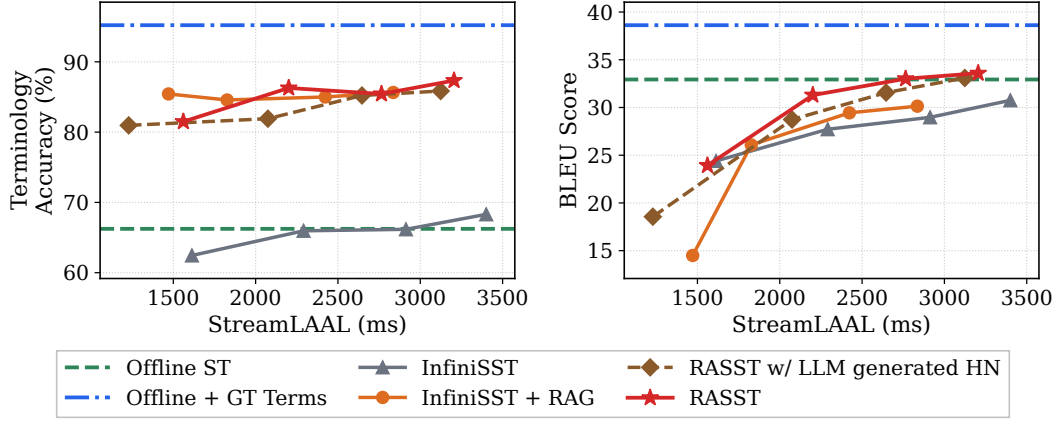


Figure 6: Quality–latency trade-off of RASST against two training-data variants: a version trained with LLM-generated distractors, and *InfiniSST* + RAG, which adds retrieval at inference to a model not trained with terms G_i . RASST achieves the best overall trade-off.

Variant	BLEU $\uparrow$	LAAL (ms) $\downarrow$	TERM_ACC $\uparrow$
Multi-scale	47.8780	1814.34	90.00
Largest-infer	47.8261	1868.37	84.72
Largest-train	45.7960	1863.20	73.93

Table 3: End-to-end ablation of retrieval strategies on ACL En–Zh at latency multiplier 2. Decoding uses temperature 0.6.

is also trained on the largest window.

Speech LLM Training Data We ablate the distractor-sampling strategy used to construct G_i during Speech LLM training. Besides retriever mined distractors in RASST, we train two variants: one with LLM-generated distractors, and one with no retrieved terms at all (equivalent to *InfiniSST*). At inference, all variants use the same retriever to populate G_i from the speech chunk before the Speech LLM generates the translation. Results are shown in Figure 6. The model trained with LLM-generated distractors achieves slightly lower terminology accuracy and BLEU than RASST. More surprisingly, *InfiniSST* + RAG, which sees no retrieved terms during training but is given them at inference, matches RASST’s terminology accuracy but has much worse BLEU score. Inspecting its outputs, we find that it tends to copy retrieved terms verbatim without checking whether they actually appear in the source or inserting them at the right time.

More ablations on retriever training data, encoder choice, context length, and tag supervision can be found in Appendix E.

6.3 Robustness Analysis

We evaluate the robustness of RASST to variations in glossary and retrieval quality in Appendix G.

6.4 Qualitative Analysis

We additionally present a case study examining the behavior of RASST in Appendix H.

7 Conclusion

We presented RASST, a retrieval-augmented framework for simultaneous speech translation that improves domain-specific terminology translation. RASST couples a lightweight cross-modal retriever with a Speech LLM and is trained with noise-robust supervision so that it remains accurate under realistic retrieval errors at inference time. Across all 12 ACL language–latency settings, RASST improves exact-form terminology accuracy over *InfiniSST* by 12.3–21.4 percentage points; on the human-reference ESO En–De condition, the gains are 26.0–31.2 points. Form-aware scoring, end-to-end retrieval ablations, corrupted-retrieval tests, and qualitative analysis support the same terminology benefit. BLEU and XCOMET effects are smaller and mixed, so the evidence supports reliable terminology translation rather than a universal improvement in general translation quality.

Limitations

This work has several limitations. First, we focus exclusively on English as the source language. Whether RASST’s cross-modal retriever and multi-scale inference transfer to other source languages

remains untested. Second, our evaluation covers only two domains (academic talks and oncology training videos), leaving RASST’s effectiveness on other terminology-rich settings such as legal and financial domains uncharacterized. Third, we rely on a fixed similarity threshold τ to suppress retrieval noise. Fourth, we do not explicitly model homophone confusion as in some prior contextual-biasing ASR work, since our focus is the integration of retrieval with an LLM-based simultaneous translation model, but it should be a natural extension. Fifth, retrieval can help only when the required entry is present in the supplied glossary. The paper-derived-glossary test covers 23.95% of the human glossary and accordingly yields much smaller gains. Finally, all experiments use Qwen3-Omni as the Speech LLM backbone, and we have not yet verified that our findings generalize to other speech LLMs.

References

- Idris Abdulmumin, Victor Agostinelli, Tanel Alumäe, Antonios Anastasopoulos, Luisa Bentivogli, Ondřej Bojar, Claudia Borg, Fethi Bougares, Roldano Cattoni, Mauro Cettolo, Lizhong Chen, William Chen, Raj Dabre, Yannick Estève, Marcello Federico, Mark Fishel, Marco Gaido, Dávid Javorský, Marek Kasztelnik, and 33 others. 2025. [Findings of the IWSLT 2025 evaluation campaign](#). In *Proceedings of the 22nd International Conference on Spoken Language Translation (IWSLT 2025)*, pages 412–481, Vienna, Austria (in-person and online). Association for Computational Linguistics.
- Ibrahim Said Ahmad, Antonios Anastasopoulos, Ondřej Bojar, Claudia Borg, Marine Carpuat, Roldano Cattoni, Mauro Cettolo, William Chen, Qianqian Dong, Marcello Federico, Barry Haddow, Dávid Javorský, Mateusz Krubiński, Tsz Kin Lam, Xutai Ma, Prashant Mathur, Evgeny Matusov, Chandresh Maurya, John P. McCrae, and 25 others. 2024. [FINDINGS OF THE IWSLT 2024 EVALUATION CAMPAIGN](#). In *Proceedings of the 21st International Conference on Spoken Language Translation (IWSLT 2024)*, pages 1–11, Bangkok, Thailand (in-person and online). Association for Computational Linguistics.
- Guoguo Chen, Shuzhou Chai, Guan-Bo Wang, Jiayu Du, Wei-Qiang Zhang, Chao Weng, Dan Su, Daniel Povey, Jan Trmal, Junbo Zhang, Mingjie Jin, Sanjeev Khudanpur, Shinji Watanabe, Shuaijiang Zhao, Wei Zou, Xiangang Li, Xuchen Yao, Yongqing Wang, Zhao You, and Zhiyong Yan. 2021. [Gigaspeech: An evolving, multi-domain asr corpus with 10,000 hours of transcribed audio](#). In *Interspeech 2021*, pages 3670–3674.
- Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. [Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation](#). Preprint, arXiv:2402.03216.
- Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian, Jian Wu, Michael Zeng, Xiangzhan Yu, and Furu Wei. 2022. [Wavlm: Large-scale self-supervised pre-training for full stack speech processing](#). *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518.
- Shanbo Cheng, Yu Bao, Zhichao Huang, Yu Lu, Ningxin Peng, Lu Xu, Runsheng Yu, Rong Cao, Yujiao Du, Ting Han, Yuxiang Hu, Zeyang Li, Sitong Liu, Shengtao Ma, Shiguang Pan, Jiongchen Xiao, Nuo Xu, Meng Yang, Rong Ye, and 9 others. 2025. [Seed liveinterpret 2.0: End-to-end simultaneous speech-to-speech translation with your voice](#). Preprint, arXiv:2507.17527.
- Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvassy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2024. [The faiss library](#). *arXiv preprint arXiv:2401.08281*.
- Zhihao Du, Changfeng Gao, Yuxuan Wang, Fan Yu, Tianyu Zhao, Hao Wang, Xiang Lv, Hui Wang, Xian Shi, Keyu An, and 1 others. 2025. [Cosyvoice 3: Towards in-the-wild speech generation via scaling-up and post-training](#). *arXiv preprint arXiv:2505.17589*.
- Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Ariavazhagan, and Wei Wang. 2022. [Language-agnostic BERT sentence embedding](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.
- Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021. [Experts, errors, and context: A large-scale study of human evaluation for machine translation](#). *Transactions of the Association for Computational Linguistics*, 9:1460–1474.
- Biao Fu, Donglei Yu, Minpeng Liao, Chengxi Li, Yidong Chen, Kai Fan, and Xiaodong Shi. 2025. [Efficient and adaptive simultaneous speech translation with fully unidirectional architecture](#). Preprint, arXiv:2504.11809.
- Li Fu, Yu Xin, Sunlu Zeng, Lu Fan, Youzheng Wu, and Xiaodong He. 2026. [Pac: Pronunciation-aware contextualized large language model-based automatic speech recognition](#). In *ICASSP 2026 - 2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 16377–16381.
- Xun Gong, Anqi Lv, Zhiming Wang, and Yanmin Qian. 2024. [Contextual Biasing Speech Recognition in](#)

- Speech-enhanced Large Language Model. In *Interspeech 2024*, pages 257–261.
- Xun Gong, Anqi Lv, Wangyou Zhang, Zhiming Wang, Huijia Zhu, and Yanmin Qian. 2025. **BR-ASR: Efficient and Scalable Bias Retrieval Framework for Contextual Biasing ASR in Speech LLM**. In *Interspeech 2025*, pages 4043–4047.
- Nuno M. Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André F. T. Martins. 2024. **xCOMET: Transparent machine translation evaluation through fine-grained error detection**. *Transactions of the Association for Computational Linguistics*, 12:979–995.
- Shoutao Guo, Xiang Li, Mengge Liu, Wei Chen, and Yang Feng. 2025. **Streamuni: Achieving streaming speech translation with a unified large speech-language model**. *Preprint*, arXiv:2507.07803.
- Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. **LoRA: Low-rank adaptation of large language models**. In *International Conference on Learning Representations*.
- Christian Huber, Juan Hussain, Sebastian Stüker, and Alexander Waibel. 2021. **Instant one-shot word-learning for context-specific neural sequence-to-sequence speech recognition**. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–7.
- Matthias Huck, Viktor Hangya, and Alexander Fraser. 2019. **Better OOV translation with bilingual terminology mining**. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5809–5815, Florence, Italy. Association for Computational Linguistics.
- Jorge Iranzo-Sánchez, Jaume Santamaría-Jordà, Gerard Mas-Mollà, Gonçal V. Garcés Díaz-Munío, Javier Iranzo-Sánchez, Javier Jorge, Joan Albert Silvestre-Cerdà, Adrià Giménez, Jorge Civera, Albert Sanchis, and Alfons Juan. 2025. **Speech translation for multilingual medical education leveraged by large language models**. *Artificial Intelligence In Medicine*, 166:103147.
- Mahaveer Jain, Gil Keren, Jay Mahadeokar, Geoffrey Zweig, Florian Metze, and Yatharth Saraf. 2020. **Contextual RNN-T for Open Domain ASR**. In *Interspeech 2020*, pages 11–15.
- Masoud Jalili Sabet, Philipp Dufter, François Yvon, and Hinrich Schütze. 2020. **SimAlign: High quality word alignments without parallel training data using static and contextualized embeddings**. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1627–1643, Online. Association for Computational Linguistics.
- Sai Muralidhar Jayanthi, Devang Kulshreshtha, Saket Dingliwal, Srikanth Ronanki, and Sravan Bodapati. 2023. **Retrieve and copy: Scaling ASR personalization to large catalogs**. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 631–639, Singapore. Association for Computational Linguistics.
- Hong Jiang. 2013. **The interpreters glossary in simultaneous interpreting: A survey**. *Interpreting*, 15(1):74–93.
- Qingnan Jiang, Mingxuan Wang, Jun Cao, Shanbo Cheng, Shujian Huang, and Lei Li. 2021. **Learning kernel-smoothed machine translation with retrieved examples**. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7280–7290, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Juraj Juraska, Daniel Deutsch, Mara Finkelstein, and Markus Freitag. 2024. **MetricX-24: The Google submission to the WMT 2024 metrics shared task**. In *Proceedings of the Ninth Conference on Machine Translation*, pages 492–504, Miami, Florida, USA. Association for Computational Linguistics.
- Urvashi Khandelwal, Angela Fan, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. 2021. **Nearest neighbor machine translation**. In *International Conference on Learning Representations*.
- Yuxiang Kong, Fan Cui, Liyong Guo, Heinrich Dinkel, Lichun Fan, Junbo Zhang, and Jian Luan. 2025. **GLCLAP: A Novel Contrastive Learning Pre-trained Model for Contextual Biasing in ASR**. In *Interspeech 2025*, pages 5173–5177.
- Roman Koshkin, Katsuhito Sudoh, and Satoshi Nakamura. 2024. **TransLLaMa: LLM-based simultaneous translation system**. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 461–476, Miami, Florida, USA. Association for Computational Linguistics.
- Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. **Efficient memory management for large language model serving with pagedattention**. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.
- Duc Le, Mahaveer Jain, Gil Keren, Suyoun Kim, Yangyang Shi, Jay Mahadeokar, Julian Chan, Yuan Shangguan, Christian Fuegen, Ozlem Kalinli, Yatharth Saraf, and Michael L. Seltzer. 2021a. **Contextualized Streaming End-to-End Speech Recognition with Trie-Based Deep Biasing and Shallow Fusion**. In *Interspeech 2021*, pages 1772–1776.
- Duc Le, Gil Keren, Julian Chan, Jay Mahadeokar, Christian Fuegen, and Michael L. Seltzer. 2021b. **Deep shallow fusion for rnn-t personalization**. In *2021 IEEE Spoken Language Technology Workshop (SLT)*, pages 251–257.

- Bryan Li, Jiaming Luo, Eleftheria Briakou, and Colin Cherry. 2025. [Leveraging domain knowledge at inference time for LLM translation: Retrieval versus generation](#). In *Proceedings of the 4th International Workshop on Knowledge-Augmented Methods for Natural Language Processing*, pages 91–106, Albuquerque, New Mexico, USA. Association for Computational Linguistics.
- Christopher Li, Pat Rondon, Diamantino Caseiro, Leonid Velikovich, Xavier Velez, and Petar Aleksic. 2021. [Improving entity recall in automatic speech recognition with neural embeddings](#). In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7353–7357.
- Siqi Li, Danni Liu, and Jan Niehues. 2024. [Optimizing rare word accuracy in direct speech translation with a retrieval-and-demonstration approach](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12703–12719, Miami, Florida, USA. Association for Computational Linguistics.
- Xutai Ma, Mohammad Javad Dousti, Changhan Wang, Jiatao Gu, and Juan Pino. 2020a. [SIMULEVAL: An evaluation toolkit for simultaneous translation](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 144–150, Online. Association for Computational Linguistics.
- Xutai Ma, Juan Pino, and Philipp Koehn. 2020b. [SimulMT to SimulST: Adapting simultaneous text translation to end-to-end simultaneous speech translation](#). In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 582–587, Suzhou, China. Association for Computational Linguistics.
- Yuxian Meng, Xiaoya Li, Xiayu Zheng, Fei Wu, Xiaofei Sun, Tianwei Zhang, and Jiwei Li. 2022. [Fast nearest neighbor machine translation](#). In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 555–565, Dublin, Ireland. Association for Computational Linguistics.
- Tsendsuren Munkhdalai, Khe Chai Sim, Angad Chandorkar, Fan Gao, Mason Chua, Trevor Strohman, and Françoise Beaufays. 2022. [Fast contextual adaptation with neural associative memory for on-device personalized speech recognition](#). In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6632–6636.
- Tsendsuren Munkhdalai, Zelin Wu, Golan Pundak, Khe Chai Sim, Jiayang Li, Pat Rondon, and Tara N. Sainath. 2023. [Nam+: Towards scalable end-to-end contextual biasing for adaptive asr](#). In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 190–196.
- Siqi Ouyang, Xi Xu, and Lei Li. 2025a. [CMU’s IWSLT 2025 simultaneous speech translation system](#). In *Proceedings of the 22nd International Conference on Spoken Language Translation (IWSLT 2025)*, pages 309–314, Vienna, Austria (in-person and online). Association for Computational Linguistics.
- Siqi Ouyang, Xi Xu, and Lei Li. 2025b. [InfiniSST: Simultaneous translation of unbounded speech with large language model](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 3032–3046, Vienna, Austria. Association for Computational Linguistics.
- Sara Papi, Marco Gaido, Matteo Negri, and Luisa Bentivogli. 2024. [StreamAtt: Direct streaming speech-to-text translation with attention-based audio history selection](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3692–3707, Bangkok, Thailand. Association for Computational Linguistics.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. [Bleu: a method for automatic evaluation of machine translation](#). In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.
- Matt Post. 2018. [A call for clarity in reporting BLEU scores](#). In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.
- Golan Pundak, Tara N. Sainath, Rohit Prabhavalkar, Anjuli Kannan, and Ding Zhao. 2018. [Deep context: End-to-end contextual speech recognition](#). In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 418–425.
- Yi Ren, Jinglin Liu, Xu Tan, Chen Zhang, Tao Qin, Zhou Zhao, and Tie-Yan Liu. 2020. [SimulSpeech: End-to-end simultaneous speech to text translation](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3787–3796, Online. Association for Computational Linguistics.
- Elizabeth Salesky, Kareem Darwish, Mohamed Al-Badrashiny, Mona Diab, and Jan Niehues. 2023. [Evaluating multilingual speech translation under realistic conditions with resegmentation and terminology](#). In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*, pages 62–78, Toronto, Canada (in-person and online). Association for Computational Linguistics.
- Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2019. [Representation learning with contrastive predictive coding](#). *Preprint*, arXiv:1807.03748.
- Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. [Multilingual e5 text embeddings: A technical report](#). *arXiv preprint arXiv:2402.05672*.

- Minghan Wang, Thuy-Trang Vu, Yuxia Wang, Ehsan Shareghi, and Gholamreza Haffari. 2025. [Conversational SimulMT: Efficient simultaneous translation with large language models](#). In *Proceedings of the 22nd International Conference on Spoken Language Translation (IWSLT 2025)*, pages 93–105, Vienna, Austria (in-person and online). Association for Computational Linguistics.
- Suhang Wu, Jialong Tang, Chengyi Yang, Pei Zhang, Baosong Yang, Junhui Li, Junfeng Yao, Min Zhang, and Jinsong Su. 2025. [Locate-and-focus: Enhancing terminology translation in speech language models](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11345–11360, Vienna, Austria. Association for Computational Linguistics.
- Zelin Wu, Tsendsuren Munkhdalai, Pat Rondon, Golan Pundak, Khe Chai Sim, and Christopher Li. 2023. [Dual-Mode NAM: Effective Top-K Context Injection for End-to-End ASR](#). In *Interspeech 2023*, pages 221–225.
- Zelin Wu, Gan Song, Christopher Li, Pat Rondon, Zhong Meng, Xavier Velez, Weiran Wang, Diamantino Caseiro, Golan Pundak, Tsendsuren Munkhdalai, Angad Chandorkar, and Rohit Prabhavalkar. 2024. [Deferred NAM: Low-latency top-k context injection via deferred context encoding for non-streaming ASR](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 6: Industry Track)*, pages 315–323, Mexico City, Mexico. Association for Computational Linguistics.
- Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, Yuanjun Lv, Yongqi Wang, Dake Guo, He Wang, Linhan Ma, Pei Zhang, Xinyu Zhang, Hongkun Hao, Zishan Guo, and 19 others. 2025. [Qwen3-omni technical report](#). *Preprint*, arXiv:2509.17765.
- 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.
- Guanrou Yang, Ziyang Ma, Zhifu Gao, Shiliang Zhang, and Xie Chen. 2024. [Ctc-assisted llm-based contextual asr](#). In *2024 IEEE Spoken Language Technology Workshop (SLT)*, pages 126–131.
- Bolaji Yusuf, Aditya Gourav, Ankur Gandhe, and Ivan Bulyko. 2023. [On-the-fly text retrieval for end-to-end asr adaptation](#). In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.
- Zixin Zeng, Rui Wang, Yichong Leng, Junliang Guo, Shufang Xie, Xu Tan, Tao Qin, and Tie-Yan Liu. 2023. [Extract and attend: Improving entity translation in neural machine translation](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1697–1710, Toronto, Canada. Association for Computational Linguistics.
- Jingyi Zhang, Masao Utiyama, Eiichiro Sumita, Graham Neubig, and Satoshi Nakamura. 2018. [Guiding neural machine translation with retrieved translation pieces](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1325–1335, New Orleans, Louisiana. Association for Computational Linguistics.
- Tong Zhang, Long Zhang, Wei Ye, Bo Li, Jinan Sun, Xiaoyu Zhu, Wen Zhao, and Shikun Zhang. 2021. [Point, disambiguate and copy: Incorporating bilingual dictionaries for neural machine translation](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3970–3979, Online. Association for Computational Linguistics.
- 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 2019*, pages 1418–1422.
- Wenhao Zhu, Jingjing Xu, Shujian Huang, Lingpeng Kong, and Jiajun Chen. 2023. [INK: Injecting kNN knowledge in nearest neighbor machine translation](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15948–15959, Toronto, Canada. Association for Computational Linguistics.

A Wikidata Processing

We normalize each label by removing parenthetical qualifiers, lowercasing for deduplication, and retaining only short alphabetic or hyphenated terms with at most five words. We further require a non-empty short description from the glossary index that does not simply repeat the term, which filters out many ambiguous or non-English labels. To avoid a corpus dominated by very frequent instance types such as people, taxa, or scholarly articles, we assign each entity a primary P31 type using its least frequent instance type and rank terms with inverse-type-frequency sampling. Terms appearing in the held-out evaluation glossaries are removed before training.

For each retained term, we prompt Gemini-2.5-Flash to generate three natural utterances of 5–15 words containing the term (Figure 15). We then

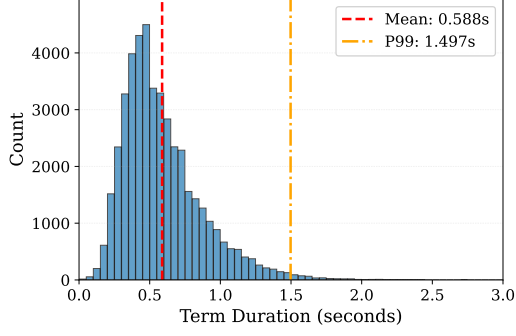


Figure 7: The distribution of term duration on GigaSpeech training set.

synthesize the utterances with CosyVoice 3¹², using speaker prompts randomly sampled from GigaSpeech.

B Speech LLM Data Synthesis

As in InfiniSST, we build robust segments of 28.8 seconds, filter out those with WER larger than 0.1 using para-keet-0.6b-tdt-v2¹³ and generate the target language translation with Qwen3-30B-A3B-Instruct-2507-FP8 (Yang et al., 2025)¹⁴ and filter out those with MetricX score (Juraska et al., 2024)¹⁵ worse than -3. Finally we build the interleaved training data using SimAlign (Jalili Sabet et al., 2020) with LaBSE (Feng et al., 2022).

Translation Prompt

You are given English document split into lines. Translate each line into Chinese. Do not include any other text.
 <begin>
 {}
 <end>

C Term Duration Distribution

We show the distribution of term duration on GigaSpeech training set in Figure 7.

D Training Details

Retriever We train the retriever with LoRA applied to both the speech and text encoders. For the audio encoder, LoRA is inserted into

¹²<https://huggingface.co/FunAudioLLM/Fun-CosyVoice3-0.5B-2512>

¹³<https://huggingface.co/nvidia/para-keet-tdt-0.6b-v2>

¹⁴<https://huggingface.co/Qwen/Qwen3-30B-A3B-Instruct-2507-FP8>

¹⁵<https://huggingface.co/google/metricx-24-hybrid-x1-v2p6>

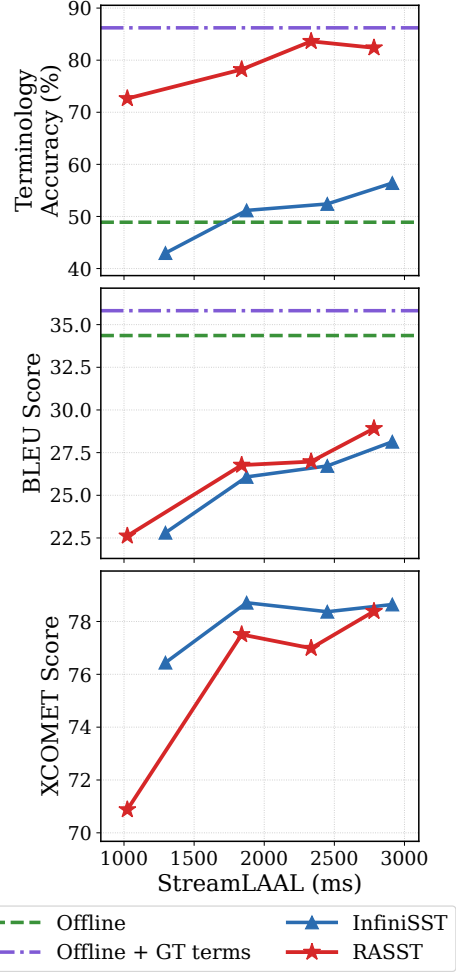


Figure 8: Quality-latency trade-off on ESO dataset End-De direction. Terminology accuracy uses strict exact-form matching. RASST improves terminology accuracy and BLEU but not XCOMET.

the query/key/value/output projections and MLP blocks with rank $r=32$ and scaling factor $\alpha=64$. For the text encoder, LoRA targets the query/key/value projections with rank $r=16$ and $\alpha=32$. We train with a multi-positive InfoNCE objective using a fixed temperature $\gamma=0.03$. We optimize with AdamW (learning rate 1×10^{-4} , weight decay 0.01), cosine annealing with 10% warmup, and gradient clipping (max norm 1.0) under BF16 mixed precision. Training is performed with Distributed Data Parallel (DDP) on eight RTX A6000 GPUs with a global batch size of 8192.

Speech LLM We fine-tune the Speech LLM using LoRA with rank $r=32$ and $\alpha=32$. For Speech LLM training, we enable sequence packing with a maximum packed sequence length of 3,072 tokens and use a global batch size of 4 packed se-

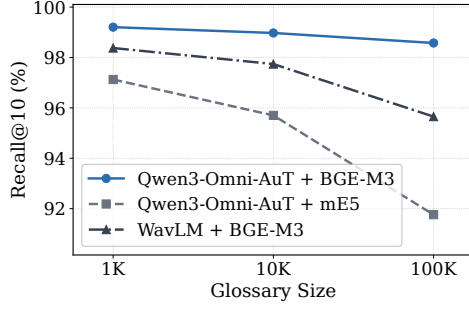


Figure 9: Recall@10 across glossary sizes for different encoder combinations. Qwen3-Omni-AuT + BGE-M3 achieves the highest recall.

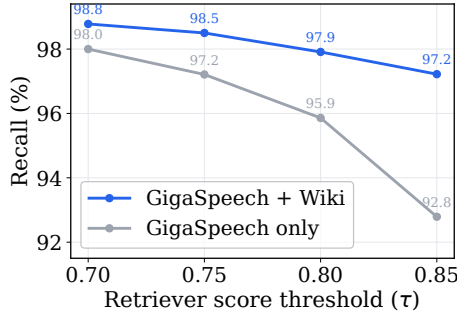


Figure 10: Recall@10 under different threshold τ on the GigaSpeech dev set for retrievers trained with and without Wikidata.

quences. We optimize with AdamW (learning rate 1×10^{-4} , weight decay 0.01) and gradient clipping (max norm 1.0). Training is performed with DDP and expert parallelism (EP) of size 2 on four NVIDIA RTX A6000 GPUs.

E Additional Ablations

This section reports additional ablations that do not fit in the main text.

Retriever training data We examine the impact of training data by comparing our default retriever, trained on both GigaSpeech and Wikidata, against a GigaSpeech-only variant. As shown in Figure 10, adding Wikidata substantially improves recall, with the largest gains at high τ , indicating that the retriever assigns higher similarity to true positives when trained with Wikidata.

Retriever encoders We test different encoder combinations here. We additionally tried mE5 text encoder (Wang et al., 2024)¹⁶ and WavLM-large speech encoder (Chen et al., 2022)¹⁷. As shown

¹⁶<https://huggingface.co/intfloat/multilingual-e5-large>

¹⁷<https://huggingface.co/microsoft/wavlm-large>

Context Length	1K	10K	100K
Fixed 1.92s	98.11	97.47	95.18
Fixed 3.84s	98.65	98.31	97.58
Fixed 5.76s	99.64	99.58	99.33
Variable 2.88–5.76s	99.20	98.97	98.58

Table 4: Recall@10 of retrievers trained with different context lengths, across glossary sizes. Longer speech contexts yield higher recall.

in Figure 9, Qwen3-Omni Audio Transformer + BGE-M3 achieves the best recall.

Retriever context length. RASST trains the retriever with variable-length speech contexts and here we compare against fixed-length alternatives. As shown in Table 4, longer fixed-length contexts generally yield higher recall across glossary sizes. The variable-length setup matches the recall of the longest fixed-length context (5.76s) while offering better flexibility at inference.

Term-Tag Supervision The `<term>` tags in Speech LLM targets explicitly identify when a retrieved translation is realized. We train an otherwise matched En-Zh model without these tags. As shown in Figure 13, the tagged model improves exact-form TERM_ACC by 3.60–5.06 points across all latency settings while achieving similar BLEU scores, supporting the use of tags as training-time supervision. The tags are stripped before evaluation and are not displayed to user.

F Form-Aware Terminology Accuracy

The primary exact-form scorer first segments system hypothesis into sentences each aligned to a reference sentence. A glossary entry is eligible when its English source form occurs in the aligned source sentence and its target form occurs in the reference. Target forms are deduplicated, and each eligible target form is counted once per sentence. Form-aware metrics uses the same setting as exact-form scorer and differs only in determining whether the target form appears in the hypothesis sentence.

For reproducibility, let $N(z)$ denote Unicode NFKC normalization followed by case folding and whitespace collapse. Let $C(z)$ concatenate only the Unicode letters and numbers in $N(z)$, thereby removing spaces, punctuation, and hyphens. For every eligible target form t and aligned hypothesis h , we apply the following ordered rules; the first successful rule determines the match type.

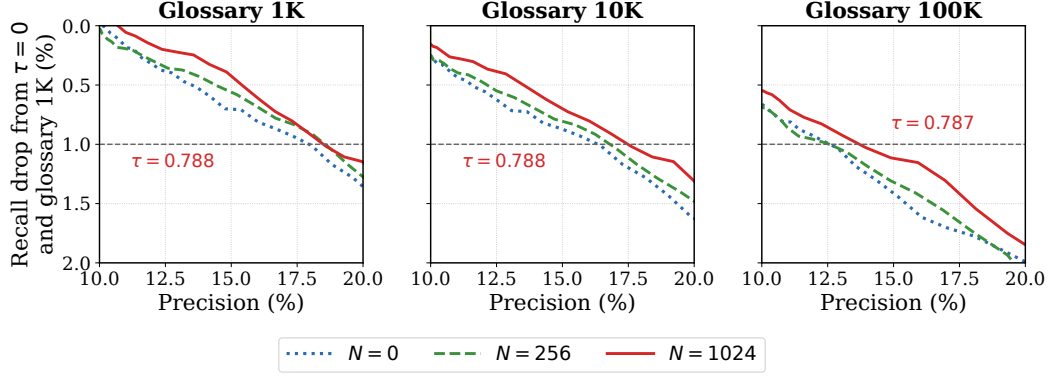


Figure 11: Precision–recall curves across glossary sizes for retrievers trained with $N=0, 256, 1024$ hard negatives, obtained by sweeping $\tau \in [0, 1]$. The dashed horizontal line marks a 1% recall drop; the red annotations show the corresponding $\tau \approx 0.78$ operating points.

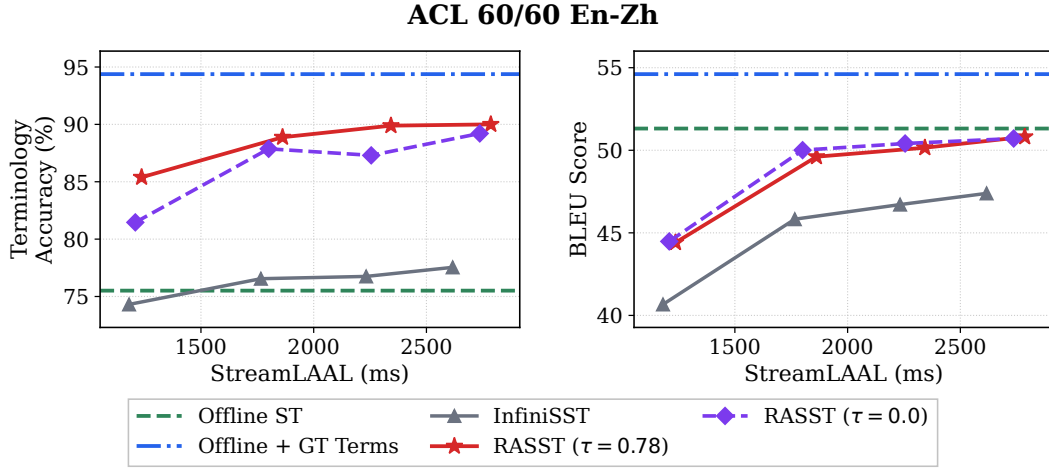


Figure 12: Quality–latency trade-off of RASST with $\tau = 0$ and $\tau = 0.78$.

1. **Exact form.** Accept if the unnormalized glossary target t is a substring of the unnormalized hypothesis h .
2. **Normalized surface form.** If t consists only of Latin letters or digits and $|C(t)| < 4$, accept only when $N(t)$ occurs in $N(h)$ with Unicode word boundaries on both sides. For every other target, accept when $N(t)$ is a substring of $N(h)$. This rule handles case and Unicode-width variation without allowing a short acronym to match inside a longer word.
3. **Spacing or hyphenation variant.** If the previous rules fail, accept when $C(t)$ is a substring of $C(h)$, provided that $|C(t)| \geq 4$ for Latin-alphanumeric targets and $|C(t)| \geq 2$ otherwise. The minimum lengths prevent permissive matches for very short strings.
4. **German inflection.** For En–De only, if the previous rules fail, tokenize t and h , discard punctuation and whitespace tokens, normalize

each lemma with $N(\cdot)$, and accept only if the complete lemma sequence of t equals a contiguous subsequence of the lemma sequence of h . We use spaCy 3.8.5 with `de_core_news_sm` 3.8.0 and disable its parser and named-entity recognizer.

The form-aware score counts a term as correct if any rule succeeds. It never accepts a synonym, paraphrase, reordered lemma sequence, or merely semantically related expression. Normalization is used only to test the hypothesis; eligibility continues to require the prescribed target surface form in the reference. The diagnostic was run on existing hypotheses, not newly decoded outputs. Across 32 system settings, it scored 27,296 eligible occurrences; every row with archived exact numerator and denominator reproduced the original exact count.

As shown in Table 5, the largest adjustment is in German, where inflection and productive com-

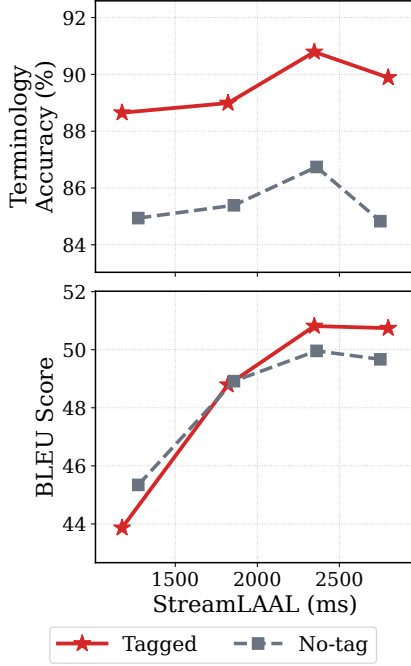


Figure 13: En-Zh ablation of training-time term tags. LAAL is in milliseconds and TERM is exact-form TERM_ACC (%).

Condition	Exact	Form-aware	Δ
ACL En-Zh	88.99	89.10	+0.11
ACL En-Ja	84.57	84.57	+0.00
ACL En-De	82.99	87.38	+4.39
ESO En-De	78.21	82.23	+4.02

Table 5: RASST exact-form and conservative form-aware TERM_ACC (%) at latency multiplier 2.

pounding create valid surface variants. At latency multiplier 2, ACL En-De InfiniSST also rises from 67.59 to 74.44, while RASST rises from 82.99 to 87.38; the form-aware margin therefore remains 12.94 points. On ESO En-De the corresponding margin is 23.03 points (82.23 versus 59.20). Thus normalization reduces both systems’ apparent error, without changing the conclusion that retrieval improves terminology realization.

G Robustness to Glossary and Retrieval Quality

G.1 Paper-Derived Glossaries

To approximate a less curated setting, we extract a separate glossary from the paper PDF corresponding to each ACL talk using Gemini, without consulting the transcripts, references, or human-

Direction	TERM ACC	BLEU
En-Zh	77.87 / 75.17	46.32 / 45.82
En-Ja	65.32 / 65.96	27.76 / 27.72
En-De	70.91 / 68.21	29.20 / 30.27

Table 6: ACL latency-multiplier-2 evaluation with paper-derived retrieval glossaries. Cells show RASST/InfiniSST.

Dir.	Replaced	TERM	BLEU	XCOMET
En-Zh	0%	90.00	47.878	83.628
En-Zh	25%	88.09	48.137	83.028
En-Zh	50%	85.73	48.331	82.608
En-De	0%	83.10	30.869	83.966
En-De	25%	79.79	31.077	83.214
En-De	50%	75.51	31.584	83.496
En-Ja	0%	84.57	29.889	70.379
En-Ja	25%	82.77	28.717	68.417
En-Ja	50%	78.40	27.717	67.137

Table 7: Sensitivity to retrieval quality at latency multiplier 2. We randomly replaced 0, 25%, 50% of the retrieved terms with irrelevant terms. Decoding uses temperature 0.6.

annotated glossary. These glossaries overlap with only 57 of 238 human-annotated glossary entries (23.95% source-term coverage). We use the paper-derived glossary for RASST retrieval but retain the human glossary as the fixed TERM_ACC denominator for both systems. Table 6 shows the latency-multiplier-2 results. As expected, the gains are much smaller because retrieval cannot supply entries absent from the glossary. Nevertheless, terminology accuracy improves for En-Zh and En-De.

G.2 Retrieval Degradation

We replace 25% or 50% of retrieved entries with irrelevant entries from the same ACL glossary, holding the number of hints and inference configuration fixed. Table 7 reports results at latency multiplier 2. Exact-form TERM_ACC decreases by at most 7.59 points under 50% replacement and remains above the corresponding InfiniSST accuracy in all three directions. All six corrupted settings have lower XCOMET than their language’s uncorrupted setting while have similar BLEU.

System	You are a professional simultaneous interpreter. You will be given chunks of English audio and you need to translate the audio into Chinese text. Use the term_map as a reference for terminology if provided.
User 1	<audio> term_map: things → 事; everything → 所有努力; right → 权利; thing → 东西; that thing → 那东西.
Assistant 1	她会做些我本意不想让她做的<term>事</term>。这样一来，之前<term>所有努力</term>都白费了，我不得不重新写一遍。
User 2	<audio> term_map: the Porch Gable → 门廊小屋; story → 传说; Anne → 安妮; seclusion → 静谧; Diana → 戴安娜; gabelle → 法国盐税; that story → 这个故事; porch → 门廊.
Assistant 2	然而，最终<term>传说</term>终于完成了，<term>安妮</term>在<term>门廊小屋</term>的<term>静谧</term>中把它读给了<term>戴安娜</term>听。她
User 3	<audio> term_map: Diana → 戴安娜; watchful eye → 密切留意; Robert Ray → 罗伯特·雷; her pathetic scene → 那个悲情场景; scene → 现场; that scene → 那场戏.
Assistant 3	成功地呈现了那个悲情场景，同时又没有牺牲<term>罗伯特·雷</term>，她一边读一边<term>密切留意</term>着<term>戴安娜</term>的反应。戴安娜

Figure 14: An example of En-Zh Speech LLM training data.

Category	Occurrences
Synonym	28
Grammar-conditioned surface change	3
Spacing or capitalization	12
StreamLAAL segmentation error	48
Term absent	14
Different concept or named entity	22
Total	127

Table 8: Manual categorization of cases where InfiniSST is exact-match correct and RASST is not.

H Qualitative and Failure Analysis

At latency multiplier 2, ACL 60/60 contains 3,266 evaluated terminology occurrences across the three directions. RASST is exact-match correct in 638 cases where InfiniSST is wrong, while the reverse occurs in 127 cases. We manually categorize these 127 apparent regressions in Table 8.

The first three categories contain the intended term but not the reference surface form, for example “Softmax” versus “Soft max” and “Transformer” versus “transformer.” They account for 43 cases and motivate the form-aware diagnostic. A further 48 are boundary artifacts from mWERSegmenter splitting Chinese or Japanese output mid-sentence.

Of the 14 genuinely absent terms, seven are not retrieved because their similarity falls just below $\tau=0.78$. The other seven are retrieved but not used: some hints arrive near the end of a sentence after the model has started generating another sentence, while others are omitted because the concept is inferable from context. The remaining 22 cases are

genuine substitutions or hallucinations, including “BETO” rendered as “BERT.” Acoustic acronym confusions also occur without retrieval, such as “BLEU” rendered as “blue.”

I ESO Glossary Construction

We construct the ESO hard-term glossary with the following seven-stage pipeline. (1) GPT-5.4 processes batches of 10 sentences with full-talk context, extracts high-precision medical candidates, and associates each candidate with exact target spans; the dataset-provided German sentence translation is kept unchanged. (2) GPT-5.4 enforces exact-span and abbreviation consistency, dropping copied or speculative acronyms. (3) Deterministic checks retain only entries whose source and target strings occur in their corresponding sentences. (4) We deduplicate and normalize source forms without changing target strings. (5) GPT-5.4 examines batches of three terms against no-retrieval outputs and identifies terms that are genuinely difficult rather than harmless variants. (6) We manually inspect and correct the glossary. (7) We serialize the final 215 rows (212 unique English terms) used to build the index. Precision is prioritized over recall throughout.

Stage 1 prompt: candidate extraction

Given the full transcript and sentence-level JSON, return one JSON item per sentence. Reuse the provided German translation exactly. Extract only genuinely domain-specific concepts such as procedures, diagnoses, biomarkers, clinical-trial terminology, anatomical terms, and acronyms. Exclude generic nouns, people, organizations, and vague phrases. For each retained English term, return a German target that is a contiguous substring of the provided German sentence. Precision is more important than recall; return an empty term list when no sufficiently technical term is present.

Stage 2 prompt: exact span and abbreviation consistency

Revise each candidate so every kept target is an exact contiguous span of the sentence translation. Do not add new terms. For abbreviations, keep the item only when its expansion or translation is clear and meaningful; drop a copied acronym, a speculative expansion, or an abbreviation that duplicates a full-form term. Preserve sentence and term identifiers and return JSON only.

Stage 5 prompt: hard-term filtering

Given the English term, prescribed target, source sentence, reference, and no-RAG prediction, judge whether the prediction conveys the term in context. Accept exact forms, faithful synonyms, and harmless spelling or spacing variants. Mark wrong or missing when the term is omitted, mistranslated, over-generalized, replaced by another concept, or merely copied when an expansion is required. Use not assessable only for alignment failure and return one JSON judgment per identifier.

Prompt template for Wikidata term utterance generation

System instruction:

You are a dataset builder for speech recognition research.

Your task: given a list of terms with their short descriptions, generate short English utterances (5–15 words each) that a speaker might say in a lecture, tutorial, or technical discussion. Use the description as context to produce accurate, meaningful sentences, but do NOT just copy the description.

Each utterance MUST contain the given term EXACTLY as written (same spelling, but case may differ). Vary sentence structure and term position (beginning, middle, end). Do NOT repeat the same pattern across terms.

Output valid JSON only – no markdown fences, no explanation.

User prompt:

Generate {variants_per_term} short utterances (5–15 words) for EACH of the following terms. Use the description as context to make the utterance accurate and meaningful. Each utterance must contain the term verbatim.

Terms with descriptions:

```
[
  {"term": "Seismology", "description": "scientific study of earthquakes"},
  {"term": "dataflow graph", "description": "graph representation of data dependencies"}
]
```

Return a JSON object mapping each term to a list of {variants_per_term} utterance strings. Example format:

```
{
  "Seismology": [
    "Seismology helps us understand earthquake patterns.",
    "The field of seismology has advanced rapidly."
  ]
}
```

IMPORTANT: output raw JSON only, no markdown code fences.

Figure 15: Prompt template for Wikidata term utterance generation.