DOA: Training-Free Decoder-Only Attention Policy for Long-Form Simultaneous Translation with SpeechLLMs

Anonymous ACL submission

Abstract

Simultaneous speech-to-text translation (SimulST) generates translations while speech is still unfolding, requiring a streaming policy that decides when to read and when to write. State-of-the-art approaches rely on attention-based encoder-decoder models where cross-attention provides explicit alignment signals. In contrast, Speech Large Language Models (SpeechLLMs) are decoder-only architectures relying solely on self-attention. This raises a central question: whether decoder self-attention contains sufficiently stable alignment signals to guide the streaming policy. Moreover, existing approaches typically rely on training-based adaptations or heuristic wait-k policies and have not been validated in long-form settings. To fill these gaps, we propose **D**ecoder-**O**nly **A**ttention (DOA), a training-free policy that enables long-form simultaneous translation with off-the-shelf SpeechLLMs by deriving a proxy alignment from self-attention. Experiments on Phi4-Multimodal and Qwen3-Omni show that DOA provides an effective alignment signal for supporting streaming decisions, enabling low-latency long-form SimulST with quality close to offline decoding without retraining.

1 Introduction

Simultaneous speech-to-text translation (SimulST) aims to generate translations while the source speech is still unfolding, balancing translation quality and latency (Fügen et al., 2007). This requires a *simultaneous policy*, i.e., a decision strategy that determines when to read additional speech input and when to write output tokens during streaming inference (Grissom II et al., 2014).

Recent advances in Speech Large Language Models (SpeechLLMs) have shifted attention toward decoder-only architectures, which unify speech and text processing within a single autoregressive LLM-powered model (Wu et al., 2023; Huang et al., 2024). While these models have shown strong performance across speech understanding and offline translation tasks (Huang et al., 2024; Gupta et al., 2024), their applicability to SimulST remains largely unexplored. Existing work on SpeechLLMs either relies on ad-hoc training or fine-tuning to induce streaming behavior (Chen et al., 2024; Guo et al., 2025a; Ouyang et al., 2024, 2025), or adapts classical training-free heuristics such as wait-k policies (Ma et al., 2019) to decoder-only architectures (Guo et al., 2025b). In contrast, state-of-the-art systems in encoder-decoder settings are typically driven by training-free attention-based policies that exploit cross-attention signals to drive streaming decisions, rather than relying on fixed schedules (Ahmad et al., 2024; Abdulmumin et al., 2025).

At the same time, previous studies have shown that evaluating SimulST systems on pre-segmented utterances can underestimate the challenges of realistic streaming conditions, where models must continuously process growing acoustic and textual contexts over long-form audio streams (Polák and Bojar, 2023). Despite these challenges, no prior work has investigated whether offline SpeechLLMs can be directly repurposed for long-form SimulST through a training-free attention-based policy.

A key obstacle is that SpeechLLMs do not expose explicit cross-attention scores. In attention-based encoder-decoder (AED) systems, cross-attention provides alignment signals that are sufficiently stable to support streaming decisions (Wang et al., 2024). However, it remains unclear whether SpeechLLMs' self-attention exhibits similar properties. This motivates the main research question of this work: *does SpeechLLMs' self-attention provide sufficiently stable alignment information to support streaming decisions as in AED models?*

To answer this question, we propose **D**ecoder-**O**nly **A**ttention (DOA), the first training-free attention-based policy for long-form simultaneous

translation with decoder-only SpeechLLMs.[1] DOA derives a proxy cross-attention matrix directly from decoder self-attention weights and exploits the resulting alignments to drive incremental generation while dynamically pruning both acoustic and textual histories during streaming inference for efficient long-form processing.

Experiments on popular off-the-shelf SpeechLLMs, Phi4-Multimodal and Qwen3-Omni, on English→German and English→Italian translation show that DOA can effectively generalize beyond encoder-decoder architectures, achieving low-latency streaming generation with quality close to offline decoding without task-specific retraining.

2 Decoder-Only Attention Policy

2.1 Simultaneous Policy

We propose a **d**ecoder-**o**nly **a**ttention-based (DOA) streaming policy for SpeechLLMs, adapting the policy originally introduced for attention-based encoder-decoder (AED) models (Papi et al., 2024) to architectures without explicit cross-attention. Unlike AED systems, decoder-only models process speech and text within a single autoregressive sequence, making source-target alignments less explicit. To address this limitation, we derive a *proxy cross-attention* signal from decoder self-attention weights. During autoregressive generation, we extract decoder self-attention weights

$$A^{(l,h)} \in \mathbb{R}^{T\times(S+T)},$$

where l is the decoder layer, h is the attention head, and S and T denote the number of audio and generated text tokens, respectively. Since audio tokens occupy the prefix of the sequence, we isolate attention directed toward the acoustic context:

$$\tilde{A}^{(l,h)} = A^{(l,h)}[:, : S],$$

obtaining a proxy cross-attention matrix

$$\tilde{A} \in \mathbb{R}^{T\times S},$$

analogous to the cross-attention matrix in AED models.

The attention matrix can be extracted from a single layer/head or averaged across selected layers and heads:

$$\bar{A} = \frac{1}{|\mathcal{L}||\mathcal{H}|}\sum_{l\in\mathcal{L}}\sum_{h\in\mathcal{H}}\tilde{A}^{(l,h)}.$$

[1]Code will be released under Apache 2.0 license upon paper acceptance.

Inspired by prior works (Papi et al., 2024; Wang et al., 2024), the resulting matrix is converted into a token-to-audio alignment by selecting, for each generated token, the audio position receiving the highest attention score:

$$a_t = \arg\max_s \bar{A}_{t,s},$$

where a_t denotes the aligned audio index for token t. The alignment sequence is then used by the streaming policy to estimate whether newly generated tokens are grounded in the currently available audio context. We introduce a cutoff hyperparameter f representing the number of most recently received audio frames considered acoustically unstable. Tokens aligned to positions falling within the last f frames are not emitted, as their supporting acoustic evidence may still evolve with future input. Therefore, only tokens satisfying

$$a_t < S - f$$

are committed to the output, where S denotes the current number of audio frames. Larger values of f increase the amount of audio treated as unstable, leading to more conservative generation and higher latency, while smaller values favor lower latency at the risk of emitting less stable hypotheses.

2.2 Long-form Adaptation

The DOA policy operates incrementally over a rolling, potentially extremely long, audio sequence. At each step, the newly received speech chunk is appended to the previously observed waveform, the *audio history*, and provided to the model together with a retained *textual history* used as prefix for the next decoding step. To avoid both histories to grow indefinitely, audio and textual history selection mechanisms are adopted.

Following prior streaming frameworks (Iranzo-Sánchez et al., 2022, 2024; Papi et al., 2024), we explore two history selection strategies for the textual part: *(i)* **fixed words**, which preserves the last N generated words/characters, and *(ii)* **punctuation**, which preserves only the text segment after the most recent strong punctuation mark.

The proxy cross-attention-based alignments are then used for the audio history selection. Specifically, we discard consecutive audio frames aligned exclusively with the discarded textual history, while retaining the frames aligned with the preserved textual prefix and the newly generated hypothesis. This mechanism progressively removes acoustic

segments already translated and no longer required by the model, enabling long-form streaming inference without unbounded context growth. As an edge case, where attention-based pruning fails to discard any audio frames and the audio history exceeds a predefined duration, the oldest frames are truncated to maintain bounded memory usage.

Our framework is completely *model-agnostic* and only requires access to decoder self-attention weights, making it applicable to any SpeechLLM.

3 Experimental Settings

Data and Metrics. Following standard evaluation settings of the IWSLT Evaluation Campaigns (Abdulmumin et al., 2025), we adopt MCIF (Papi et al., 2026) as test set for English→German and English→Italian, and ACL 60/60 (Salesky et al., 2023) English→German as dev set for the hyperparameters selection and analyses. Always following IWSLT 2026 settings, we use the SimulStream toolkit (Gaido et al., 2025) as the inference framework, and OmniST-Eval (Polák et al., 2025) to compute the LongYAAL and LongLAAL latency metrics. BLEU and COMET (Rei et al., 2022) are used as quality metrics.

Models. For the experiments, we selected open-weight SpeechLLMs based on their language support, therefore capable of translating from English into German and Italian. This resulted into the selection of Phi4-Multimodal (Microsoft et al., 2025) and Qwen3-Omni (Xu et al., 2025). Hyperparameters selection and analyses are performed on Phi4-Multimodal only, and the model-agnosticity is verified on Qwen3-Omni in the final results. We also compare with the SimulStream baseline, StreamAtt (Papi et al., 2024), applied to the AED SeamlessM4T model (Seamless Communication et al., 2023). Detailed settings are in Appendix A.

4 Results

Punctuation vs. Fixed Textual History Selection. Figure 1 shows the results of the DOA policy applied to Phi4-Multimodal with the two textual history selection methods, Fixed Words and Punctuation, with proxy cross-attention matrices obtained by averaging across layers and heads. The curves clearly show that the Punctuation method is more stable, yielding the best or near-best quality across all latency regimes, while remaining comparable in terms of latency (spanning between 1.5 and 3.5s). Interestingly, this behavior contrasts with current findings on AED-based streaming policies, where Fixed Words generally outperformed Punctuation-based history selection (Papi et al., 2024). Prior work has shown that punctuation and sentence segmentation are important for improving machine translation quality (Cho et al., 2017b,a), and that contextual continuity is beneficial in transformer-based speech sequence modeling (Żelasko et al., 2021; Huang et al., 2023). Consistent with these findings, our results suggest that decoder-only SpeechLLMs benefit from punctuation-based history selection, which preserves sentence-level textual continuity and yields a more stable autoregressive decoding context than retaining a fixed number of words. These results motivate the adoption of the Punctuation strategy throughout the remainder of the paper.

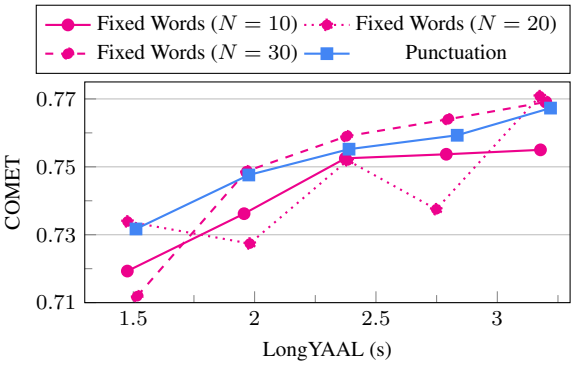
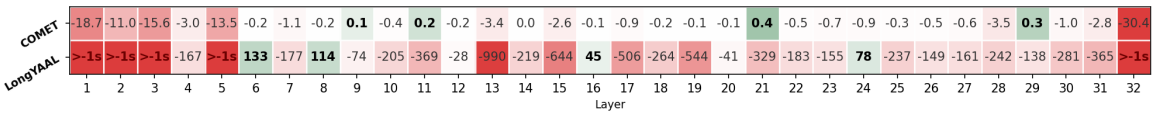


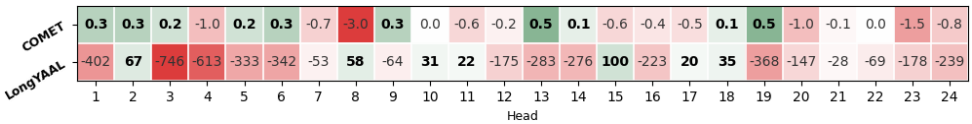
Figure 1: Latency (LongLAAL↓) - Quality (COMET↑) curves of Punctuation and Fixed Words methods applied to Phi4-Multimodal on ACL 60/60 en-de dev set. Numerical results are in Appendix B.

Layers and Heads Analysis. Figure 2 shows the performance difference between averaging the proxy cross-attention matrix across both layers and heads, and selecting a specific layer or head. Notably, the performance across layers is more unstable than across heads, with six layers leading to a complete failure (latency degradation close or more than 1s). No layer leads to gains on both latency and quality with respect to averaging, in contrast with previous works that found specific layer selection outperforms the average (Papi et al., 2023). Conversely, some heads improves in both quality and latency (e.g., heads 2 and 10) but gains are minimal. Therefore, the average across layers and heads represents the easiest and best-performing choice, and it is used for final results.

Final Results. Figure 3 reports the performance of the proposed DOA policy applied to both off-

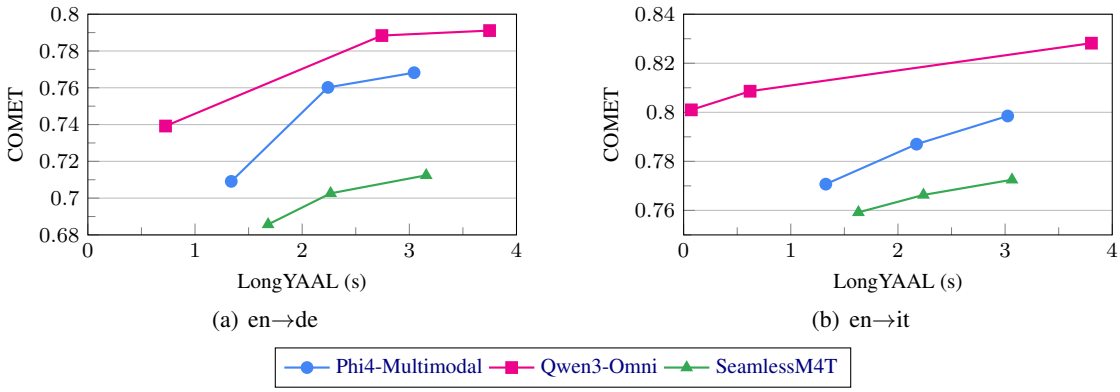


(a) Layer-wise Performance



(b) Head-wise Performance

Figure 2: Layer- and Head-wise performance difference compared to the average. Green squares indicate improvement, and red degradation, with their magnitude (COMET $\times 100$ and LongYAAL in ms for readability).



(a) en→de (b) en→it

Figure 3: Latency (LongYAAL↓) - Quality (COMET↑) curves on MCIF of DOA policy on Phi4-Multimodal and Qwen3-Omni, and of StreamAtt baseline on SeamlessM4T. Numerical results are in Appendix B.

the-shelf Phi4-Multimodal and Qwen3-Omni using attention averaged across layers and heads. Both models achieve highly competitive translation quality, approaching the performance of offline systems reported on the same benchmark (Papi et al., 2026), where the best SpeechLLM (Phi4-Multimodal) reaches COMET scores of 0.78 on English-German and 0.81 on English-Italian. In terms of latency, Qwen3-Omni covers a broad operating range, spanning from 400ms to 3.8s average LongYAAL across languages, while also achieving the best overall translation quality. Phi4-Multimodal, although at lower COMET scores (with a difference of 0.02 on average) compared to Qwen3-Omni, is easier to manage in terms of latency, as increasing the cut-off frame by 10 always corresponds to 800ms latency increase. Notably, both DOA-equipped SpeechLLMs consistently outperform the StreamAtt AED baseline in terms of latency-quality trade-off, demonstrating the effectiveness of leveraging decoder self-attention for streaming decisions. Finally, the strong performance obtained by both Phi- and Qwen-based models highlights the robustness and model-agnostic nature of DOA. Despite the substantial architectural differences between the two families (dense vs. MoE), the proposed policy generalizes effectively without requiring any adaptation.

5 Conclusions

We presented DOA, a training-free policy that enables long-form simultaneous speech-to-text translation with off-the-shelf decoder-only SpeechLLMs by exploiting self-attention as a proxy alignment signal. Our results show that **decoder self-attention provides sufficiently stable information to drive effective streaming decisions**, making it possible to repurpose offline SpeechLLMs for SimulST without retraining. We also find that punctuation-based history selection is consistently more effective than fixed-word strategies, and that averaging across layers and heads leads to the best results. These findings suggest that attention-based alignment can generalize beyond encoder-decoder models and serve as a simple yet effective mechanism for streaming inference in SpeechLLMs.

Limitations

Our evaluation is limited to English source speech, primarily due to the scarcity of publicly available speech translation benchmarks with continuous audio of several minutes of duration. While this setting is representative of most long-form SimulST works (Papi et al., 2024; Ouyang et al., 2025), it does not fully capture multilingual or cross-domain variability in real-world deployments.

In addition, we focus on output languages using Latin scripts (German, Italian). It remains an open question whether the proposed proxy alignment derived from decoder self-attention generalizes equally well to languages with different scripts or tokenization characteristics, such as logographic (e.g., Chinese or Japanese) or morphologically rich languages (e.g., Turkish or Finnish).

Finally, we do not report computationally aware latency due to a heterogeneous GPU execution environment across experiments. While we provide *ideal* latency comparisons by following standard best practices (Abdulmumin et al., 2025), computational overhead and hardware requirements is substantially different across models (especially of Qwen3-Omni versus Phi4-Multimodal and SeamlessM4T), and therefore absolute timing comparisons across architectures should be interpreted with caution. Besides hardware inhomogeneity, other factors such as codebase optimization and different HuggingFace repository versions required by each model (see Appendix A) can lead to significant computationally aware latency differences (Chitty-Venkata et al., 2024), and we believe that comprehensively analyzing these aspects is out of scope for this work.

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, Fortuné Kponou, Mateusz Krubiński, Tsz Kin Lam, Danni Liu, Evgeny Matusov, Chandresh Kumar Maurya, John P. McCrae, Salima Mdhaffar, Yasmin Moslem, Kenton Murray, Satoshi Nakamura, Matteo Negri, Jan Niehues, Atul Kr. Ojha, John E. Ortega, Sara Papi, Pavel Pecina, Peter Polák, Piotr Połeć, Ashwin Sankar, Beatrice Savoldi, Nivedita Sethiya, Claytone Sikasote, Matthias Sperber, Sebastian Stüker, Katsuhito Sudoh, Brian Thompson, Marco Turchi, Alex Waibel, Patrick Wilken, Rodolfo Zevallos, Vilém Zouhar, and Maike Züfle. 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, Kenton Murray, Satoshi Nakamura, Matteo Negri, Jan Niehues, Xing Niu, Atul Kr. Ojha, John Ortega, Sara Papi, Peter Polák, Adam Pospíšil, Pavel Pecina, Elizabeth Salesky, Nivedita Sethiya, Balaram Sarkar, Jiatong Shi, Claytone Sikasote, Matthias Sperber, Sebastian Stüker, Katsuhito Sudoh, Brian Thompson, Alex Waibel, Shinji Watanabe, Patrick Wilken, Petr Zemánek, and Rodolfo Zevallos. 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.

Zhehuai Chen, He Huang, Oleksii Hrinchuk, Krishna C. Puvvada, Nithin Rao Koluguri, Piotr Żelasko, Jagadeesh Balam, and Boris Ginsburg. 2024. Bestow: Efficient and streamable speech language model with the best of two worlds in gpt and t5. In *2024 IEEE Spoken Language Technology Workshop (SLT)*, pages 147–154.

Krishna Teja Chitty-Venkata, Siddhisanket Raskar, Bharat Kale, Farah Ferdaus, Aditya Tanikanti, Ken Raffenetti, Valerie Taylor, Murali Emani, and Venkatram Vishwanath. 2024. Llm-inference-bench: Inference benchmarking of large language models on ai accelerators. In *SC24-W: Workshops of the International Conference for High Performance Computing, Networking, Storage and Analysis*, pages 1362–1379. IEEE.

Eunah Cho, Jan Niehues, and Alex Waibel. 2017a. Domain-independent punctuation and segmentation insertion. In *Proceedings of the 14th International Conference on Spoken Language Translation*, pages 74–81, Tokyo, Japan. International Workshop on Spoken Language Translation.

Eunah Cho, Jan Niehues, and Alex Waibel. 2017b. NMT-Based Segmentation and Punctuation Insertion for Real-Time Spoken Language Translation. In *Interspeech 2017*, pages 2645–2649.

Christian Fügen, Alex Waibel, and Muntsin Kolss. 2007. Simultaneous translation of lectures and speeches. *Machine translation*, 21:209–252.

Marco Gaido, Sara Papi, Mauro Cettolo, Matteo Negri, and Luisa Bentivogli. 2025. Simulstream:

Open-source toolkit for evaluation and demonstration of streaming speech-to-text translation systems. *Preprint*, arXiv:2512.17648.

Alvin Grissom II, He He, Jordan Boyd-Graber, John Morgan, and Hal Daumé III. 2014. Don't until the final verb wait: Reinforcement learning for simultaneous machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1342–1352, Doha, Qatar. Association for Computational Linguistics.

Shoutao Guo, Xiang Li, Mengge Liu, Wei Chen, and Yang Feng. 2025a. Streamuni: Achieving streaming speech translation with a unified large speech-language model. *Preprint*, arXiv:2507.07803.

Shoutao Guo, Shaolei Zhang, Zhengrui Ma, and Yang Feng. 2025b. Large language models are read/write policy-makers for simultaneous generation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(22):23969–23977.

Ankit Gupta, George Saon, and Brian Kingsbury. 2024. Exploring the limits of decoder-only models trained on public speech recognition corpora. In *Interspeech 2024*, pages 252–256.

Chao-Wei Huang, Hui Lu, Hongyu Gong, Hirofumi Inaguma, Ilia Kulikov, Ruslan Mavlyutov, and Sravya Popuri. 2024. Investigating Decoder-only Large Language Models for Speech-to-text Translation. In *Interspeech 2024*, pages 832–836.

W. Ronny Huang, Hao Zhang, Shankar Kumar, Shuo-Yiin Chang, and Tara Sainath. 2023. Semantic Segmentation with Bidirectional Language Models Improves Long-form ASR. In *Interspeech 2023*, pages 2778–2782.

Javier Iranzo-Sánchez, Jorge Civera, and Alfons Juan. 2022. From simultaneous to streaming machine translation by leveraging streaming history. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6972–6985, Dublin, Ireland. Association for Computational Linguistics.

Javier Iranzo-Sánchez, Jorge Iranzo-Sánchez, Adrià Giménez, Jorge Civera, and Alfons Juan. 2024. Segmentation-free streaming machine translation. *Transactions of the Association for Computational Linguistics*, 12:1104–1121.

Mingbo Ma, Liang Huang, Hao Xiong, Renjie Zheng, Kaibo Liu, Baigong Zheng, Chuanqiang Zhang, Zhongjun He, Hairong Liu, Xing Li, Hua Wu, and Haifeng Wang. 2019. STACL: Simultaneous translation with implicit anticipation and controllable latency using prefix-to-prefix framework. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3025–3036, Florence, Italy. Association for Computational Linguistics.

Xutai Ma, Juan Pino, and Philipp Koehn. 2020. 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.

Microsoft, Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, Jianmin Bao, Alon Benhaim, Martin Cai, Vishrav Chaudhary, Congcong Chen, Dong Chen, Dongdong Chen, Junkun Chen, Weizhu Chen, Yen-Chun Chen, Yi ling Chen, Qi Dai, Xiyang Dai, Ruchao Fan, Mei Gao, Min Gao, Amit Garg, Abhishek Goswami, Junheng Hao, Amr Hendy, Yuxuan Hu, Xin Jin, Mahmoud Khademi, Dongwoo Kim, Young Jin Kim, Gina Lee, Jinyu Li, Yunsheng Li, Chen Liang, Xihui Lin, Zeqi Lin, Mengchen Liu, Yang Liu, Gilsinia Lopez, Chong Luo, Piyush Madan, Vadim Mazalov, Arindam Mitra, Ali Mousavi, Anh Nguyen, Jing Pan, Daniel Perez-Becker, Jacob Platin, Thomas Portet, Kai Qiu, Bo Ren, Liliang Ren, Sambuddha Roy, Ning Shang, Yelong Shen, Saksham Singhal, Subhojit Som, Xia Song, Tetyana Sych, Praneetha Vaddamanu, Shuohang Wang, Yiming Wang, Zhenghao Wang, Haibin Wu, Haoran Xu, Weijian Xu, Yifan Yang, Ziyi Yang, Donghan Yu, Ishmam Zabir, Jianwen Zhang, Li Lyna Zhang, Yunan Zhang, and Xiren Zhou. 2025. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras. *Preprint*, arXiv:2503.01743.

Siqi Ouyang, Xi Xu, Chinmay Dandekar, and Lei Li. 2024. Fasst: Fast llm-based simultaneous speech translation. *arXiv preprint arXiv:2408.09430*.

Siqi Ouyang, Xi Xu, and Lei Li. 2025. 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.

Sara Papi, Matteo Negri, and Marco Turchi. 2023. Attention as a guide for simultaneous speech translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13340–13356, Toronto, Canada. Association for Computational Linguistics.

Sara Papi, Maike Züfle, Marco Gaido, Beatrice Savoldi, Danni Liu, Ioannis Douros, Luisa Bentivogli, and Jan Niehues. 2026. MCIF: Multimodal crosslingual instruction-following benchmark from scientific

talks. In *The Fourteenth International Conference on Learning Representations*.

Peter Polák and Ondřej Bojar. 2023. Long-form end-to-end speech translation via latent alignment segmentation. *arXiv preprint arXiv:2309.11384*.

Peter Polák, Sara Papi, Luisa Bentivogli, and Ondřej Bojar. 2025. Better late than never: Evaluation of latency metrics for simultaneous speech-to-text translation. *arXiv preprint arXiv:2509.17349*.

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). 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.

Seamless Communication, Loïc Barrault, Yu-An Chung, Mariano Cora Meglioli, David Dale, Ning Dong, Paul-Ambroise Duquenne, Hady Elsahar, Hongyu Gong, Kevin Heffernan, John Hoffman, Christopher Klaiber, Pengwei Li, Daniel Licht, Jean Maillard, Alice Rakotoarison, Kaushik Ram Sadagopan, Guillaume Wenzek, Ethan Ye, Bapi Akula, Peng-Jen Chen, Naji El Hachem, Brian Ellis, Gabriel Mejia Gonzalez, Justin Haaheim, Prangthip Hansanti, Russ Howes, Bernie Huang, Min-Jae Hwang, Hirofumi Inaguma, Somya Jain, Elahe Kalbassi, Amanda Kallet, Ilia Kulikov, Janice Lam, Daniel Li, Xutai Ma, Ruslan Mavlyutov, Benjamin Peloquin, Mohamed Ramadan, Abinesh Ramakrishnan, Anna Sun, Kevin Tran, Tuan Tran, Igor Tufanov, Vish Vogeti, Carleigh Wood, Yilin Yang, Bokai Yu, Pierre Andrews, Can Balioglu, Marta R. Costa-jussà, Onur Celebi, Maha Elbayad, Cynthia Gao, Francisco Guzmán, Justine Kao, Ann Lee, Alexandre Mourachko, Juan Pino, Sravya Popuri, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, Paden Tomasello, Changhan Wang, Jeff Wang, and Skyler Wang. 2023. Seamlessm4t: Massively multilingual & multimodal machine translation. *Preprint*, arXiv:2308.11596.

Haoyu Wang, Guoqiang Hu, Guodong Lin, Wei-Qiang Zhang, and Jian Li. 2024. Simul-Whisper: Attention-Guided Streaming Whisper with Truncation Detection. In *Interspeech 2024*, pages 4483–4487.

Jian Wu, Yashesh Gaur, Zhuo Chen, Long Zhou, Yimeng Zhu, Tianrui Wang, Jinyu Li, Shujie Liu, Bo Ren, Linquan Liu, et al. 2023. On decoder-only architecture for speech-to-text and large language model integration. In *2023 IEEE automatic speech recognition and understanding workshop (ASRU)*, pages 1–8. IEEE.

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, Baosong Yang, Bin Zhang, Ziyang Ma, Xipin Wei, Shuai Bai, Keqin Chen, Xuejing Liu, Peng Wang, Mingkun Yang, Dayiheng Liu, Xingzhang Ren, Bo Zheng, Rui Men, Fan Zhou, Bowen Yu, Jianxin Yang, Le Yu, Jingren Zhou, and Junyang Lin. 2025. Qwen3-omni technical report. *Preprint*, arXiv:2509.17765.

Piotr Żelasko, Raghavendra Pappagari, and Najim Dehak. 2021. What helps transformers recognize conversational structure? importance of context, punctuation, and labels in dialog act recognition. *Transactions of the Association for Computational Linguistics*, 9:1163–1179.

A Detailed Experimental Settings

Information about model version and weights are provided in Table 1.

For SpeechLLMs, the incremental speech chunk size received at each step is set to 1 second, the default of SimulStream. To obtain quality-latency curves, typical of SimulST evaluation (Ma et al., 2020), we vary the cutoff frame f in $\{5, 10, 15, 20, 25\}$ for Phi4-Multimodal during parameter selection. For the final results, both Phi4-Multimodal and Qwen3-Omni cut-off frame is varied in $\{5, 15, 25\}$ for both target languages to obtain the three latency regimes: low, medium, and high latency. For the AED baseline, the SeamlessM4T v1 medium model, default settings from SimulStream are adoped and the cut-off frame is varied in $\{4, 8, 12\}$.

The maximum audio length is set to $120s$ for Phi4-Multimodal and SeamlessM4T, and $90s$ Qwen3-Omni as a fallback in the edge cases in which attention does not discard enough frames and the history risks to exceed the available memory. The maximum number of new tokens allowed to be generated at each step is set to 32, and the maximum textual history length (in tokens) is 128.

Following the specific model cards, the prompts used for inference are the following(a simpler one for Phi4-Multimodal,[2] a more complex one for Qwen3-Omni):

Phi4-Multimodal

```
Translate the audio to {tgt_lang}.
```

Qwen3-Omni

```
You are a professional {src_lang}-to-{tgt_lang} translator. Your goal is to accurately convey the meaning and nuances of the original {src_lang} speech while adhering to {tgt_lang} grammar, vocabulary, and cultural sensitivities. Use precise terminology and a tone appropriate for academic or instructional materials. Produce only the {tgt_lang} translation, without any additional explanations or commentary. Please translate the provided {src_lang} speech into {tgt_lang}:
```

The inference is conducted on a mixed environment with NVIDIA A40 40GB, and NVIDIA L40S 48GB. A single GPU is used. On average, a single run takes ∼1-2 hours for SeamlessM4T, ∼4-5 hours for Phi4-Multimodal, ∼25-26 hours for Qwen3-Omni.

[2]Preliminary experiments on Phi4-Multimodal with more complex prompts led to significant degradation in the results.

B Numerical Results

Tables 2 and 3 show numerical results for Figures 1 and 3, together with quality and latency complementary metrics (BLEU score and StreamLAAL, respectively).

C AI Use Statement

AI tools such as ChatGPT have been used for polishing the writing of the paper, and Codex has been used only for debugging purposes.

Model	Param.	Weights	HFv
Phi4-Multimodal (Microsoft et al., 2025)	5.6B	`microsoft/Phi-4-multimodal-instruct`	4.48.2
Qwen3-Omni (Xu et al., 2025)	30B	`Qwen/Qwen3-Omni-30B-A3B-Instruct`	5.0.0
SeamlessM4T (Seamless Communication et al., 2023)	1B	`facebook/hf-seamless-m4t-medium`	4.48.2

Table 1: Details of the analyzed models, including the number of parameters, their public weights release, and the HuggingFace Transformer version (HFv) used for the experiments.

Method	f	BLEU↑	COMET↑	LongYAAL↓	LongLAAL↓	empty%↓
punctuation	5	30.12	0.7317	1511	1599	0.43%
	10	32.23	0.7476	1976	2150	0.21%
	15	32.80	0.7552	2389	2487	0.43%
	20	33.98	0.7593	2835	2947	0.43%
	25	34.70	0.7673	3219	3460	0.21%
fixed words ($N = 10$)	5	29.02	0.7193	1476	1560	0.00%
	10	31.29	0.7362	1957	2050	0.00%
	15	32.95	0.7525	2374	2471	0.00%
	20	33.74	0.7537	2790	2891	0.00%
	25	33.97	0.755	3178	3305	0.43%
fixed words ($N = 20$)	5	29.37	0.7339	1476	1619	0.21%
	10	27.01	0.7274	1980	3825	0.21%
	15	31.13	0.7519	2383	2927	0.00%
	20	30.30	0.7375	2748	3511	0.21%
	25	34.36	0.7708	3175	3317	0.43%
fixed words ($N = 30$)	5	16.74	0.7119	1516	7791	0.21%
	10	32.95	0.7487	1966	2049	0.43%
	15	33.74	0.759	2379	2476	0.43%
	20	34.08	0.764	2797	2897	0.43%
	25	34.57	0.7691	3200	3312	0.21%

Table 2: Numerical results for systems reported in Figure 1 on ACL 60/60.

Model	f	BLEU↑	COMET↑	LongYAAL↓	LongLAAL↓	empty%↓
en-de						
SeamlessM4T	4	21.95	0.6857	1682	1825	1.09
	8	23.77	0.7026	2265	2379	1.20
	12	24.58	0.7124	3157	3292	0.98
Phi4Multimodal	5	15.45	0.7091	1338	7739	0.44
	15	29.40	0.7602	2240	2334	0.33
	25	30.88	0.7682	3044	3161	0.22
Qwen3-Omni	5	24.59	0.7392	725	896	9.25
	15	26.49	0.7884	2744	3959	3.70
	25	28.18	0.7911	3749	4889	0.44
en-it						
SeamlessM4T	4	33.37	0.7592	1631	1735	0.76
	8	34.58	0.7663	2238	2319	0.65
	12	36.48	0.7725	3064	3209	0.65
Phi4Multimodal	5	29.73	0.7707	1327	2950	1.20
	15	32.26	0.787	2174	4051	1.20
	25	33.68	0.7985	3024	4891	0.11
Qwen3-Omni	5	34.78	0.801	72	998	3.81
	15	37.26	0.8086	620	1926	3.70
	25	38.14	0.8282	3806	5466	0.33

Table 3: Numerical results for systems reported in Figure 3 on MCIF.