Towards Dynamic Attention Masking for Simultaneous Speech Translation

Benjamin Pong
University of Washington
Seattle WA, USA
{benpong}@uw.edu

Abstract

We present a proof-of-concept system for simultaneous speech translation based on dynamic attention masking. Our approach builds on SeamlessM4T by injecting lightweight per-layer schedulers into the conformer-encoder, training each scheduler to predict the number of future frames needed for translation. The schedulers are trained jointly with LoRA adapters across three language directions: English to German, Italian, and Chinese. At inference time, we evaluate our system using sliding window retranslation inference regime (Sen et al., 2022), and an adapted version of StreamAtt (Papi et al., 2024) that replaces the fixed cutoff with a content-aware threshold derived from the learnt representations from the scheduler outputs.

1 Introduction

This system paper describes a system submission[1] to the IWSLT2026 Simultaneous Speech Translation Track (Adelani et al., 2026). Our system is built on top of SeamlessM4T, with a dynamic masking attention mechanism where each frame learns how much future context is needed. The motivation behind the approach is inspired by ANCAT (Strimel et al., 2023) which was implemented for real-time ASR systems. To our knowledge, this approach has not been applied to real-time speech translation. We then ran experiments with the dynamic mask model on two simultaneous policies; the sliding window retranslation policy (Sen et al., 2022) and StreamAtt (Papi et al., 2024). In an attempt to make the dynamic mask model compatible with StreamAtt, we further adapted the original fixed cutoff policy by replacing the cutoff threshold with learnt representations from the dynamic mask model.

[1]Code is available at https://github.com/Benjamin-Pong/Simultaneous-Speech-Translation.git

Our goal is to prototype this dynamic masking attention mechanism that was originally designed for simultaneous ASR task to multilingual Speech Translation task, and test out its effects on the state-of-the-art simultaneous policies.

2 Related Work

Simultaneous speech-to-text translation (SST) requires systems to produce translated segments incrementally while receiving audio input. A key challenge in any streaming task is finding out how much future context frames the encoder should attend to at each frame. A fundamental tension in streaming speech processing lies between causal and non-causal attention. In causal models, each frame can only attend to past frames, permitting low-latency emission with limited context. On the other hand, in non-causal models frames attend to both past and future frames, improving translation quality at the cost of increased latency. Several approaches have been proposed to find the balance between this tradeoff. For example, chunking approaches have been designed for streaming ASR (Shi et al., 2020; Chen et al., 2021) and have been recently been applied to Speech Translation (Ouyang et al., 2025). More recently, ANCAT (Strimel et al., 2023) proposed an adaptive lookahead for streaming Automatic Speech Recognition task. They proposed schedulers that dynamically determine how many future frames to attend to at each step, conditioned on the previous layer's hidden states. This approach has yet to be adapted for simultaneous speech translation.

Two recent simultaneous speech translation inference paradigms that have been proposed are relevant for the current proposed work. Streamatt (Papi et al., 2024) is an inference paradigm that uses cross-attention scores to decide when tokens should be emitted. The intuition is that if a token attends to recent frames, the model may not have suf-

Proceedings of the 23rd International Conference on Spoken Language Translation (IWSLT 2026), pages 183–188

ficient context to generate a decent translation, and is therefore with-held until more audio is present. StreamAtt also introduces a history selection mechanism that discards audio frames no longer attended by the current textual history, allowing continuous streaming without unbounded memory. The second paradigm is the sliding window retranslation inference regime (Sen et al., 2022) where the input audio is processed in overlapping fixed-length windows. At each time step, the model retranslates the entire window and a deduplication mechanism based on longest common subsequence between contiguous windows.

3 Model Architecture

Our work adapts ANCAT to multilingual simultaneous speech translation, applied to Transformer-based encoder-decoder model, which is a first, to the best of our knowledge. For inference, we apply both the sliding window retranslation regime, and a modified version of StreamAtt that uses the learnt representations from the model to decide the emission policy.

Our system builds on Seamlessm4t-medium (Communication et al., 2023), a multilingual multimodal translation model. It employs a conformer-based speech encoder consisting 12 layers, combining multi-head self-attention mechanism with convolutional module to capture both long-range and local acoustic dependencies.

We implement a dynamic attention masking mechanism to constrain the conformer encoder's future lookahead during training. Specifically, we incorporate a scheduler into each of the 12 conformer layers. Each scheduler consists a two-layer feedforward network that maps the hidden states of the previous frame to a scalar score o_i for each frame i. To reduce overfitting, the input hidden state is projected from the full hidden size d (1024 for Seamlessm4T) to $d/8$ (64).

This o_i score represents the model's learned assessment of how many future frames are needed at the current frame i. It is then used to compute a soft attention mask which is applied to the self-attention weights additively to suppress the attention to future frames. The maximum lookahead budget K_{max} is a hyperparameter that controls the strictness of the constraint. We use a $K_{max} = 24$ for English $\rightarrow German, Italian$, and $K_{max} = 16$ for English $\rightarrow Chinese$. These lookahead budgets were selected based on quality performance on the development dataset.

These schedulers undergo full parameter finetuning, alongside LoRA adapters (Hu et al., 2021) applied to all output projection matrices of the self-attention layers in both the speech encoder and decoder. All other pretrained weights remain frozen. The best checkpoint is selected based on validation BLEU score. The training parameters used can be found in Appendix A.

4 Experiments

4.1 Data

We finetuned the en-de model on a data mixture of Europarl (Iranzo-Sánchez et al., 2020) and CoVoST2 (Wang et al., 2021), the en-it model on Europarl and the en-zh model on CoVoST2. Only the train-dev splits were used for finetuning.

The systems were evaluated on the Multimodal Crosslingual Instruction Following (MCIF) dataset which is a multilingual human-annotated benchmark (Papi et al., 2026). It contains audio recordings of conference talks from the Association of Computational Linguistics (ACL), both in long-form and short-form formats. For our purposes, we evaluate our systems on the long-form format of MCIF.

4.2 Inference Strategy

4.2.1 Sliding Window Retranslation

We evaluated our system using the sliding window retranslation approach designed by (Sen et al., 2022) implemented in the Simulstream toolkit (Gaido et al., 2025). This approach ingests a fixed-length of audio that overlaps in length. Each window is translated independently. To account for duplicate words translated due to the overlaps, a deduplication approach using the longest common subsequence between consecutive windows is applied. Across all experiments, we fix a window length of 12s which is biased towards quality since a larger window provides more context for retranslation, but lowers latency. We also set the minimum threshold for the minimum number of overlapping tokens between windows to 0.1. This parameter is used to control when tokens get emitted.

4.2.2 Adapting StreamAtt for Dynamic Mask Model

Additionally, we also applied the StreamAtt inference regime by adapting StreamAtt for our dynamic mask model. In StreamAtt, there is a cutoff

Lang	Chunk	Dynamic SeamlessM4T Model chrF	COMET	YAAL	SeamlessM4T chrF	COMET	YAAL
En→De	2s	37.17	0.506	-14548	51.62	0.690	3542
	4s	45.85	0.574	**3249**	51.59	0.702	3633
	6s	44.94	0.592	3624	49.73	0.698	3748
	8s	44.03	0.598	3586	48.13	0.694	3359
En→Zh	2s	17.95	0.474	4157	22.29	0.543	3576
	4s	17.24	0.484	4079	21.81	0.546	3499
	6s	15.97	0.482	3712	21.47	0.539	3356
	8s	15.43	0.496	3598	20.00	0.542	3144
En→It	2s	45.55	0.460	3483	58.07	0.681	3337
	4s	44.27	0.496	3506	57.72	0.678	3507
	6s	44.15	0.510	3652	55.65	0.664	3310
	8s	40.96	0.499	3233	54.23	0.647	3411

Table 1: Results for sliding window policy on MCIF dev set. YAAL = LongYAAL (CU) in ms. Bold indicates best latency per language direction. German 2s dynamic results are excluded from analysis due to alignment issues.

Lang	Chunk	Dynamic SeamlessM4T Model chrF	COMET	YAAL	SeamlessM4T chrF	COMET	YAAL
En→De	2s	33.81	0.4558	3643.0	52.74	0.8057	927.0
	4s	31.17	0.4433	1976.0	52.56	0.8090	2793.0
	6s	29.98	0.4442	2012.0	46.08	0.7616	-102.0
	8s	28.65	0.4507	4500.0	45.41	0.7610	-1458.0
En→It	2s	27.70	0.3957	1634.0	61.67	0.7504	2270.0
	4s	25.6	0.4229	-557.0	57.40	0.7353	1612.0
	6s	24.22	0.4019	-1748.3	59.89	0.7533	3515.0
	8s	23.85	0.4038	-10609.0	54.22	0.7164	3914.0

Table 2: Results for modified StreamAtt policy on MCIF dev set. YAAL = LongYAAL (CU) in ms. Bold indicates best latency per language direction.

property that determines which generated tokens are safe to emit. A token is safe to emit only if its most-attended encoder frame is within the cutoff value. This cutoff value is defined as $T - f$, where T is the total number of encoded frames for an audio segment, and f is a fixed integer representing the number of frames from the end of the audio segment that are unreliable for emission. With a larger f-value, more tokens are suppressed, increasing latency but improves translation quality.

However, as the dynamic model uses partial causal masking, the StreamAtt's fixed cutoff implementation is incompatible. As part of an experimentation, we replaced this fixed cutoff with a context-aware cutoff derived from the o-scores produced by the trained schedulers in the dynamic mask model. We extracted the scheduler from the final layer of the encoder. The cutoff is set to the final frame whose o-score falls below a language specific threshold. The threshold is computed based on the mean of the distribution of o-scores per locale. Hence, any number of frames that fall outside of the o-score will not be emitted. Experimentations for the modified StreamAtt are only done on English-German and English-Italian.

4.3 Training and Inference toolkits

Training was done using Google Colab with one A100 GPU. Model development and finetuning were implemented using PyTorch, by inheriting and extending the SeamlessM4T architecture in HuggingFace Transformers library. The simultaneous inference pipeline was built by extending the SimulStream toolkit (Gaido et al., 2025), inheriting and modifying the StreamAtt and sliding window retranslation classes for integration with our dynamic attention masking architecture

4.4 Evaluation Metrics

Three evaluation metrics were used for evaluation. For translation quality, we report chrF (Popović, 2015) and COMET (Guerreiro et al., 2023) using the Unbabel/XCOMET-XL model. For latency, we report LongYAAL (Polák et al., 2026), the non-computation-aware variant which measures the average lagging between of unsegmented long form speech by using a soft alignment tool to match

reference and predicted segments. LongYAAL includes words beyond the segment boundary and excludes final tailed words towards the end of the stream to avoid the segmentation bias of short-form latency metrics.

5 Results

Table 1 presents chrF, COMET, and LongYAAL (CU) scores for both systems across all language directions and chunk sizes under the sliding window retranslation inference strategy. In general, under the sliding window regime, translation quality decreases as chunk size increases for chrF but generally increases for COMET. Latency generally increases as speech chunks increase for en-de and en-it, but the converse occurs for en-zh. Across all language directions, both systems operate within the high latency regime given the hyperparameters chosen. The results also show that dynamic masking does not show latency advantage over the baseline. We also attribute the high negative latency to alignment issues at training time.

Notably, while chrF declines for both systems with larger chunk sizes, the dynamic mask model shows some improvement in COMET scores from 4s to 8s across all language directions (German: + 0.024, Chinese: +0.012, Italian: +0.003), whereas the baseline consistently degrades (German: -0.008, Chinese: -0.004, Italian: -0.031). This suggests that the dynamic mask model produces semantically more consistent translations at larger chunks.

As for the modified StreamAtt, the results are shown in Table 2. The o-score emission policy achieves lower latency than the StreamAtt baseline across chunk sizes, with English→Italian at 4s and 6s chunks falling into the Low latency regime (LongYAAL < 0), indicating the system emits translations ahead of the reference segment boundaries. English→German achieves 1976ms at 4s chunks, 817ms faster than the StreamAtt baseline.

6 Discussion

This section evaluates the results for both simultaneous policies across three dimensions; translation quality, robustness of translation quality to speech chunk sizes, and latency.

6.1 Translation Quality

With regards to translation quality, for both simultaneous policies, the dynamic mask model consistently underperformed the baseline. We attribute this to two potential issues. Primarily, the introduction of dynamic mask imposed attention constraints, resulting in a loss in translation quality from the base model. Secondly, translation quality could be compromised by a domain mismatch between the training data and evaluation data. The training data comes from general domain and/or political domain but the evaluation data falls within the academic/scientific domain with specialized vocabulary and style. More thorough hyperparameter search or data synthesis approaches can be used to recover translation quality. This shall be left for future work

6.2 Robustness of Translation Quality to speech chunk sizes

With reference to the results for the sliding window simultaneous policy, despite lower absolute scores for COMET compared to the baseline, the dynamic mask model show higher improvements in COMET scores from 4s to 8s speech chunks across all language directions compared to the baseline. This shows that while the dynamic mask model produces translations with lower surface-level alignment with the reference, the semantic quality of its translations improves with longer context.

Unlike the results for sliding window policy, the results for the modified StreamAtt show consistent rate or decline in translation quality for larger speech chunks.

6.3 Latency

For the sliding window policy, both systems operate within the high latency regime based on LongYAAL, under the sliding-window retranslation inference regime. The fact that the dynamic mask model does not show latency advantage over the base model is expected since the learned attention restriction is constrained to the encoder's representations and does not interact with the emission policy. In other words, the sliding window does not exploit the learnt dynamic lookahead budgets.

However, with the modified StreamAtt that is sensitive to attention mechanism, we are able to extract mixed signals. The o-score cutoff consistently achieves lower latency than the StreamAtt baseline for English→Italian but not for English→German. For English→German, the dynamic model achieves 1976ms at 4s chunks, 817ms faster than the baseline (2793ms). For English→Italian, the system falls into the Low la-

tency regime at 4s and 6s chunks, with LongYAAL of -557ms and -1748ms respectively, indicating that translations are emitted ahead of the reference segment boundaries. This demonstrates that the learned o-scores provide a meaningful adaptive signal for emission timing. It is unclear why there is more variability for German translations, this investigation will be left for future work.

7 Conclusion

To our knowledge this system is the first to apply a dynamic masking approach to simultaneous speech translation. While we do see a severe loss in translation quality due to the changes in architecture, we do see partial results in latency advantages when we incorporate the learnt representations of the scheduler with StreamAtt simultaneous policy. As such we view this work as a first step towards training-time lookahead control for simulteneous speech translation, showing that encoder schedulers can learn meaningful lookahead budgets.

8 Limitations Future Work

A significant limitaton of our system is the quality gap between the dynamic mask model and the unmodified Seamlessm4t base model. Several directions may help to close this gap in future work. More thorough hyperparameter search or data synthesis approaches can be used to recover translation quality. Another promising direction for recovering translation quality is knowledge distillation from either the unmodified base model or Large Language Models by applying Bidirectional SeqKD (Inaguma et al., 2021) or Word-KD (Gaido et al., 2020).

Our training objective is limited to the base model's standard cross entropy loss. Future work may include incorporating latency regularization term similar to ANCAT (Strimel et al., 2023) by jointly optimizing translation and lookahead budget during finetuning, rather than imposing a hard constraint on the number of lookahead frames.

References

David Ifeoluwa Adelani, Victor Agostinelli, Antonios Anastasopoulos, Luisa Bentivogli, Ondřej Bojar, Sebastien Bratières, Marine Carpuat, Roldano Cattoni, Mauro Cettolo, Lizhong Chen, Marcello Federico, Marco Gaido, Mahendra Gupta, HyoJung Han, Ali Hatami, David Javorský, Yejin Jeon, Marek Kasztelnik, Antoine Laurent, and 33 others. 2026. Speech translation and metrics in 2026: Findings of the iwslt campaign. In *Proceedings of the 23rd International Conference on Spoken Language Translation (IWSLT 2026)*, San Diego, California, US. Association for Computational Linguistics.

Xie Chen, Yu Wu, Zhenghao Wang, Shujie Liu, and Jinyu Li. 2021. Developing real-time streaming transformer transducer for speech recognition on large-scale dataset. *Preprint*, arXiv:2010.11395.

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, and 49 others. 2023. Seamlessm4t: Massively multilingual multimodal machine translation. *Preprint*, arXiv:2308.11596.

Marco Gaido, Mattia A Di Gangi, Matteo Negri, and Marco Turchi. 2020. End-to-end speech-translation with knowledge distillation: Fbk@ iwslt2020. In *Proceedings of the 17th International Conference on Spoken Language Translation*, pages 80–88.

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. *arXiv*.

Nuno M. Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André F. T. Martins. 2023. xcomet: Transparent machine translation evaluation through fine-grained error detection. *Preprint*, arXiv:2310.10482.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

Hirofumi Inaguma, Tatsuya Kawahara, and Shinji Watanabe. 2021. Source and target bidirectional knowledge distillation for end-to-end speech translation. In *Proceedings of the 2021 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 1872–1881.

Javier Iranzo-Sánchez, Joan Albert Silvestre-Cerdà, Javier Jorge, Nahuel Roselló, Adrià Giménez, Albert Sanchis, Jorge Civera, and Alfons Juan. 2020. Europarl-st: A multilingual corpus for speech translation of parliamentary debates. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8229–8233.

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

Sara Papi, Marco Gaido, Matteo Negri, and Luisa Bentivogli. 2024. Streamatt: Direct streaming speech-to-text translation with attention-based audio history selection. *Preprint*, arXiv:2406.06097.

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. *Preprint*, arXiv:2507.19634.

Peter Polák, Sara Papi, Luisa Bentivogli, and Ondřej Bojar. 2026. Better late than never: Meta-evaluation of latency metrics for simultaneous speech-to-text translation. *Preprint*, arXiv:2509.17349.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Sukanta Sen, Ondřej Bojar, and Barry Haddow. 2022. Simultaneous translation for unsegmented input: A sliding window approach. *Preprint*, arXiv:2210.09754.

Yangyang Shi, Yongqiang Wang, Chunyang Wu, Ching-Feng Yeh, Julian Chan, Frank Zhang, Duc Le, and Mike Seltzer. 2020. Emformer: Efficient memory transformer based acoustic model for low latency streaming speech recognition. *Preprint*, arXiv:2010.10759.

Grant Strimel, Yi Xie, Brian King, Martin Radfar, Ariya Rastrow, and Thanasis Mouchtaris. 2023. Lookahead when it matters: Adaptive non-causal transformers for streaming neural transducers.

Changhan Wang, Anne Wu, Jiatao Gu, and Juan Pino. 2021. CoVoST 2 and Massively Multilingual Speech Translation. In *Interspeech 2021*, pages 2247–2251.

A Training Parameters

Hyperparameter	**Value**
LoRA	
Rank r	8
α	16
Dropout	0.15
Target modules	q,k,v linear modules
Optimizer	
Optimizer	AdamW
Scheduler MLP learning rate	5×10^{-6}
LoRA learning rate	3×10^{-6}
Weight decay	0.01
Gradient clipping	1.0
Training	
Batch size (per device)	16
Gradient accumulation steps	4
Effective batch size	64
Max epochs	20
Warmup steps	500
LR schedule	Cosine
Mixed precision	`bf16`
Text decoder	Frozen
Checkpoint selection	Best validation BLEU
Dynamic Masking Parameters	
K_{max} (En→De, En→It)	24
K_{max} (En→Zh)	16
Scheduler hidden size	64

Table 3: Training hyperparameters.