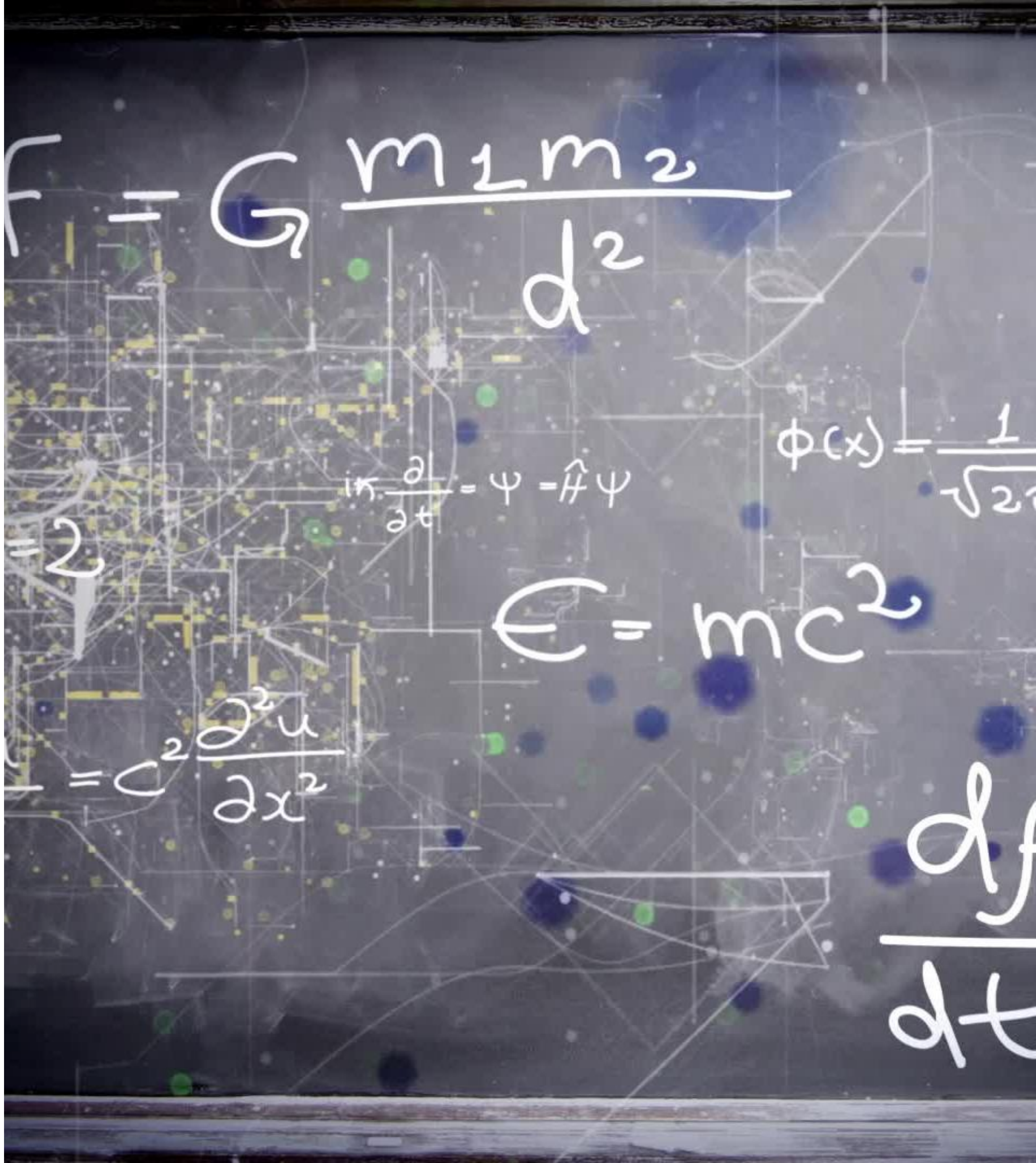
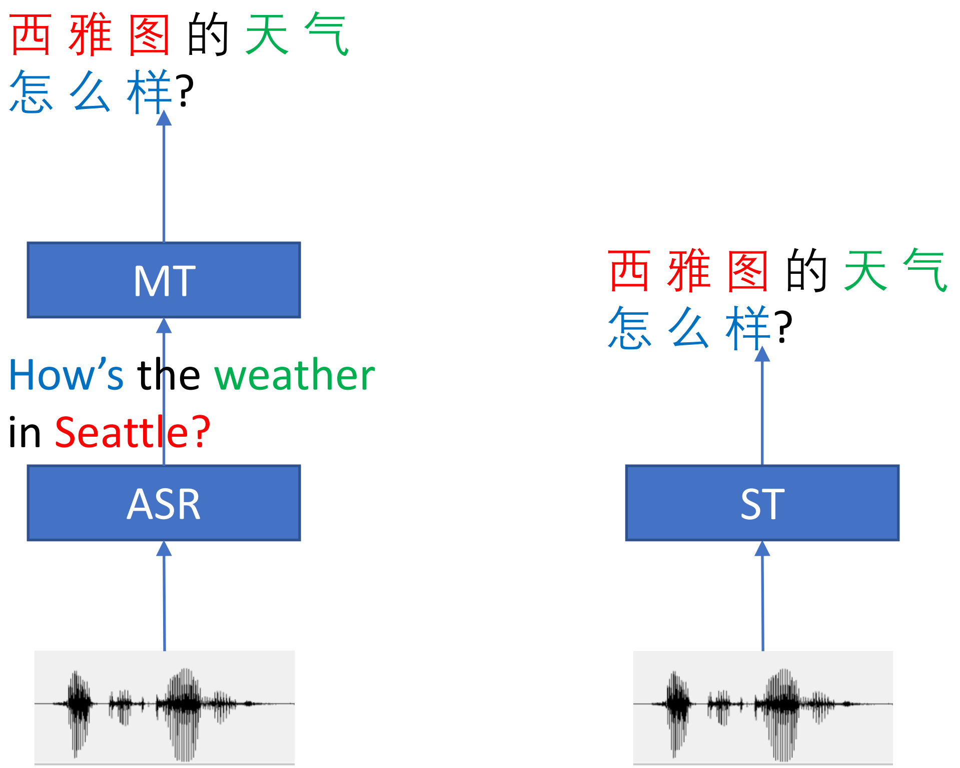
Large-Scale Streaming End-to-End Speech Translation

Terminologies

- Machine translation (MT)
- Speech translation (ST)
- Automatic speech recognition (ASR)
- End-to-end (E2E)
- Direct ST = E2E ST
- Simultaneous ST = Streaming ST



Cascaded vs. E2E



	Cascaded	End-to-end
Model Size	❌	✅
Latency	❌	✅
Error Propagation	❌	✅
Data	✅	❌
Quality	✅	?

Wait-K for Simultaneous Translation

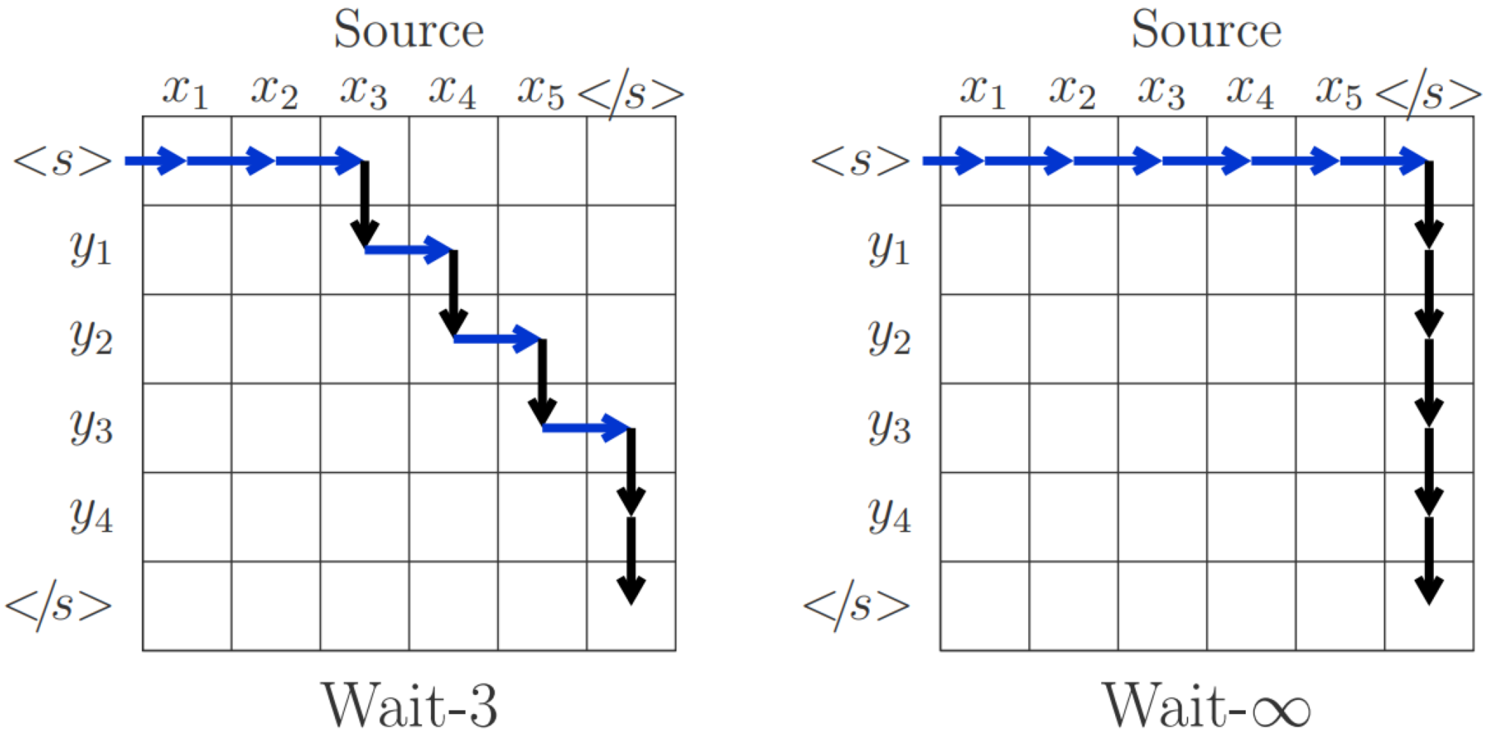


Image: M. Elbayad, L. Besacier, and J. Verbeek, Efficient wait-k models for simultaneous machine translation," arXiv preprint arXiv:2005.08595, 2020.

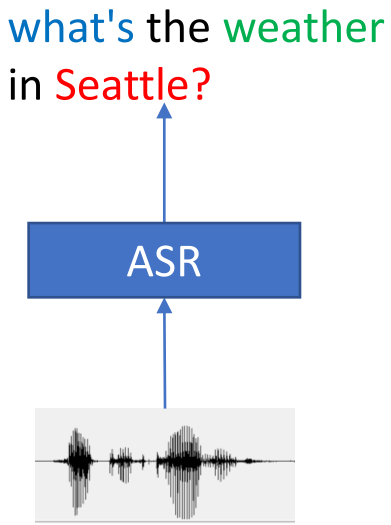
The Challenge of Wait-K

- Not flexible
  - The read-write operation is interleaving
  - K is pre-determined

- More works need to be done for direct speech translation because the step rates of speech and transcription are different.

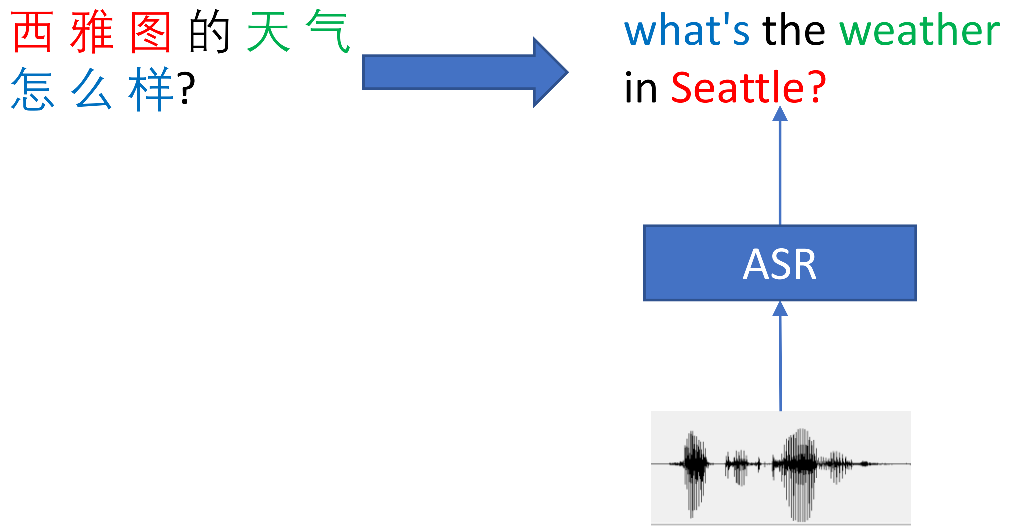
Can We Build a Simultaneous E2E ST System?

- Treating ST as an ASR problem – we already have the success in streaming E2E ASR.



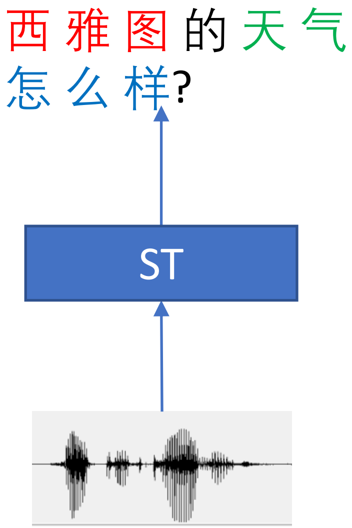
Can We Build a Simultaneous Direct ST System?

- Treating ST as an ASR problem – we already have the success in streaming E2E ASR.



Can We Build a Simultaneous Direct ST System?

- Treating ST as an ASR problem – we already have the success in streaming E2E ASR.

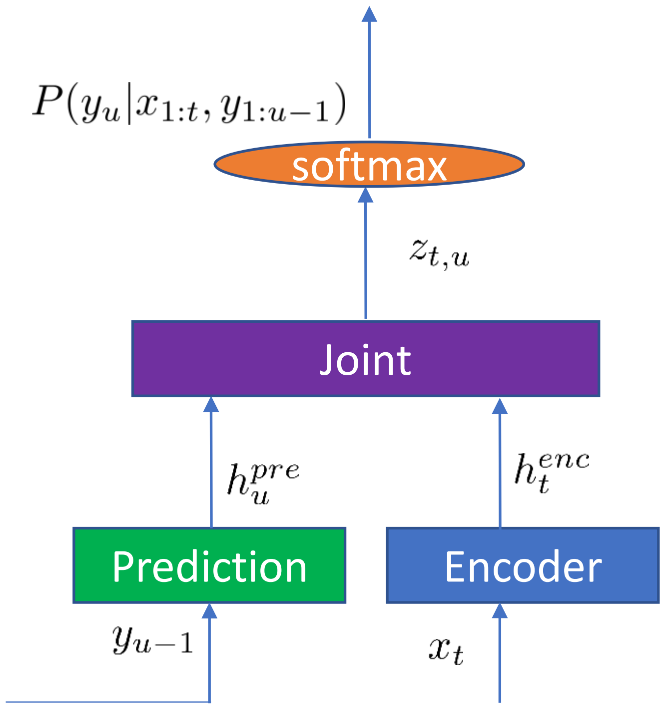


Innovating Streaming ST Method

- Most existing streaming ST methods either rely on wait-k style solution or use MOCHA style solution which has been almost discarded in ASR.
- We first proposed to use RNN Transducer (RNN-T) which is the dominating streaming E2E method in ASR as the solution for streaming ST.

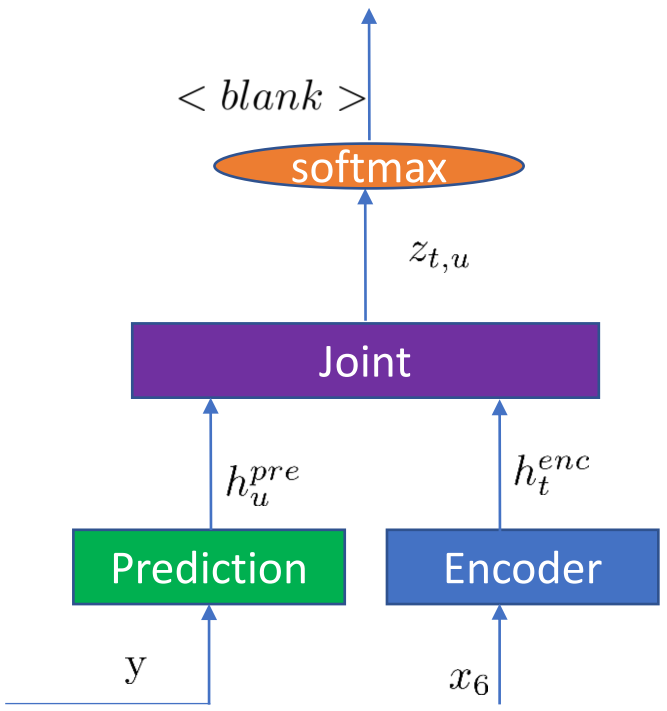
RNN-T: Streaming E2E ASR

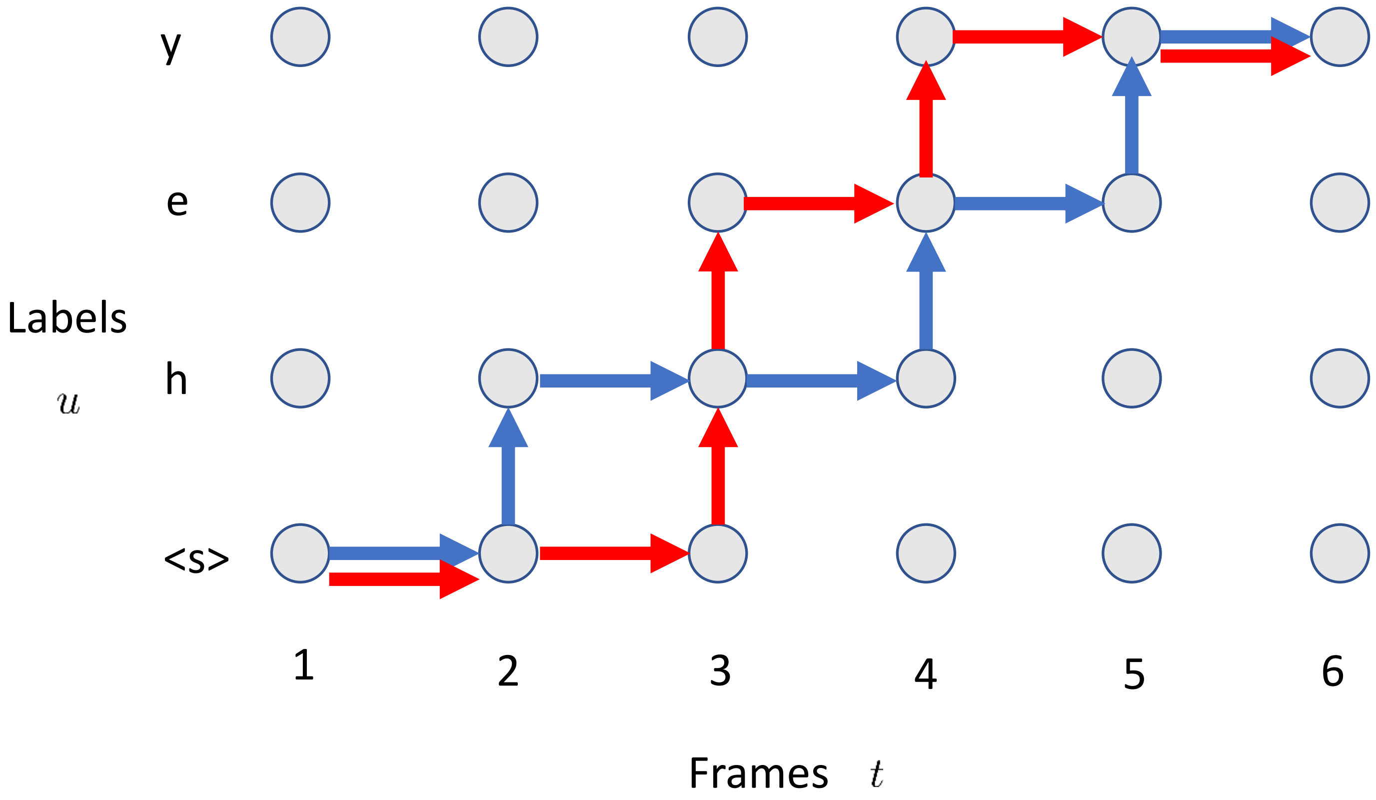
- Encoder: converts input feature sequences into high-level hidden feature sequences.
- Prediction network: producing a high-level representation based on previous label.
- Joint network: combines the outputs from encoder and prediction network.



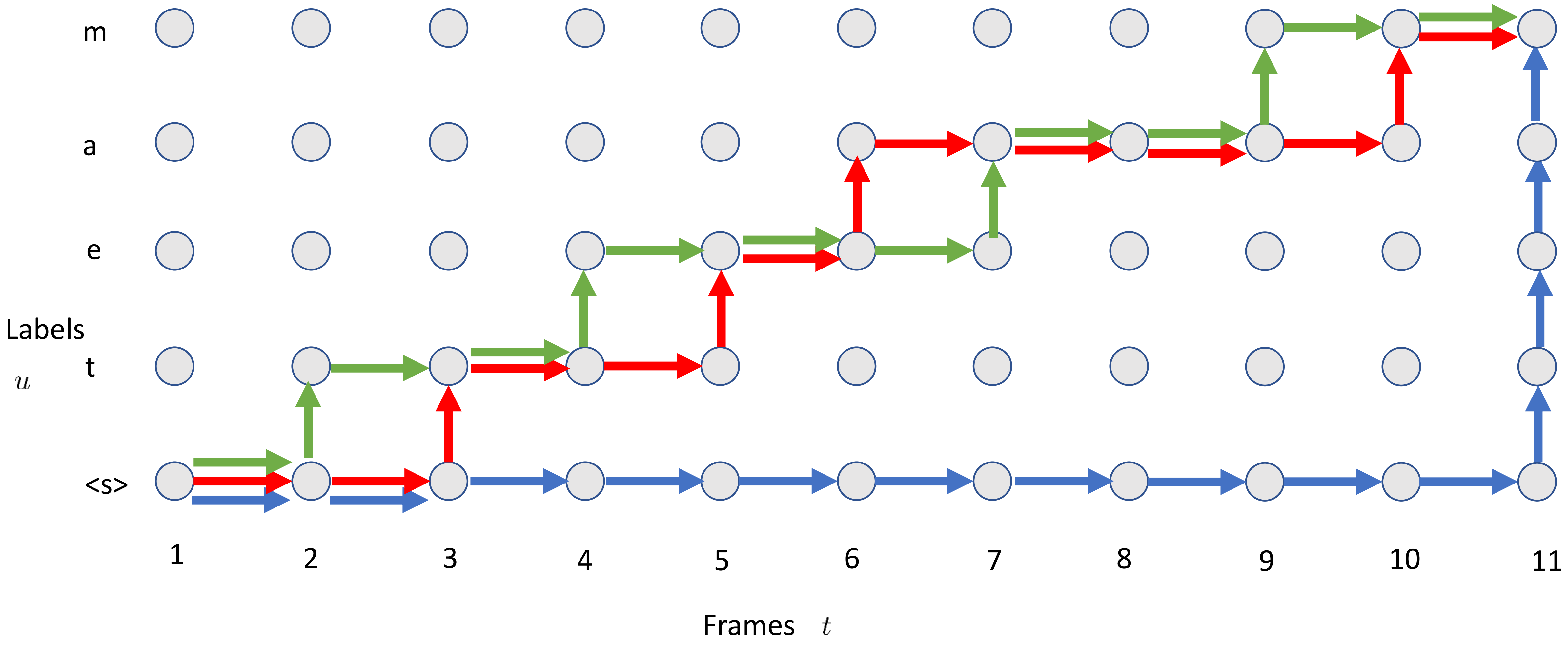
RNN-T Training

Given a label sequence of length U and acoustic frames T, we generate UxT softmax. The training maximizes the probabilities of all RNN-T paths.

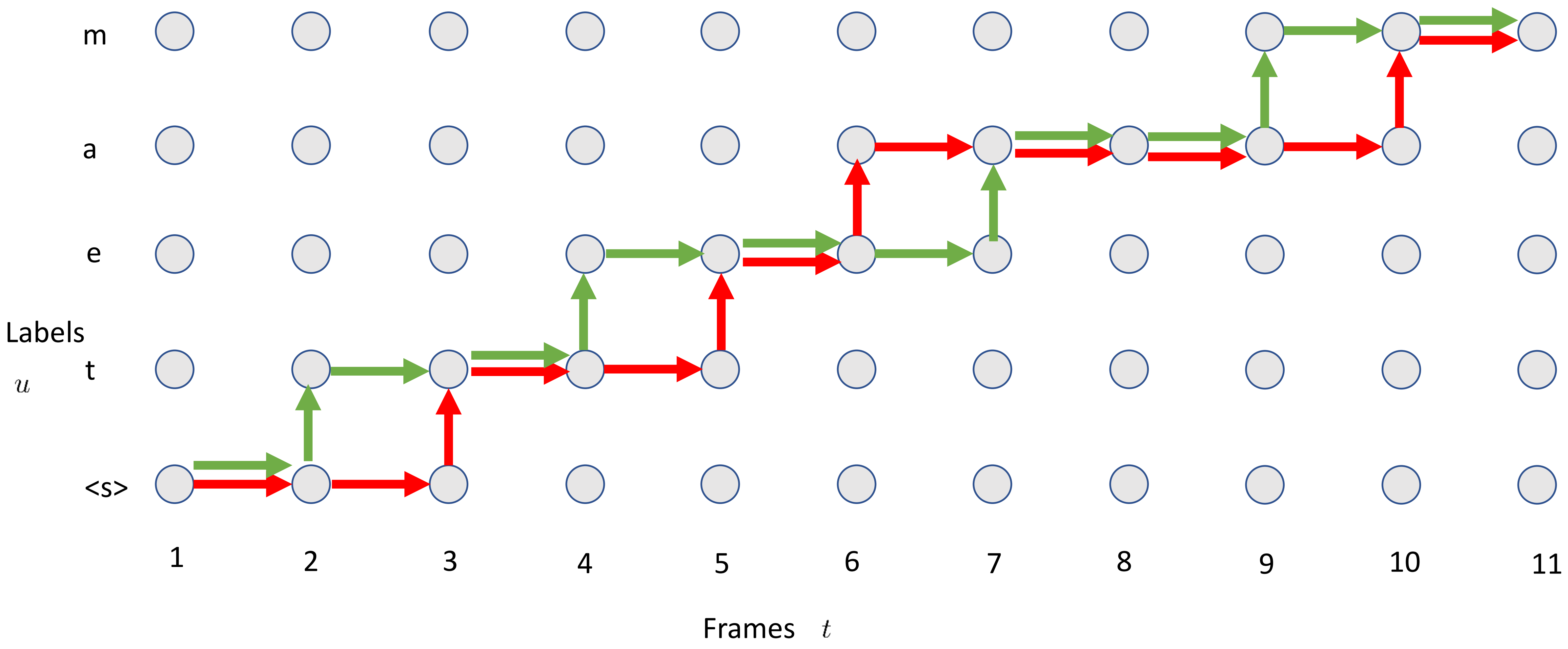




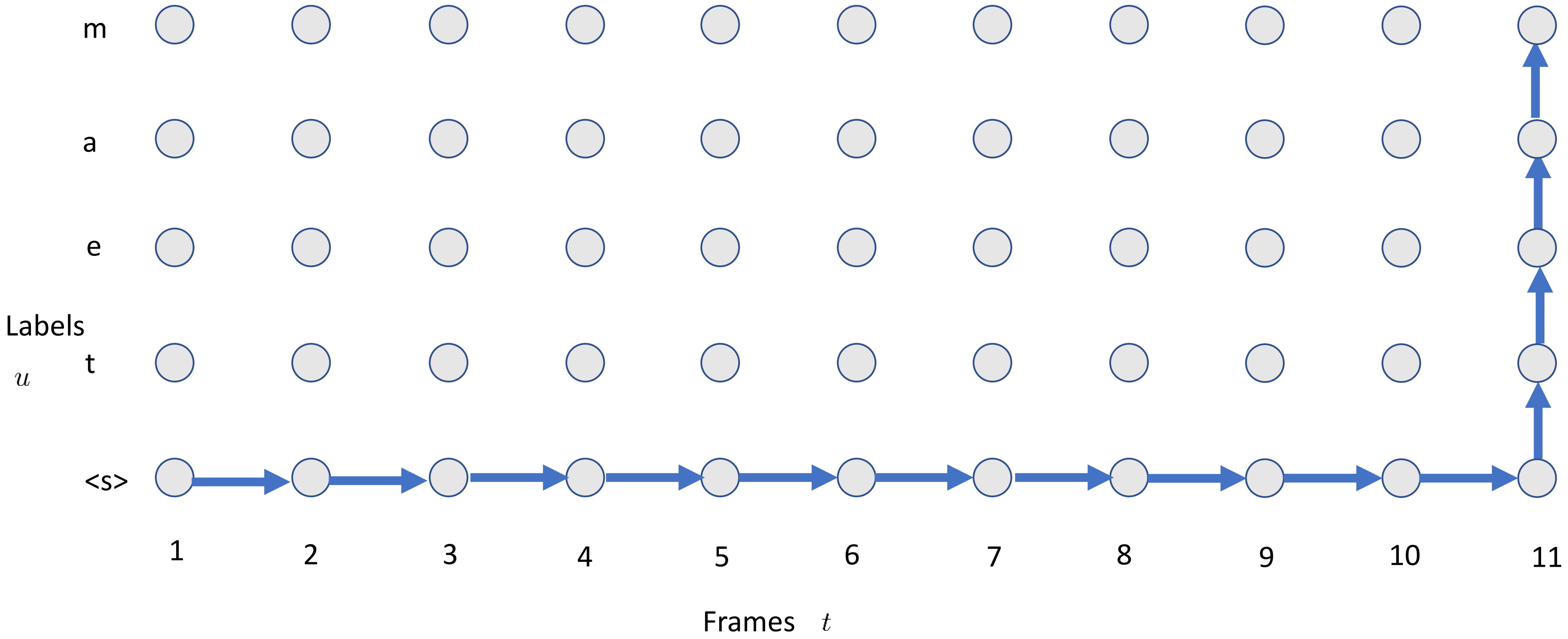
Flexible RNN-T Path



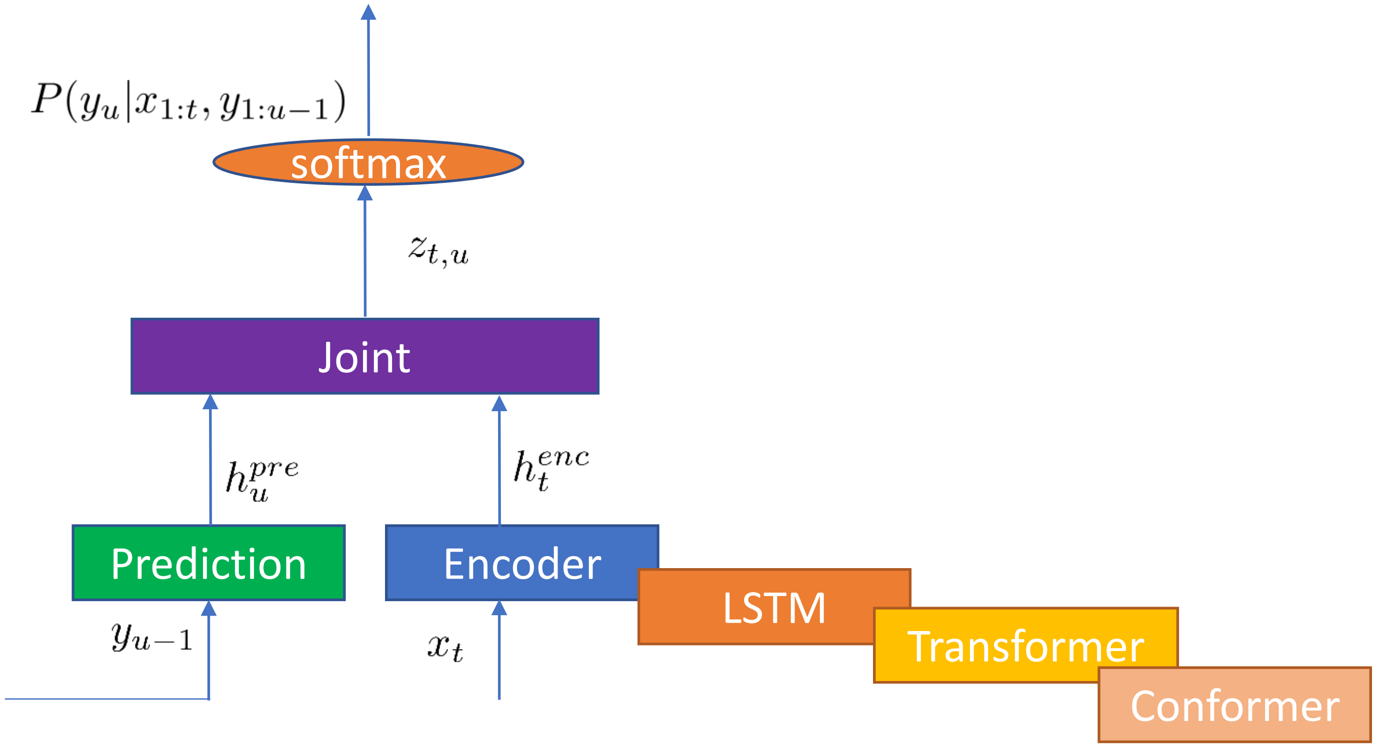
No Word-Reordering



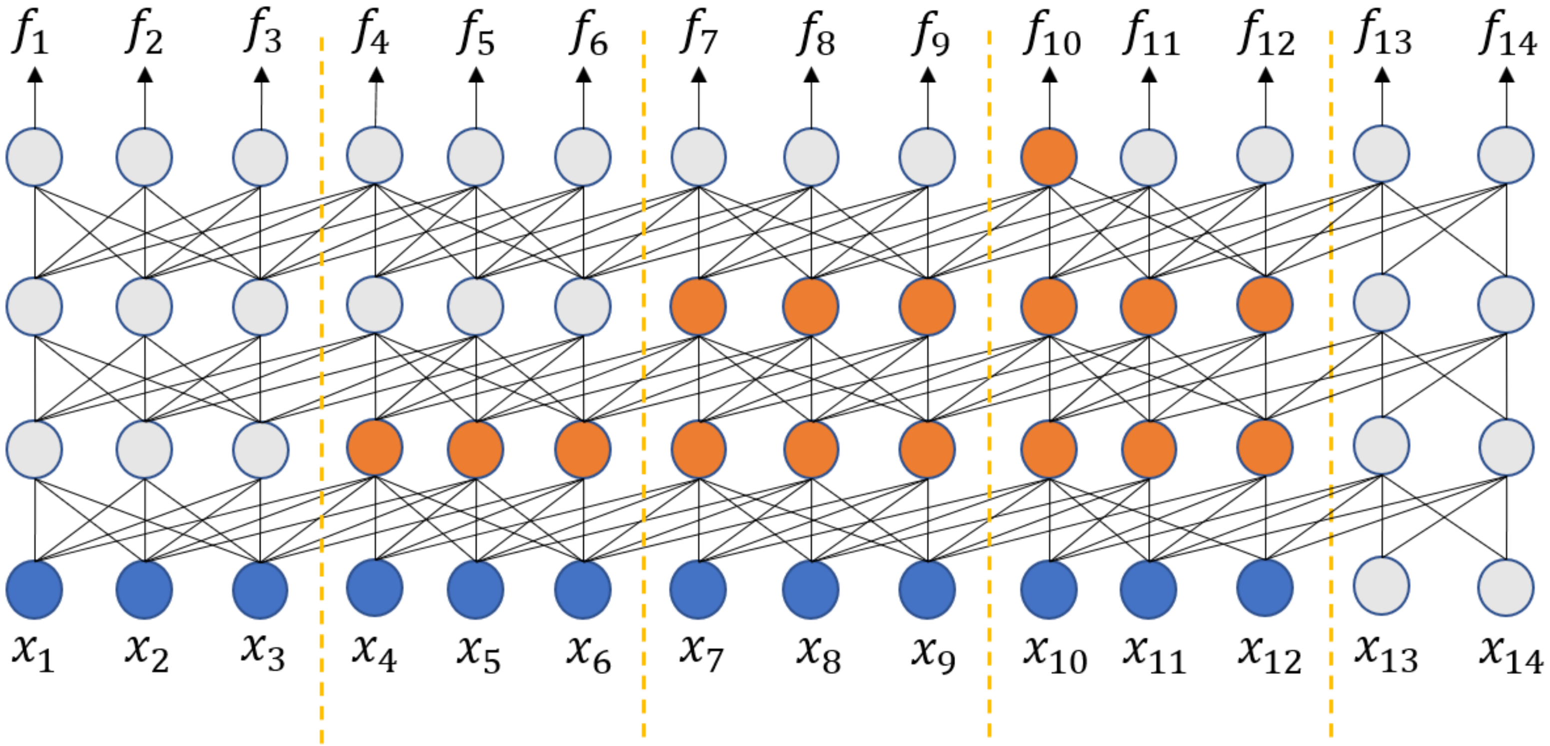
Word-Reordering at the End of Utterance



Encoder for RNN-T



Streaming Transformer



Chen, X., et al. Developing real-time streaming transformer transducer for speech recognition on large-scale dataset. in *Proc. ICASSP,* 2021.

Evaluation Metrics

- Accuracy evaluation: BLEU score
- Latency evaluation:

1) AP (average proportion; Cho & Esipova, 2016): Average of proportion of source input read when generating a target prediction, approaches 0.5.

$$\mathrm{AP} = \frac{1}{|\boldsymbol{X}||\boldsymbol{Y}|}\sum_{i=1}^{|\boldsymbol{Y}|} d_i$$, where *d*i = number of input features when output *y*i (delay of *y*i)

2) AL (average lagging; Ma et all, 2019): Number of words behind the optimal path.

$$\mathrm{AL} = \frac{1}{\tau(|\boldsymbol{X}|)}\sum_{i=1}^{\tau(|\boldsymbol{X}|)} d_i - \frac{(i-1)}{\gamma} ,$$

γ = |*Y*|/|*X*|, *T*(|*X*|) = index of the output sequence when first reaches the end of input

3) DAL (differentiable average lagging; Cherry and Foster, 2019)

$$\mathrm{DAL} = \frac{1}{|\boldsymbol{Y}|}\sum_{i=1}^{|\boldsymbol{Y}|} d'_i - \frac{i-1}{\gamma}, \quad \text{where} \quad d'_i = \begin{cases} d_i & i = 0 \\ \max(d_i, d'_{i-1} + \gamma) & i > 0 \end{cases}$$

Experimental Results

- En-Zh:

BLEUs:

	MSLT_v1.1_dev	MSLT_v1.1_test
Cascaded	37.5	40.0
TT_3.2s	34.5	35.7
TT_160ms	32.9	34.7
TT_160ms	34.3	36.3

Latency measurements on MSLT_v1.1_test set:

	AP ↓	AL ↓	DAL ↓
Cascaded	1	∞	∞
TT_3.2s	0.74	2151	1886
TT_160ms	0.61	841	834

Experimental Results

- En-DE

  BLEUs

	MSLT_v1.0_dev	MSLT_v1.0_test
Cascaded	29.4	29.3
TT_3.2s	31.6	30.8
TT_160ms	30.2	29.4

Latency measurements on MSLT_v1.0_test set:

	AP ↓	AL ↓	DAL ↓
Cascaded	1	∞	∞
TT_3.2s	0.74	2152	1890
TT_160ms	0.61	828	828

Streaming Multilingual Speech Model (SM^2)

- Multilingual data is pooled together to train a streaming model to perform both ST and ASR functions.

- ST training is totally weakly supervised without using any human labeled parallel corpus.

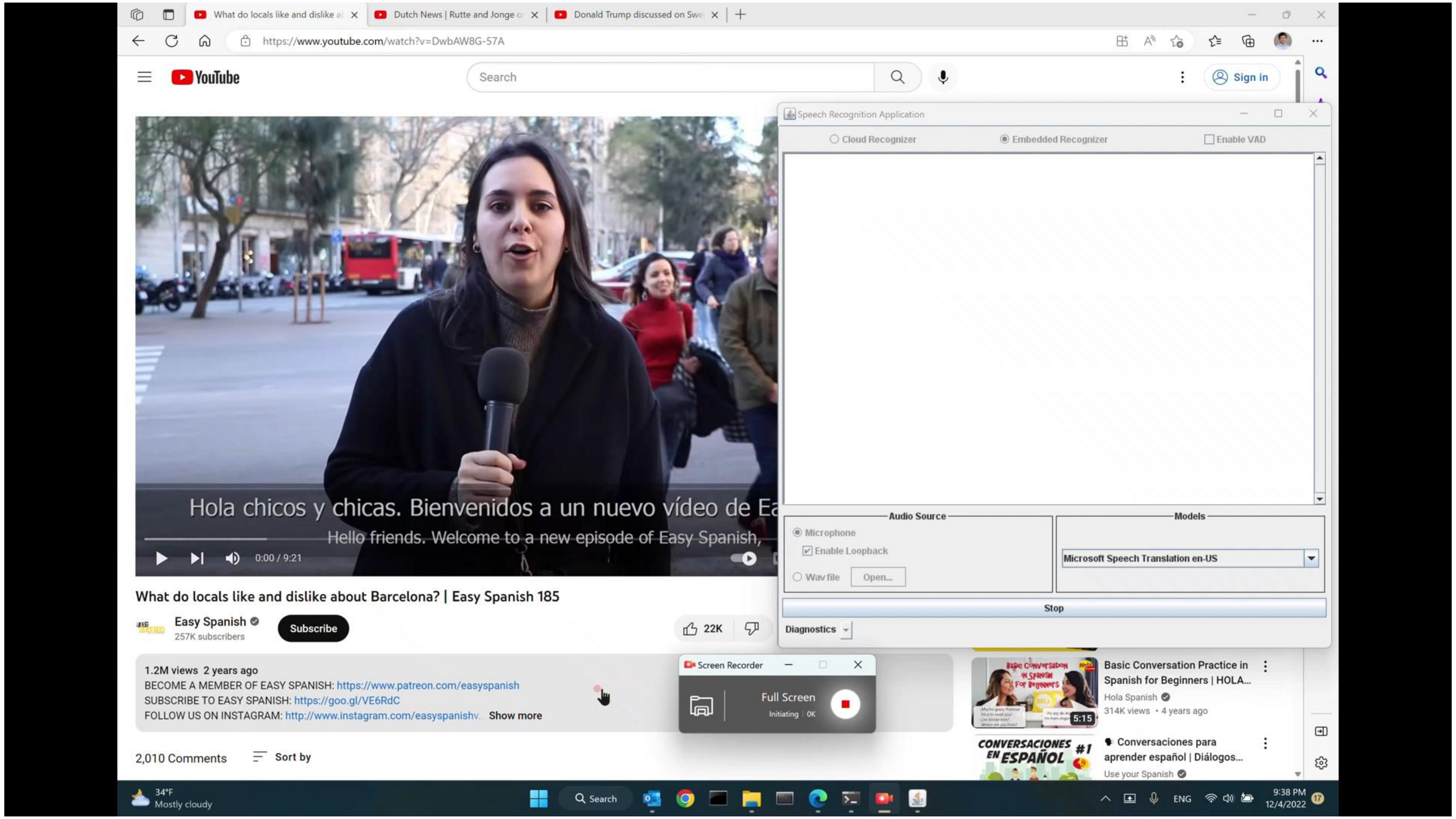
- The model is very small, running on devices.

Xue, J., et al. A Weakly-Supervised Streaming Multilingual Speech Model with Truly Zero-Shot Capability. In *Proc. ASRU,* 2023.

BLEU evaluation on CoVoST 2 test sets

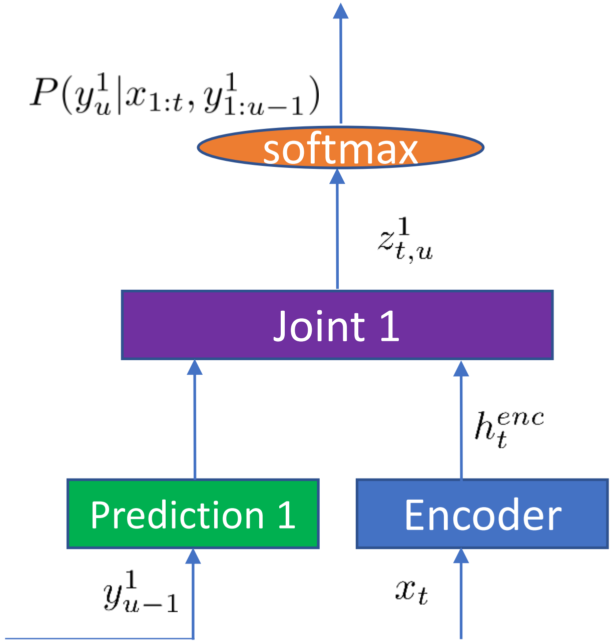
	Whisper [25]		SM^2			
model size	244M	1550M	211M			343M
chunk size	30s	30s	0.32s	1s	30s	30s
DE→EN	25.3	36.3	32.3	34.0	36.4	**37.8**
ZH→EN	6.8	18.0	15.9	18.0	19.8	**21.6**
JA→EN	17.3	**26.1**	20.1	21.6	23.5	25.4
RU→EN	30.9	43.3	36.8	39.8	43.3	**44.8**
NL→EN	28.1	41.2	36.1	38.5	42.2	**43.4**
ET→EN	2.4	15.0	15.3	17.9	21.3	**22.3**
SV→EN	29.9	**42.9**	33.6	37.1	36.5	33.8
SL→EN	9.2	21.6	15.3	**22.4**	18.1	20.4
ES→EN	33.0	**40.1**	32.9	34.7	36.8	37.3
FR→EN	27.3	**36.4**	31.5	33.0	34.9	35.9
IT→EN	24.0	30.9	31.7	33.4	35.0	**36.1**
PT→EN	40.6	**51.6**	42.4	44.7	45.6	45.8
Average	22.9	33.6	28.7	31.3	32.8	**33.7**

SM^2 Trained with 25 Languages->English



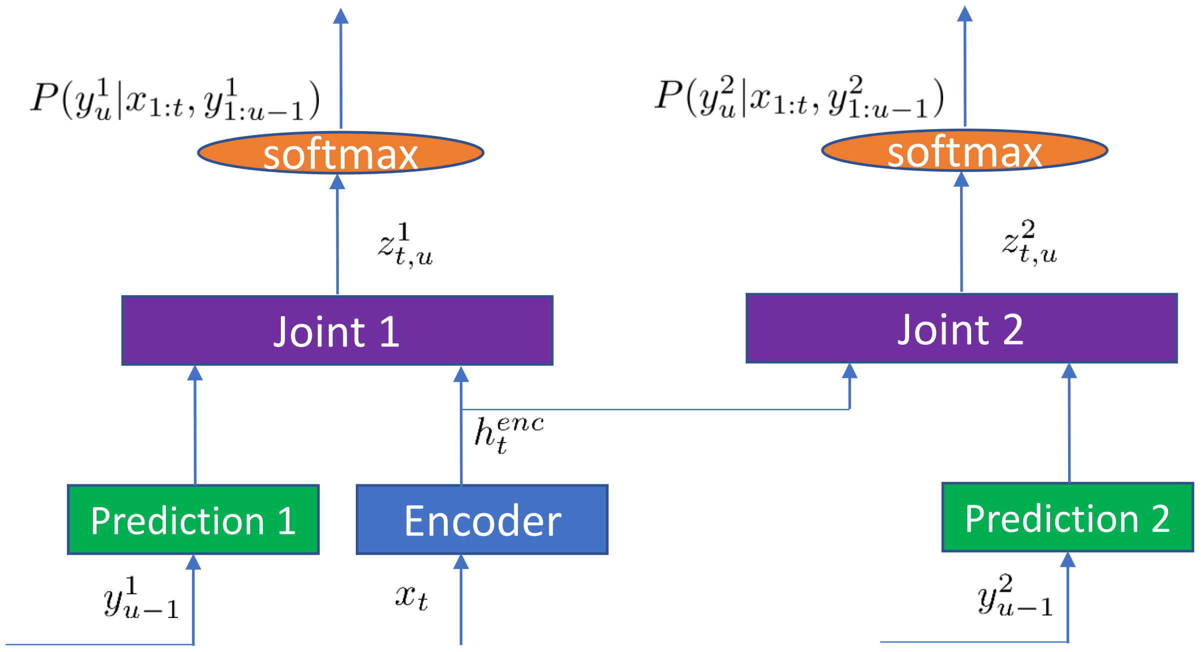
Language Expansion

- Every language has its own prediction and joint network, sharing the same encoder



Language Expansion

- Every language has its own prediction and joint network, sharing the same encoder



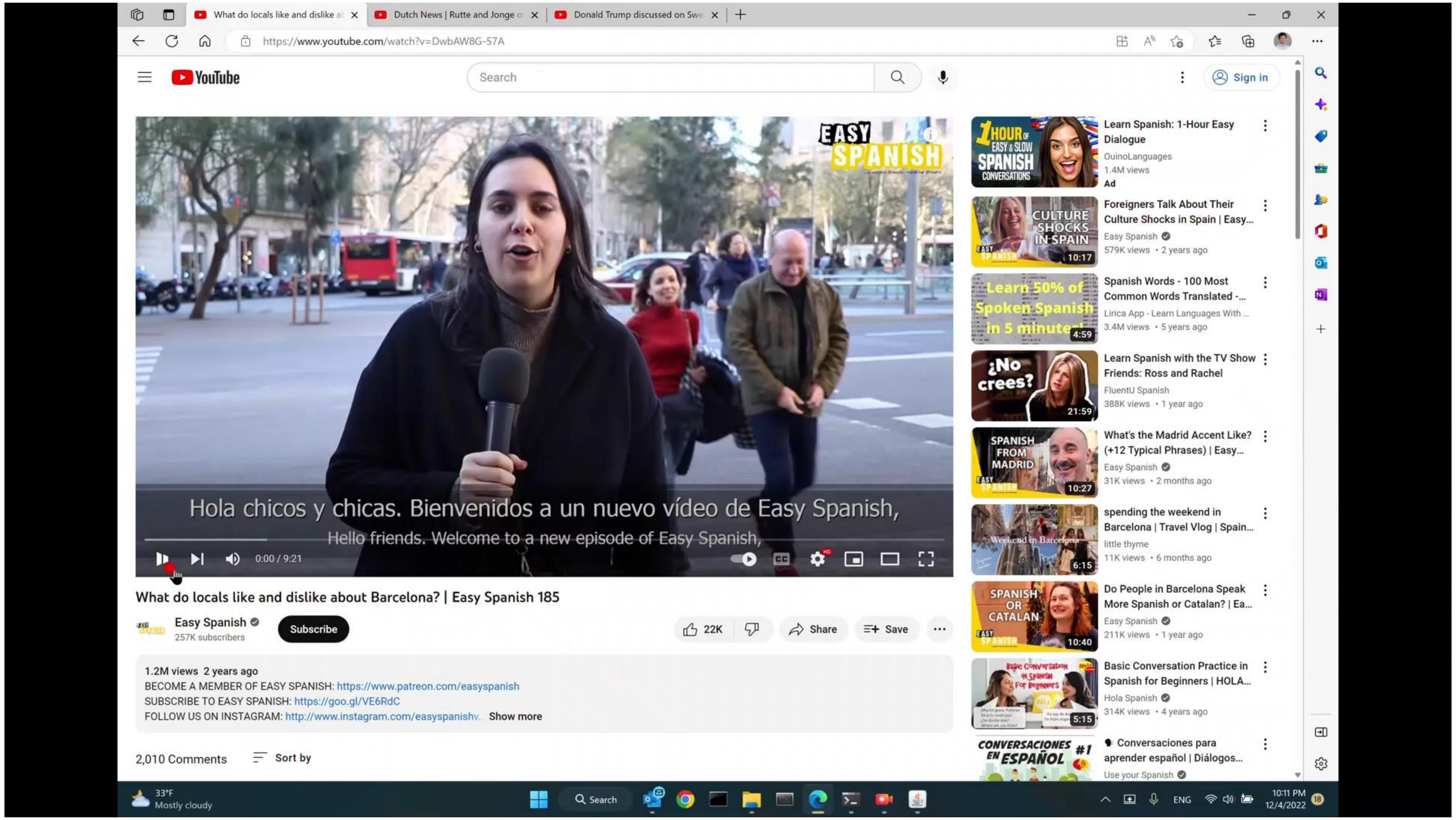
BLEU comparison among different X->ZH models

# source languages	1	3	12	21	25
DE→ZH	**2.2**	21.0	21.8	22.5	21.3
EN→ZH	**0.1**	28.9	29.2	29.3	28.2
JA→ZH	**4.5**	**11.4**	20.0	20.2	20.2
RU→ZH	**8.9**	**20.1**	27.8	28.3	26.8
NL→ZH	**3.5**	**18.4**	**22.6**	24.5	23.9
ET→ZH	**3.9**	**9.7**	**12.4**	14.0	13.1
SV→ZH	**5.8**	**19.3**	**22.4**	23.4	23.1
SL→ZH	**2.1**	**6.3**	**8.1**	8.5	8.7
ES→ZH	**2.0**	**17.3**	**22.3**	**22.8**	25.0
FR→ZH	**2.9**	**16.0**	**20.7**	**21.7**	23.8
IT→ZH	**2.3**	**16.4**	**21.0**	**22.2**	24.2
PT→ZH	**5.1**	**21.6**	**26.4**	**27.0**	28.8
Average	**3.6**	17.2	21.2	22.0	22.3

Bold numbers indicate zero-shot evaluations

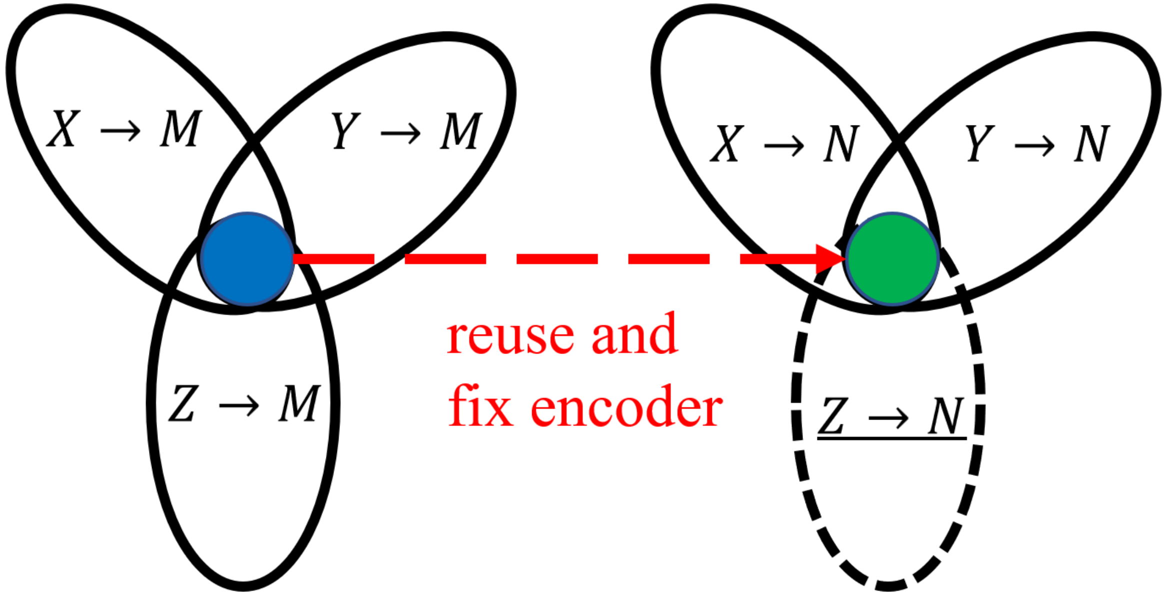
Zero-Shot Speech Translation

Trained only with English/German/Chinese->Chinese data, without observing any other language to Chinese.



Why Can SM^2 Do the Zero-Shot Translation?

- The utterances in the interlingua space (circle) have the same semantic meaning.
- Encoder is frozen for a new language output.
- Utterances in the interlingua space learn to translate to the new target language even if the pair is not observed.
- Because of the calibration inside the language, the learning can be extended to other utterances in the unseen language (dashed area).



Erase-Free Decoding

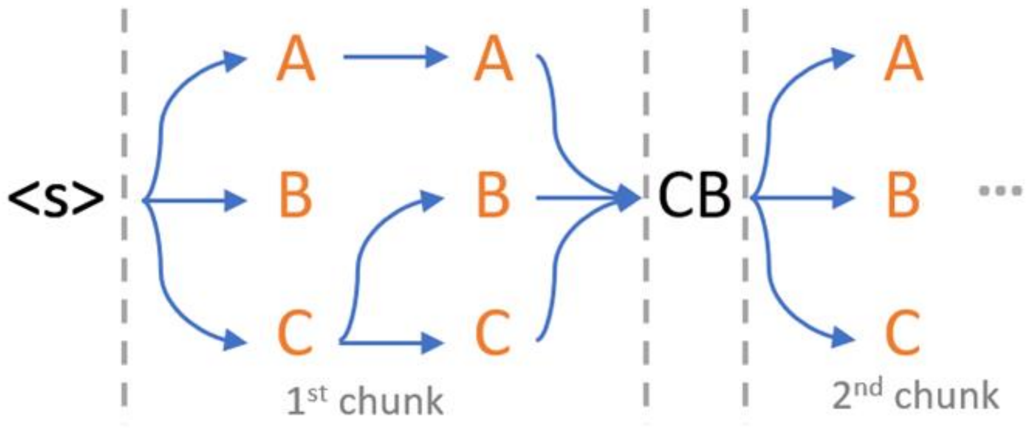
Streaming ST does NOT favor Flickering

- Flickering causes discomfort among audience members, who might consequently lose track of the content.
- Flickering poses significant challenges for incremental synthesis of speech in the target language

Source Transcription	*měiguó de zhōng xī bù yǒu hěnduō gāo shān* 美国 的 中 西 部 有 很多 高 山 *USA 's central west area have many big mountain*	
Translation-Ref	**there are many big mountains in west central US**	
(a) E2E Streaming Translation	*(audio and segment start)* [t_1] **American** [t_2] **West central US** [t_3] **West central US has many** [t_3] **there are many big mountains in west central US** *(audio and segment end)*	☹
(b) Revision-Free Decoding	*(audio and segment start)* [t_1] **American** [t_2] **American midwest** [t_3] **American midwest has many** [t_3] **American midwest has many big mountains** *(audio and segment end)*	☺

Erase-Free Decoding

- Beam Search in Chunks
  - Standard Beam Search within each chunk window.
- Stability-oriented pruning between Chunks
  - Prune the Beam based on different stability requirements, e.g., prune the beam to 1 to prevent erasing.
  - commit the best hypothesis.
- Able to achieve no erasing during inference.



Chen, J., et al. Improving Stability in Simultaneous Speech Translation: A Revision-Controllable Decoding Approach. In *Proc. ASRU,* 2023.

Controllable Decoding

- At end of each chunk, pruning the beam based on a Revision Window (RW).
- Candidates that might cause revision beyond the window will be pruned.
- Trade-off the decoding quality and stability.
- When RW=0, there is no erasing.

Src: 美国 西部 有 很多 国家公园

USA west have many national parks

				RW=1	
Beam	American	west	has	many	*(top candidate)*
	American	west	has	much	✅
	Western	US	has	much	❌

Experiment

- We evaluate our method on CoVoST2 dataset with Streaming T-T model.

	DE->EN			ES->EN			IT->EN		
	BELU	AL	NE	BELU	AL	NE	BELU	AL	NE
Greedy	19.55	1317	0.00	18.96	1239	0.00	17.94	1270	0.00
Standard Beam	26.28	1057	1.49	26.68	1054	1.74	26.50	1052	1.59
Ours (RW=0)	25.13	689	**<u>0.00</u>**	24.28	549	**<u>0.00</u>**	25.18	648	**<u>0.00</u>**
Ours (RW=3)	26.33	800	0.11	26.61	730	0.11	26.55	768	0.11

Joint Output of ASR and ST

Joint Simultaneous Speech Recognition and Translation

- Motivation
  - Help users' understanding: when users have partial knowledge of the spoken language and better understanding of the translation language;
  - Easy to synchronize: one model produces both outputs;
  - Consistency: similar and coherent transcriptions and translations;
  - Explainability: provides insights on the model behavior.

- We propose a novel joint token-level serialized output training (**joint t-SOT**) method to learn how to generate transcription and translation words in an interleaving way

Papi, S., et al. Token-Level Serialized Output Training for Joint Streaming ASR and ST Leveraging Textual Alignments. In *Proc. ASRU,* 2023.

Novel Interleaving Methods

We introduce two novel interleaving methods:

1. **Alignment-based Interleaving**: ASR and ST references are aligned with an alignment tool and words are interleaved based on the obtained alignments
2. **Timestamp-based Interleaving**: the timestamps of the ASR and ST references are estimated through ASR/ST models and this information is used to decide the interleaving

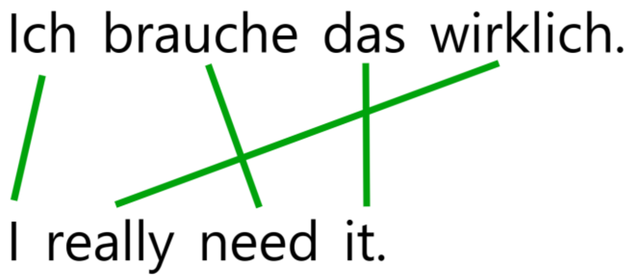
Joint t-SOT INTER ALIGN

- We leverage an off-the-shelf neural textual aligner **`awesome-align`** (Dou et al., 2021) to predict the alignment between transcription and translation texts

Joint t-SOT INTER ALIGN

- We leverage an off-the-shelf neural textual aligner `awesome-align` (Dou et al., 2021) to predict the alignment between transcription and translation texts

Transcription: Ich brauche das wirklich.
Translation: I really need it.



- We interleave the aligned transcription and translation words

INTER ALIGN: #ASR# Ich #ST# I #ASR# brauche das wirklich. #ST# really need it.

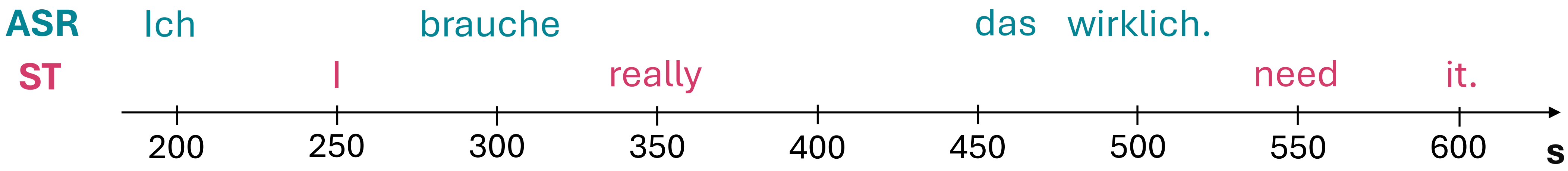
Joint t-SOT INTER TIME

- We leverage the word-level timestamps obtained by applying the Viterbi algorithm on streaming ASR and ST models starting from the reference transcriptions or translations

Transcription: Ich brauche das wirklich.
Translation: I really need it.

Timestamps (ms): 200, 300, 460, 500
250, 350, 550, 600

- We interleave ASR and ST words based on their timestamps in ascending order



INTER TIME: #ASR# Ich #ST# I #ASR# brauche #ST# really #ASR# das wirklich. #ST# need it.

Evaluation Benchmark and Metrics Setup

- **Evaluation Benchmark**
  - CoVoST 2 for the Many-To-English Scenario ({it, es, de}→en)

- **Metrics**
  - WER for the transcription quality ↓
  - BLEU for the translation quality ↑
  - LAAL for the latency (in milliseconds) ↓

Many To English Results

	# inf. steps	it-en				es-en				de-en			
		WER	**LAAL**	**BLEU**	**LAAL**	**WER**	**LAAL**	**BLEU**	**LAAL**	**WER**	**LAAL**	**BLEU**	**LAAL**
Separate ASR & ST	2	25.83	1191	16.41	1844	**22.69**	1149	19.24	1682	23.11	**1071**	19.11	**1613**
Multilingual ASR & ST	2	**23.48**	**1181**	**21.06**	**1663**	22.84	**1147**	**22.76**	**1622**	**21.82**	1133	**21.51**	1642

Many To English Results

	# inf. steps	it-en				es-en				de-en			
		WER	**LAAL**	**BLEU**	**LAAL**	**WER**	**LAAL**	**BLEU**	**LAAL**	**WER**	**LAAL**	**BLEU**	**LAAL**
Separate ASR & ST	2	25.83	1191	16.41	1844	**22.69**	1149	19.24	1682	23.11	**1071**	19.11	**1613**
Multilingual ASR & ST	2	**23.48**	**1181**	**21.06**	**1663**	22.84	**1147**	**22.76**	**1622**	**21.82**	1133	**21.51**	1642

→ Multilingual models are overall better than mono/bilingual models

Many To English Results

	# inf. steps	it-en				es-en				de-en			
		WER	**LAAL**	**BLEU**	**LAAL**	**WER**	**LAAL**	**BLEU**	**LAAL**	**WER**	**LAAL**	**BLEU**	**LAAL**
Multilingual ASR & ST	2	23.48	1181	21.06	1663	22.84	1147	22.76	1622	21.82	1133	21.51	1642
Joint t-SOT INTER 0.0	1	21.81	1228	20.42	3894	20.76	1196	23.26	3752	20.82	1168	21.53	3647
Joint t-SOT INTER 1.0	1	26.05	3389	22.17	1743	23.45	2172	23.99	1683	26.88	3234	21.85	1964
Joint t-SOT INTER 0.5	1	22.35	1110	20.22	1515	21.19	1126	22.25	1468	21.25	1051	20.19	1547

Many To English Results

	# inf. steps	it-en				es-en				de-en			
		WER	**LAAL**	**BLEU**	**LAAL**	**WER**	**LAAL**	**BLEU**	**LAAL**	**WER**	**LAAL**	**BLEU**	**LAAL**
Multilingual ASR & ST	2	23.48	1181	21.06	1663	22.84	1147	22.76	1622	21.82	1133	21.51	1642
Joint t-SOT INTER 0.0	1	21.81	1228	20.42	3894	20.76	1196	23.26	3752	20.82	1168	21.53	3647
Joint t-SOT INTER 1.0	1	26.05	3389	22.17	1743	23.45	2172	23.99	1683	26.88	3234	21.85	1964
Joint t-SOT INTER 0.5	1	22.35	1110	20.22	1515	21.19	1126	22.25	1468	21.25	1051	20.19	1547

→ INTER 0.0 and 1.0 show **high latency** for one of the two modalities

Many To English Results

	# inf. steps	it-en				es-en				de-en			
		WER	**LAAL**	**BLEU**	**LAAL**	**WER**	**LAAL**	**BLEU**	**LAAL**	**WER**	**LAAL**	**BLEU**	**LAAL**
Multilingual ASR & ST	2	23.48	1181	21.06	1663	22.84	1147	22.76	1622	21.82	1133	21.51	1642
Joint t-SOT INTER 0.5	1	22.35	1110	20.22	1515	21.19	1126	22.25	1468	21.25	1051	20.19	1547

→ **Joint t-SOT INTER 0.5** achieves similar or better results compared to multilingual ASR and ST

Many To English Results

	# inf. steps	it-en				es-en				de-en			
		WER	**LAAL**	**BLEU**	**LAAL**	**WER**	**LAAL**	**BLEU**	**LAAL**	**WER**	**LAAL**	**BLEU**	**LAAL**
Multilingual ASR & ST	2	23.48	1181	21.06	1663	22.84	1147	22.76	1622	21.82	1133	21.51	1642
Joint t-SOT INTER 0.5	1	22.35	1110	20.22	1515	21.19	1126	22.25	1468	21.25	1051	20.19	1547
Joint t-SOT INTER ALIGN	1	21.74	**1092**	21.80	**1355**	21.04	**1094**	23.42	**1341**	22.07	1043	21.36	**1335**
Joint t-SOT INTER TIME	1	**21.11**	1141	21.70	1442	19.79	1143	23.38	1452	**21.16**	1112	19.96	1719

Many To English Results

	# inf. steps	it-en				es-en				de-en			
		WER	**LAAL**	**BLEU**	**LAAL**	**WER**	**LAAL**	**BLEU**	**LAAL**	**WER**	**LAAL**	**BLEU**	**LAAL**
Multilingual ASR & ST	2	23.48	1181	21.06	1663	22.84	1147	22.76	1622	21.82	1133	21.51	1642
Joint t-SOT INTER 0.5	1	22.35	1110	20.22	1515	21.19	1126	22.25	1468	21.25	1051	20.19	1547
Joint t-SOT INTER ALIGN	1	21.74	**1092**	21.80	**1355**	21.04	**1094**	23.42	**1341**	22.07	1043	<u>21.36</u>	**<u>1335</u>**
Joint t-SOT INTER TIME	1	**<u>21.11</u>**	<u>1141</u>	21.70	1442	19.79	1143	23.38	1452	**<u>21.16</u>**	<u>1112</u>	19.96	1719

→ **INTER TIME** shows improvements on ASR while being comparable on ST (except for de-en) when compared with INTER ALIGN

Conclusions

- We proposed to use T-T for streaming E2E speech translation, with low latency/computation cost.
- We built a multilingual E2E speech translation model, which can be easily extended with zero-shot capability.
- We proposed an erase-free decoding method to improve the stability of translation results.
- We proposed joint t-SOT model can jointly output ASR and ST results.