/ kyutai

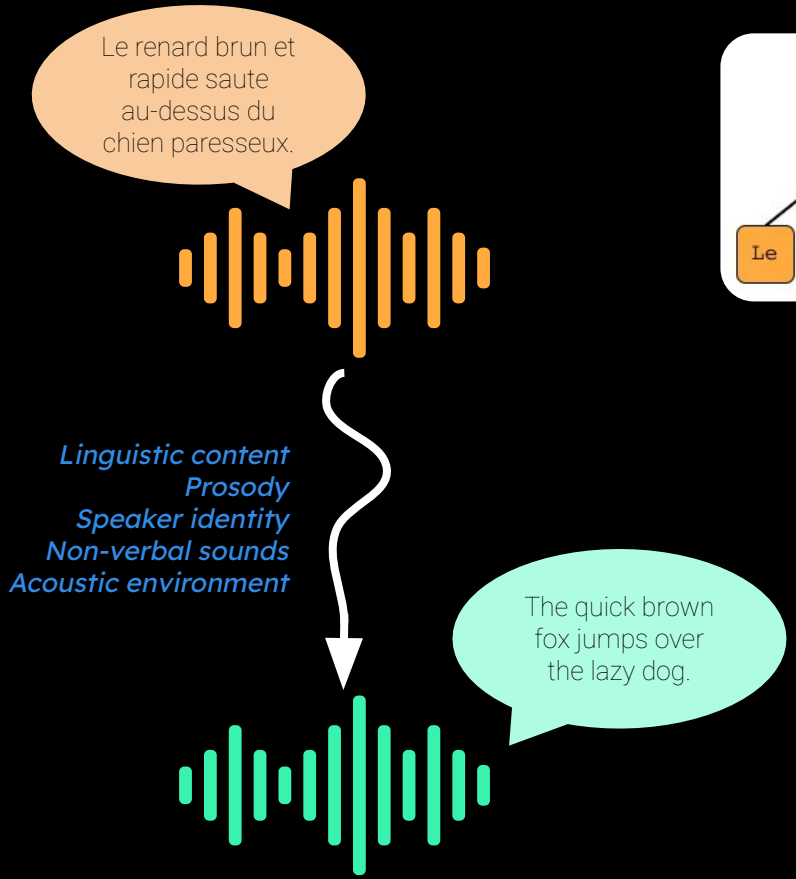
OPEN-SCIENCE AI LAB

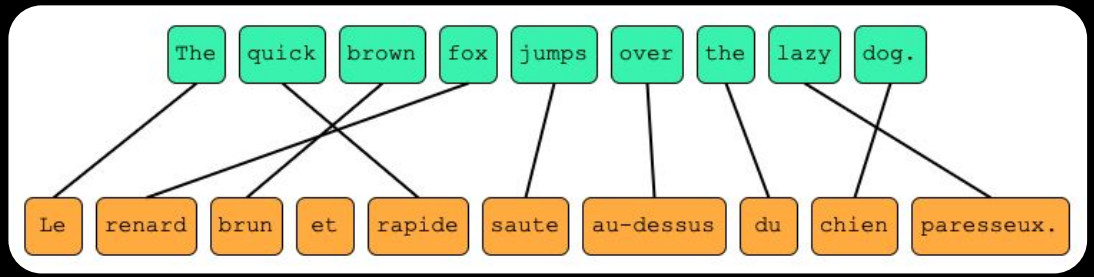
Simultaneous Speech-To-Speech Translation Without Aligned Data

Hibiki-Zero @ ICML2026

Tom LABIAUSSE, Romain FABRE, Yannick ESTÈVE, Alexandre DÉFOSSEZ, Neil ZEGHIDOUR

/ Simultaneous speech-to-speech translation





Objectives:

- Accurate, general simultaneous translations
- Intelligible speech and natural flow
- Expressive translation with voice transfer
- Long durations consistency (>1min)

/ Simultaneous S2ST without aligned data

Full-duplex multilingual translation system that can HEAR, SPEAK and WRITE at the same time.

- Model = Decoder-only transformer + streaming neural audio codec.
- Synthetic training targets to learn sentence-level simultaneous translation.
- Reinforcement Learning to optimize translation latency while retaining quality.
- Natively transfers non-linguistic information (speaker identity, prosody, intonation…)

/ Mimi: a semantic-acoustic audio codec

- Neural audio codec producing semantic and acoustic quantized representation of speech.
- Causal encoder/decoder architecture with a 80ms frame rate and a 1.1kbps bitrate.

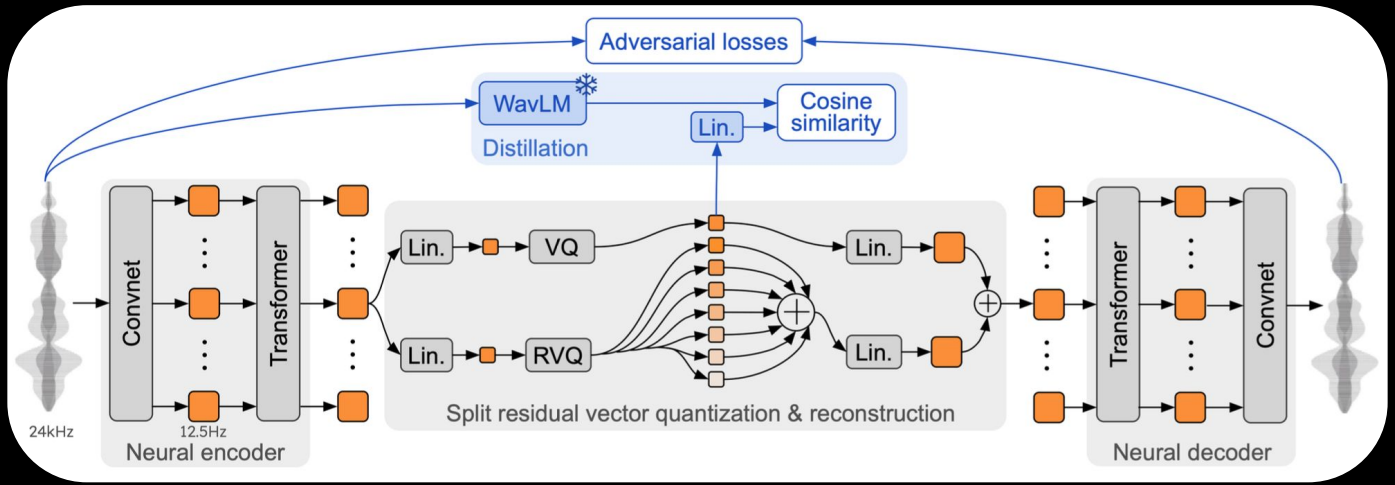
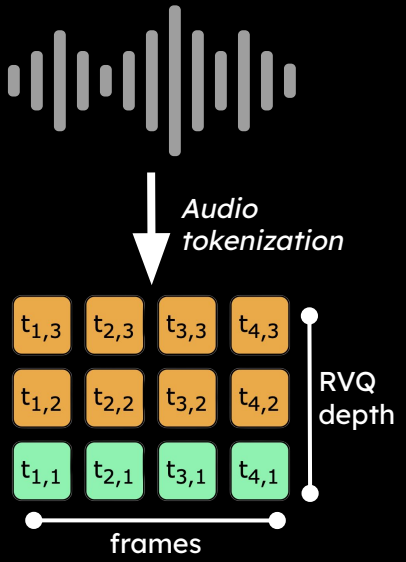
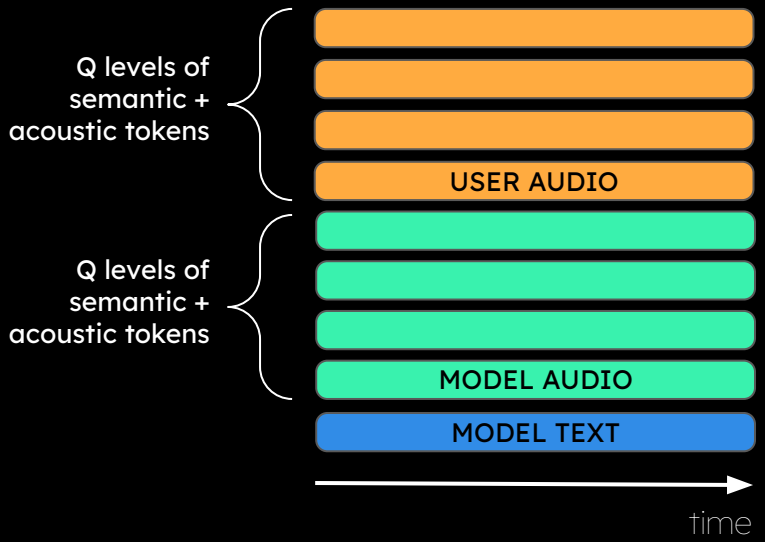


Figure from *Moshi: a speech-text foundation model for real-time dialogue - Défossez et al. (2024)*



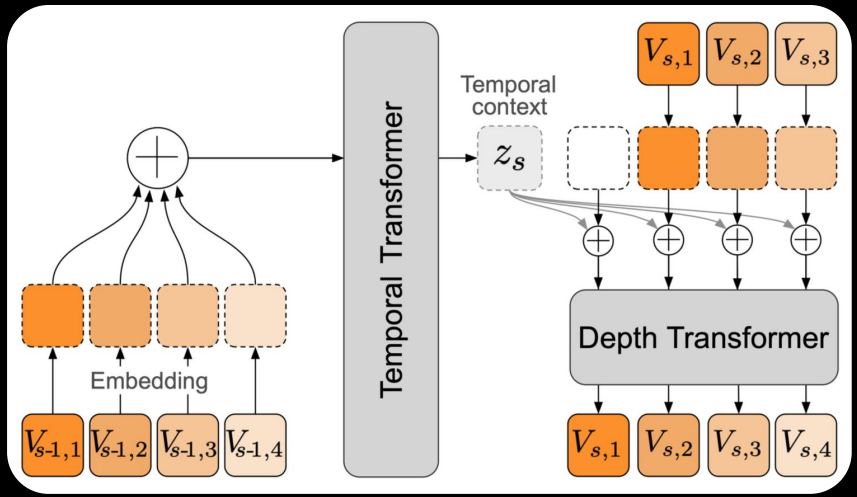
/ Model and architecture



Multistream: Process input audio in parallel of generating an output audio and text streams.

Inner Monologue: Align output text tokens with output audio tokens at the word level by inserting text padding tokens.

Architecture: RQ-Transformer (*Temporal* + *Depth*) continuously predicts output text/audio tokens while receiving user audio.



Inference mechanism:

1) Sum the output text, output audio and input audio embeddings of the last frame tokens as input to the *Temporal Transformer*.

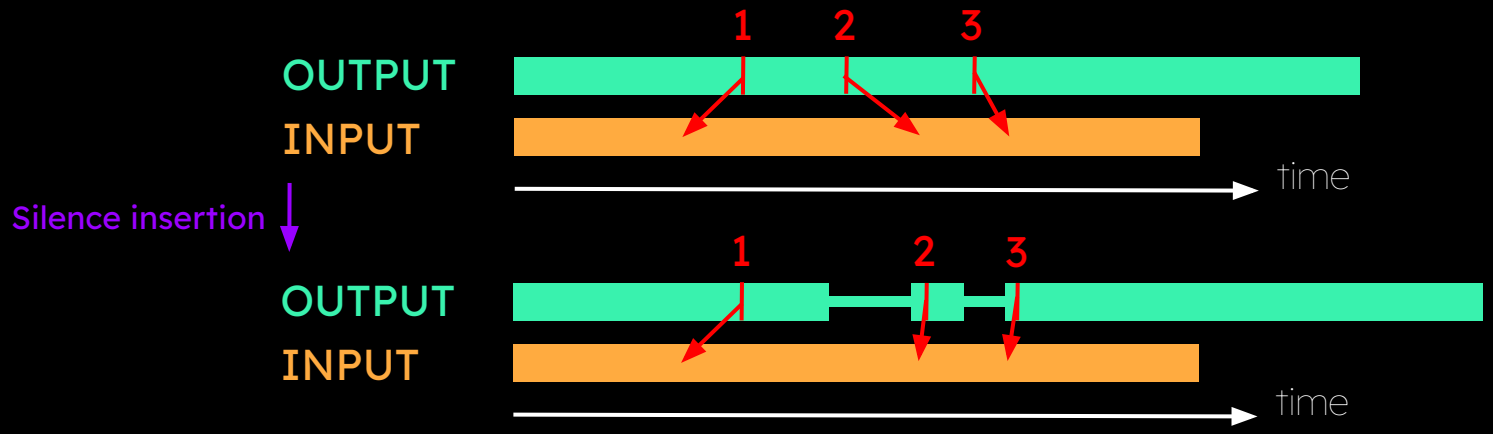
2) Run one step of the *Temporal Transformer* to predict a Temporal context embedding.

3) Use the Temporal context embedding to predict a text token and as a conditioning of the *Depth Transformer* which runs multiple steps to predict the model's audio tokens for the current frame.

/ Training data for simultaneous S2ST

❌ Real and high-quality simultaneous S2ST data doesn't exist at scale.

✅ It is possible to build synthetic simultaneous S2ST data as it was done to train Hibiki[1]:



→ Hibiki's translation latency is fully determined by the silence patterns in the training data.

[1] *High-Fidelity Simultaneous Speech-to-Speech Translation*, Labiausse et al. (2025)

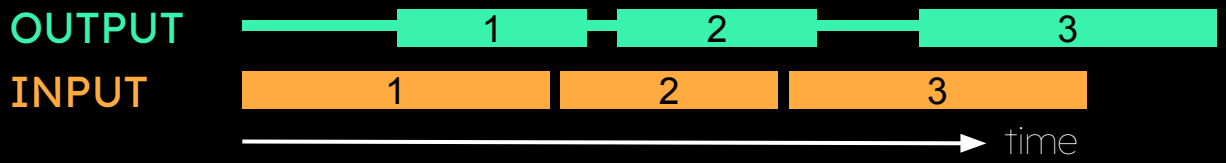
/ New training method for simultaneous S2ST

❌ Building low latency synthetic S2ST data is hard and language-dependent…

✅ What if the model could **optimize translation latency by itself** during post-training ?

Our approach *decouples* the training of a simultaneous S2ST model:

I. Quality training with Supervised Learning:
Learn to translate accurately with high latency by inserting sentence-level silences.

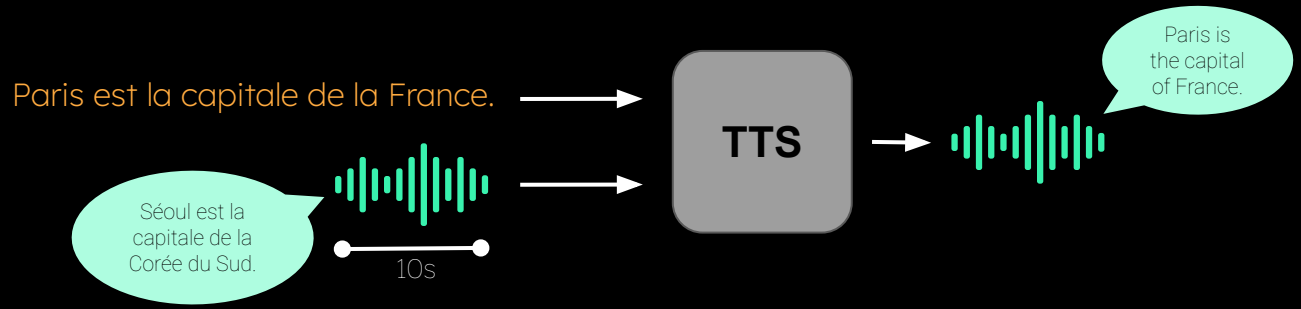


II. Latency training with Reinforcement Learning:
Reduce translation latency as much as possible while retaining quality.
Reach an optimal quality/latency trade-off.

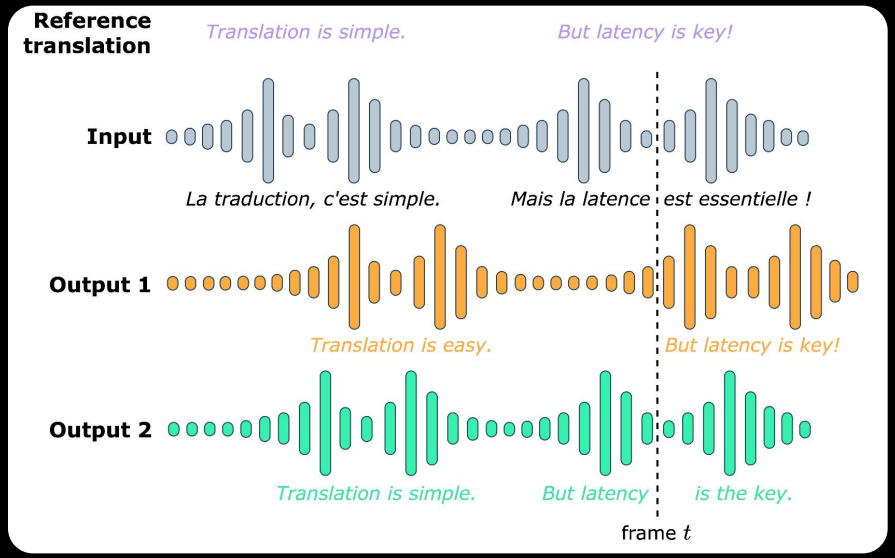
/ S2ST ($L_A \rightarrow L_B$) synthetic data creation pipeline

Data requirements: a large, diverse speech dataset in L_A (various speakers, accents, contexts…).

1) Extract single speaker audio chunks (<2min) and perform ASR to obtain transcripts in language L_A.
2) Translate the transcripts in a *sentence-per-sentence* manner into language L_B with a MT model.
3) Create translated speech in language L_B with a TTS copying the voice of the original speaker.



/ Improve the quality-latency trade-off with RL



- At instant *t*, reward the output with the most complete content before *t*.

$$\text{BLEU}^{(1)}_{\text{total}} = \text{BLEU}\left[\begin{matrix}\textit{Translation is easy. But latency is key!}\\ \textit{Translation is simple. But latency is key !}\end{matrix}\right]$$

$$\text{BLEU}^{(2)}_{\text{total}} = \text{BLEU}\left[\begin{matrix}\textit{Translation is simple. But latency is the key.}\\ \textit{Translation is simple. But latency is key !}\end{matrix}\right]$$

$$r_t^{(1)} = (1-\alpha)\text{BLEU}\left[\begin{matrix}\textit{Translation is easy.}\\ \textit{Translation is simple. But latency is key !}\end{matrix}\right] + \alpha\,\text{BLEU}^{(1)}_{\text{total}}$$

$$r_t^{(2)} = (1-\alpha)\text{BLEU}\left[\begin{matrix}\textit{Translation is simple. But latency}\\ \textit{Translation is simple. But latency is key !}\end{matrix}\right] + \alpha\,\text{BLEU}^{(2)}_{\text{total}}$$

$$r_t^{(i)} = (1-\alpha)\text{BLEU}\left(\hat{y}_t^{(i)}, y_t\right) + \alpha\text{BLEU}\left(\hat{y}_T^{(i)}, y_T\right)$$

- Compute BLEU[1] scores to evaluate the intermediate translation content.
- Use GRPO[2] to update the model given the combinations of process and outcome BLEU rewards.

[1] *BLEU: a Method for Automatic Evaluation of Machine Translation*, Papineni et al. (2002)
[2] *DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models*, Shao et al. (2024)

/ Objective evaluations

Table 1. Objective comparison of Hibiki-Zero with Seamless (Barrault et al., 2023) and Hibiki (Labiausse et al., 2025) on short-form (Europarl-ST) and long-form (Audio-NTREX-4L) test data introduced in Section 4.3.

	Short-form						Long-form					
	BLEU (↑)	ASR BLEU (↑)	ASR COMET (↑)	Speaker Sim. (↑)	End Offset (↓)	LAAL (↓)	BLEU (↑)	ASR BLEU (↑)	ASR COMET (↑)	Speaker Sim. (↑)	End Offset (↓)	LAAL (↓)
Seamless												
French	33.8	32.8	76.6	19.1	2.4	**2.8**	27.8	23.9	33.7	44.4	3.2	6.2
Spanish	**34.4**	33.6	79.1	21.9	2.6	2.7	29.9	25.2	36.1	42.6	2.8	6.5
Portuguese	**34.1**	**33.6**	78.9	23.9	2.8	3.1	29.0	25.6	35.0	35.7	3.2	6.6
German	27.8	27.3	**82.3**	20.6	2.4	3.0	27.8	24.0	40.6	47.8	2.5	7.3
Hibiki												
French	32.4	31.8	**81.5**	35.7	2.5	3.5	29.5	26.4	42.0	52.8	2.6	6.8
Hibiki-Zero												
French	**35.0**	**34.6**	80.3	**49.5**	**2.1**	**2.8**	**30.6**	**28.7**	**43.7**	**61.3**	**2.3**	**6.1**
Spanish	33.8	**33.9**	**80.3**	**57.0**	**2.3**	3.1	**32.3**	**31.5**	**42.3**	**64.6**	**2.6**	**5.6**
Portuguese	33.6	**33.6**	**78.9**	**51.4**	**2.4**	**3.0**	**33.2**	**31.3**	**42.6**	**62.1**	**2.3**	**6.3**
German	**28.7**	**28.6**	82.0	**51.5**	**1.9**	**2.8**	**29.1**	**28.3**	**42.3**	**66.0**	**2.0**	**5.9**

- Short-form: 2 to 20 seconds samples from Europarl-ST: www.mllp.upv.es/europarl-st
- Long-form: Multi-sentences text translations synthesized with TTS: huggingface.co/datasets/kyutai/Audio-NTREX-4L

/ Subjective evaluations

Table 2. **Human evaluation.** Raters report Mean Opinion Scores (MOS) on a scale ranging from 0 to 100 for each audio sample.

Input language	Model	Audio Quality	Speaker Similarity	Speech Naturalness
French	Seamless	11.4 ± 3.1	21.1 ± 4.9	21.2 ± 3.8
	Hibiki	62.9 ± 4.8	44.7 ± 5.1	57.0 ± 4.2
	Hibiki-Zero	**64.5** ± 4.2	**70.0** ± 5.1	**67.2** ± 4.1
Spanish	Seamless	10.7 ± 2.6	21.2 ± 4.5	26.5 ± 4.4
	Hibiki-Zero	**66.8** ± 3.9	**69.0** ± 3.9	**66.2** ± 4.9
Portuguese	Seamless	11.8 ± 3.1	32.5 ± 6.0	22.8 ± 3.9
	Hibiki-Zero	**62.0** ± 4.1	**60.7** ± 4.2	**75.6** ± 3.4
German	Seamless	15.6 ± 2.7	25.2 ± 4.9	26.4 ± 4.8
	Hibiki-Zero	**73.5** ± 3.4	**65.3** ± 4.3	**69.9** ± 3.9

Human evaluations setup:

- 20 raters
- 50 samples per model
- 5 comparisons per rater

/ New language adaptation (Italian-to-English)

Table 3. Objective results of model adaptation to input Italian speech with 850 hours of finetuning data on short-form evaluation.

	BLEU (↑)	ASR BLEU (↑)	SPEAKER SIM. (↑)	END OFFSET (↓)	LAAL (↓)
SEAMLESS	**32.5**	**32.0**	22.2	**3.0**	**3.5**
OURS					
BASE	14.3	14.3	50.6	3.9	4.3
FINETUNED	31.4	31.0	**55.2**	3.7	4.5
FINETUNED + RL	32.1	31.9	54.2	**3.0**	**3.5**

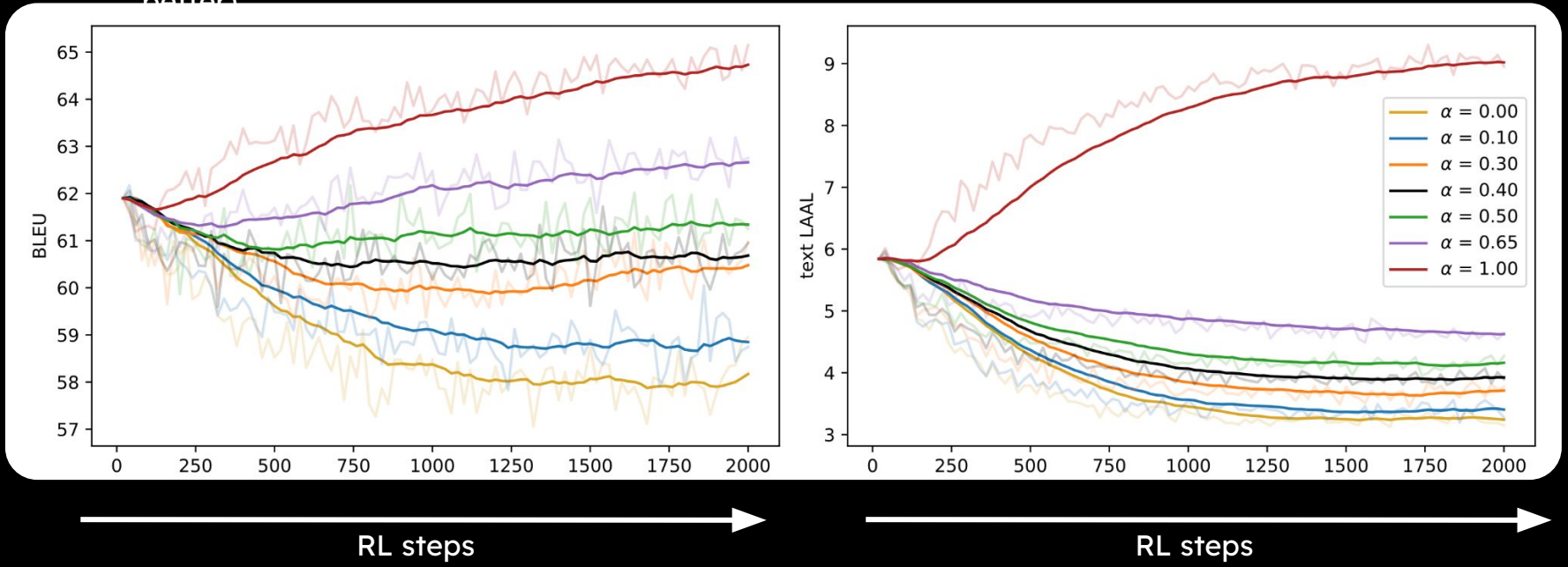
- Supervised fine-tuning + RL on a dataset with less than 1000h of Italian-to-English data.

/ Control the quality-latency trade-off during RL

- We introduce a training mechanism to control latency optimization with a single hyperparameter α.

$$r_t^{(i)} = (1-\alpha)\mathrm{BLEU}\left(\hat{y}_t^{(i)}, y_t\right) + \alpha\mathrm{BLEU}\left(\hat{y}_T^{(i)}, y_T\right)$$

Translation quality ⬆️ (higher is better)



Translation latency ⬇️ (lower is better)

Simultaneous Speech-To-Speech Translation Without Aligned Data

ICML
International Conference
On Machine Learning

/k

/ kyutai
OPEN-SCIENCE AI LAB

kyutai.org/blog/2026-02-12-hibiki-zero