The Silent Thought: Modeling Internal Cognition in Full-Duplex Spoken Dialogue Models via Latent Reasoning

ICML 2026 Presentation

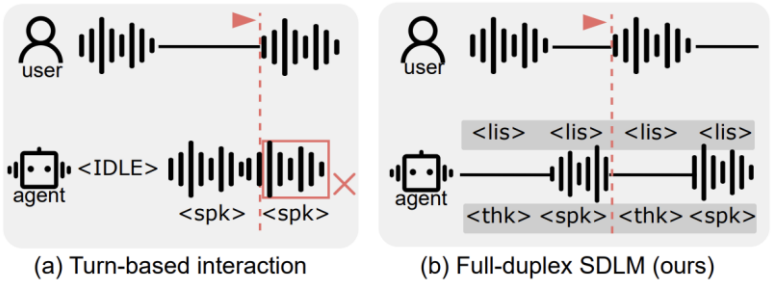
Donghang Wu*, Tianyu Zhang*, Yuxin Li, Hexin Liu, Chen Chen, Eng Siong Chng, Yoshua Bengio

Core thesis

Use the user's speaking time for latent reasoning, **improving response quality and reasoning while preserving full-duplex interaction.**

Background

Full-duplex changes spoken dialogue from turn-taking to overlap. The agent continuously listen while deciding whether to think, speak, or yield.



(a) Turn-based interaction (b) Full-duplex SDLM (ours)

Full-duplex enables streaming input, barge-in, and proactive turn-taking, but it also creates a new question: **what should the model compute while the user is still speaking?**

- Turn-based models wait until the end of the user turn.
- Full-duplex SDLMs keep listening and must switch states online.

Motivation

Full-duplex gives an always-on model; the open question is how to spend computation during user speech.

01 **Existing full-duplex systems**

Often output non-informative **padding or silence tokens** while listening.

02 **Explicit CoT is not a clean fit**

Explicit textual CoT creates causal problems in streaming speech.

03 **FLAIR (Our method)**

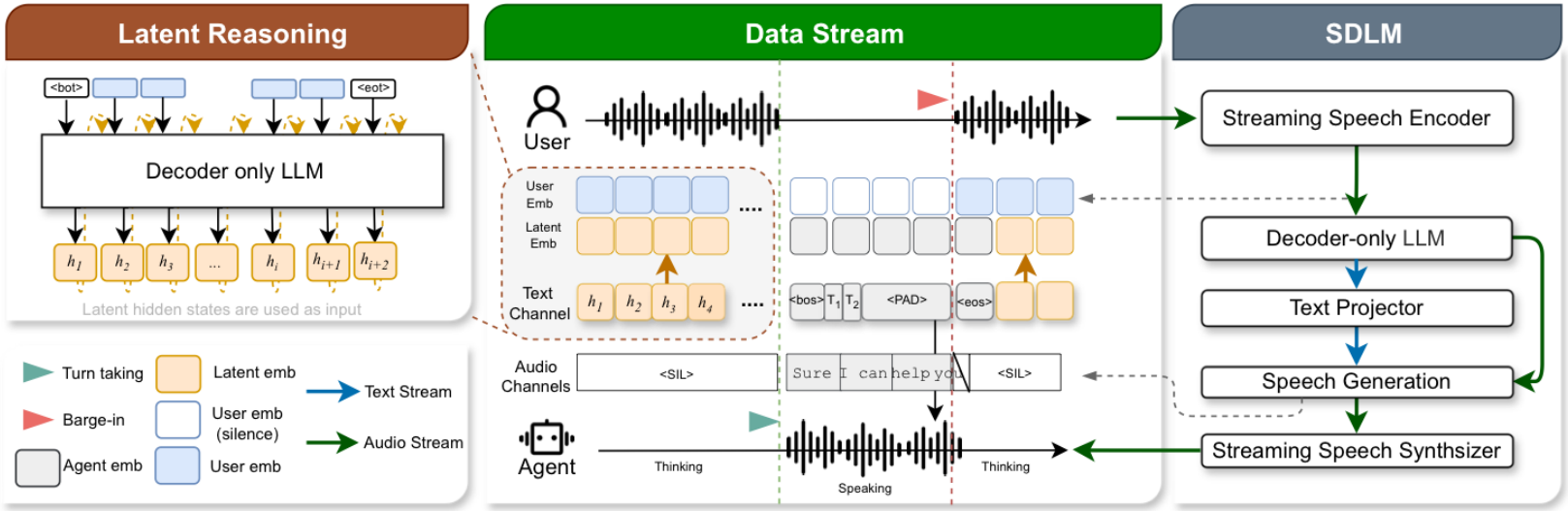
Use an **implicit latent state** that evolves with audio and switches to response generation instantly.

Research objective:

Make a full-duplex SDLM think in latent space while listening, to improve the response quality without explicit reasoning annotations or added inference latency.

Method overview

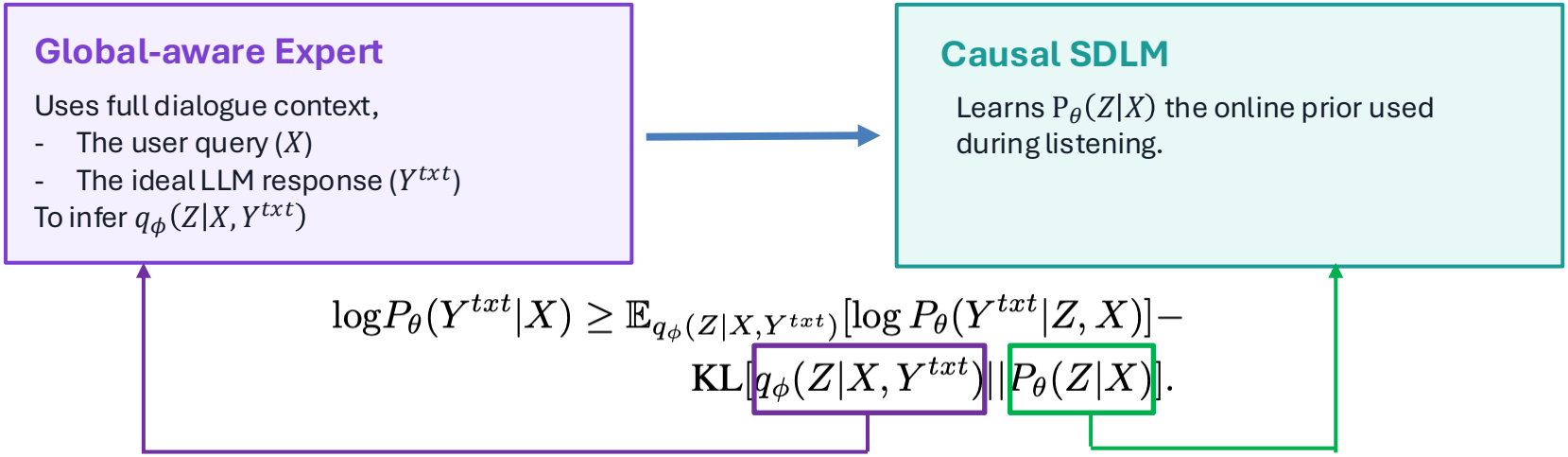
FLAIR replaces silence-token listening with latent reasoning



During user speech, latent embeddings are fed back into the decoder-only LLM; after turn-taking fires, text tokens drive response and speech synthesis.

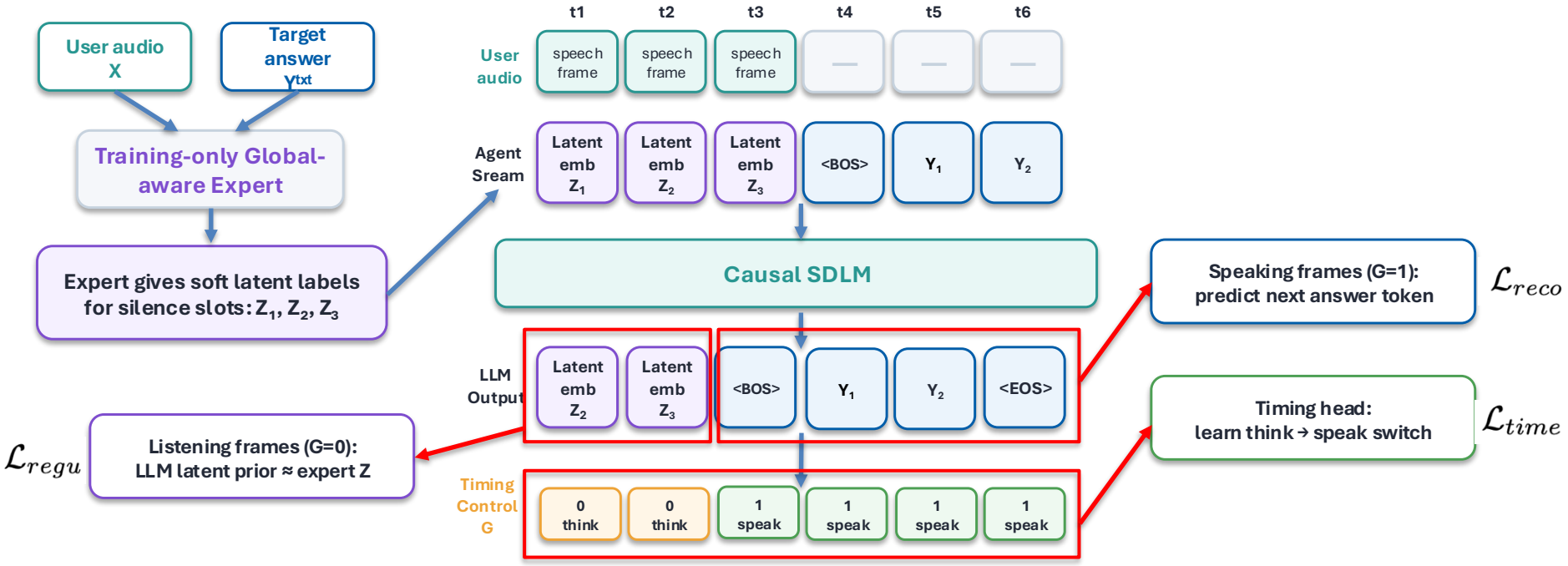
SFT for Latent Reasoning: ELBO makes latent reasoning trainable

A non-causal Global-aware expert is used only during SFT to provide an approximate posterior over latent states $\boldsymbol{Z}$.



Training transfers global reasoning from the expert posterior into a causal prior that survives after the expert is removed.

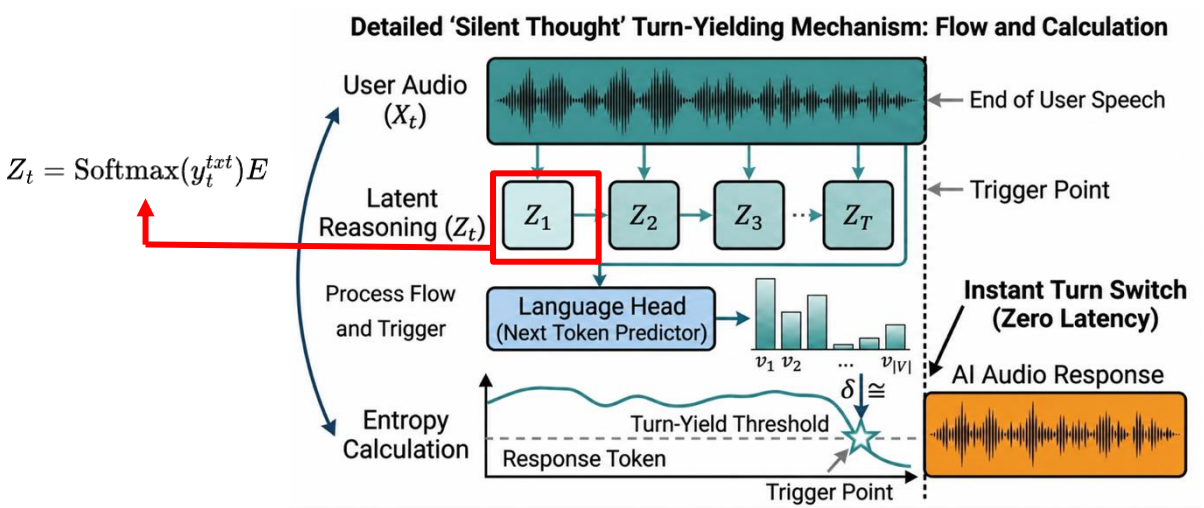
SFT for Latent Reasoning: input construction and output supervision



$$\mathcal{L}_{elbo} = \mathcal{L}_{reco} + \alpha \cdot \mathcal{L}_{regu} + \beta \cdot \mathcal{L}_{time}$$

Inference: G_t controls the online state machine.

The Global-aware Expert is discarded; the causal model decides step-by-step whether to think or speak.



G_t= 0: latent reasoning

Use text logits to compute a weighted vocabulary embedding Z_t and feed it into the next step; pass silence to the speech generator.

G_t :1 = response generation

Output a text token and feed its token embedding into the next step; speech synthesis follows the response stream.

Barge-in

If the user interrupts, the model predicts when to stop speaking and returns to the latent reasoning state for the new user speech.

Training and Evaluation

- Training:
  - Pre-training: 530K hours of **speech continuation** data.
  - SFT: instruction following on 70K hours of **instruction QA**, and 20K hours of **ASR-QA data**
    - Latent Reasoning SFT: Train expert with L_{reco} first, then optimize full L_{elbo} jointly.
    - Speech Synthesizing SFT: Freeze other modules and train streaming speech generation.
- Evaluation
  - Response quality and reasoning
    - Llama Questions, WebQuestions, TriviaQA, SDQA, AlpacaEval, CommonEval, OpenbookQA, MMSU.
  - Full-duplex interaction
    - Full-Duplex-Bench
  - **What we want to show: latent reasoning improves answer quality and reasoning while keeping the full-duplex behavior essentially intact.**

Main result

- Response Quality:

Method	FD	LlamaQ	WebQ	TriQA	SDQA	AlpacaE	ComE	OBQA	MMSU
Moshi (Défossez et al., 2024)	✓	54.5	22.1	16.7	15.6	2.01	1.60	25.9	24.0
Freeze-Omni (Wang et al., 2024b)	✓	56.2	27.9	28.5	53.5	4.03	3.46	31.0	28.1
SALMONN-omni (Yu et al., 2025)	✓	73.6	43.7	56.0	-	3.22	-	-	30.0
SALM-Duplex (Hu et al., 2025)	✓	51.3	25.0	16.9	26.0	2.99	2.50	39.6	26.3
GLM-4-Voice (Zeng et al., 2024)	✗	65.7	37.0	47.5	37.0	3.97	3.42	53.4	39.8
Qwen2-Audio (Chu et al., 2024)	✗	69.7	45.2	40.3	35.7	3.74	3.43	49.5	35.7
Kimi-Audio (Ding et al., 2025)	✗	68.3	37.3	51.2	63.1	4.46	3.97	83.5	62.2
Baichuan-Audio (Li et al., 2025)	✗	74.0	40.7	53.0	45.8	4.41	4.08	71.7	53.2
STITCH-R (Chiang et al., 2025b)	✗	70.0	50.3	49.6	-	2.70	-	-	-
FLAIR w/o *thk*	✓	73.0	41.7	**53.8**	54.4	3.80	3.54	72.9	50.2
FLAIR w/ *thk*	✓	**78.0**	**43.0**	51.2	**56.2**	**3.85**	**3.65**	**74.2**	**56.2**

- Full-duplex interaction

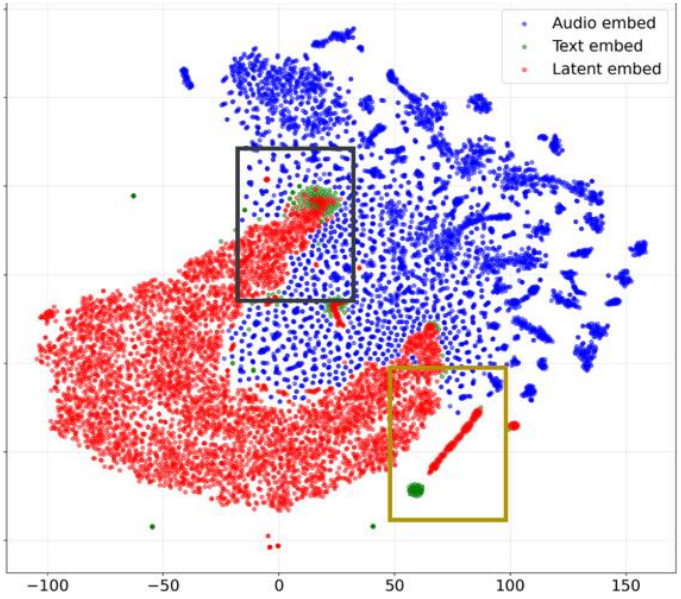
Method	Turn-taking		Barge-in		
	TOR (↑)	Latency (↓)	TOR (↑)	GPT-4o (↑)	Latency (↓)
Freeze-Omni (Wang et al., 2024b)	33.6	0.95	86.7	3.62	1.41
dGSLM (Nguyen et al., 2023)	97.5	0.35	91.7	0.20	2.53
Moshi (Défossez et al., 2024)	94.1	0.27	100	0.77	0.26
Gemini Live [1]	65.5	1.30	89.1	3.38	1.18
FLAIR w/o *thk*	**94.1**	**0.37**	89.0	4.08	**0.35**
FLAIR w/ *thk*	93.0	0.43	**92.0**	**4.22**	0.36

[1] https://ai.google.dev/gemini-api/docs/live

Takeaway

Latent reasoning **improves response quality** while maintaining **interaction behavior** at roughly the same level as the non-thinking full-duplex baseline.

Visualization: latent embeddings bridge audio embeddings and target text.



What the plot shows

Latent embeddings form trajectories from the user audio embedding space toward target text embeddings, suggesting an internal bridge for response generation.

Why it matters

This gives qualitative evidence that the listening-phase latent states are not arbitrary padding; they carry response-oriented information.

Conclusion

1) Proposed FLAIR Framework: the "think-while-listening" framework for full-duplex SDLMs, enabling simultaneous speech perception and implicit latent reasoning.
2) ELBO-Based Training Strategy: an ELBO-based objective using a Global-aware Expert to guide the causal model, effectively bypassing the need for explicit, manually annotated reasoning datasets.
3) Achieved significant improvements on response quality. Maintained optimal conversational dynamics with zero additional inference latency.
4) FLAIR marks a pivotal transition from simple "perceive-respond" patterns to human-like synchronous perception-cognition-expression in spoken AI agents.