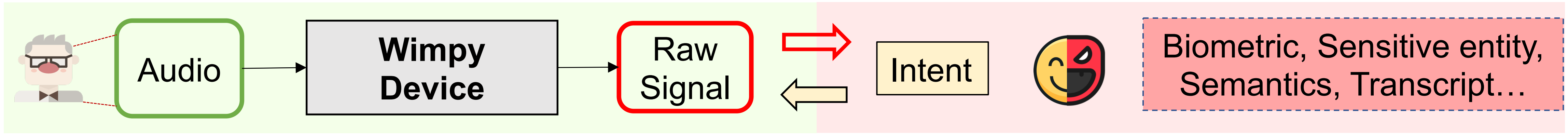
The 38th Annual Conference on Neural Information Processing Systems (NeurIPS 2024)

SILENCE: Protecting privacy in offloaded speech understanding on resource-constrained devices

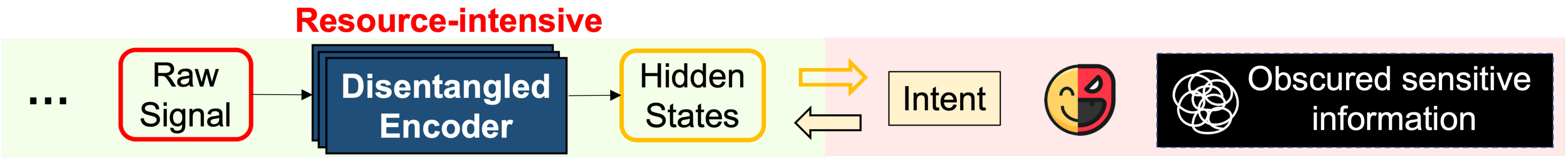
Dongqi Cai[1], Shangguang Wang[1], Zeling Zhang[1],
Felix Xiaozhu Lin[2], Mengwei Xu[1]

[1] Beiyou Shenzhen Institute
[2] University of Virginia

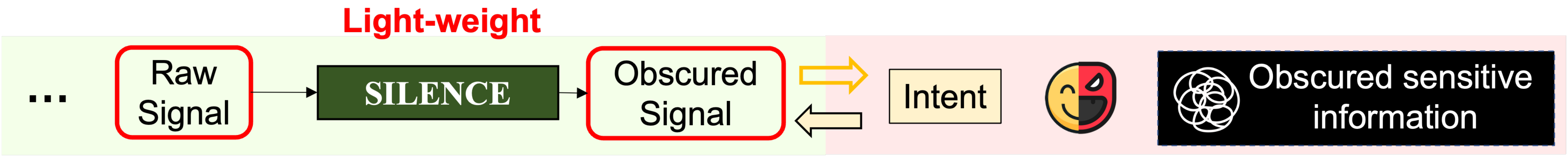
Privacy concern for cloud speech service



(a). Offloaded speech understanding on wimpy devices

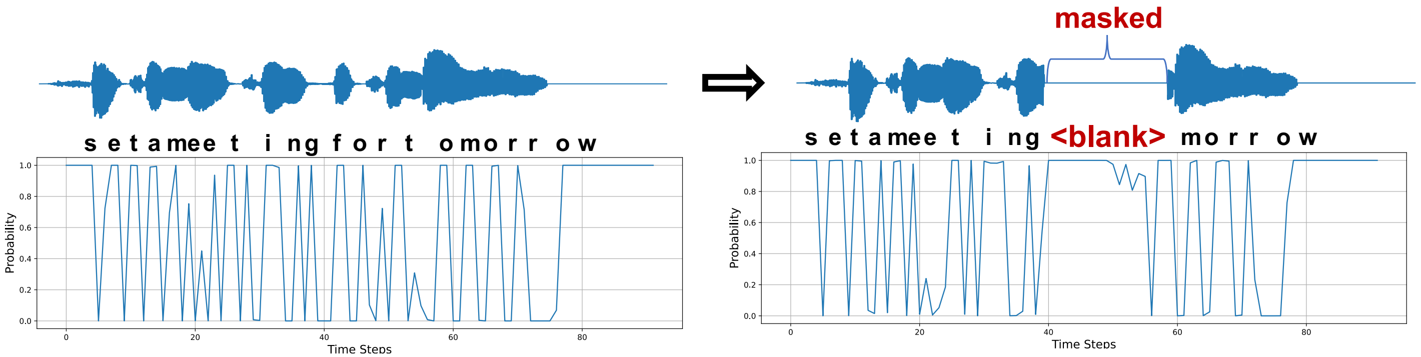


(b). Previous approaches to protect speech privacy

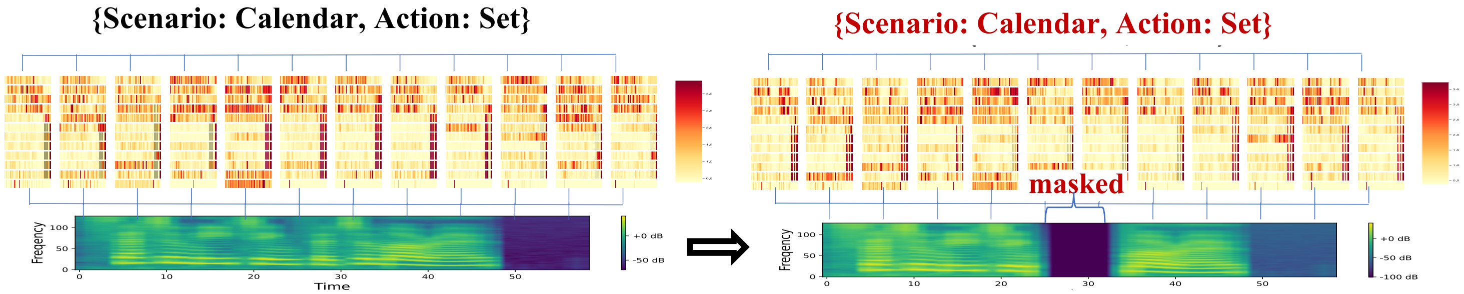


(c). Our SILENCE: a novel asymmetric dependency-based encoder

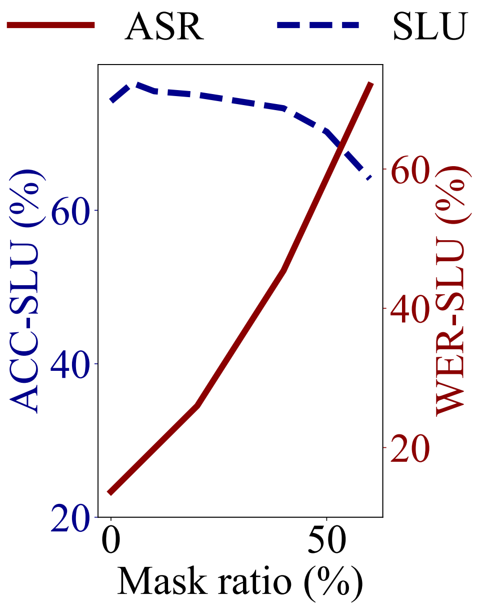
Observation: Asymmetric dependency



(a) Peaky phoneme is short-dependent

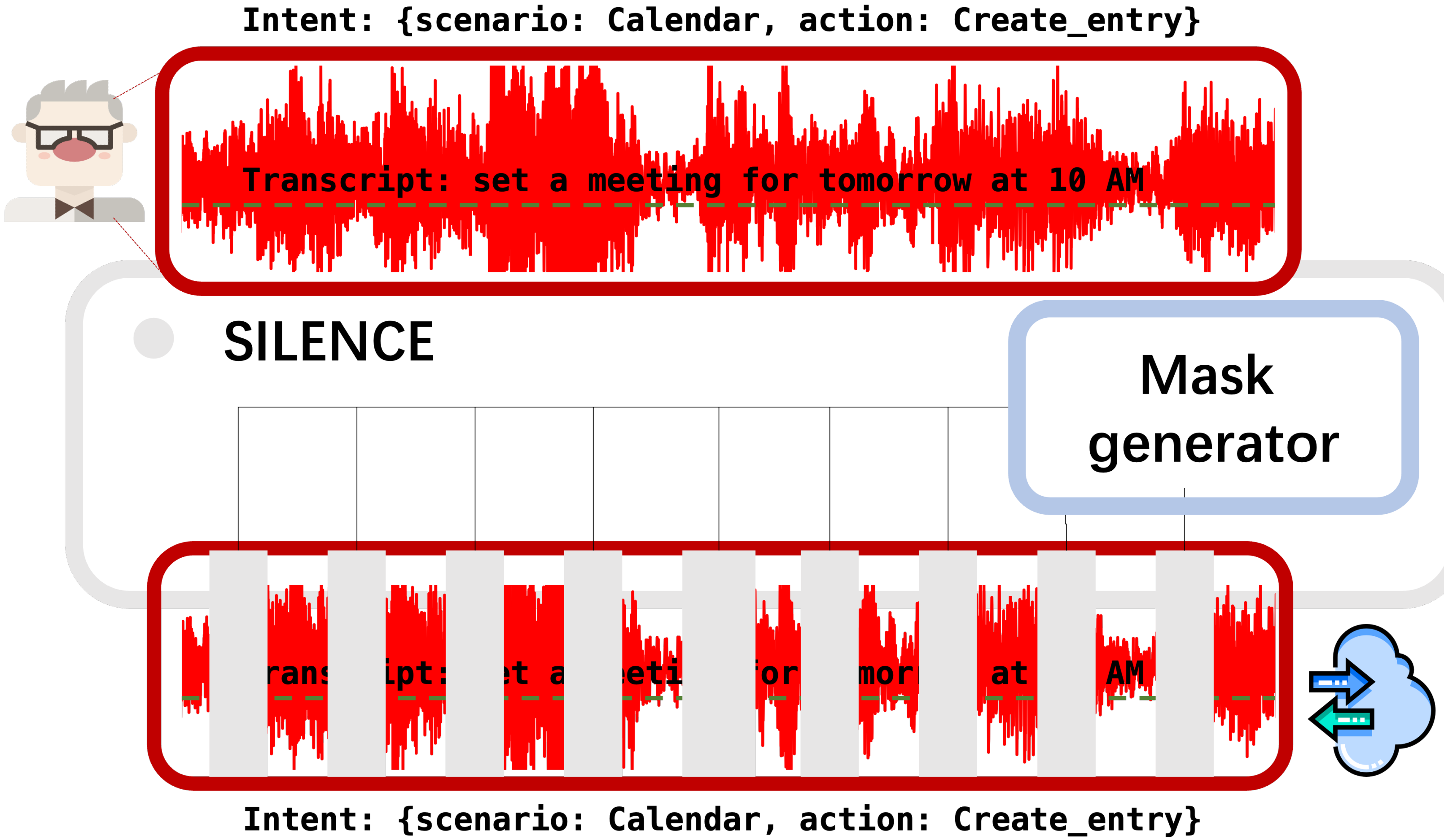


(b) Attention seeks intent globally



(c) Empirical performance under different ratios of masked portion.

System overview



Intent: {scenario: Calendar, action: Create_entry}

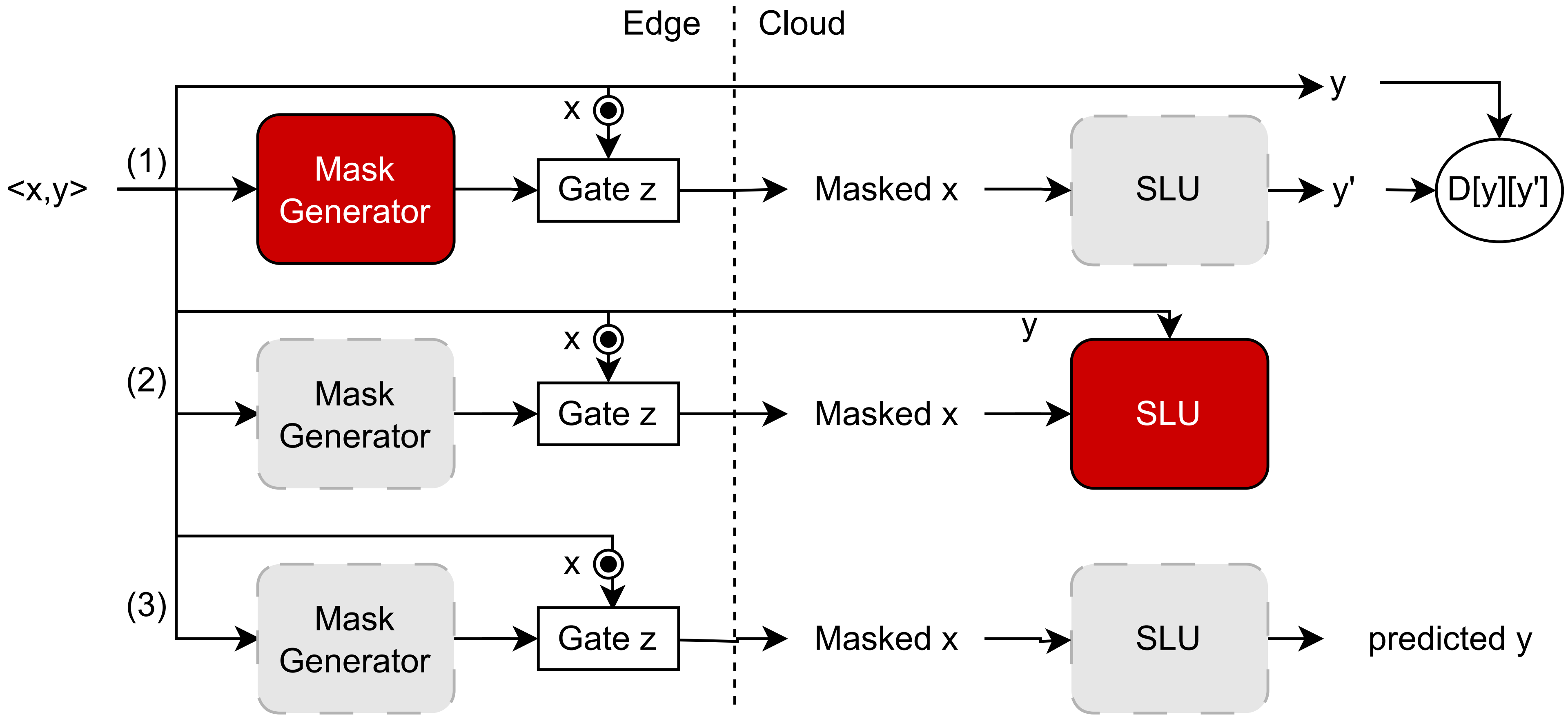
Transcript: set a meeting for tomorrow at 10 AM

SILENCE

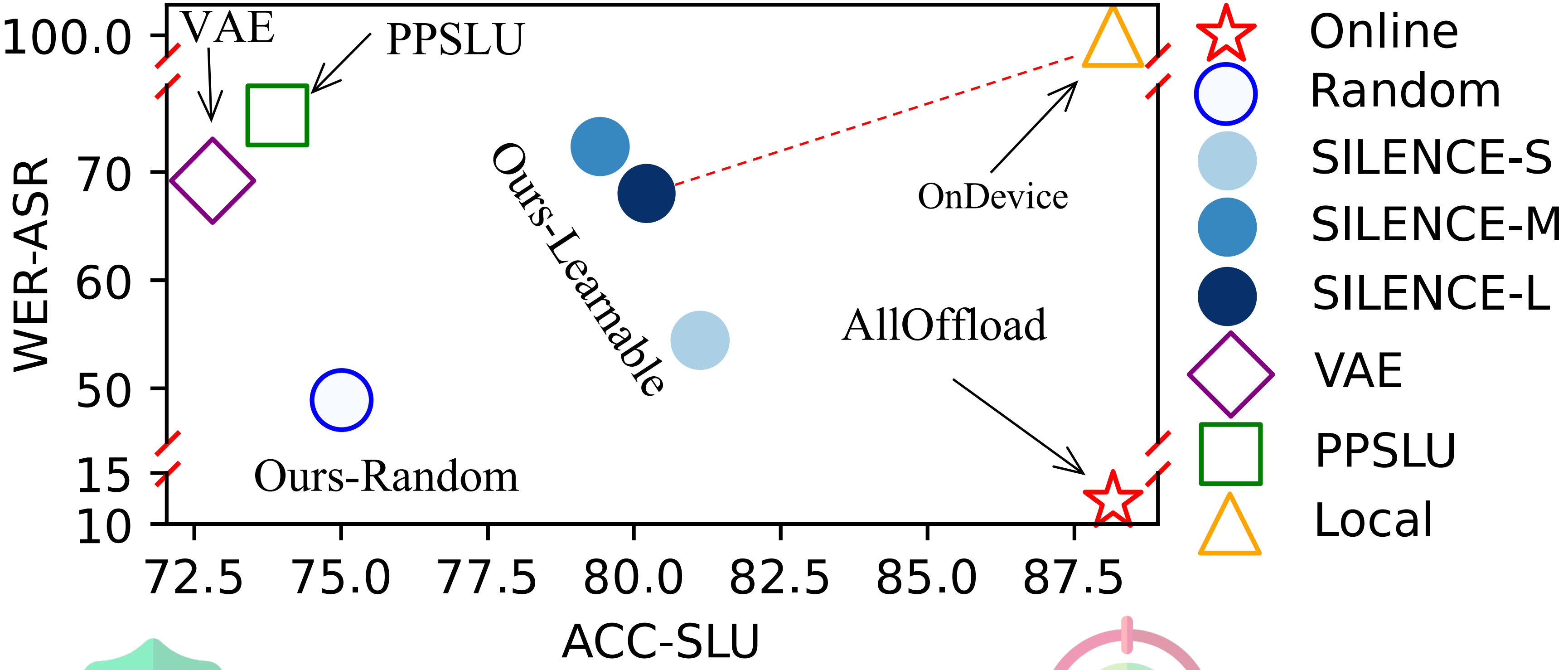
Mask generator

Intent: {scenario: Calendar, action: Create_entry}

Concrete design: interpretable mask

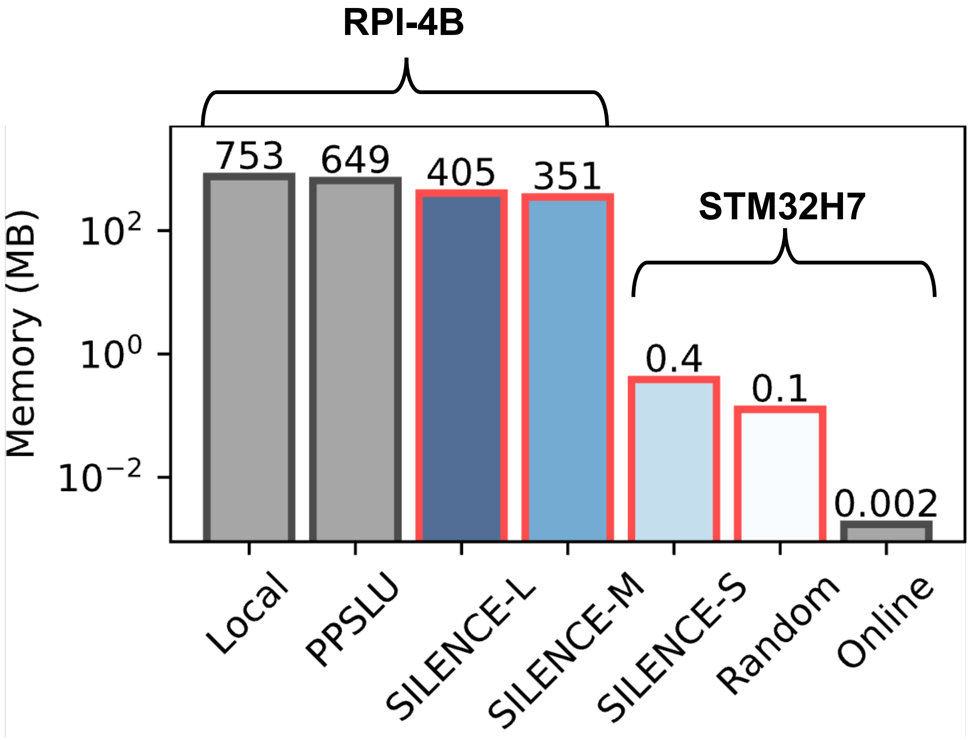


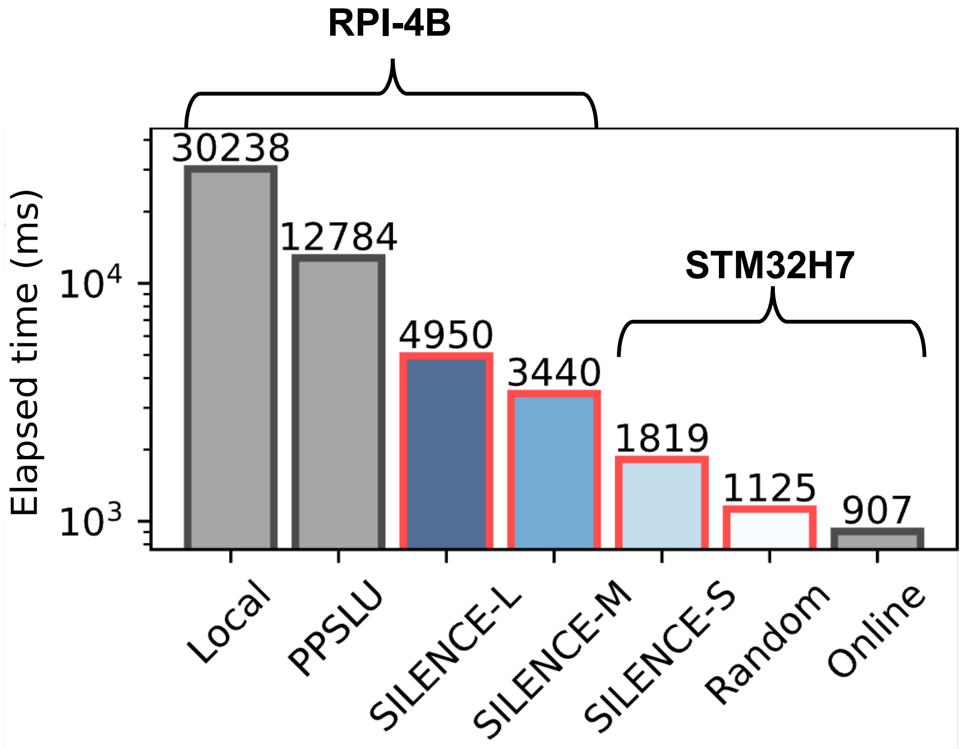
Results: attack protection



As safe as entangled encoder, with good SLU performance.

Results: on-device efficiency





134.1x less memory, making it runnable on MCU!
With up to **53.3x** speedup.

The 38th Annual Conference on Neural Information Processing Systems (NeurIPS 2024)

SILENCE: Protecting privacy in offloaded speech understanding on resource-constrained devices

Dongqi Cai[1], Shangguang Wang[1], Zeling Zhang[1],
Felix Xiaozhu Lin[2], Mengwei Xu[1]

[1] Beiyou Shenzhen Institute
[2] University of Virginia