



Northeastern University
Institute for Intelligent
Networked Systems



Low-Latency Semantic Communications

Francesca Meneghello

Principal Research Scientist

Northeastern University, Boston (MA), United States

E-mail: fr.meneghello@northeastern.edu

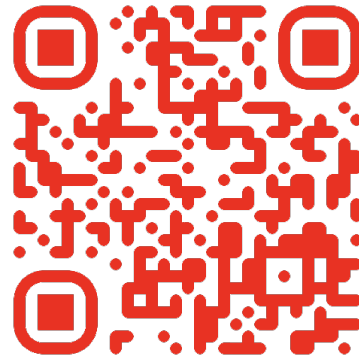
Website: <https://francescamen.github.io/>

Research Group Website: <http://mentis.info>

KTH Royal Institute of Technology – Stockholm, Sweden – June 5, 2026

Francesca Meneghello

- Principal Research Scientist at the Institute for the Institute for Intelligent Networked Systems, Northeastern University, USA
- PhD in Telecommunications Engineering 2022 from the Department of Information Engineering, University of Padova, Italy
- Fulbright-Schuman Scholar, Northeastern University, USA 2023



MENTIS Laboratory

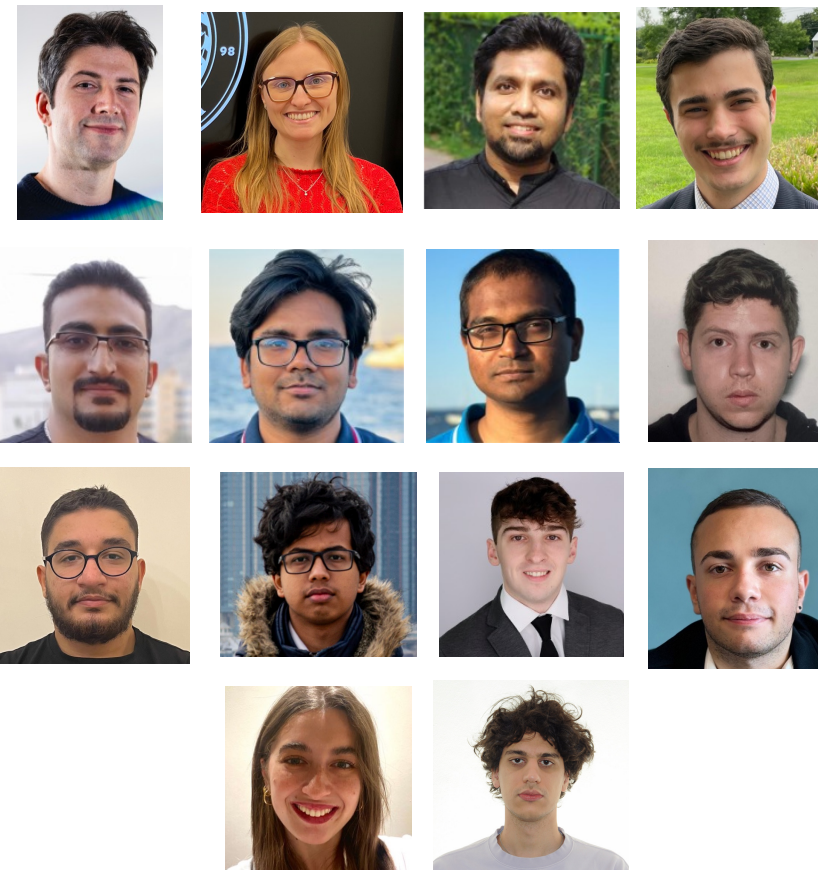
- **Research Group**

- 12 PhD students, 1 Research Scientist
- Mainly funded by NSF and DoW



- **Research Focus**

- Cyber-physical AI-driven systems
 - **networking, learning**
 - ... and their **hardware** interfaces



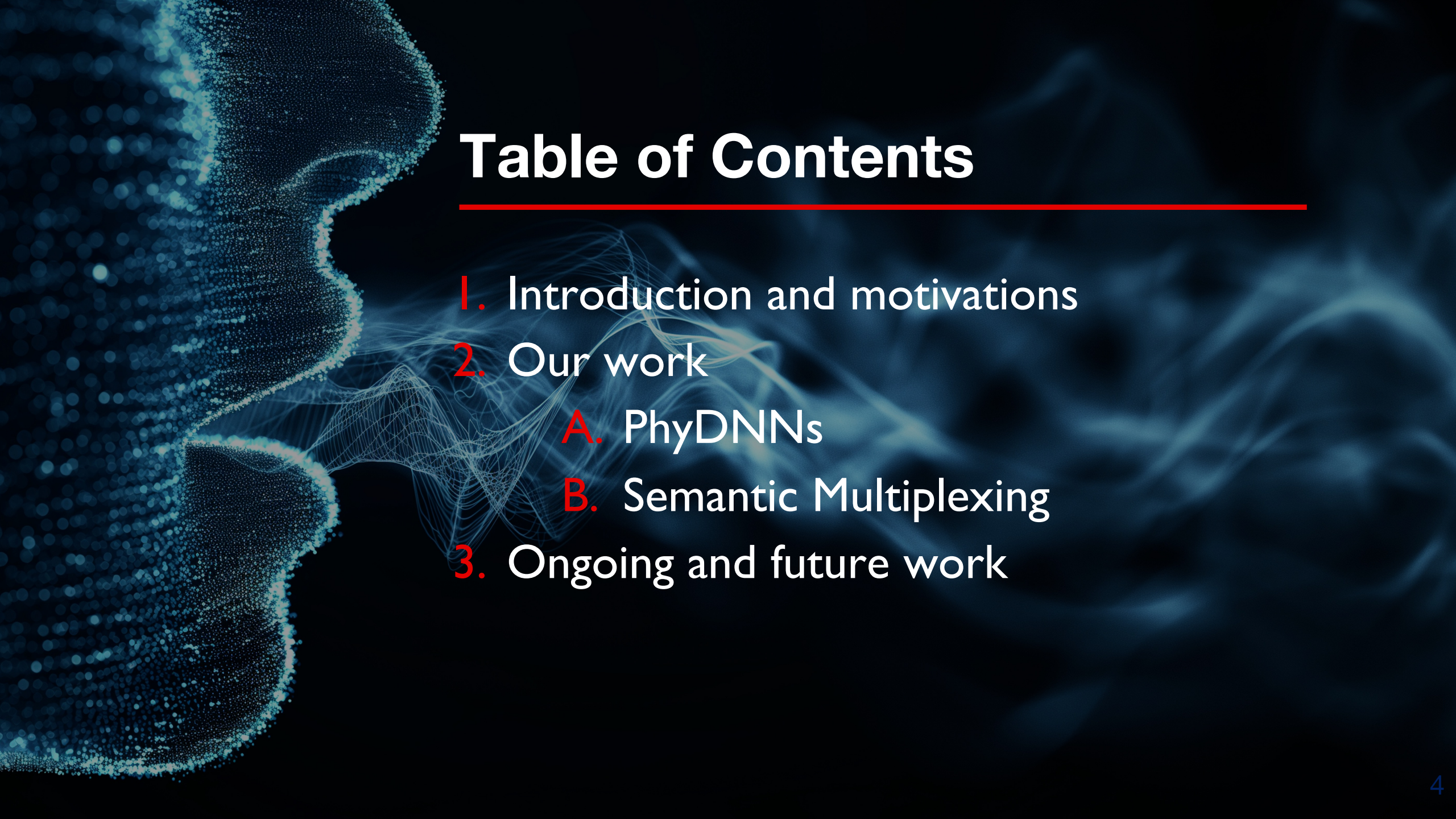


Table of Contents

1. Introduction and motivations
2. Our work
 - A. PhyDNNs
 - B. Semantic Multiplexing
3. Ongoing and future work



Intro and Motivations

- What does semantic communication mean?
- Why low-latency?

Physical world



Digital world



New applications



Virtual reality

Telemedicine

Augmented reality

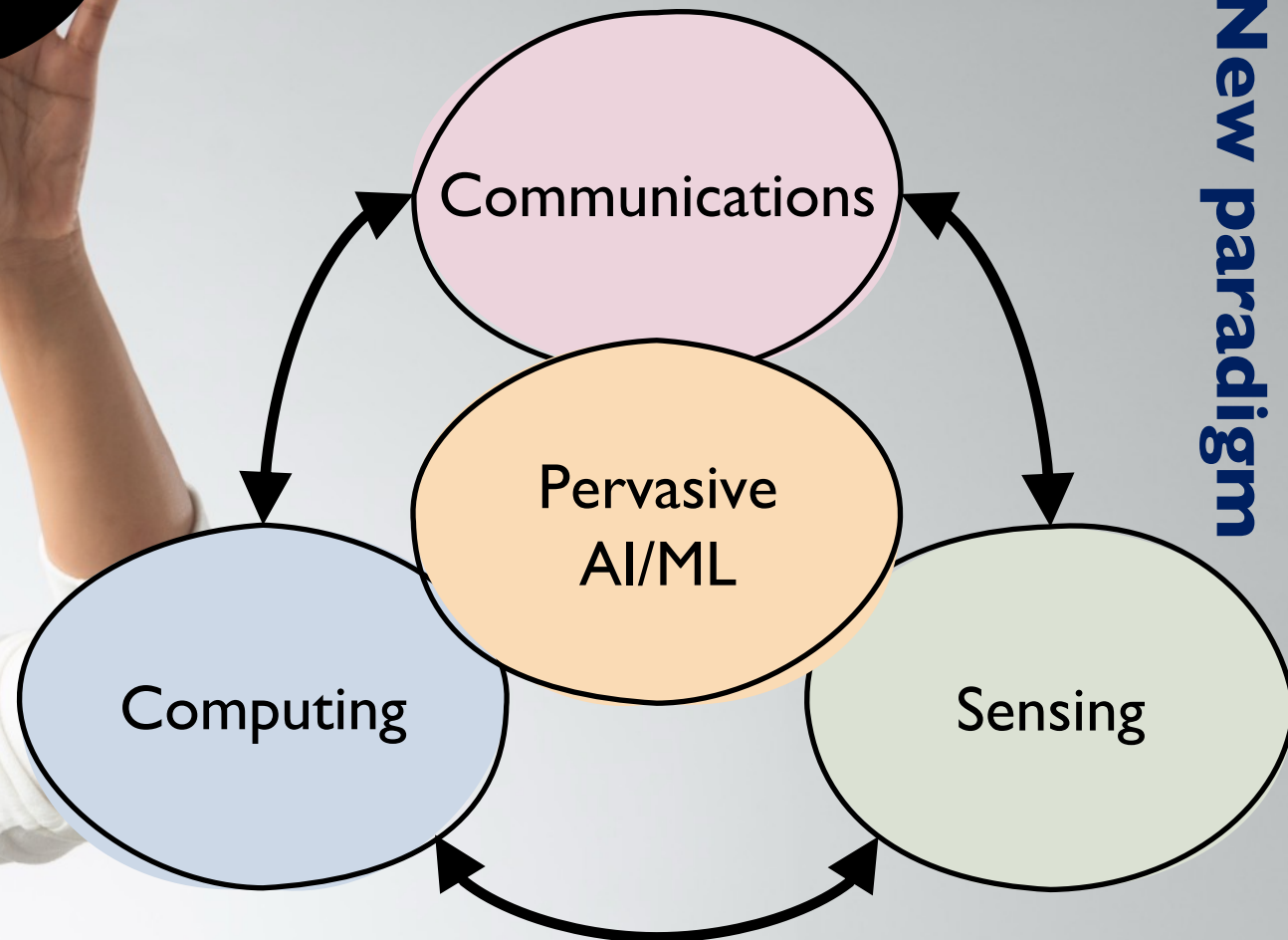
Digital twin

Metaverse

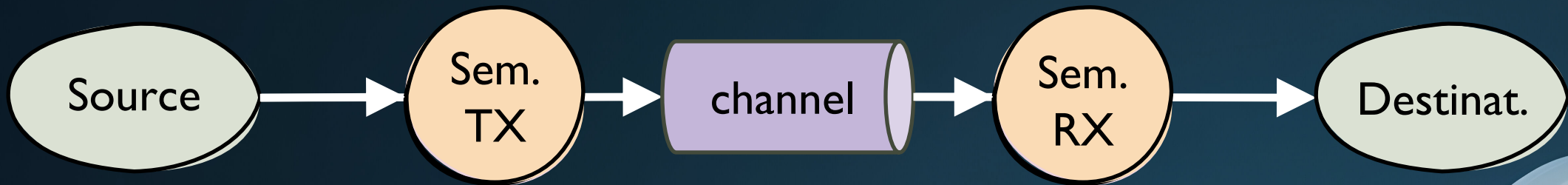
Indoor monitoring

Telepresence

New paradigm



Semantic Communications



Weaver's three levels of communication

Level A — Technical

How accurately are **bits** transmitted?

Level B — Semantic

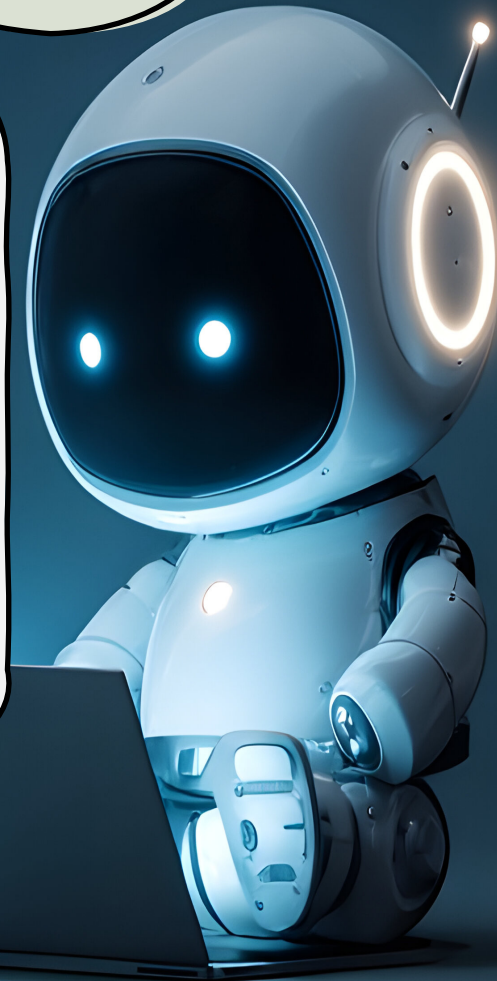
How precisely is **meaning** conveyed?

Level C — Effectiveness

How effectively does **meaning** affect behavior?

Traditional Comms

Semantic Comms

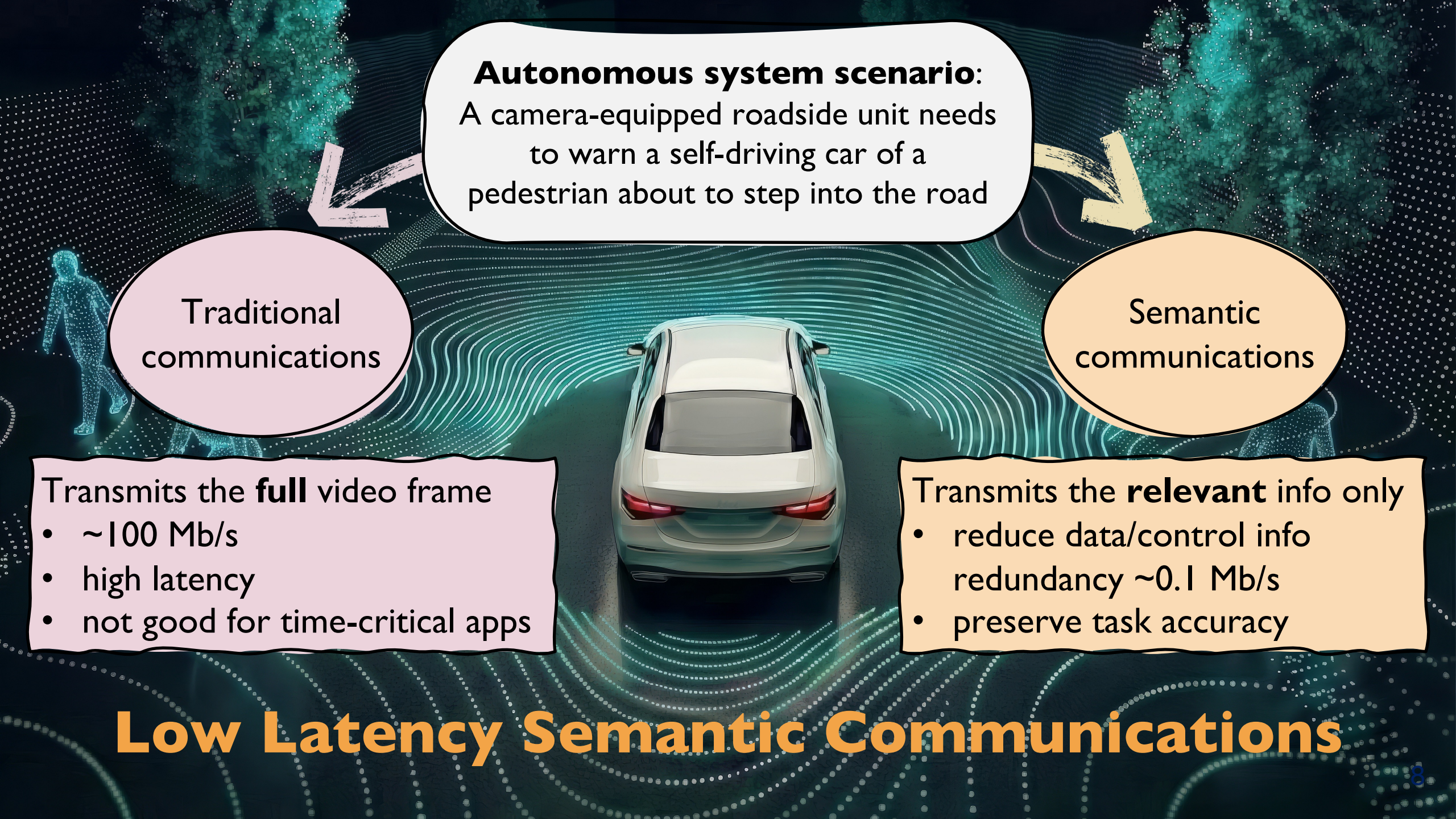


Warren Weaver, "Recent Contributions to the Mathematical Theory of Communication," in Claude E. Shannon and Warren Weaver, The Mathematical Theory of Communication, University of Illinois Press, 1949.

Nguyen, Loc X., et al. "A contemporary survey on semantic communications: Theory of mind, generative ai, and deep joint source-channel coding," IEEE Communications Surveys & Tutorials, 2025.

Cabrera, Juan A., et al. "6G and the Post-Shannon Theory," Shaping Future 6G Networks: Needs, Impacts, and Technologies, Wiley 2021.

Chaccour, Christina, et al. "Less data, more knowledge: Building next-generation semantic communication networks," IEEE Communications Surveys & Tutorials, 2024.



Autonomous system scenario:
A camera-equipped roadside unit needs to warn a self-driving car of a pedestrian about to step into the road

The diagram illustrates a self-driving car (a white sedan) on a road, with green concentric circles representing its sensor range. To the left, a blue wireframe pedestrian is shown stepping into the road. Two large arrows point from a central white box towards two colored boxes: a pink one on the left and an orange one on the right. The background is dark with green particle effects.

Traditional
communications

Transmits the **full** video frame

- ~100 Mb/s
- high latency
- not good for time-critical apps

Semantic
communications

Transmits the **relevant** info only

- reduce data/control info redundancy ~0.1 Mb/s
- preserve task accuracy

Low Latency Semantic Communications

2-A

PhyDNNs

Mohammad Abdi, Khandaker Foysal Haque,
Francesca Meneghello, Jonathan Ashdown,
and Francesco Restuccia

**PhyDNNs: Bringing Deep Neural
Networks to the Physical Layer**

in Proceedings of IEEE INFOCOM 2025

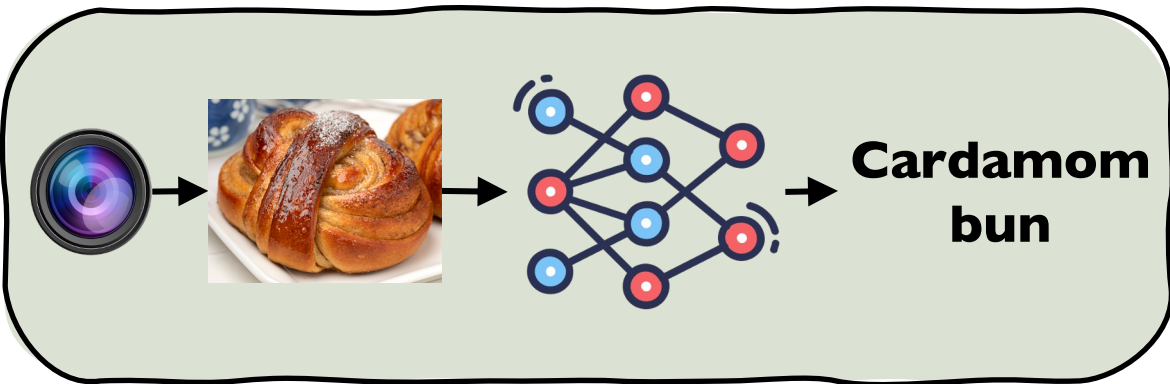


**Cardamom
bun**



Traditional Solutions

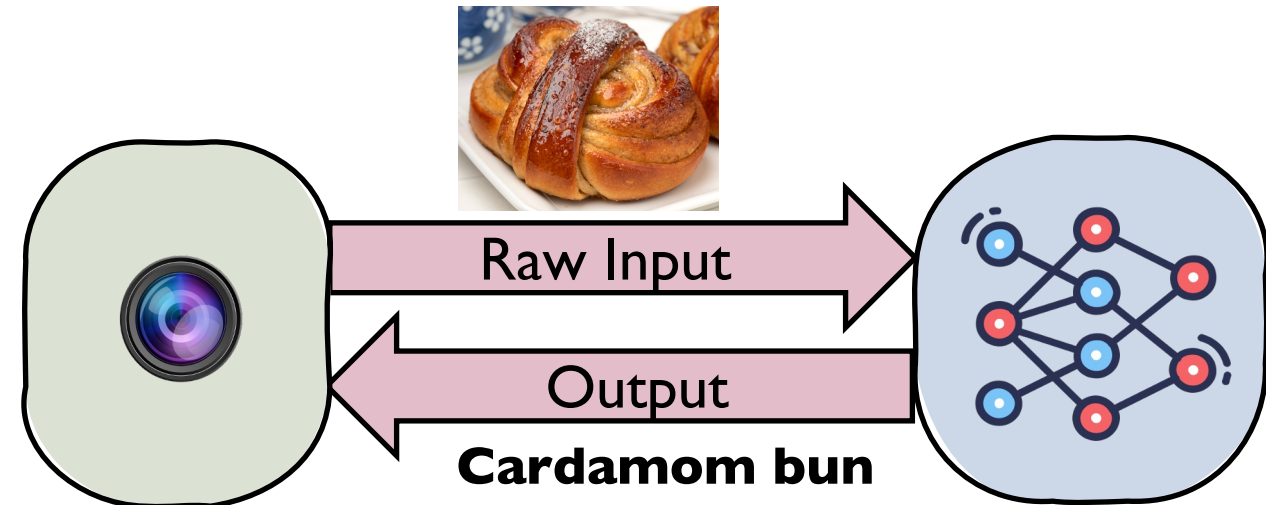
Local Computing



Mobile Device



Full Edge Offloading



Mobile Device



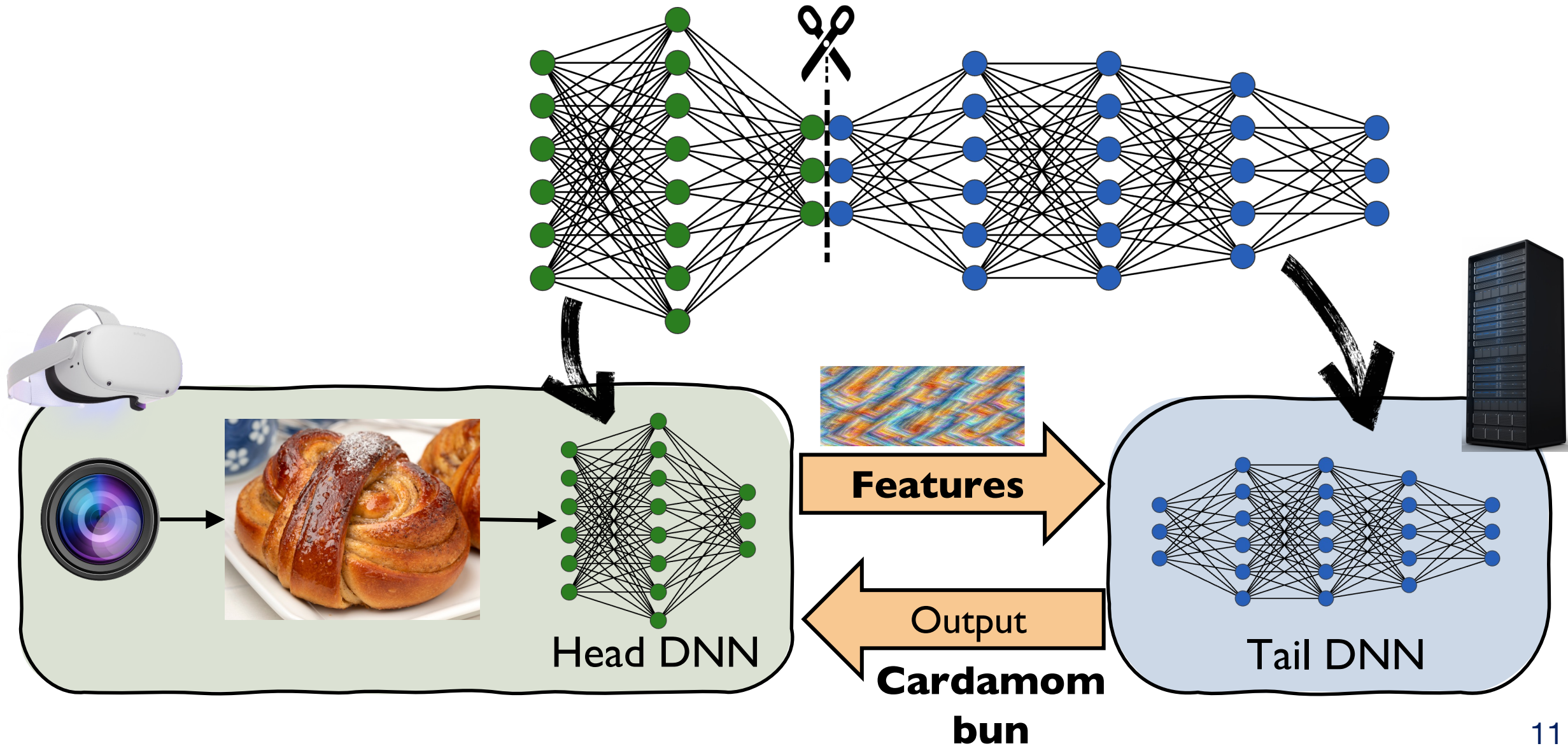
Communication network



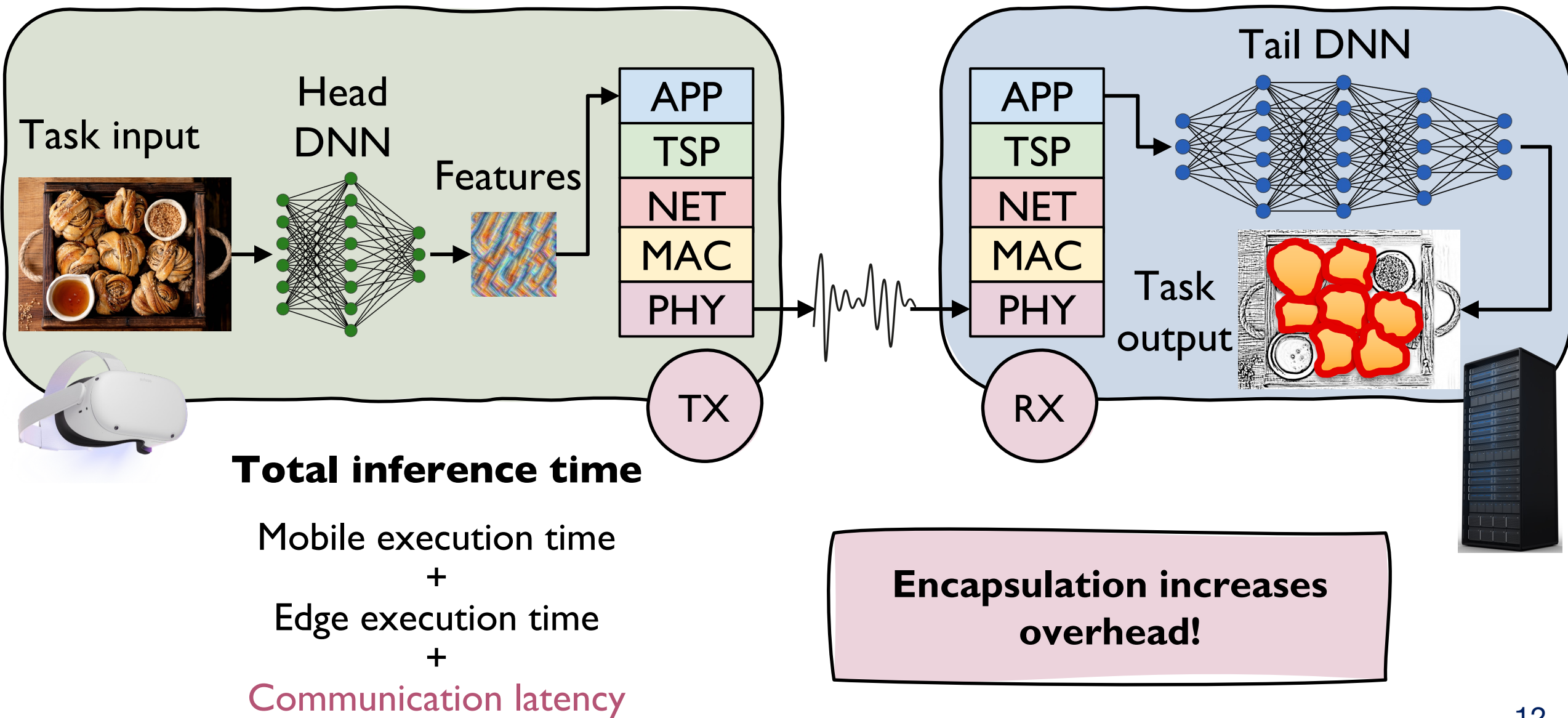
Edge Server



Split the Computation and Transmit Semantics

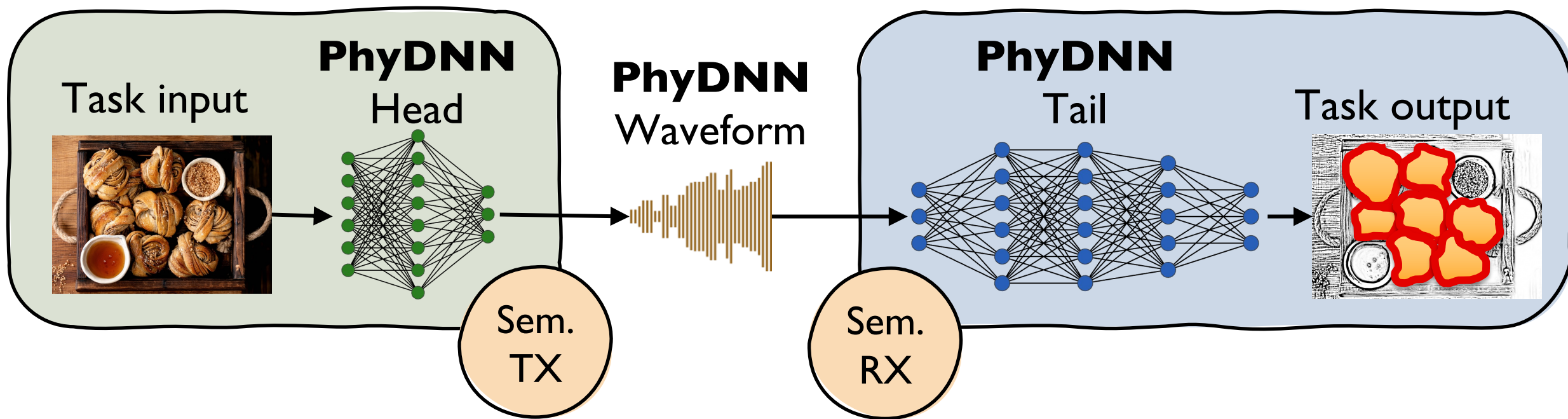


Existing Semantic Comms Challenge



PhyDNNs Approach

PhyDNNs are DNNs that can be executed directly at the **waveform** layer

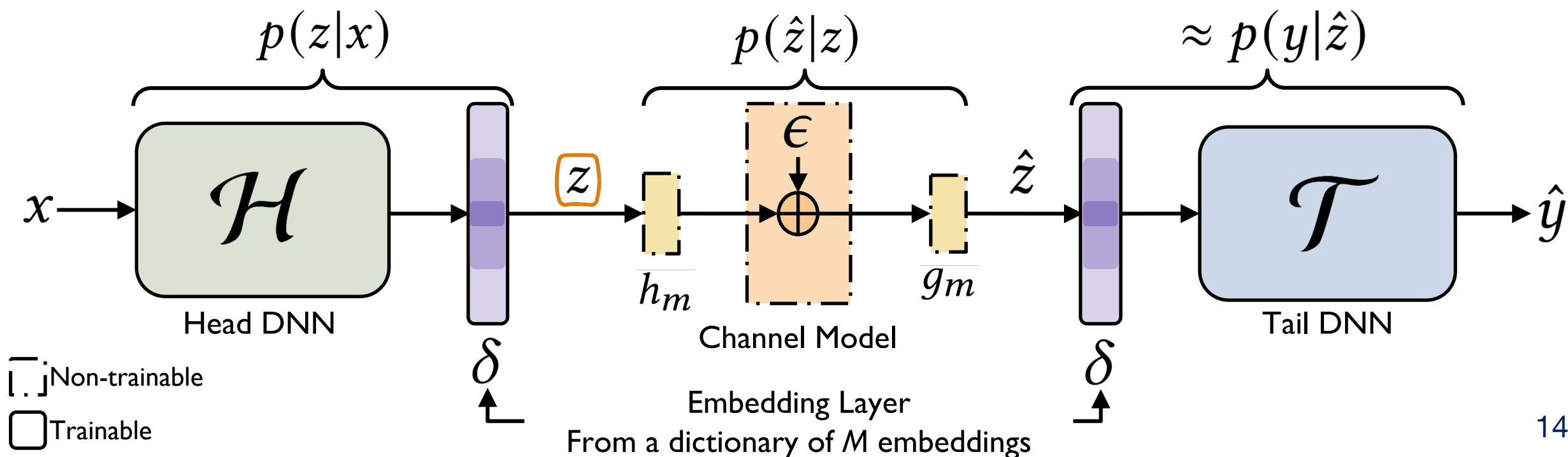


Challenge:

Need to learn channel coding on top of the main task

PhyDNNs Probabilistic System Model

1. Take a pre-trained arbitrary DNN
2. Divide it into head and tail DNNs based on mobile/edge processing power
3. Add **embedding layers** (stochastic) as trainable layers $\Rightarrow$ Extract **latent symbols** $\bar{z}$
4. Add the **modulation block** as non-trainable layers $\Rightarrow$ Convert to **waveform**
5. Add the **wireless channel model** as a non-trainable layer $\Rightarrow$ Increase **robustness**



Loss Function Formulation

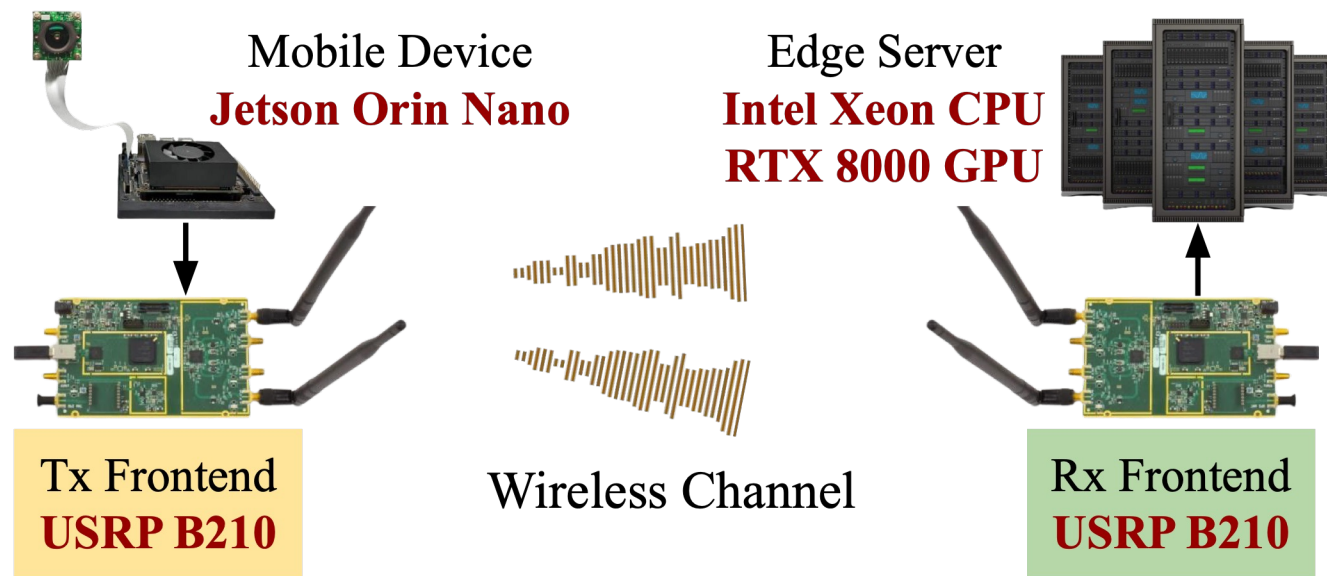
The latent symbols $\hat{Z}$ should carry only **task-specific information**

Maximally predictive about output Y

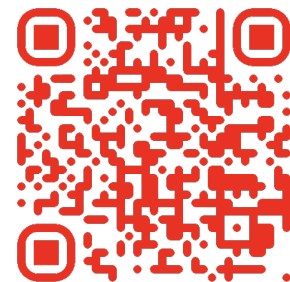
Maximally compressive about input X

$$\begin{aligned}\mathcal{L}_{IB} &= \boxed{-I(\hat{Z}; Y)} + \beta \cdot \boxed{I(X; \hat{Z})} \\ &= \mathbb{E}_{p(x, y)} \left\{ \underbrace{\mathbb{E}_{p(\hat{z}|x)} [-\log p(y|\hat{z})]}_{\text{Task Performance}} \right. \\ &\quad \left. + \beta \cdot \underbrace{D_{KL}(p(\hat{z}|x) || p(\hat{z}))}_{\text{Communication Efficiency}} \right\} - H(Y),\end{aligned}$$

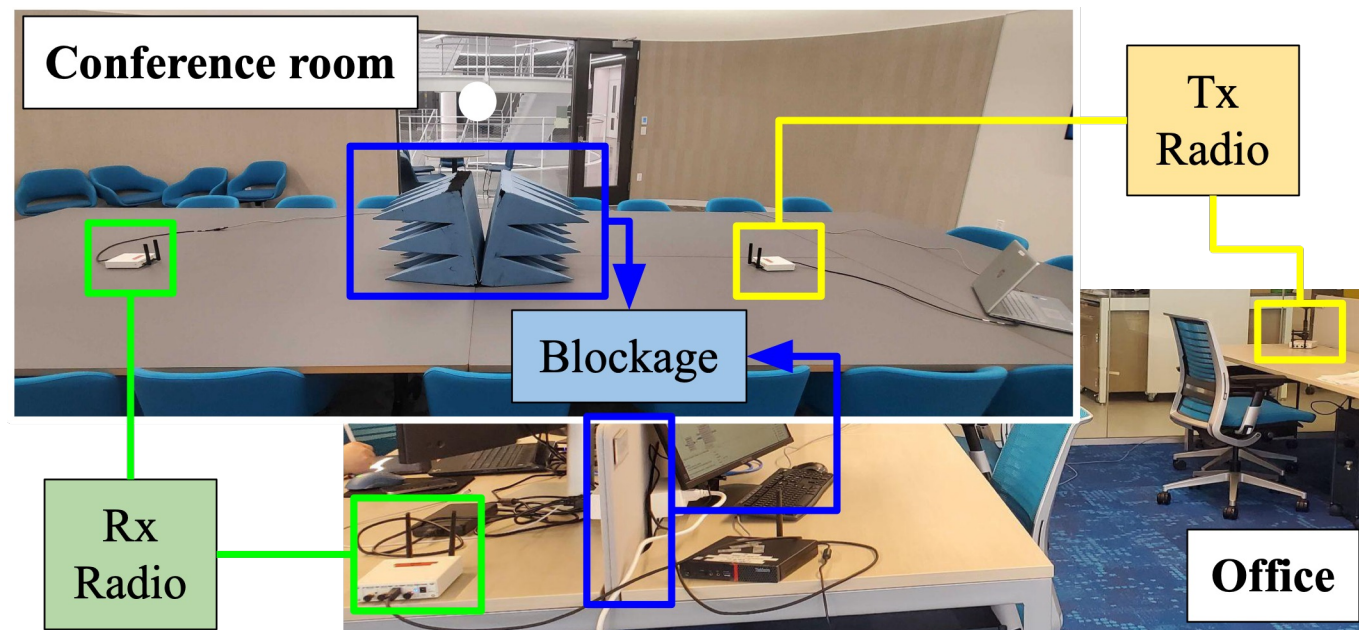
Experimental Setup



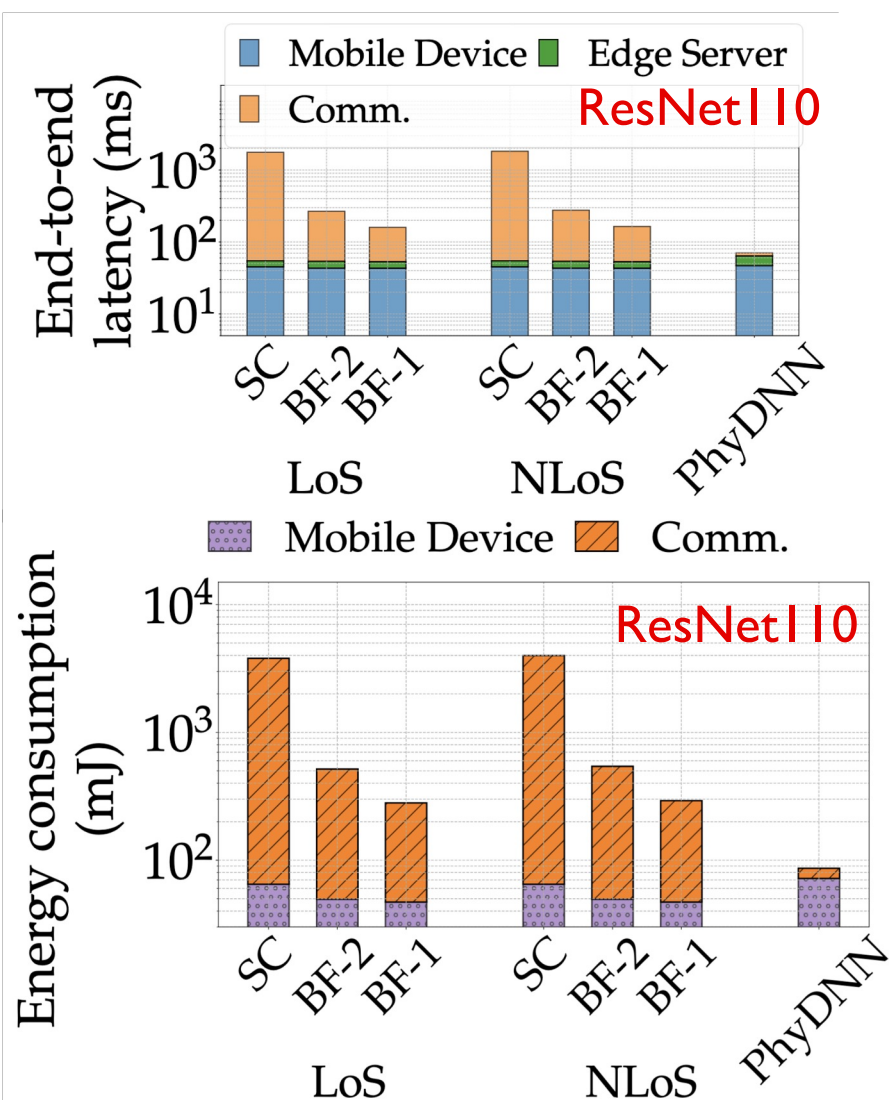
Scan me to
get the code!



- **Task:** Image classification
- **Datasets:** CIFAR-10/CIFAR-100
- **DNN architectures:** ResNet20, ResNet56, and ResNet110



Experimental Results



Baselines:

- **SC**: split computing [1] uses the first ResNet block as head
- **BF**: BottleFit [2] reduces width of the last conv in the first ResNet block of 1-2 kernels

Results (average across different DNNs):

- Reduces by 2.3x **latency** than BF
- Reduces by 17x **communication latency** than BF (17x smaller payload)
- Reduces by 70% **energy consumption**
- Reduces by 92% **communication energy consumption**

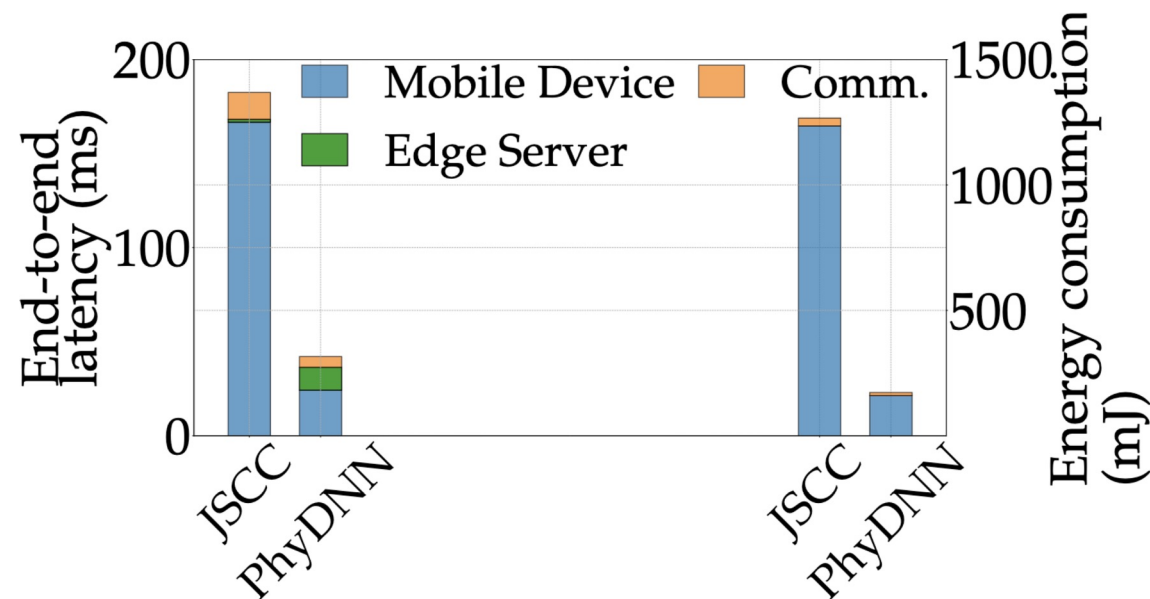
	ResNet-20	ResNet-56	ResNet-110
PhyDNNs			
LOS	78.23 %	82.34 %	84.85 %
NLOS	74.64 %	78.51 %	80.74 %
Full stack	Baseline	87.52%	90.6%
		92.26%	

[1] Y. Kang, et al., "Neurosurgeon: Collaborative Intelligence Between the Cloud and Mobile Edge," ACM SIGARCH Computer Architecture News, 2017

[2] Y. Matsubara, et al., "BottleFit: Learning Compressed Representations in Deep Neural Networks for Effective and Efficient Split Computing," in Proc. of IEEE WoWMoM, 2022.

Comparison with SOTA DT-JSCC

	DT-JSCC	PhyDNNs
#Params (Head)	4.36 M	42.5 K
#FLOPs (Head) [Mac]	618.4 M	291.65 M
Accuracy [%]	82.22	80.43



- Encoder is 100x larger than PhyDNNs
- FLOPS are 2x greater than PhyDNNs
- Only 1.79% loss in accuracy

- DT-JSCC increases the latency by $\sim 7\times$
- DT-JSCC consumes $\sim 8\times$ more energy

2-B

Semantic Multiplexing

Mohammad Abdi, Francesca Meneghello
and Francesco Restuccia

Semantic Multiplexing

arXiv preprint 2026



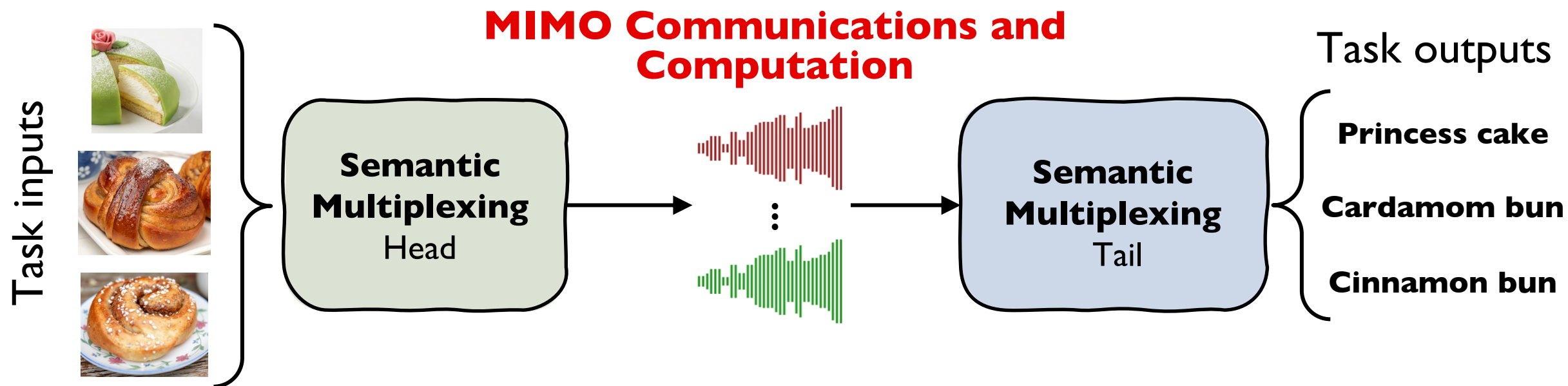
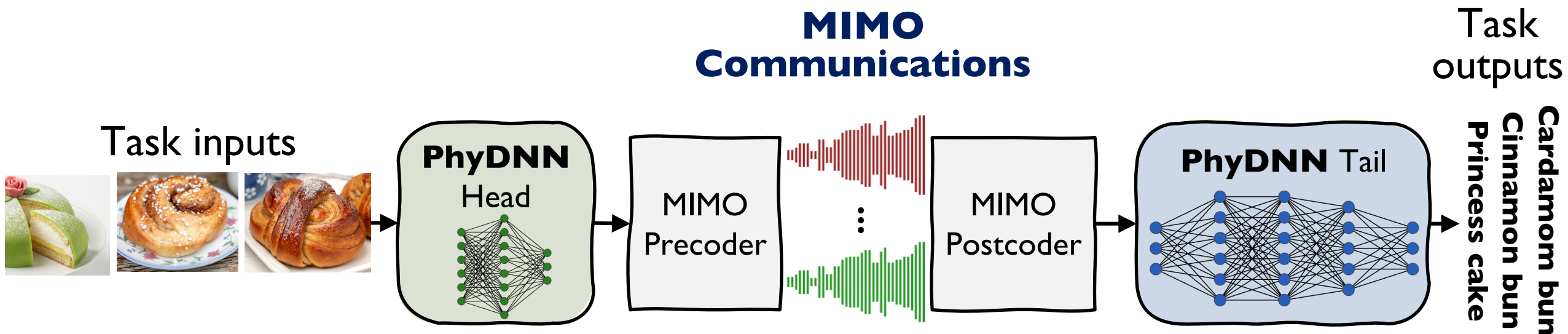
Cardamom bun

Cinnamon bun

Princess cake



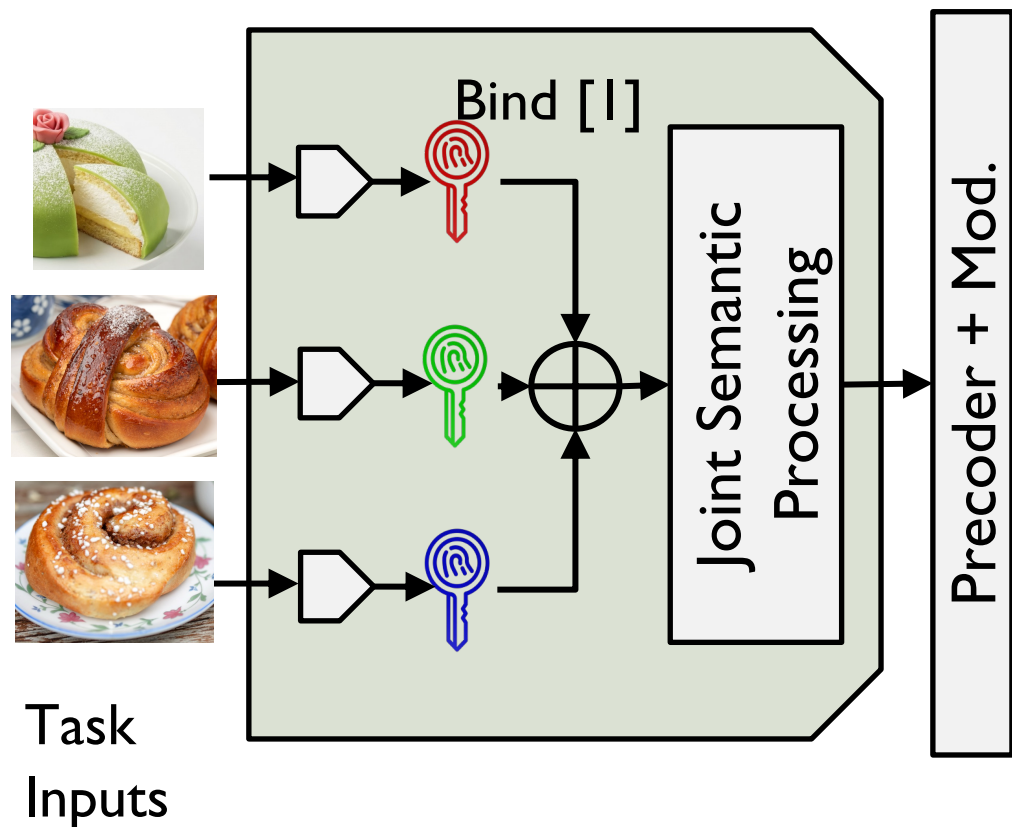
Main Idea



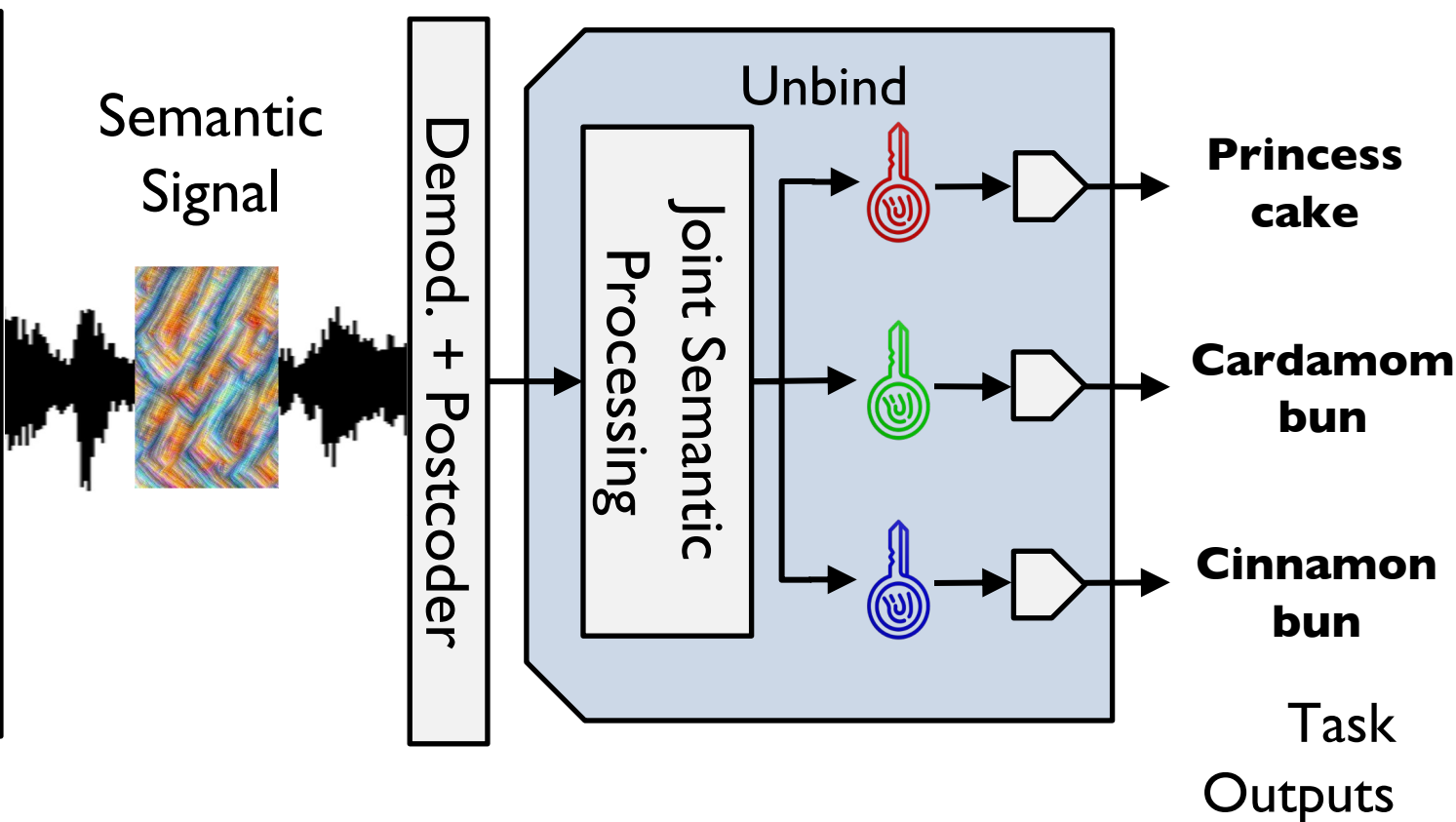
Semantic Multiplexing

Objective: Transmit **multiple semantic streams** independently of the number of antennas

Semantic Multiplexing Transmitter



Semantic Multiplexing Receiver



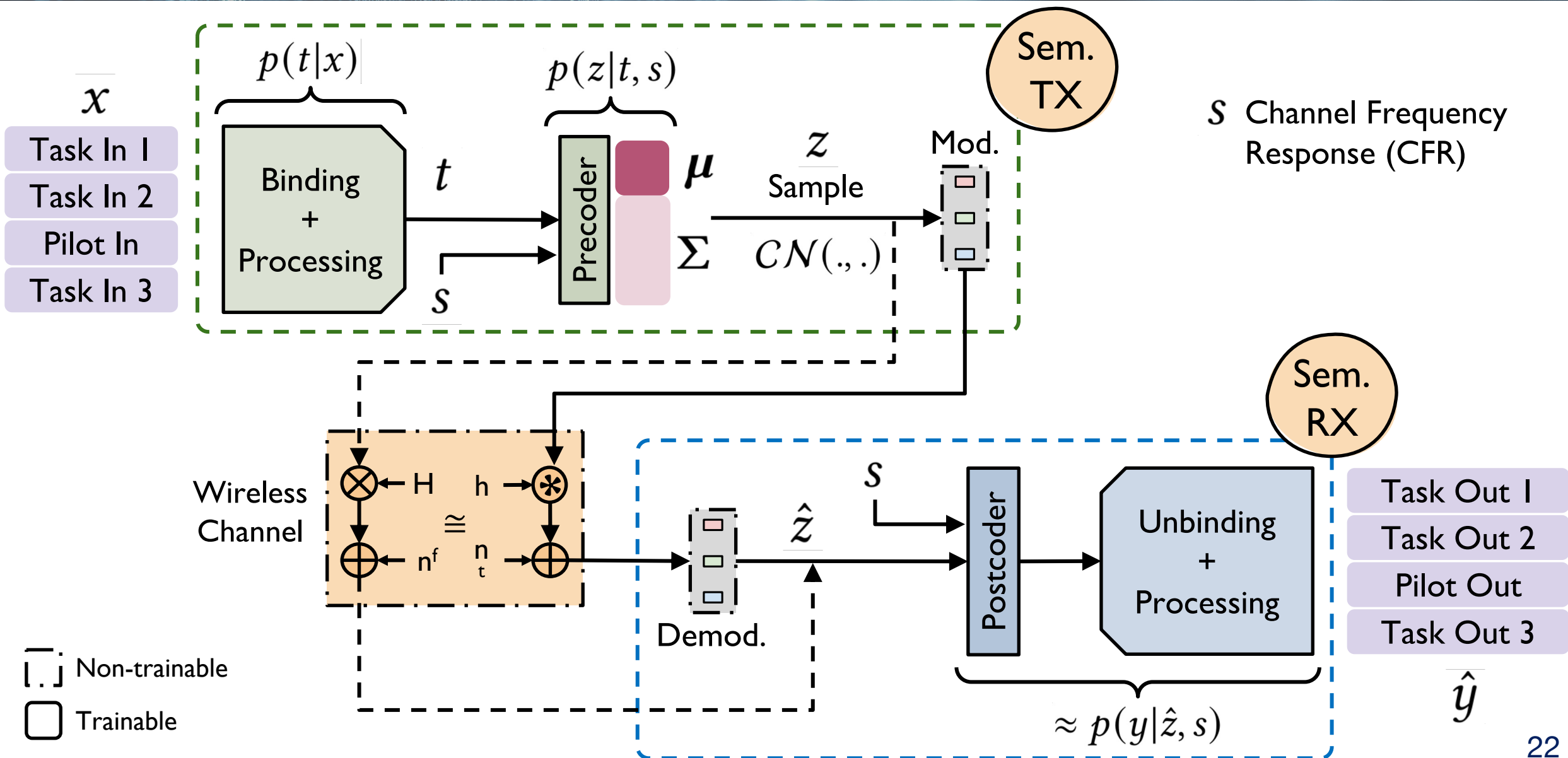
Binding key



Disjoint processing

[1] Tony A Plate, "Holographic Reduced Representations," IEEE Transactions on Neural networks, 1995.

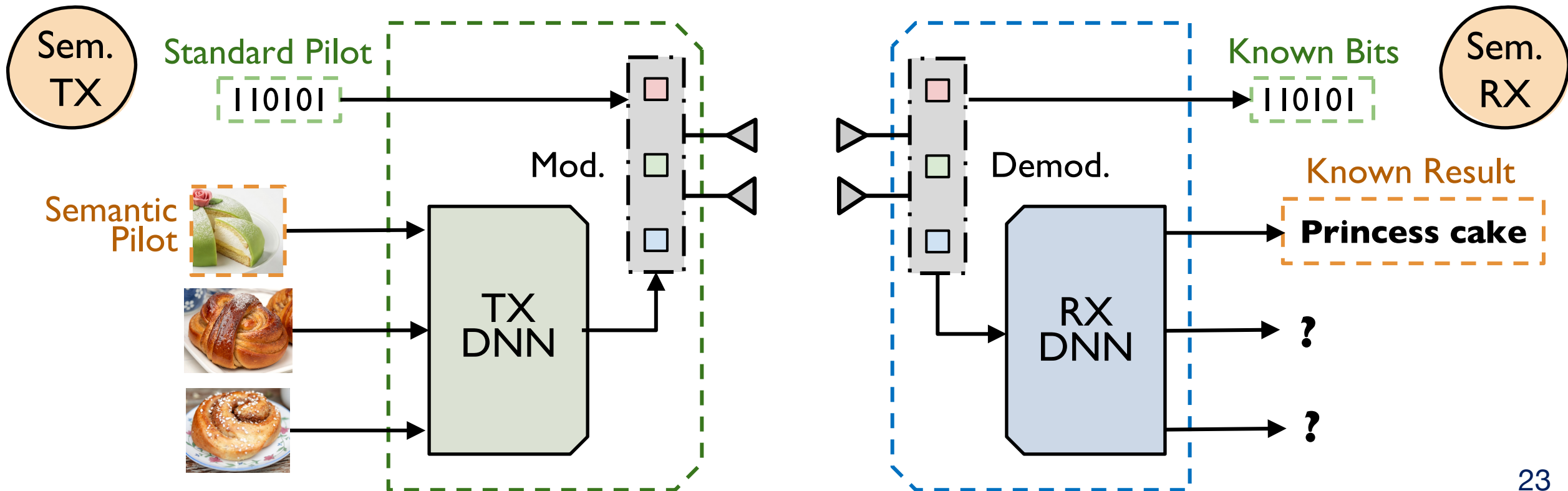
Probabilistic System Model



Dynamic Adaptation of Semantic Multiplexing

Two types of pilots:

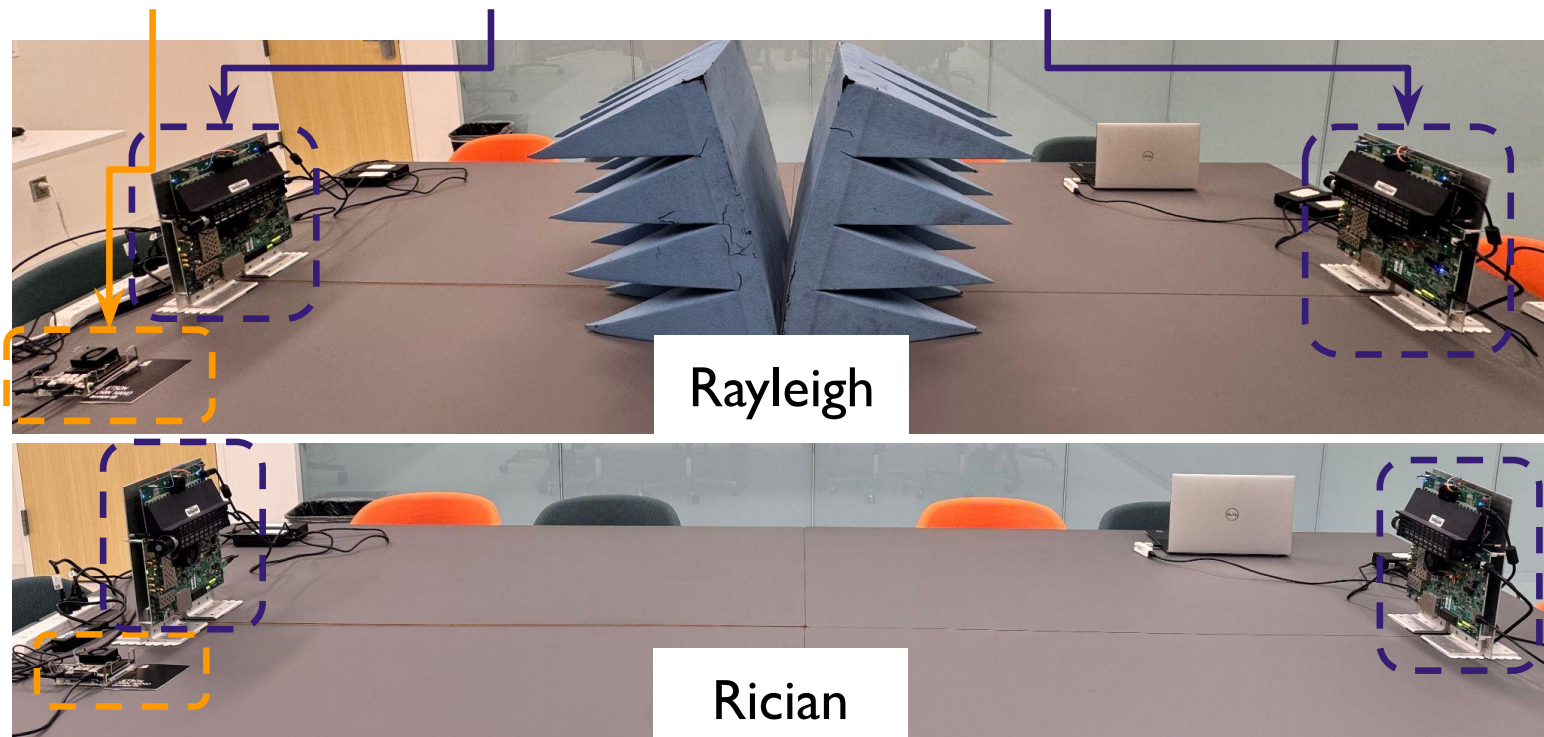
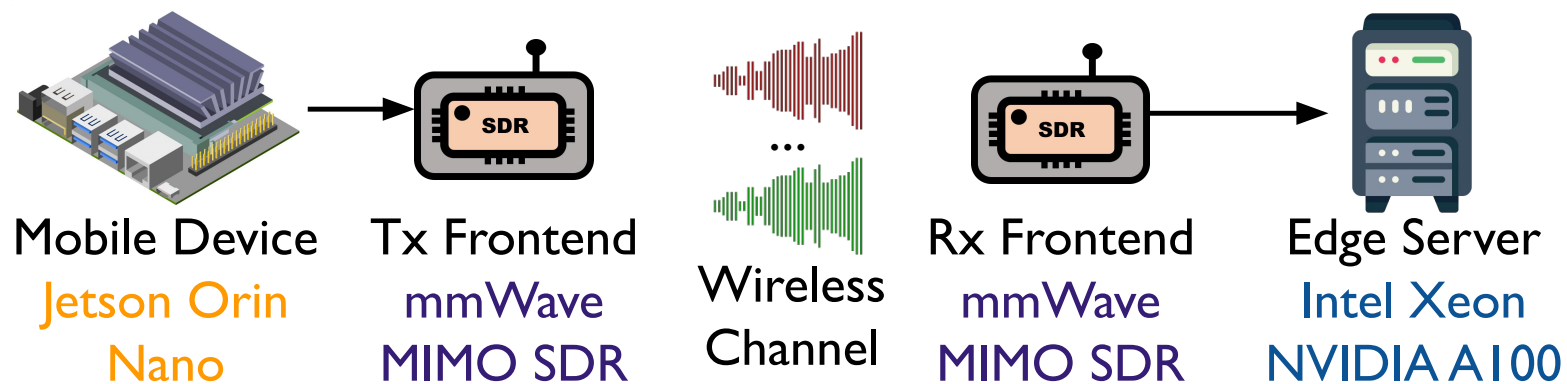
- **Communication Pilots** – as in standard transmissions, used to estimate the channel response
- **Task-Oriented Pilots** – transmitted in one of the computing channels and used to adapt the SSM TX and RX blocks



Experimental Evaluation

Testbed:

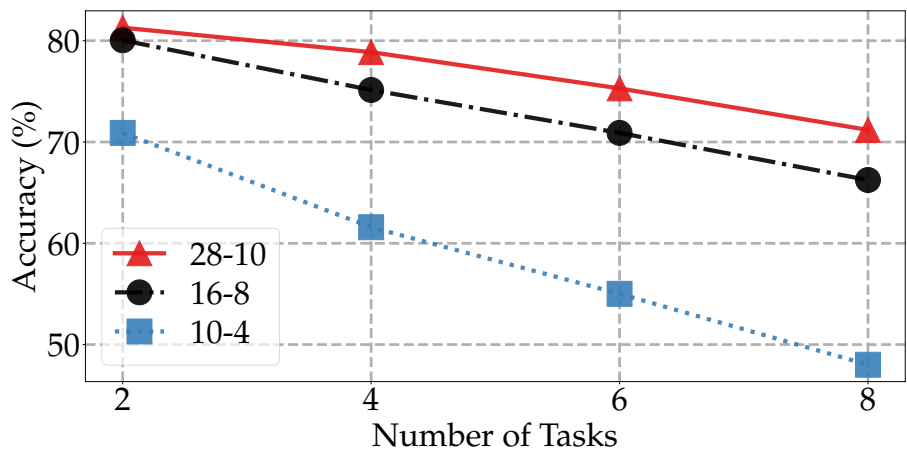
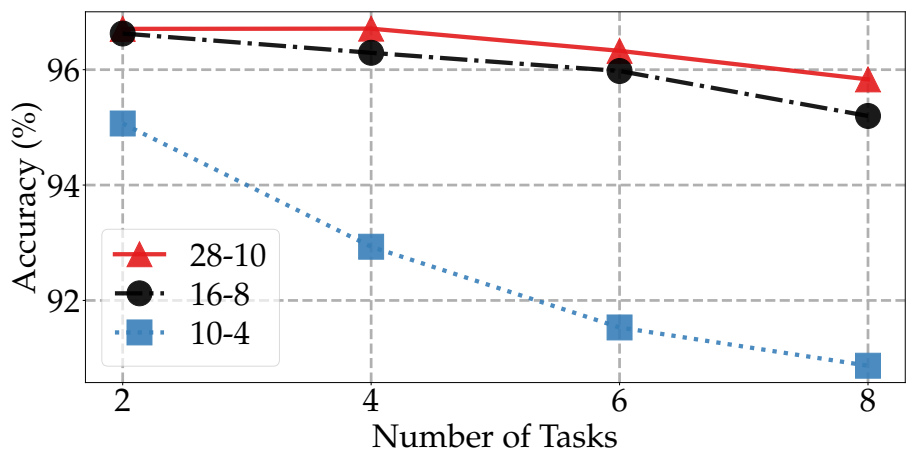
- Fully-digital SDR built from two Zynq UltraScale+ System-on-Chip with Pi-Radio front-ends
- 57–64 GHz band (mmWave)
- 8×8 multiplexing
- 1 GHz bandwidth



Scalability Analysis

Image Classification, 4 comms channels

- WideResNet DNNs (WRN): 28-10, 16-8, 10-4 [1]
- CIFAR-10, CIFAR-100



Results:

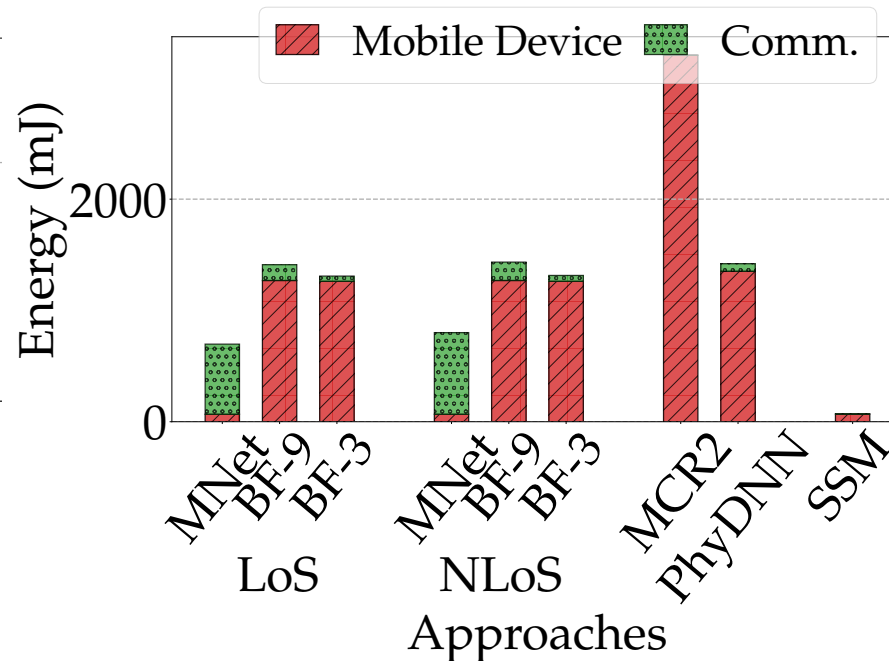
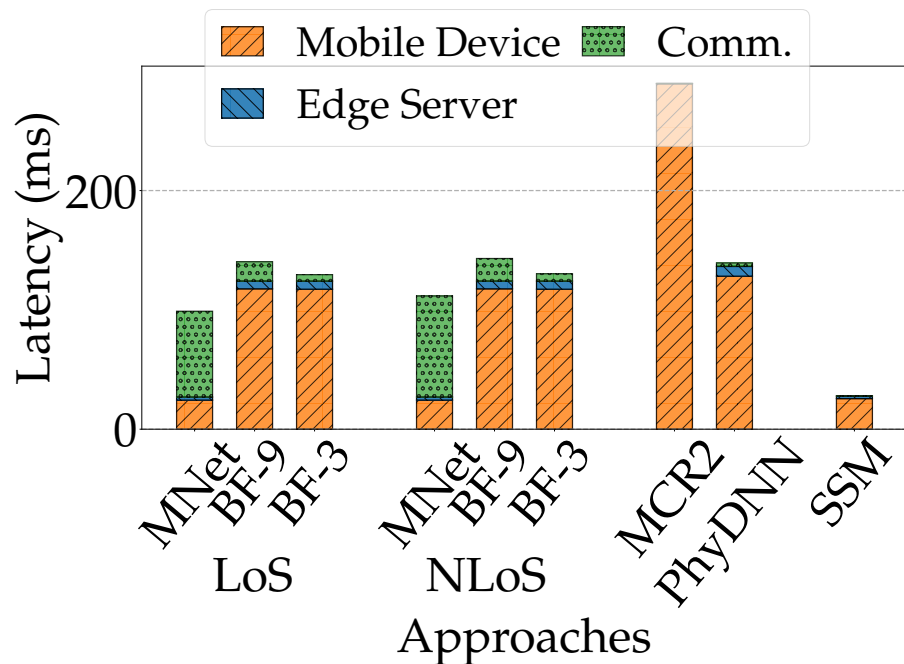
- **WRN-28-10 CIFAR-10**: 93.09% 8-task accuracy, only 4% less than 2-task accuracy while reducing latency by 4×

CIFAR-10		MNet	BF-9	BF-3	MCR2	PhyDNN	SM
WRN-28-10	LoS				89.87 %	92.50 %	94.48 %
	NLoS	94.74 %	96.88 %	96.42 %	86.76 %	88.23 %	90.92 %
WRN-16-8	LoS				89.17 %	91.85 %	93.69 %
	NLoS	93.86 %	95.96 %	95.44 %	84.97 %	87.05 %	89.46 %
WRN-10-4	LoS				85.36 %	85.25 %	87.29 %
	NLoS	87.71 %	91.38 %	90.13 %	83.40 %	83.34 %	85.76 %

CIFAR-100		MNet	BF-9	BF-3	MCR2	PhyDNN	SM
WRN-28-10	LoS				74.68 %	75.75 %	78.87 %
	NLoS	79.20 %	81.32 %	80.63 %	71.46 %	72.52 %	76.62 %
WRN-16-8	LoS				71.87 %	72.74 %	75.13 %
	NLoS	76.84 %	77.58 %	77.12 %	68.14 %	69.30 %	72.31 %
WRN-10-4	LoS				60.55 %	61.43 %	63.58 %
	NLoS	66.23 %	69.08 %	67.38 %	58.42 %	59.88 %	61.14 %

[1] Sergey Zagoruyko and Nikos Komodakis, “Wide Residual Networks,” in British Machine Vision Conference 2016.

Latency and Energy Consumption



Setup:

- Image classification, WRN-28-10
- 8 communication channels and 4 multiplexed tasks

Baselines:

- **MNet/ms**: MIMO-DNNs [1] + split computing [2]
- **BF/ss**: [3]
- **MCR2/sm**: precoding/postcoding similar to MIMO comms [4]
- **PhyDNN/ss**

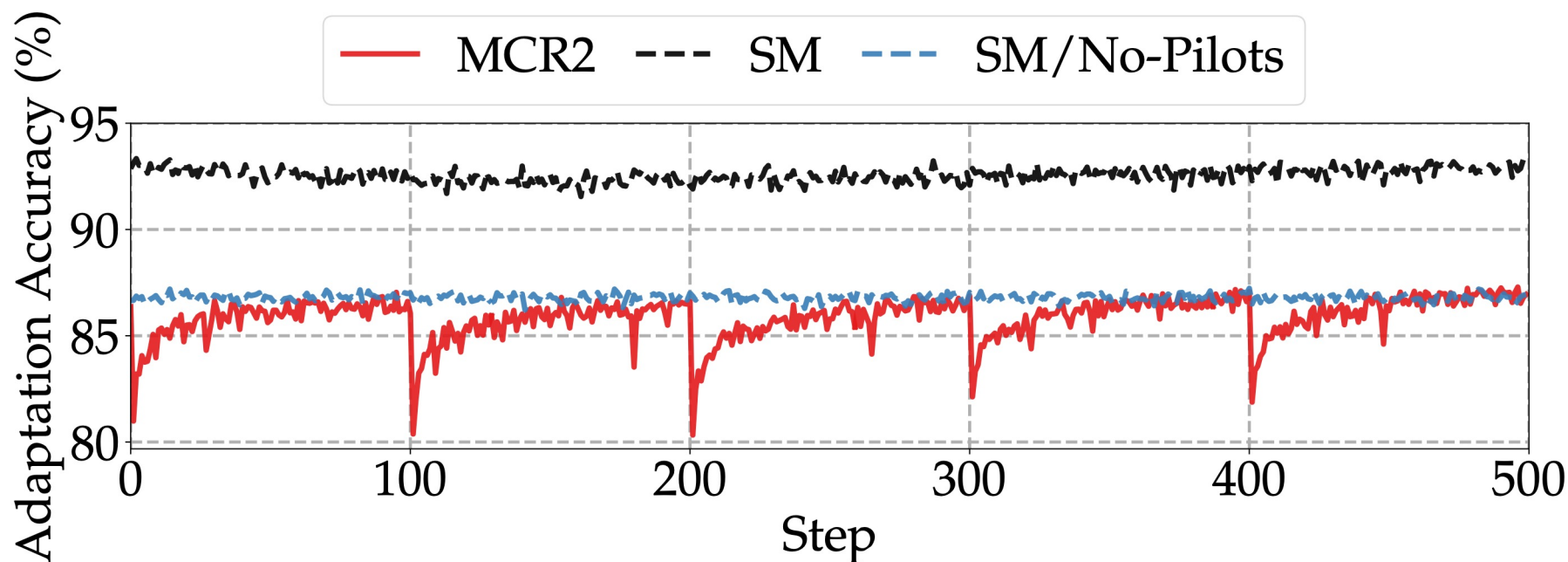
Results:

- Reduces **latency** by $\sim 4\times$ over MNet, BF-9/ss, BF-3/ss, PhyDNN, and $8\times$ over MCR2 (average across models)
- Reduces **computation energy** by $\sim 34\times$ compared to task-oriented SemCom approaches with WRN-28-10, and reduces **comms energy consumption** by $\sim 187\times$ over full-stack methods on average with WRN-10-4

[1] Yiping Kang, et al., "Neurosurgeon: Collaborative Intelligence Between the Cloud and Mobile Edge," ACM SIGARCH Computer Architecture News 45, 2017.
[2] Nicolas Menet, et al., "MIMONets: Multiple-Input-Multiple-Output Neural Networks Exploiting Computation in Superposition". Advances in Neural Inf. Proc. Systems," 2023.
[3] Y. Matsubara, et al., "BottleFit: Learning Compressed Representations in Deep Neural Networks for Effective and Efficient Split Computing," in Proc. of IEEE WoWMoM, 2022.
[4] Chang Cai, et al., "Multi-Device Task-Oriented Communication via Maximal Coding Rate Reduction," IEEE Transactions on Wireless Communications, 2024.
[5] Chang Cai, et al., "End-to-End Learning for Task-Oriented Semantic Communications Over MIMO Channels: An Information-Theoretic Framework," IEEE Journal on Selected Areas in Communications, 2025.

Inference-time Adaptation

- Dynamic channel: positions of the transmitter and receiver changed every 100 inference steps
- Comparison with the SOTA MCR2 approach



3

Ongoing and Future Work



Semantic communications

3 key research questions

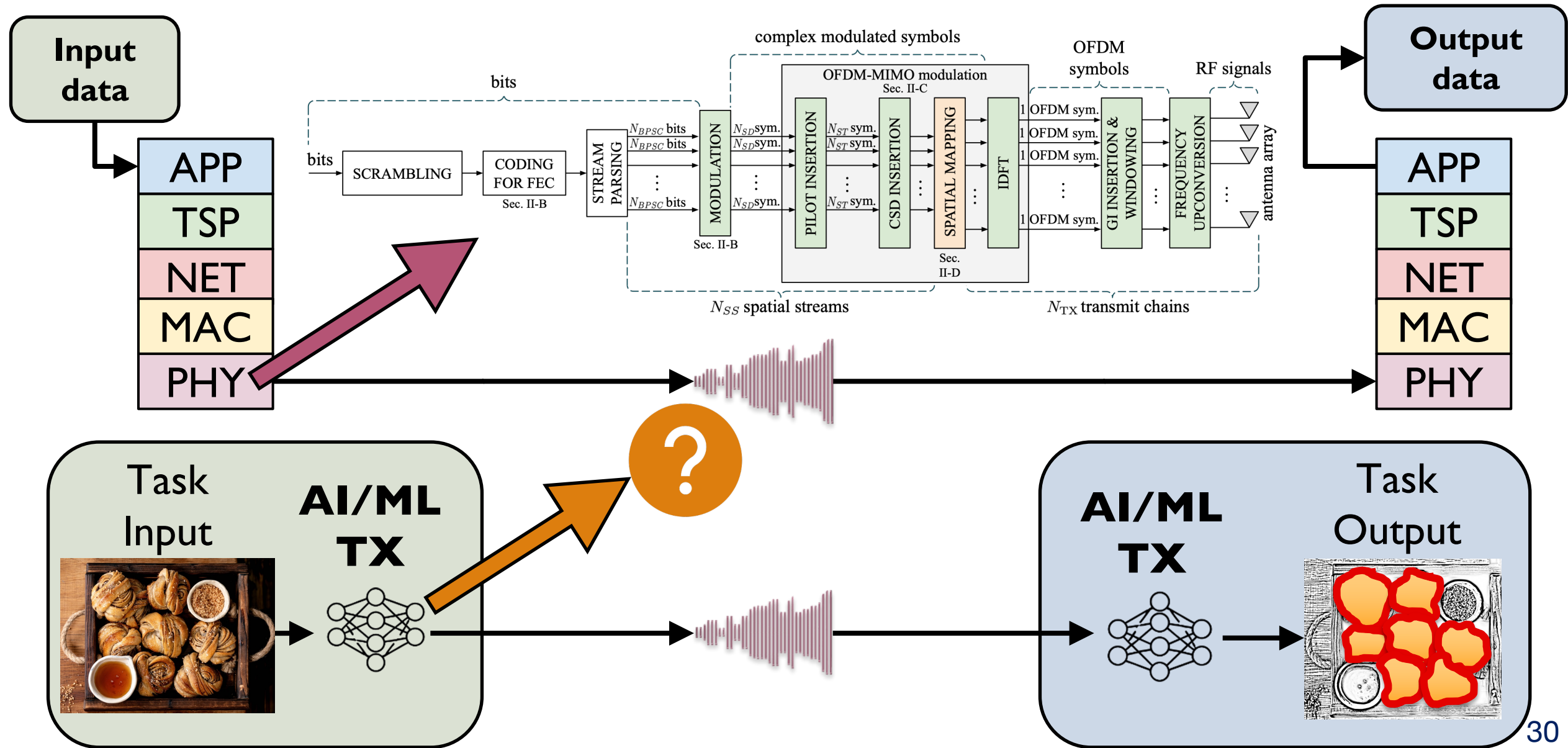


1) How to reduce latency?

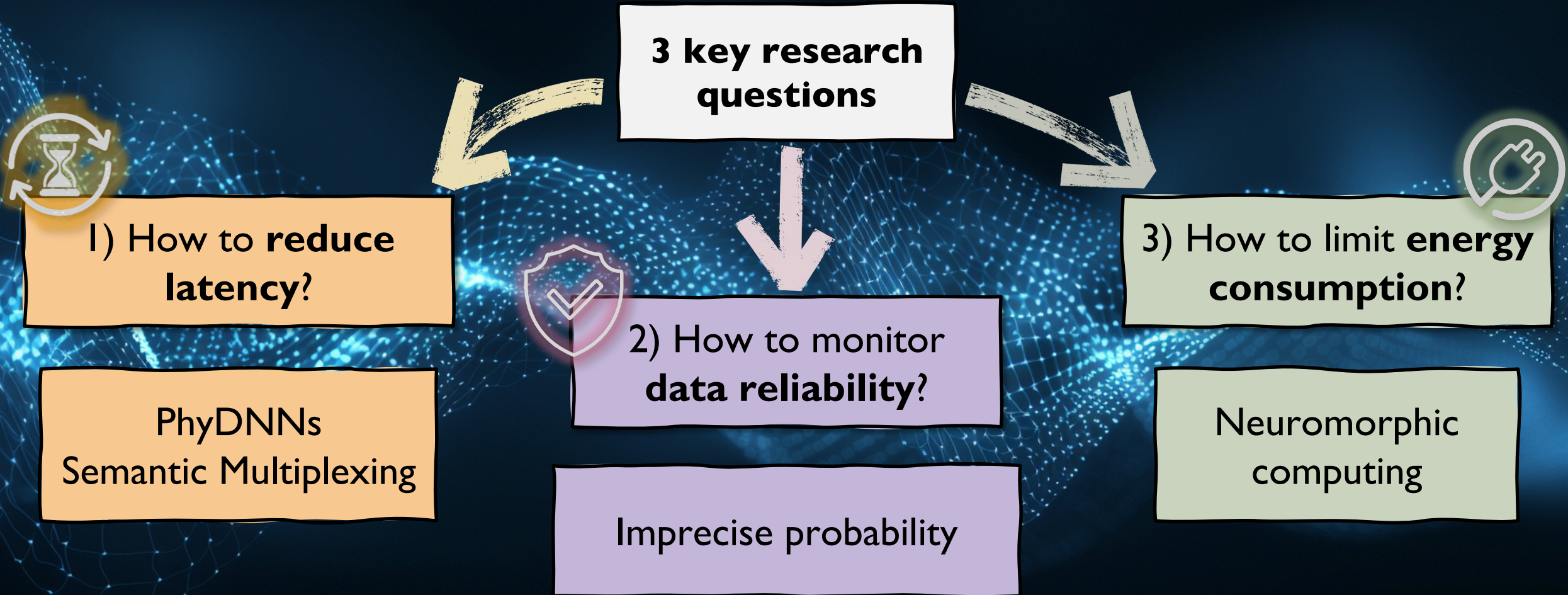


2) How to monitor data reliability?

Traditional VS Semantic Communications

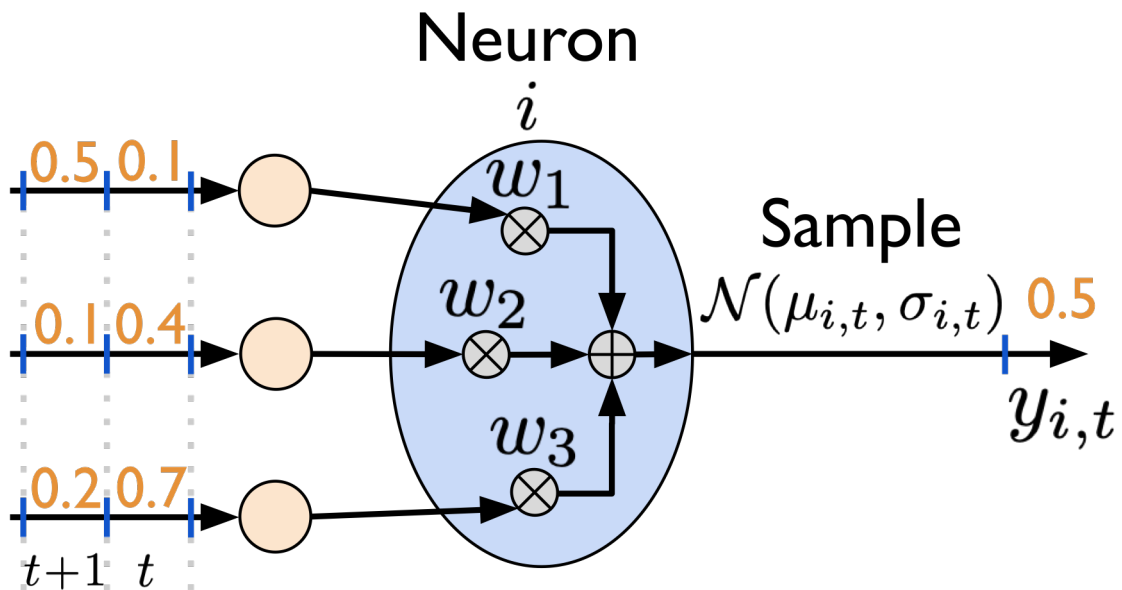


Semantic communications

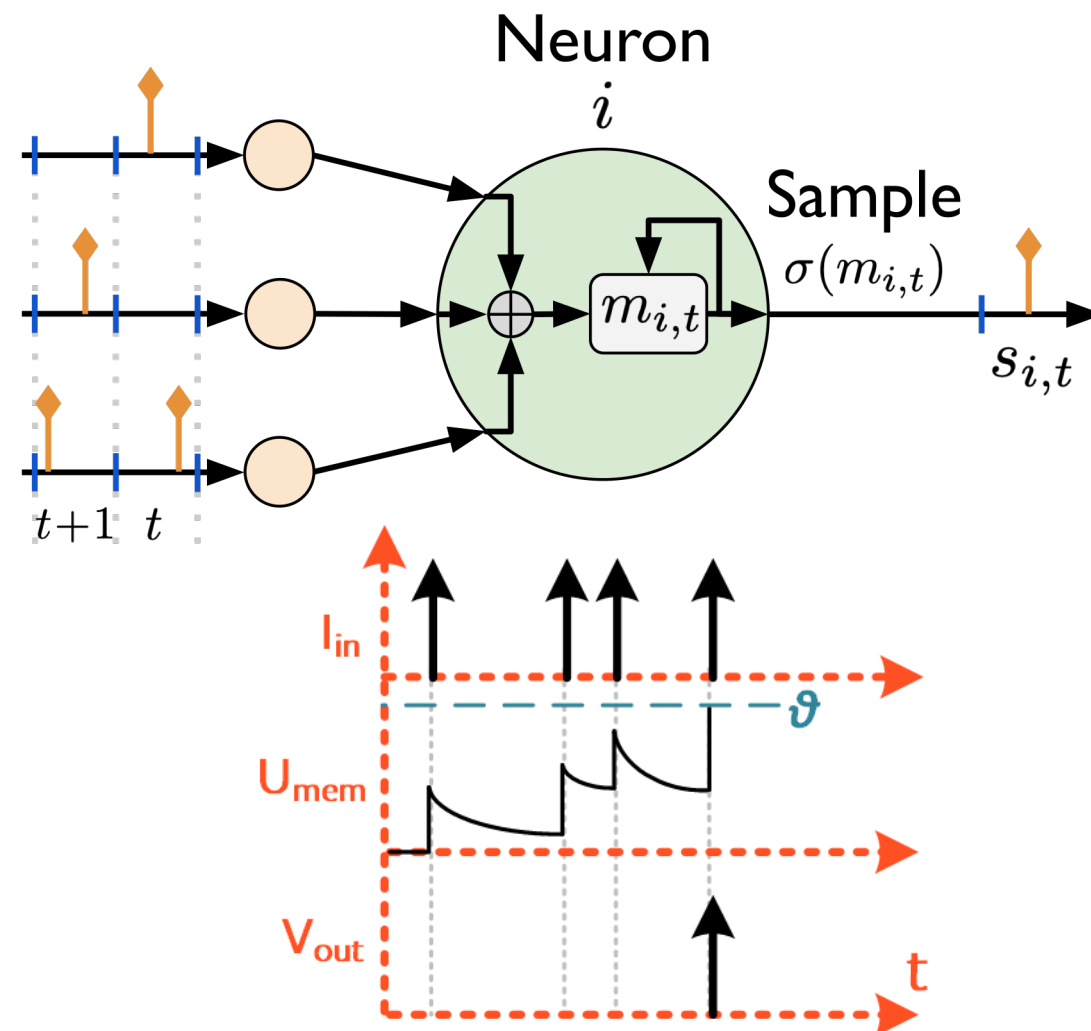


Neuromorphic Computing

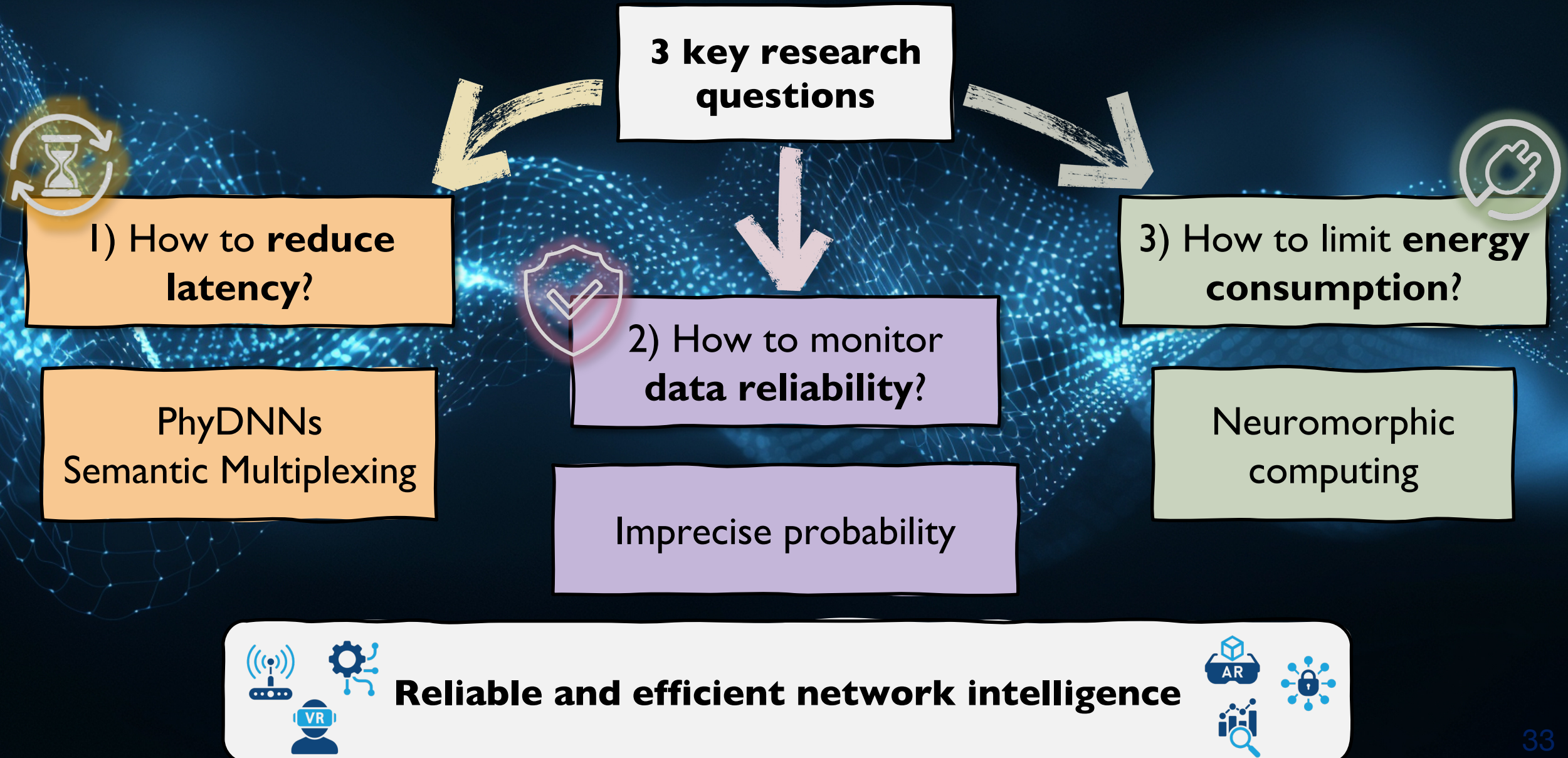
Artificial Neuron



Spiking Neuron



Semantic communications





Northeastern University
Institute for Intelligent
Networked Systems



Thank you!

Francesca Meneghello

Principal Research Scientist

Northeastern University, Boston (MA), United States

E-mail: fr.meneghello@northeastern.edu

Website: <https://francescamen.github.io/>

Research Group Website: <http://mentis.info>

