When Learning Meets the Channel: Edge AI through Split and Semantic Design
Part II

Vukan Ninković

October 28, 2025

Overview

1. Introduction & Motivation

2. Background & System Model

3. Vanilla Approach – Does It Work?

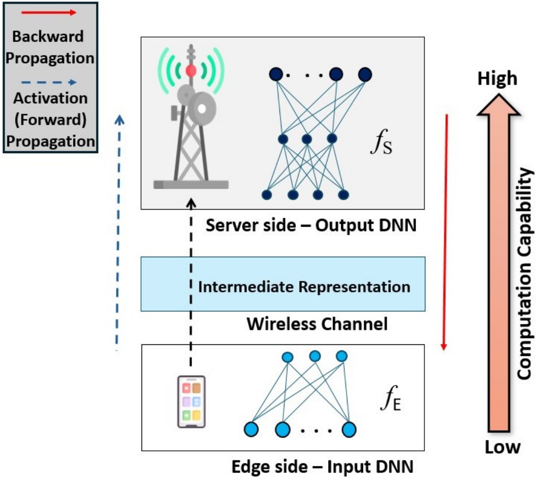
4. Extension 1 – Are All Symbols Equally Important?

5. Extension 2 – Semantic & AE-PHY
in UAV-Assisted IoT

6. Discussion

Introduction & Motivation

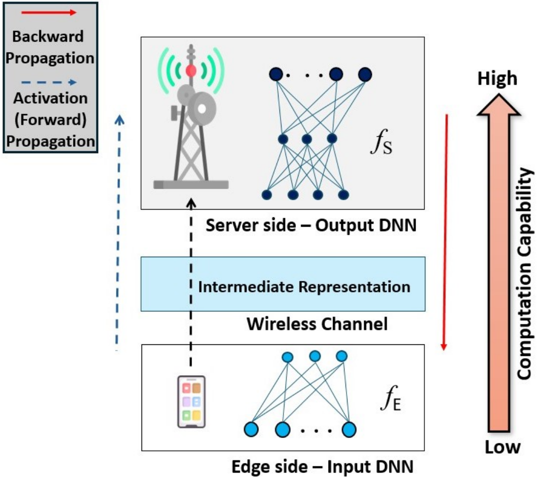
(Quick) Recap



Backward Propagation
Activation (Forward) Propagation
f_S
Server side – Output DNN
Intermediate Representation
Wireless Channel
f_E
Edge side – Input DNN
High
Computation Capability
Low

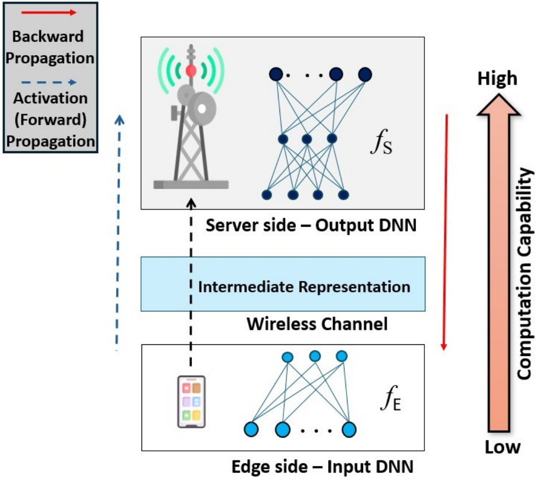
(Quick) Recap

- **Focus:** Communications for AI-driven systems.



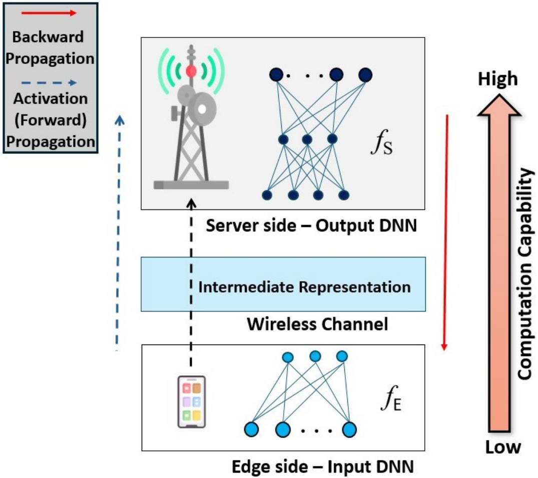
(Quick) Recap

- **Focus:** Communications for AI-driven systems.
- **Key concepts:**
  - Channel-aware training of neural networks.
  - Designing robust latent representation.
  - Promising performances in controlled settings.



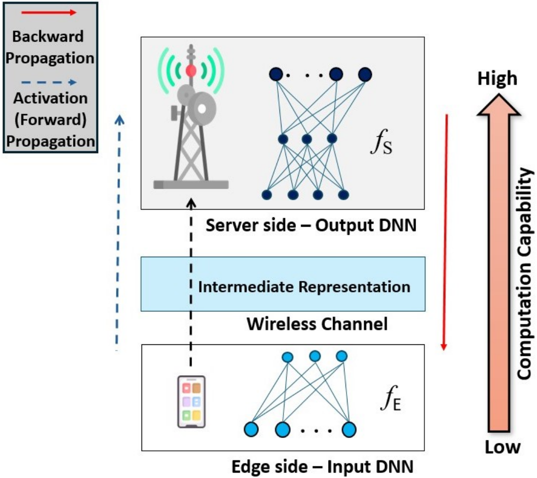
(Quick) Recap

- **Focus:** Communications for AI-driven systems.
- **Key concepts:**
  - Channel-aware training of neural networks.
  - Designing robust latent representation.
  - Promising performances in controlled settings.
- **Goal:** Toward real-world implementation.



(Quick) Recap

- **Focus:** Communications for AI-driven systems.
- **Key concepts:**
  - Channel-aware training of neural networks.
  - Designing robust latent representation.
  - Promising performances in controlled settings.
- **Goal:** Toward real-world implementation.
- **Challenges:**
  - Channel-coded latent transmission?
  - Max-info latent under bandwidth constraints?



Introduction

- Why are these challenges crucial?

Introduction

- Why are these challenges crucial?
- Beyond-5G/6G must support intelligent IoT devices with with strict constraints:
  - Latency.
  - Energy consumption.
  - Bandwidth usage.

Introduction

- Why are these challenges crucial?
- Beyond-5G/6G must support intelligent IoT devices with with strict constraints:
  - Latency.
  - Energy consumption.
  - Bandwidth usage.
- Shift from accurate signal reconstruction to **task-oriented communication**.

Introduction

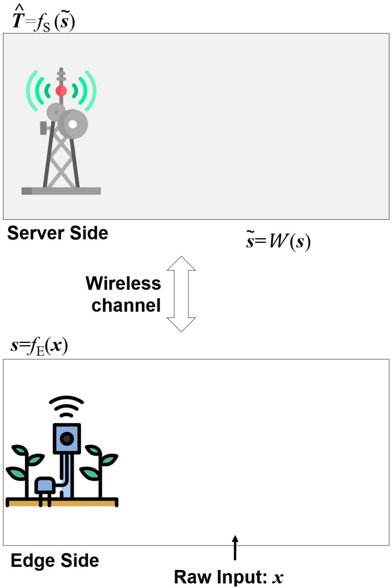
- Why are these challenges crucial?
- Beyond-5G/6G must support intelligent IoT devices with with strict constraints:
  - Latency.
  - Energy consumption.
  - Bandwidth usage.
- Shift from accurate signal reconstruction to **task-oriented communication**.
- **Semantic communication:** Send only information necessary for the task.

Introduction

- Why are these challenges crucial?
- Beyond-5G/6G must support intelligent IoT devices with with strict constraints:
  - Latency.
  - Energy consumption.
  - Bandwidth usage.
- Shift from accurate signal reconstruction to **task-oriented communication**.
- **Semantic communication:** Send only information necessary for the task.
- **Solution:** Synergy of SL, AE–based PHY layer design and semantic communication:
  - Semantic- and channel-aware learning architecture for IoT edge-to-cloud forecasting.

Background & System Model

System Overview

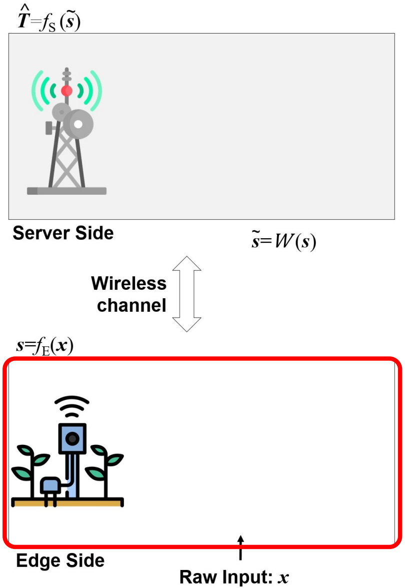


System Overview

$$\boldsymbol{s} = f_{\mathrm{E}}(\boldsymbol{x})$$

$$f_{\mathrm{E}} : \mathbb{R}^N \to \mathbb{R}^n, \quad n < N$$

$$\boldsymbol{x} \in \mathbb{R}^N$$



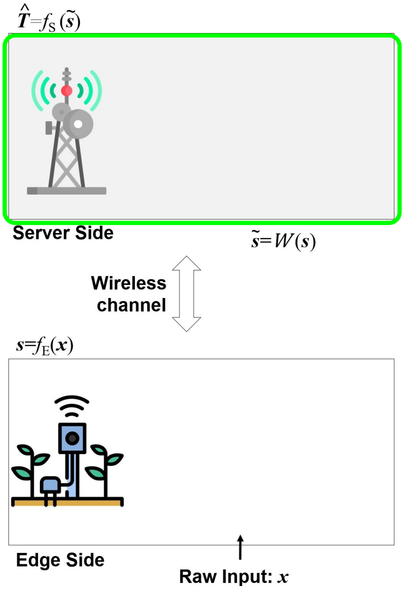
System Overview

$$\hat{T} = f_{\mathrm{S}}(\tilde{\boldsymbol{s}})$$

$$f_{\mathrm{S}} = \mathbb{R}^n \to \mathbb{R}^*$$
$$\hat{\boldsymbol{s}} = \mathcal{W}(\boldsymbol{s})$$

$$\boldsymbol{s} = f_{\mathrm{E}}(\boldsymbol{x})$$

$$f_{\mathrm{E}} : \mathbb{R}^N \to \mathbb{R}^n, \quad n < N$$
$$\boldsymbol{x} \in \mathbb{R}^N$$



System Overview

$$\hat{T} = f_{\mathrm{S}}(\tilde{\boldsymbol{s}})$$

$$f_{\mathrm{S}} = \mathbb{R}^n \to \mathbb{R}^*$$

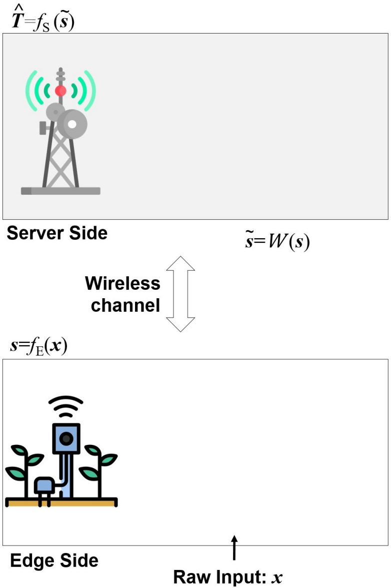
$$\hat{\boldsymbol{s}} = \mathcal{W}(\boldsymbol{s})$$

Compact and informative representation?

$$\boldsymbol{s} = f_{\mathrm{E}}(\boldsymbol{x})$$

$$f_{\mathrm{E}} : \mathbb{R}^N \to \mathbb{R}^n, \quad n < N$$

$$\boldsymbol{x} \in \mathbb{R}^N$$



System Overview - Semantic Part

$$\hat{T} = f_S(\tilde{\boldsymbol{s}})$$

$$\hat{T} = g_{\text{sem}}(\hat{\boldsymbol{z}}), \quad g_{\text{sem}} : \mathbb{R}^K \to \mathbb{R}^*$$

$$f_S = \mathbb{R}^n \to \mathbb{R}^*$$

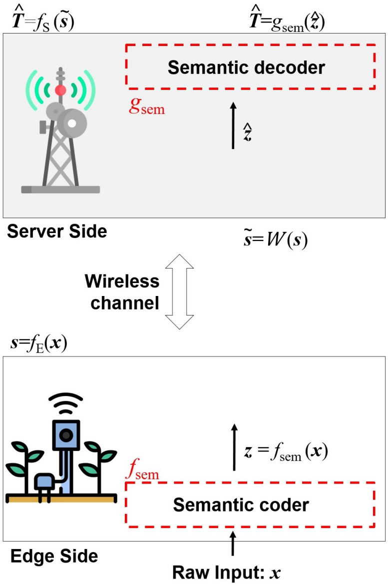
$$\hat{\boldsymbol{s}} = \mathcal{W}(\boldsymbol{s})$$

$$\boldsymbol{s} = f_E(\boldsymbol{x})$$

$$\boldsymbol{z} = f_{\text{sem}}(\boldsymbol{x}), \quad f_{\text{sem}} : \mathbb{R}^N \to \mathbb{R}^K$$

$$f_E : \mathbb{R}^N \to \mathbb{R}^n, \quad n < N$$

$$\boldsymbol{x} \in \mathbb{R}^N$$



Semantic Communication

- Redefines objectives of communication systems:
  - **Focus:** From signal–level accuracy to the successful completion of different tasks.
  - Instead to recover the input ($\boldsymbol{x}$), task-relevant output is approximated ($\hat{T}(\boldsymbol{x})$).
- Transmission of only task-essential information:
  - Improves spectral and computational efficiency.
  - Enhances robustness to channel impairments.
  - Strengthens data privacy.

Semantic Communication

- Redefines objectives of communication systems:
  - **Focus:** From signal–level accuracy to the successful completion of different tasks.
  - Instead to recover the input ($\boldsymbol{x}$), task-relevant output is approximated ($\hat{T}(\boldsymbol{x})$).
- Transmission of only task-essential information:
  - Improves spectral and computational efficiency.
  - Enhances robustness to channel impairments.
  - Strengthens data privacy.
- **Mathematical and theoretical foundation** - Information Bottleneck (IB) principle:

$$\min I(\boldsymbol{x}; \boldsymbol{z}) \quad \text{subject to} \quad I(\boldsymbol{z}; T(\boldsymbol{x})) \geq \epsilon, \tag{1}$$

Semantic Communication

- Redefines objectives of communication systems:
  - **Focus:** From signal–level accuracy to the successful completion of different tasks.
  - Instead to recover the input ($\boldsymbol{x}$), task-relevant output is approximated ($\hat{T}(\boldsymbol{x})$).
- Transmission of only task-essential information:
  - Improves spectral and computational efficiency.
  - Enhances robustness to channel impairments.
  - Strengthens data privacy.
- **Mathematical and theoretical foundation** - Information Bottleneck (IB) principle:

$$\min I(\boldsymbol{x}; \boldsymbol{z}) \quad \text{subject to} \quad I(\boldsymbol{z}; T(\boldsymbol{x})) \geq \epsilon, \tag{1}$$

- $I(\cdot, \cdot)$ - Mutual Information (MI).

Semantic Communication

- Redefines objectives of communication systems:
  - **Focus:** From signal–level accuracy to the successful completion of different tasks.
  - Instead to recover the input ($\boldsymbol{x}$), task-relevant output is approximated ($\hat{T}(\boldsymbol{x})$).
- Transmission of only task-essential information:
  - Improves spectral and computational efficiency.
  - Enhances robustness to channel impairments.
  - Strengthens data privacy.
- **Mathematical and theoretical foundation** - Information Bottleneck (IB) principle:

$$\min I(\boldsymbol{x}; \boldsymbol{z}) \quad \text{subject to} \quad I(\boldsymbol{z}; T(\boldsymbol{x})) \geq \epsilon, \tag{1}$$

- $I(\cdot, \cdot)$ - Mutual Information (MI).
- **IB framework** - Extract compact $\boldsymbol{z}$ that preserves maximal information for predicting $T(\boldsymbol{x})$.

System Overview - Semantic Part

$$\hat{T} = f_S(\tilde{\boldsymbol{s}})$$

$$\hat{T} = g_{\text{sem}}(\hat{\boldsymbol{z}}), \quad g_{\text{sem}} : \mathbb{R}^K \rightarrow \mathbb{R}^*$$

$$f_S = \mathbb{R}^n \rightarrow \mathbb{R}^*$$

$$\hat{\boldsymbol{s}} = \mathcal{W}(\boldsymbol{s})$$

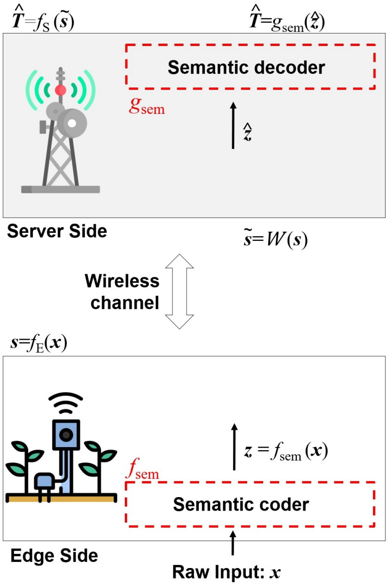
Wireless channel influence?

$$\boldsymbol{s} = f_E(\boldsymbol{x})$$

$$\boldsymbol{z} = f_{\text{sem}}(\boldsymbol{x}), f_{\text{sem}} : \mathbb{R}^N \rightarrow \mathbb{R}^K$$

$$f_E : \mathbb{R}^N \rightarrow \mathbb{R}^n, \quad n < N$$

$$\boldsymbol{x} \in \mathbb{R}^N$$



System Overview - Channel Part

$$\hat{T} = f_{\mathrm{S}}(\tilde{\boldsymbol{s}})$$

$$\hat{T} = g_{\mathrm{sem}}(\hat{\boldsymbol{z}}), \quad g_{\mathrm{sem}} : \mathbb{R}^K \to \mathbb{R}^*$$

$$\hat{\boldsymbol{z}} = g_{\mathrm{ch}}(\tilde{\boldsymbol{s}}), \quad g_{\mathrm{ch}} : \mathbb{R}^n \to \mathbb{R}^K$$

$$f_{\mathrm{S}} = \mathbb{R}^n \to \mathbb{R}^*$$

$$\hat{\boldsymbol{s}} = \mathcal{W}(\boldsymbol{s})$$

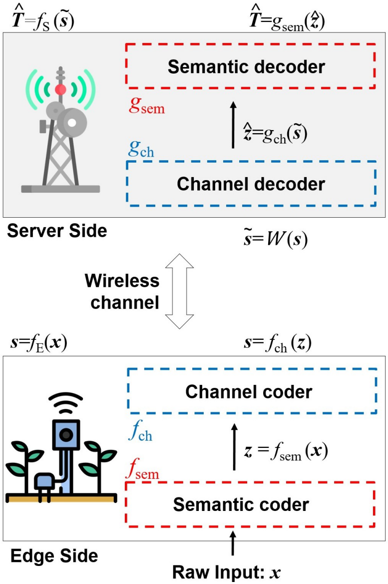
$$\boldsymbol{s} = f_{\mathrm{E}}(\boldsymbol{x})$$

$$\boldsymbol{s} = f_{\mathrm{ch}}(\boldsymbol{z}), \quad f_{\mathrm{ch}} : \mathbb{R}^K \to \mathbb{R}^n$$

$$\boldsymbol{z} = f_{\mathrm{sem}}(\boldsymbol{x}), \quad f_{\mathrm{sem}} : \mathbb{R}^N \to \mathbb{R}^K$$

$$f_{\mathrm{E}} : \mathbb{R}^N \to \mathbb{R}^n, \quad n < N$$

$$\boldsymbol{x} \in \mathbb{R}^N$$



AE-Based PHY Modeling

- Flexible framework for designing the PHY layer as a differentiable neural network:
  - Separate modulation and coding blocks $\rightarrow$ Single end-to- end trainable system[1].
- f_{ch} and g_{ch} realized as symmetric AE neural network:
  - Semantic representation $\boldsymbol{z} \in \mathbb{R}^K \rightarrow$ Transmitted signal $\boldsymbol{s} \in \mathbb{R}^n$

[1] T. O'Shea and J. Hoydis, "An introduction to deep learning for the physical layer," *IEEE Trans. Cogn. Commun. Netw.,* vol. 3, no. 4, pp. 563-575, Dec. 2017.

AE-Based PHY Modeling

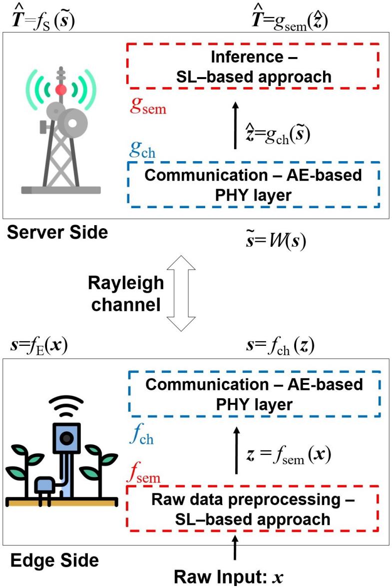
- Flexible framework for designing the PHY layer as a differentiable neural network:
  - Separate modulation and coding blocks $\rightarrow$ Single end-to- end trainable system[1].
- f_{ch} and g_{ch} realized as symmetric AE neural network:
  - Semantic representation $\boldsymbol{z} \in \mathbb{R}^K \rightarrow$ Transmitted signal $\boldsymbol{s} \in \mathbb{R}^n$
- Redefinition of the original training procedure - MSE loss:

$$\mathcal{L}_{\mathrm{AE}} = \sum_{k \in K} \|\hat{z}_k - z_k\|^2 \tag{2}$$

[1] T. O'Shea and J. Hoydis, "An introduction to deep learning for the physical layer," *IEEE Trans. Cogn. Commun. Netw.,* vol. 3, no. 4, pp. 563-575, Dec. 2017.

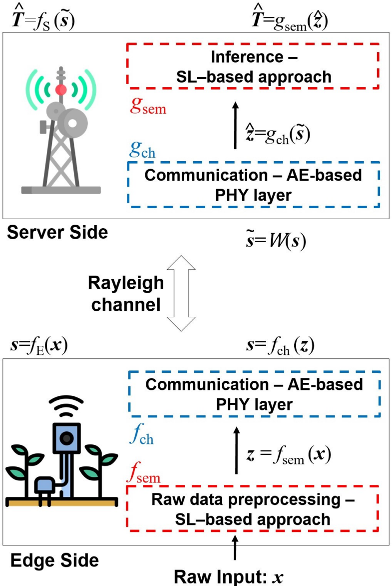
System Model

- Decomposition of f_E and f_S:
  - $f_E = f_{sem} \circ f_{ch}$
  - $f_S = g_{ch} \circ g_{sem}$
- SL and semantic communication connection(s):



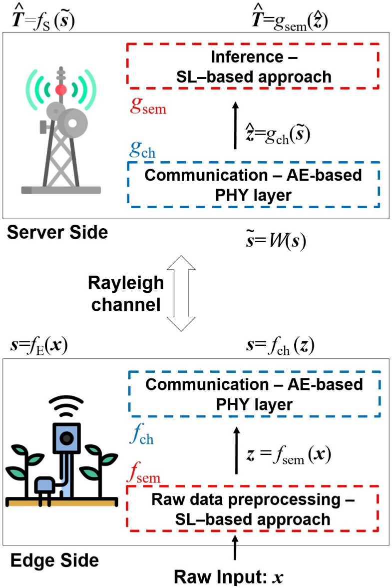
System Model

- Decomposition of f_{E} and f_{S}:
  - $f_{\mathrm{E}} = f_{\mathrm{sem}} \circ f_{\mathrm{ch}}$
  - $f_{\mathrm{S}} = g_{\mathrm{ch}} \circ g_{\mathrm{sem}}$
- SL and semantic communication connection(s):
  - **Meaningful compression:** $\boldsymbol{x} \in \mathbb{R}^N \rightarrow \boldsymbol{z} \in \mathbb{R}^K, K < N.$



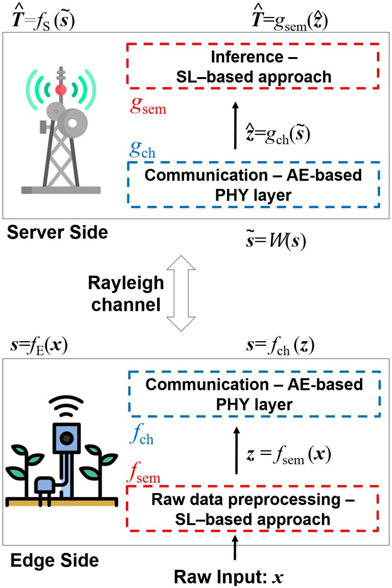
System Model

- Decomposition of f_{E} and f_{S}:
  - $f_{\mathrm{E}} = f_{\mathrm{sem}} \circ f_{\mathrm{ch}}$
  - $f_{\mathrm{S}} = g_{\mathrm{ch}} \circ g_{\mathrm{sem}}$
- SL and semantic communication connection(s):
  - **Meaningful compression:** $\boldsymbol{x} \in \mathbb{R}^N \to \boldsymbol{z} \in \mathbb{R}^K, K < N.$
  - **Task-oriented:** $\boldsymbol{z}$ encodes only task-relevant semantics.



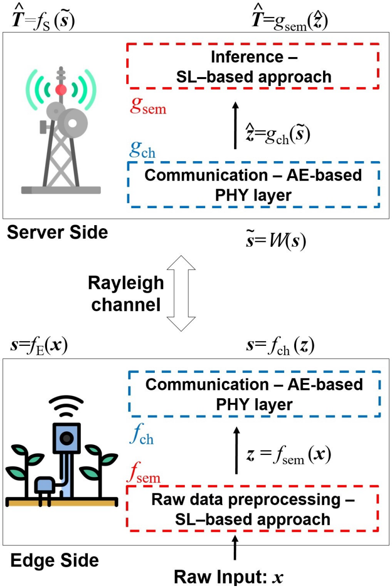
System Model

- Decomposition of f_E and f_S:
  - $f_E = f_{sem} \circ f_{ch}$
  - $f_S = g_{ch} \circ g_{sem}$
- SL and semantic communication connection(s):
  - **Meaningful compression:** $\boldsymbol{x} \in \mathbb{R}^N \rightarrow \boldsymbol{z} \in \mathbb{R}^K, K < N$.
  - **Task-oriented:** $\boldsymbol{z}$ encodes only task-relevant semantics.
  - **Decoder at server:** server-side network acts as the semantic decoder.



System Model

- Decomposition of f_E and f_S:
  - $f_E = f_{sem} \circ f_{ch}$
  - $f_S = g_{ch} \circ g_{sem}$
- SL and semantic communication connection(s):
  - **Meaningful compression:** $\boldsymbol{x} \in \mathbb{R}^N \rightarrow \boldsymbol{z} \in \mathbb{R}^K, K < N$.
  - **Task-oriented:** $\boldsymbol{z}$ encodes only task-relevant semantics.
  - **Decoder at server:** server-side network acts as the semantic decoder.
- AE–based PHY modeling included.

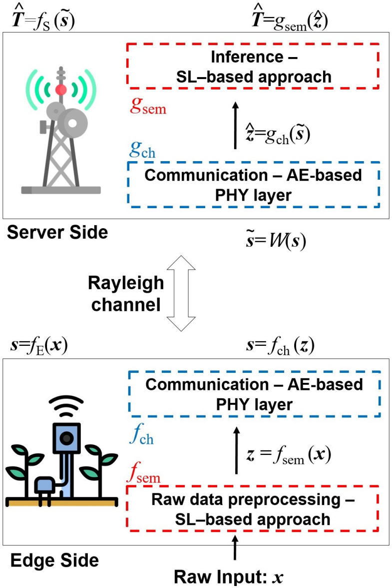


System Model

- Overall inference system:

$$\boldsymbol{x} \xrightarrow{f_{\text{sem}}} \boldsymbol{z} \xrightarrow{f_{\text{ch}}} \boldsymbol{s} \xrightarrow{\mathcal{W}(\cdot)} \tilde{\boldsymbol{s}} \xrightarrow{g_{\text{ch}}} \hat{\boldsymbol{z}} \xrightarrow{g_{\text{sem}}} \hat{T}. \tag{3}$$

- All components of the system - $f_{\text{sem}}, f_{\text{ch}}, g_{\text{ch}}, g_{\text{sem}}$:
  - Jointly optimized to minimize composite loss $\mathcal{L}(\hat{T}, T(\boldsymbol{x}))$[a].



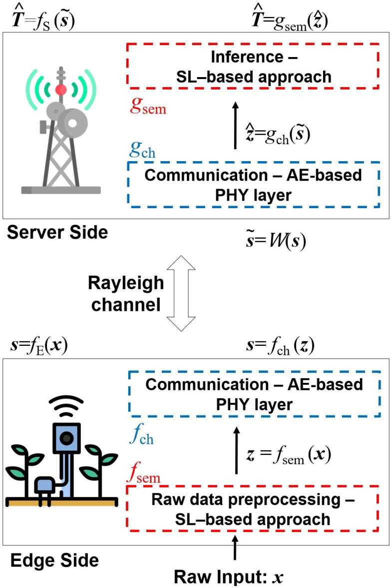
[a] The specific loss components depend on the considered scenario.

System Model

- Overall inference system:

$$\boldsymbol{x} \xrightarrow{f_{\text{sem}}} \boldsymbol{z} \xrightarrow{f_{\text{ch}}} \boldsymbol{s} \xrightarrow{\mathcal{W}(\cdot)} \tilde{\boldsymbol{s}} \xrightarrow{g_{\text{ch}}} \hat{\boldsymbol{z}} \xrightarrow{g_{\text{sem}}} \hat{T}. \quad (3)$$

- All components of the system - $f_{\text{sem}}, f_{\text{ch}}, g_{\text{ch}}, g_{\text{sem}}$:
  - Jointly optimized to minimize composite loss $\mathcal{L}(\hat{T}, T(\boldsymbol{x}))^{a}$.



[a] The specific loss components depend on the considered scenario.

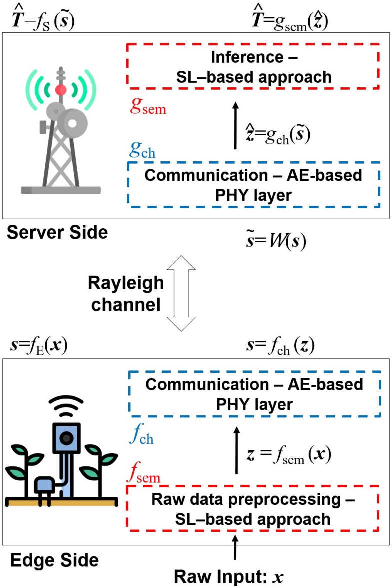
System Model

- Overall inference system:

$$\boldsymbol{x} \xrightarrow{f_{\text{sem}}} \boldsymbol{z} \xrightarrow{f_{\text{ch}}} \boldsymbol{s} \xrightarrow{\mathcal{W}(\cdot)} \tilde{\boldsymbol{s}} \xrightarrow{g_{\text{ch}}} \hat{\boldsymbol{z}} \xrightarrow{g_{\text{sem}}} \hat{T}. \tag{3}$$

- All components of the system - $f_{\text{sem}}, f_{\text{ch}}, g_{\text{ch}}, g_{\text{sem}}$:
  - Jointly optimized to minimize composite loss $\mathcal{L}(\hat{T}, T(\boldsymbol{x}))$[a].

- **Goal:** Balance between semantic fidelity, channel resilience, and task reliability.



[a] The specific loss components depend on the considered scenario.

Vanilla Approach – Does It Work?

Experimental Setup & Training Procedure

- Amazon Stock Dataset:
  - 18 years of data.
  - Predict the subsequent *Open* ($T(\boldsymbol{x}) = x_{t+1}$) value from previous $N = 20$ *Close* values.

Experimental Setup & Training Procedure

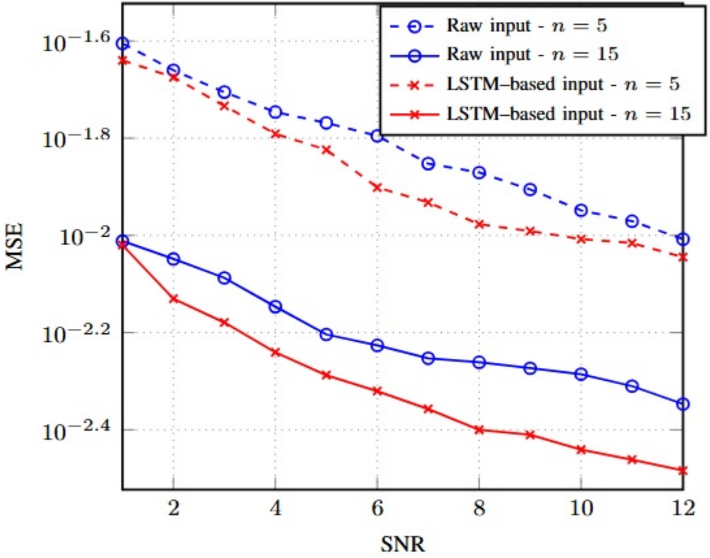
- Amazon Stock Dataset:
  - 18 years of data.
  - Predict the subsequent *Open* ($T(\boldsymbol{x}) = x_{t+1}$) value from previous $N = 20$ *Close* values.
- Neural network architecture:
  - **Semantic part** - Two–layer LSTM network ($K = 10$ hidden states) followed by a FC layer:
    - f_{sem} - LSTM layer.
    - g_{sem} - LSTM layer + FC layer.
  - **Channel part** - Simple AE architecture (with symmetric FC layers):
    - f_{ch} and g_{ch} comprise a single hidden layer with 10 neurons.
    - Latent dimension n controls the compression level and bandwidth usage - $n = 5$ or $n = 15$.
  - Centralized baseline scenario – Semantic part solely on the server, for comparison.

Experimental Setup & Training Procedure

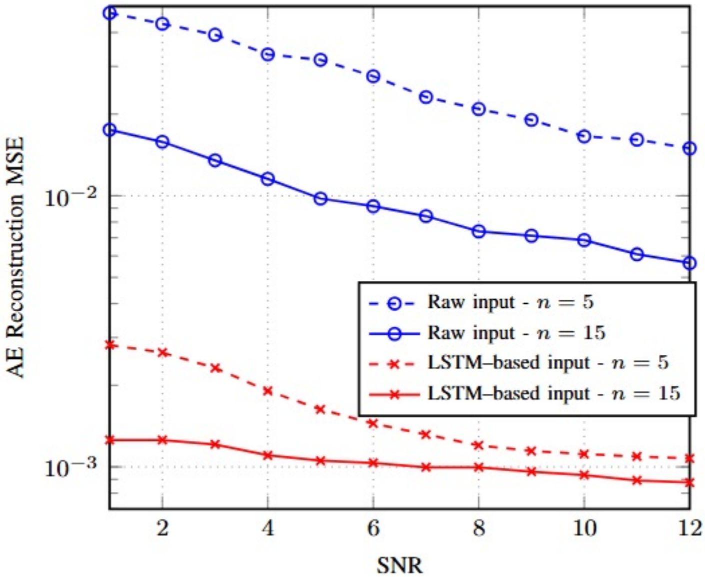
- Amazon Stock Dataset:
  - 18 years of data.
  - Predict the subsequent *Open* ($T(\boldsymbol{x}) = x_{t+1}$) value from previous $N = 20$ *Close* values.
- Neural network architecture:
  - **Semantic part** - Two–layer LSTM network ($K = 10$ hidden states) followed by a FC layer:
    - f_{sem} - LSTM layer.
    - g_{sem} - LSTM layer + FC layer.
  - **Channel part** - Simple AE architecture (with symmetric FC layers):
    - f_{ch} and g_{ch} comprise a single hidden layer with 10 neurons.
    - Latent dimension n controls the compression level and bandwidth usage - $n = 5$ or $n = 15$.
  - Centralized baseline scenario – Semantic part solely on the server, for comparison.
- Loss function:

$$\mathcal{L} = \underbrace{\mathcal{L}(T, \hat{T}(\boldsymbol{x}))}_{\text{Task Loss}} + \underbrace{\mathcal{L}_{\text{AE}}(\boldsymbol{z}, \hat{\boldsymbol{z}})}_{\text{AE Rec. Loss}} = \mathbb{E}\left[(\hat{T} - x_{t+1})^2\right] + \mathbb{E}\left[||\hat{\boldsymbol{z}} - \boldsymbol{z}||_2^2\right] \quad (4)$$

Performance Evaluation - Task MSE



Performance Evaluation - AE Reconstruction MSE



Extension 1 – Are All Symbols Equally Important?

Experimental Setup & Training procedure

- We reuse dataset and NN architecture from the vanilla approach.

[2] V. Ninković, D. Vukobratović, C. Häger, H. Wymeersch and A. Graell i Amat, "Autoencoder-Based Unequal Error Protection Codes," *IEEE Commun. Lett.*, vol. 25, no. 11, pp. 3575–3579, Nov. 2021

Experimental Setup & Training procedure

- We reuse dataset and NN architecture from the vanilla approach.
- **Key question:** Are all symbols of $\boldsymbol{z}$ equally informative for the final performance?

[2] V. Ninković, D. Vukobratović, C. Häger, H. Wymeersch and A. Graell i Amat, "Autoencoder-Based Unequal Error Protection Codes," *IEEE Commun. Lett.*, vol. 25, no. 11, pp. 3575–3579, Nov. 2021

Experimental Setup & Training procedure

- We reuse dataset and NN architecture from the vanilla approach.
- **Key question:** Are all symbols of $\boldsymbol{z}$ equally informative for the final performance?
- **Solution:** AE-based unequal error protection (UEP) codes[2].
  - Not all parts of the message are equally protected.
  - Training procedure – composite loss function:

$$\mathcal{L}_{\mathrm{UEP}} = \lambda \sum_{k \in \mathcal{K}_{\mathrm{imp}}} \|\hat{z}_k - z_k\|^2 + (1-\lambda) \sum_{k \notin \mathcal{K}_{\mathrm{imp}}} \|\hat{z}_k - z_k\|^2, \tag{5}$$

[2] V. Ninković, D. Vukobratović, C. Häger, H. Wymeersch and A. Graell i Amat, "Autoencoder-Based Unequal Error Protection Codes," *IEEE Commun. Lett.*, vol. 25, no. 11, pp. 3575–3579, Nov. 2021

Experimental Setup & Training procedure

- We reuse dataset and NN architecture from the vanilla approach.
- **Key question:** Are all symbols of ***z*** equally informative for the final performance?
- **Solution:** AE-based unequal error protection (UEP) codes[2].
  - Not all parts of the message are equally protected.
  - Training procedure – composite loss function:

$$\mathcal{L}_{\mathrm{UEP}} = \lambda \sum_{k \in \mathcal{K}_{\mathrm{imp}}} \|\hat{z}_k - z_k\|^2 + (1-\lambda) \sum_{k \notin \mathcal{K}_{\mathrm{imp}}} \|\hat{z}_k - z_k\|^2, \tag{5}$$

- $1 \geq \lambda > 0.5$ – weighting parameter (flexible trade-off between classes).

[2]V. Ninković, D. Vukobratović, C. Häger, H. Wymeersch and A. Graell i Amat, "Autoencoder-Based Unequal Error Protection Codes," *IEEE Commun. Lett.*, vol. 25, no. 11, pp. 3575–3579, Nov. 2021

Experimental Setup & Training procedure

- We reuse dataset and NN architecture from the vanilla approach.
- **Key question:** Are all symbols of $\boldsymbol{z}$ equally informative for the final performance?
- **Solution:** AE-based unequal error protection (UEP) codes[2].
  - Not all parts of the message are equally protected.
  - Training procedure – composite loss function:

$$\mathcal{L}_{\mathrm{UEP}} = \lambda \sum_{k \in \mathcal{K}_{\mathrm{imp}}} \|\hat{z}_k - z_k\|^2 + (1-\lambda) \sum_{k \notin \mathcal{K}_{\mathrm{imp}}} \|\hat{z}_k - z_k\|^2, \quad (5)$$

- $1 \geq \lambda > 0.5$ – weighting parameter (flexible trade-off between classes).
- **Problem:** $\mathcal{K}_{\mathrm{imp}}$ is fixed in conventional UEP schemes, but in a dynamic IoT system symbol importance evolves over time.

[2] V. Ninković, D. Vukobratović, C. Häger, H. Wymeersch and A. Graell i Amat, "Autoencoder-Based Unequal Error Protection Codes," *IEEE Commun. Lett.*, vol. 25, no. 11, pp. 3575–3579, Nov. 2021

Training Procedure - Dynamic UEP

- **Proposed solution:**
  - Recall the IB principle:

$$\min I(\boldsymbol{x};\boldsymbol{z}) \quad \text{subject to} \quad I(\boldsymbol{z};T(\boldsymbol{x})) \geq \epsilon, \tag{6}$$

  - L most task-relevant latent symbols ($L < K$) are identified dynamically via MI estimation.

Training Procedure - Dynamic UEP

- **Proposed solution:**
  - Recall the IB principle:

$$\min I(\boldsymbol{x};\boldsymbol{z}) \quad \text{subject to} \quad I(\boldsymbol{z};T(\boldsymbol{x})) \geq \epsilon, \tag{6}$$

  - L most task-relevant latent symbols ($L < K$) are identified dynamically via MI estimation.
  - $\hat{I}(\boldsymbol{z};T(\boldsymbol{x}))$ estimated via the Donsker–Varadhan (DV) lower bound:

$$\hat{I}(\boldsymbol{z};T(\boldsymbol{x})) = \mathbb{E}_{P(\boldsymbol{z},T(\boldsymbol{x}))}[f_\theta(\boldsymbol{z},T(\boldsymbol{x}))] - \log \mathbb{E}_{P(\boldsymbol{z})P(T(\boldsymbol{x}))}[e^{f_\theta(\boldsymbol{z},T'(\boldsymbol{x}))}], \tag{7}$$

Training Procedure - Dynamic UEP

- **Proposed solution:**
  - Recall the IB principle:

$$\min I(\boldsymbol{x}; \boldsymbol{z}) \quad \text{subject to} \quad I(\boldsymbol{z}; T(\boldsymbol{x})) \geq \epsilon, \tag{6}$$

  - L most task-relevant latent symbols ($L < K$) are identified dynamically via MI estimation.
  - $\hat{I}(\boldsymbol{z}; T(\boldsymbol{x}))$ estimated via the Donsker–Varadhan (DV) lower bound:

$$\hat{I}(\boldsymbol{z}; T(\boldsymbol{x})) = \mathbb{E}_{P(\boldsymbol{z}, T(\boldsymbol{x}))}[f_\theta(\boldsymbol{z}, T(\boldsymbol{x}))] - \log \mathbb{E}_{P(\boldsymbol{z})P(T(\boldsymbol{x}))}[e^{f_\theta(\boldsymbol{z}, T'(\boldsymbol{x}))}], \tag{7}$$

  - $f_\theta(\boldsymbol{z}, T(\boldsymbol{x}))$ - Neural discriminator implemented as a three-layer FC + ReLU network.

Training Procedure - Dynamic UEP

- **Proposed solution:**
  - Recall the IB principle:

$$\min I(\boldsymbol{x}; \boldsymbol{z}) \quad \text{subject to} \quad I(\boldsymbol{z}; T(\boldsymbol{x})) \geq \epsilon, \tag{6}$$

  - L most task-relevant latent symbols ($L < K$) are identified dynamically via MI estimation.
  - $\hat{I}(\boldsymbol{z}; T(\boldsymbol{x}))$ estimated via the Donsker–Varadhan (DV) lower bound:

$$\hat{I}(\boldsymbol{z}; T(\boldsymbol{x})) = \mathbb{E}_{P(\boldsymbol{z}, T(\boldsymbol{x}))}[f_\theta(\boldsymbol{z}, T(\boldsymbol{x}))] - \log \mathbb{E}_{P(\boldsymbol{z})P(T(\boldsymbol{x}))}[e^{f_\theta(\boldsymbol{z}, T'(\boldsymbol{x}))}], \tag{7}$$

  - $f_\theta(\boldsymbol{z}, T(\boldsymbol{x}))$ - Neural discriminator implemented as a three-layer FC + ReLU network.
  - **Training batch:** Compute gradients of $\hat{I}(\boldsymbol{z}; T(\boldsymbol{x}))$ w.r.t. $\boldsymbol{z}$, select the L most task-relevant latent dimensions, and update the ranking dynamically.

Training Procedure - Dynamic UEP

- **Proposed solution (continued)[3]:**

[3] V. Ninkovic, D. Vukobratovic, D. Miskovic, and C. Wang, "Adaptive Unequal Error Protection for Semantic Split Learning over Wireless Channels", submitted at *2026 IEEE ICC.*

Training Procedure - Dynamic UEP

- **Proposed solution (continued)**[3]**:**
  - AE reconstruction loss (Eq. 5) remains compound, with $\mathcal{K}_{\mathrm{imp}}$ ($|\mathcal{K}_{\mathrm{imp}}| = L$) determined adaptively.

[3] V. Ninkovic, D. Vukobratovic, D. Miskovic, and C. Wang, "Adaptive Unequal Error Protection for Semantic Split Learning over Wireless Channels", submitted at *2026 IEEE ICC.*

Training Procedure - Dynamic UEP

- **Proposed solution (continued)[3]:**
  - AE reconstruction loss (Eq. 5) remains compound, with $\mathcal{K}_{\mathrm{imp}}$ ($|\mathcal{K}_{\mathrm{imp}}| = L$) determined adaptively.
  - Overall loss function:

$$\mathcal{L} = \mathcal{L}_{\text{task}} + \mathcal{L}_{\text{UEP}} + \mathcal{L}_{\text{MI}}. \tag{8}$$

[3] V. Ninkovic, D. Vukobratovic, D. Miskovic, and C. Wang, "Adaptive Unequal Error Protection for Semantic Split Learning over Wireless Channels", submitted at *2026 IEEE ICC.*

Training Procedure - Dynamic UEP

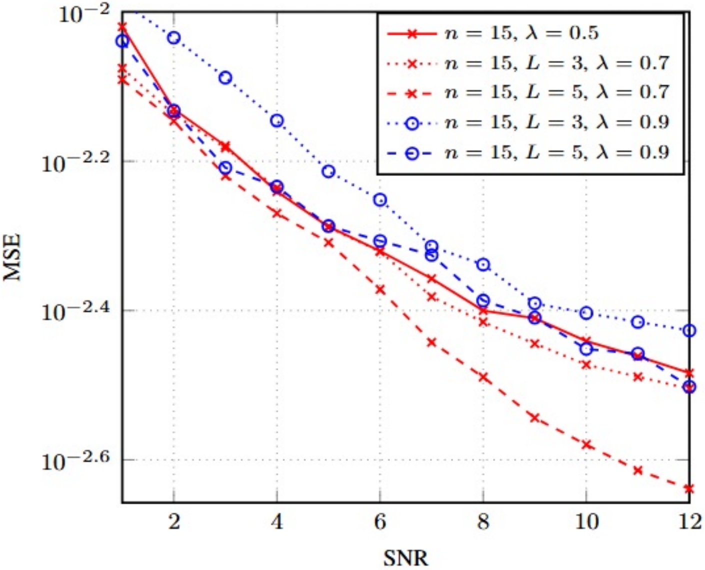
- **Proposed solution (continued)[3]:**
  - AE reconstruction loss (Eq. 5) remains compound, with $\mathcal{K}_{\mathrm{imp}}$ ($|\mathcal{K}_{\mathrm{imp}}| = L$) determined adaptively.
  - Overall loss function:

$$\mathcal{L} = \mathcal{L}_{\mathsf{task}} + \mathcal{L}_{\mathsf{UEP}} + \mathcal{L}_{\mathsf{MI}}. \tag{8}$$

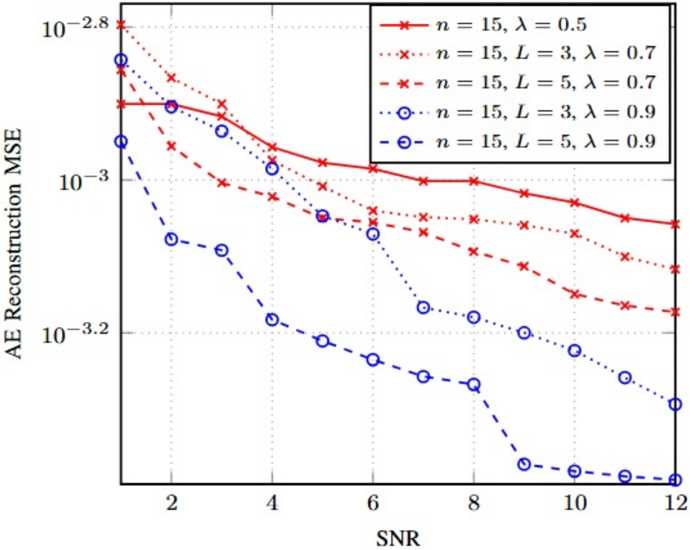
  - $\mathcal{L}_{\mathsf{MI}}$ — negative DV bound (Eq. (7)) used to train f_θ:
    - Enables stable estimation of $I(\boldsymbol{z}; T(\boldsymbol{x}))$.
    - Allows reliable identification of $\mathcal{K}_{\mathrm{imp}}$.

[3] V. Ninkovic, D. Vukobratovic, D. Miskovic, and C. Wang, "Adaptive Unequal Error Protection for Semantic Split Learning over Wireless Channels", submitted at *2026 IEEE ICC.*

Impact of Dynamic UEP on Performance - Task MSE



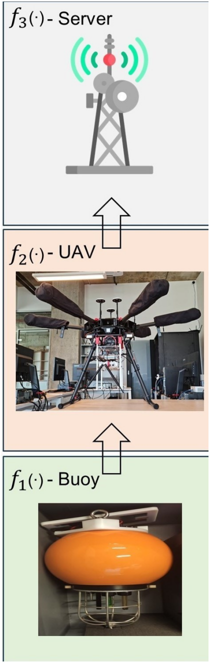
Impact of Dynamic UEP on Performance - AE Reconstruction MSE



Extension 2 – Semantic & AE-PHY in UAV-Assisted IoT

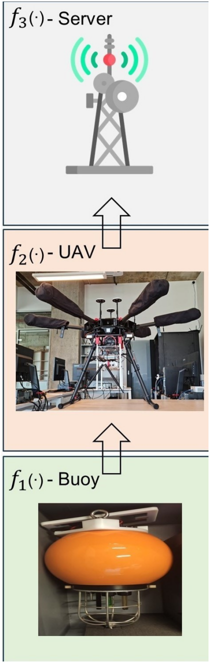
System Model

- Push toward real-world deployment.



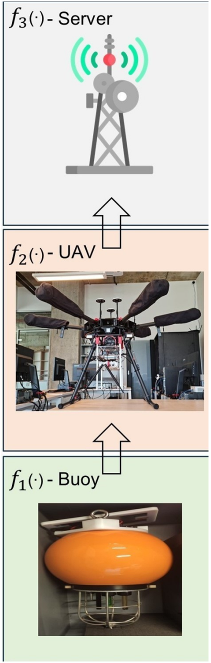
System Model

- Push toward real-world deployment.
- UAV-assisted IoT system:
  - UAV acts both as a relay and as part of the inference system.
  - Buoy $\rightarrow$ UAV $\rightarrow$ Server.



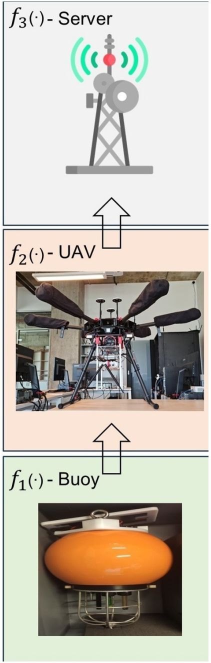
System Model

- Push toward real-world deployment.
- UAV-assisted IoT system:
  - UAV acts both as a relay and as part of the inference system.
  - Buoy $\rightarrow$ UAV $\rightarrow$ Server.
- Integration of multiple principles:
  - SL and semantic communications for information (pre)processing.
  - AE-based physical-layer/channel coding.
  - Early Exit mechanism.

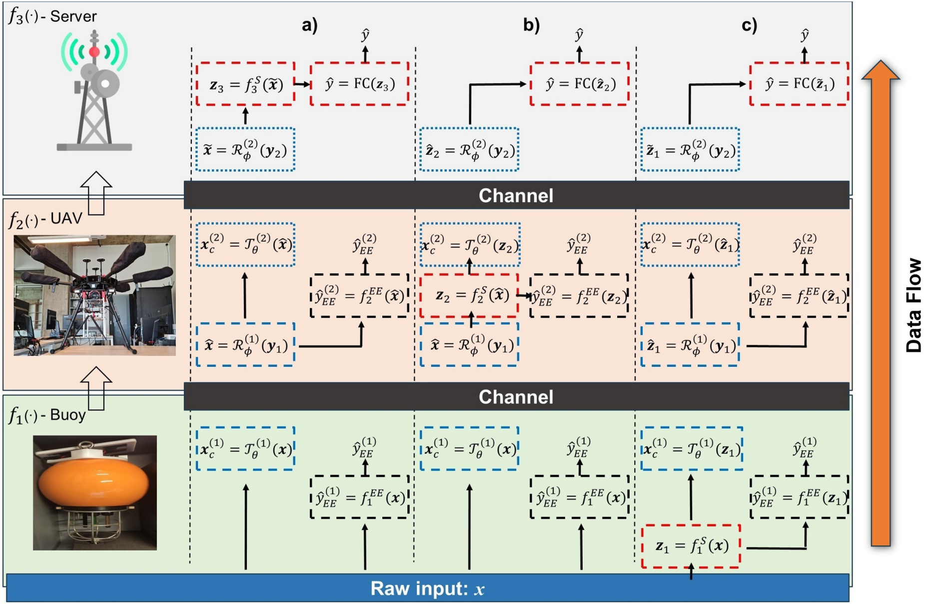


System Model

- Push toward real-world deployment.
- UAV-assisted IoT system:
  - UAV acts both as a relay and as part of the inference system.
  - Buoy $\rightarrow$ UAV $\rightarrow$ Server.
- Integration of multiple principles:
  - SL and semantic communications for information (pre)processing.
  - AE-based physical-layer/channel coding.
  - Early Exit mechanism.
- Three deployment scenarios:
  - Semantics processed at the server.
  - Semantics processed at the UAV.
  - Semantics processed at the edge (IoT device).



System Model

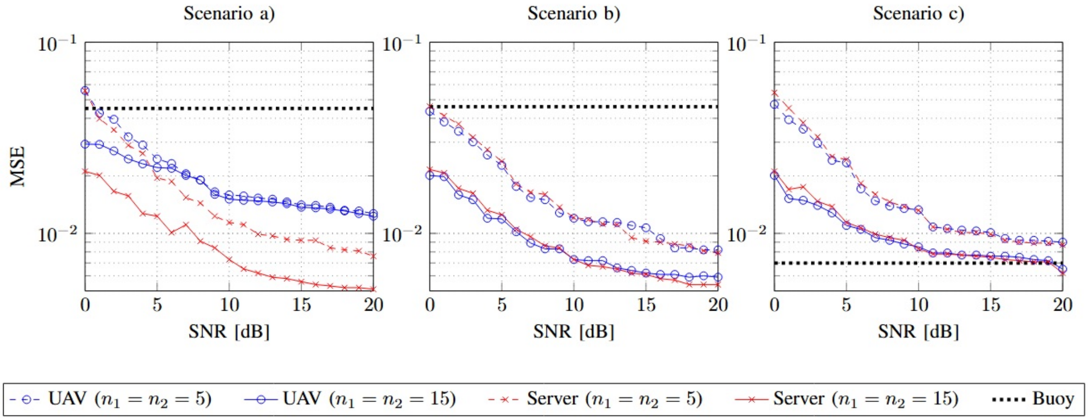


Experimental Setup & Training Procedure

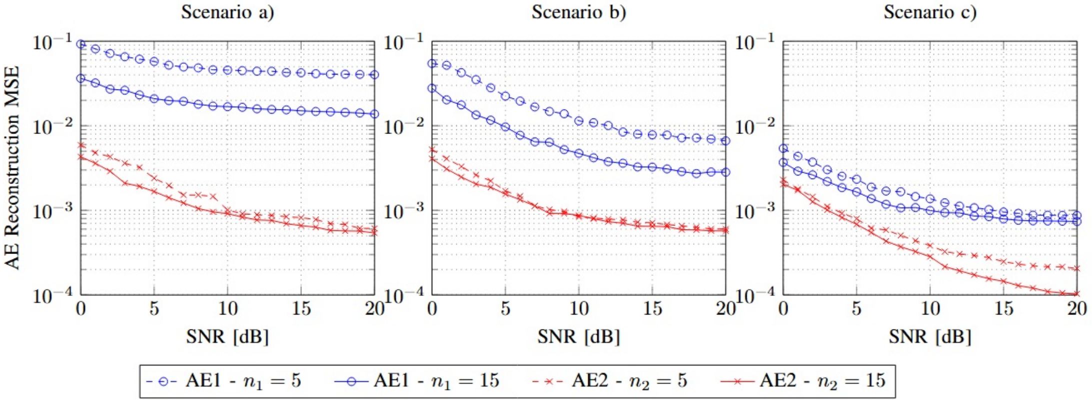
- Dataset: Pollution of the Danube river near Novi Sad.
  - 3,264 instances - Each instance represents a daily measurement from 2013 to 2022
  - Estimation of dissolved oxygen based on previous $N = 10$ days.
- Neural network architecture:
  - **Semantic part -** LSTM layer (with 10 hidden states) + FC layer:
    - Server side - Always FC layer.
    - LSTM layer position depends on implementation scenario.
  - **Channel part:**
    - Two AE modules (AE1 between buoy and UAV, AE2 between UAV and server).
    - Symmetric AE architecture with one hidden FC layer.
  - **Early Exit** - Additional FC layer.
- Loss function:

$$\mathcal{L} = \underbrace{\mathcal{L}_{\text{task}}}_{\text{Task loss}} + \underbrace{\mathcal{L}_{\text{EE}_1}}_{\text{EE at buoy}} + \underbrace{\mathcal{L}_{\text{EE}_2}}_{\text{EE at UAV}} + \underbrace{\mathcal{L}_{\mathrm{AE}_1}}_{\text{AE1 recon.}} + \underbrace{\mathcal{L}_{\mathrm{AE}_2}}_{\text{AE2 recon.}} \tag{9}$$

Performance Evaluation - Task MSE



Performance Evaluation - AE Reconstruction MSE[4]



[4] V. Ninkovic, R. S. Molina, et al. "Semantic IoT Framework for Environmental Monitoring: FPGA-Accelerated Distributed Learning at the Edge," manuscript in preparation

Discussion

Next Steps & Open Problems

- Step toward practical task-oriented IoT monitoring.
- Further move toward real-world implementation — FPGA deployment[5].

[5] Already implemented; details will be presented by Romina Soledad Molina

Next Steps & Open Problems

- Step toward practical task-oriented IoT monitoring.
- Further move toward real-world implementation — FPGA deployment[5].
- Addressing real wireless channel complexity:
  - Controlled lab environment as an initial stage.
  - Real-world field deployment as the next step.

[5] Already implemented; details will be presented by Romina Soledad Molina

Next Steps & Open Problems

- Step toward practical task-oriented IoT monitoring.
- Further move toward real-world implementation — FPGA deployment[5].
- Addressing real wireless channel complexity:
  - Controlled lab environment as an initial stage.
  - Real-world field deployment as the next step.
- Fits within the broader vision of 6G communication architectures.

[5] Already implemented; details will be presented by Romina Soledad Molina

Next Steps & Open Problems

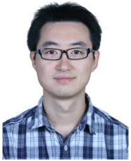
- Step toward practical task-oriented IoT monitoring.
- Further move toward real-world implementation — FPGA deployment[5].
- Addressing real wireless channel complexity:
  - Controlled lab environment as an initial stage.
  - Real-world field deployment as the next step.
- Fits within the broader vision of 6G communication architectures.
- Next steps — multimodal data processing, multi-task learning, and adaptive prioritization across heterogeneous devices.

[5] Already implemented; details will be presented by Romina Soledad Molina

Acknowledgment[6]

Dejan Vukobratovic, PhD

Dragisa Miskovic, PhD



Chao Wang, PhD

Romina Soledad Molina, PhD

Maria Liz Crespo, PhD

Marco Zennaro, PhD

[6]This work has received funding from: the Horizon 2020 grant agreement No 101086387 - REMARKABLE; the Serbian Ministry of Science, Technological Development and Innovation (No. 00101957 2025 13440 003 000 620 021) & Intergovernmental International Science and Technology Innovation Cooperation of National Key Research & Development Program of China under Grant 2024YFE0197400.

References - Edge AI & UAV–Assisted IoT Monitoring I

- J. Roostaei *et al.*, “IoT-based edge computing (IoTEC) for improved environmental monitoring,” *Sustain. Comput.: Inform. Syst.*, vol. 38, Apr. 2023, Art. no. 100870.
- H. Xing, G. Zhu, D. Liu, H. Wen, K. Huang, and K. Wu, “Task-oriented integrated sensing, computation and communication for wireless edge AI,” *IEEE Netw.*, vol. 37, no. 4, pp. 135–144, Jul./Aug. 2023.
- O. Gupta and R. Raskar, “Distributed learning of deep neural network over multiple agents,” *J. Netw. Comput. Appl.*, vol. 116, pp. 1–8, Aug. 2018.
- L. Cheng *et al.*, “Advancements in accelerating deep neural network inference on AIoT devices: A survey,” *IEEE Trans. Sustain. Comput.*, vol. 9, no. 6, pp. 830–847, Nov./Dec. 2024.
- Z. Zhou, X. Chen, E. Li, L. Zeng, K. Luo, and J. Zhang, “Edge intelligence: Paving the last mile of artificial intelligence with edge computing,” *Proc. IEEE*, vol. 107, no. 8, pp. 1738–1762, Aug. 2019.
- X. Hou *et al.*, “Split federated learning for UAV-enabled integrated sensing, computation, and communication,” *arXiv preprint* arXiv:2504.01443, 2025.

References - Edge AI & UAV–Assisted IoT Monitoring II

- J. Suo, X. Zhang, W. Shi, and W. Zhou, “E3-UAV: An edge-based energy-efficient object detection system for unmanned aerial vehicles,” *IEEE Internet Things J.*, vol. 11, no. 3, pp. 4398–4413, Feb. 2024.
- J. Tang *et al.*, “Multi-UAV-assisted federated learning for energy-aware distributed edge training,” *IEEE Trans. Netw. Serv. Manage.*, vol. 21, no. 1, pp. 280–294, Feb. 2024.
- X. Hu, K.-K. Wong, and Z. Zheng, “Wireless-powered mobile edge computing with cooperated UAV,” in *Proc. IEEE 20th Int. Workshop Signal Process. Adv. Wireless Commun. (SPAWC)*, Cannes, France, 2019, pp. 1–5.
- Z. Ma *et al.*, “An UAVs-assisted edge computing network with multi-agent reinforcement learning,” in *IoTCIT 2023, LN Eng. Electron.*, vol. 1197, pp. 115–130, 2024.
- V. Ninkovic, D. Vukobratovic, and D. Miskovic, “UAV-assisted distributed learning for environmental monitoring in rural environments,” in *Proc. 7th Int. Balkan Conf. Commun. Netw. (BalkanCom)*, Ljubljana, Slovenia, 2024, pp. 296–300.

References - Semmantic Communcations I

- D. Gündüz *et al.*, "Beyond Transmitting Bits: Context, Semantics, and Task-Oriented Communications," *IEEE J. Sel. Areas Commun.*, vol. 41, no. 1, pp. 5-41, 2023.
- W. Weaver, “Recent contributions to the mathematical theory of communication,” *ETC: Rev. Gen. Semantics,* pp. 261–281, 1953.
- M. Kountouris and N. Pappas, “Semantics-empowered communication for networked intelligent systems,” *IEEE Commun. Mag.,* vol. 59, no. 6, pp. 96–102, 2021.
- W. Yang *et al.*, “Semantic communication meets edge intelligence,” *IEEE Wireless Commun.*, vol. 29, no. 5, pp. 28–35, Oct. 2022.
- W. Xu *et al.*, “Edge learning for B5G networks with distributed signal processing: Semantic communication, edge computing, and wireless sensing,” *IEEE J. Sel. Topics Signal Process.*, vol. 17, no. 1, pp. 9–39, Jan. 2023.
- H. Xie, Z. Qin, X. Tao, and K. B. Letaief, “Task-oriented multi-user semantic communications,” *IEEE J. Sel. Areas Commun.*, vol. 40, no. 9, pp. 2584–2597, Sept. 2022.

References - Semmantic Communcations II

- L. X. Nguyen *et al.*, “An efficient federated learning framework for training semantic communication systems,” *IEEE Trans. Veh. Technol.*, vol. 73, no. 10, pp. 15872–15877, Oct. 2024.
- W. Wu *et al.*, “Split learning over wireless networks: Parallel design and resource management,” *IEEE J. Sel. Areas Commun.*, vol. 41, no. 4, pp. 1051–1066, Apr. 2023.
- L. Wang, W. Wu, F. Zhou, Z. Yang, Z. Qin, and Q. Wu, “Adaptive resource allocation for semantic communication networks,” *IEEE Trans. Commun.*, vol. 72, no. 11, pp. 6900–6916, Nov. 2024.
- J. Huang *et al.*, “Dynamic UAV-assisted cooperative edge AI inference,” *IEEE Trans. Wireless Commun.*, vol. 24, no. 1, pp. 615–628, Jan. 2025.
- Z. Cao, H. Zhang, L. Liang, H. Wang, S. Jin, and G. Y. Li, “Task-oriented semantic communication for stereo-vision 3D object detection,” *IEEE Trans. Commun.*, to appear, 2025.
- C.-H. Lee, J.-W. Lin, P.-H. Chen, and Y.-C. Chang, “Deep learning-constructed joint transmission-recognition for Internet of Things,” *IEEE Access*, vol. 7, pp. 76547–76561, 2019.

References - AE–Based PHY Modeling I

- T. O’Shea and J. Hoydis, “An introduction to deep learning for the physical layer,” *IEEE Trans. Cogn. Commun. Netw.,* vol. 3, no. 4, pp. 563-575, 2017.
- V. Ninkovic *et al.*, "Autoencoder-Based Unequal Error Protection Codes," *IEEE Commun. Lett.,* vol. 25, no. 11, pp.3575-3579, 2021.
- V. Ninkovic and D. Vukobratovic, "Structured superposition of autoencoders for UEP codes at intermediate blocklengths," *IEEE Commun. Lett.,* early access, 2025.
- F. Wang *et al.*, "Explicit semantic-base-empowered communications for 6G mobile networks," arXiv preprint arXiv:2408.05596v2, 2025.
- J. Xu, T.-Y. Tung, B. Ai, W. Chen, Y. Sun, and D. Gündüz, “Deep joint source-channel coding for semantic communications,” *IEEE Commun. Mag.*, vol. 61, no. 11, pp. 42–48, Nov. 2023.

Thank you for your attention!

vukan.ninkovic@ivi.ac.rs

ninkovic@uns.ac.rs