



北京交通大学
BEIJING JIAOTONG UNIVERSITY

RISTA前沿大讲堂

高效可信的无线基座模型

Efficient and Trustworthy Wireless Foundation Models

章嘉懿

北京交通大学

2026年7月29日

汇报提纲

- I. 研究背景与意义**
- II. 无线尺度规律：从规模扩展到能力演进**
- III. 高效模型架构：从多维表征到轻量推理**
- IV. 可信智能表征：从物理先验到泛化决策**
- V. 总结与展望**

汇报提纲

I. 研究背景与意义

II. 无线尺度规律：从规模扩展到能力演进

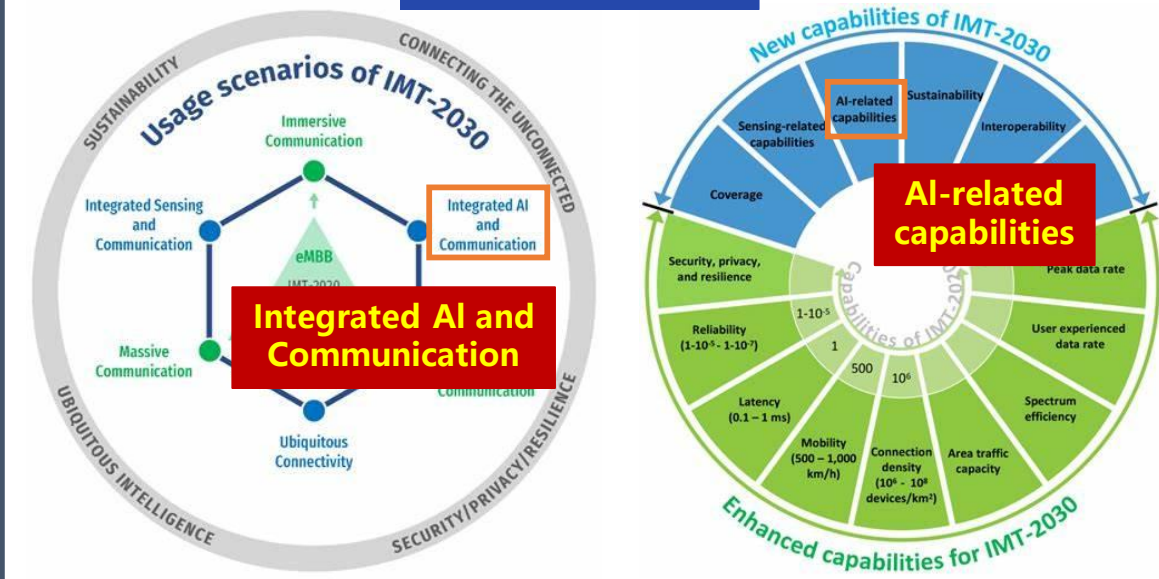
III. 高效模型架构：从多维表征到轻量推理

IV. 可信智能表征：从物理先验到泛化决策

V. 总结与展望

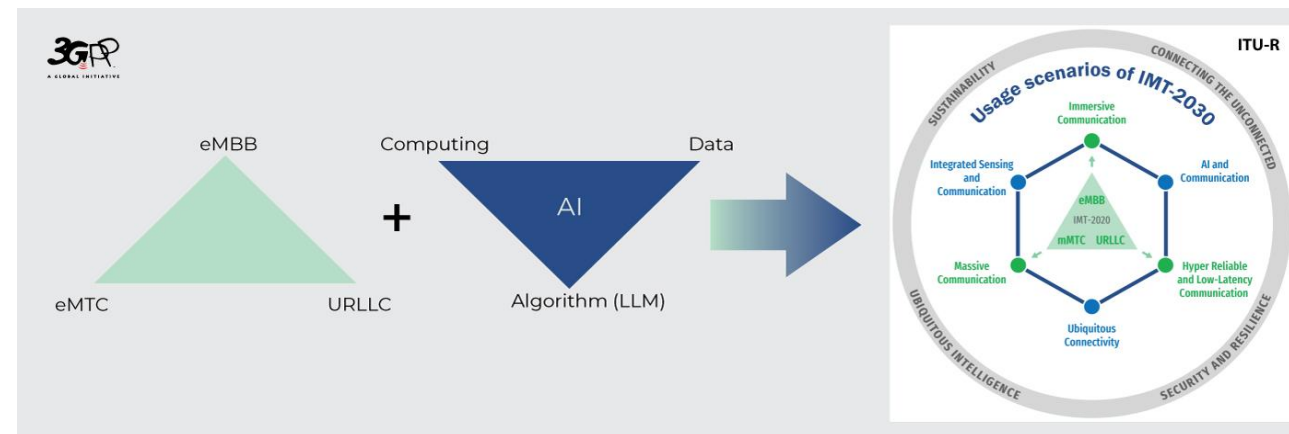
1.1 6G愿景与应用场景

6G需求



技术背景

- 美国、欧盟联合开展**6G-XCEL项目**，旨在嵌入人工智能以增强6G网络的功能
- 欧盟启动一系列**Horizon-Europe项目**，推动 AI-native RAN架构，实现高效、多场景6G网络系统
- 3GPP持续推进 R20 增强研究，启动 **6G AI原生架构、AI智能体、跨域AI协同**等议题讨论



1.2 基座模型赋能6G通信

特性	传统神经网络	无线基座模型
模型规模	参数量相对 较小	参数量 巨大
技术架构	基于规则、模板或浅层 神经网络	主要基于 Transformer 架构
训练数据	需要标注的 小规模 数据集	海量、多样化、无标注 的文本和数据
学习范式	端到端 训练	预训练+微调
处理能力	单一任务	多任务 ，强大的泛化能力和通用性

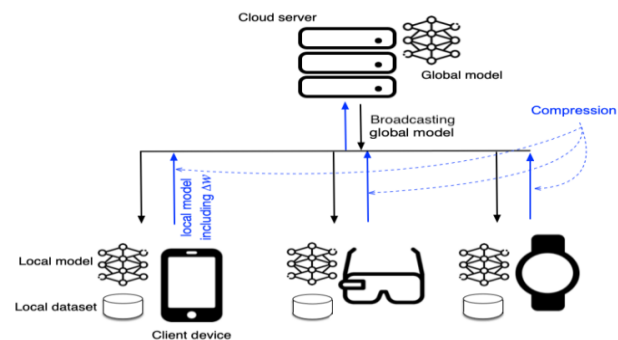
智能体：自感知、自决策、自进化

[1] Ziheng Liu, **Jiayi Zhang**, Yiyang Zhu, Enyu Shi, Bokai Xu, Dusit Niyato, Shi Jin, and Bo Ai, “MaLAM4Com: Multi-Agent Cooperative Large AI Models for Wireless Communications,” *IEEE JSAC*, vol. 44, pp. 2889-2903, 2026.

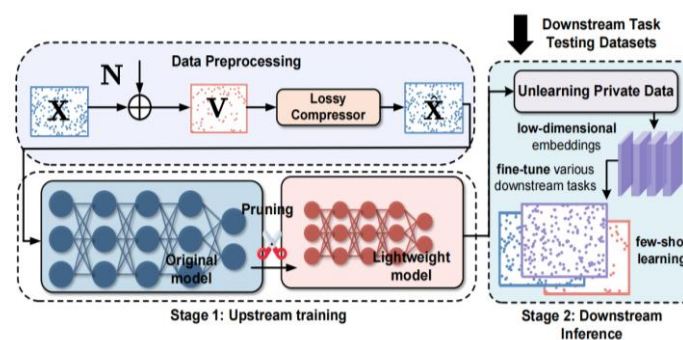
1.3 高效可信的无线基座模型

现状与需求

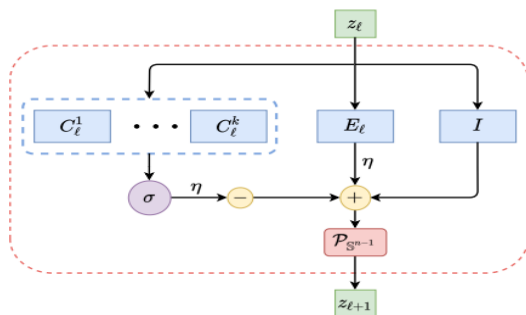
➤ 解决性能与资源矛盾



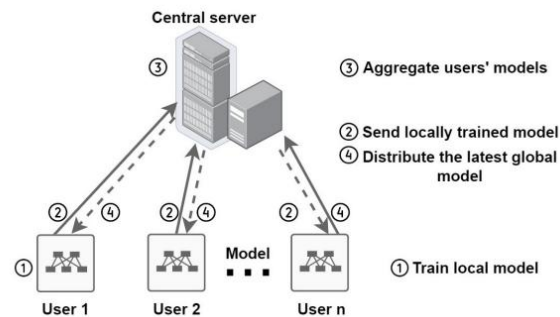
➤ 满足通信低开销要求



➤ 构建可解释的智能系统



➤ 保障数据安全与合规



优势:

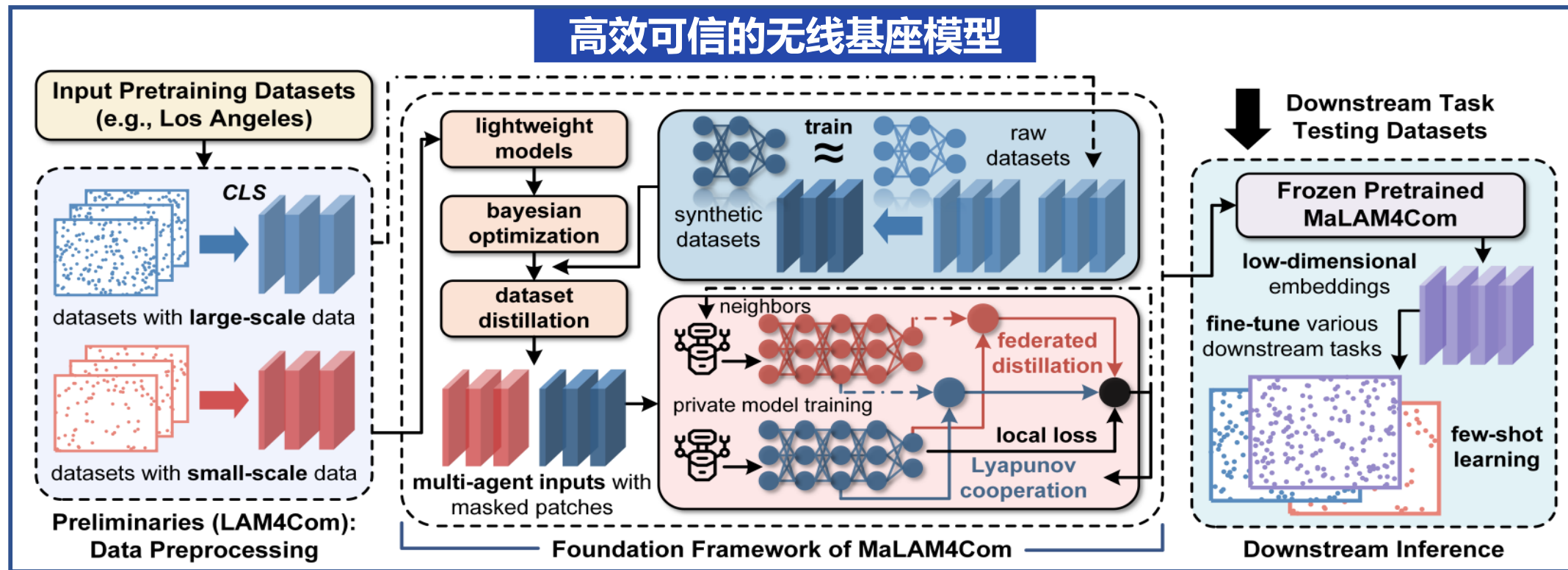
- 毫秒级响应速度
- 数据本地隐私保护
- 个性化训练能力

挑战:

- 部署资源限制
- 部署设备多样

1.3 高效可信的无线基座模型

无线基座模型突破了传统**特定任务**模型的局限性，
通过生成**通用**的信道表征，能够在**各种环境**和**下游任务**中实现无缝适应，**无需重新训练**



数据/模型预处理
无线尺度规律
从规模扩展到能力演进

高效可信探索
高效模型架构: 多维表征到轻量推理
可信智能表征: 物理先验到泛化决策

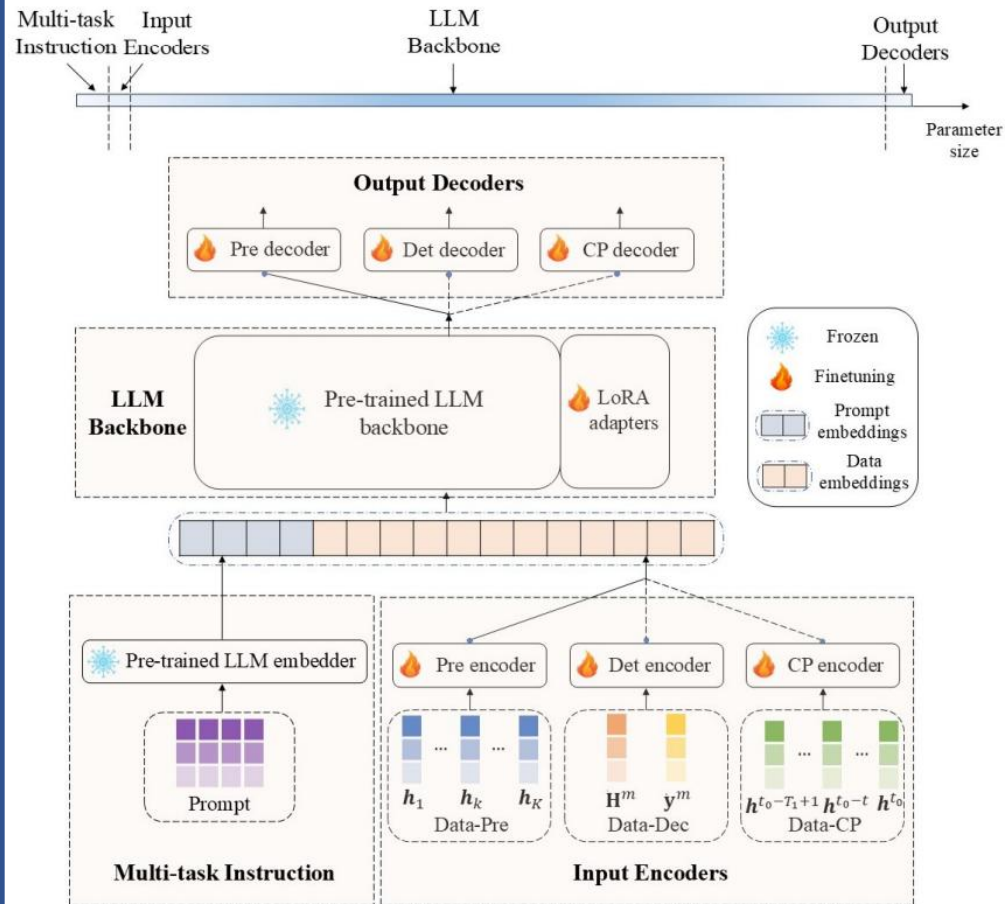
核心信道表征输出
高效赋能下游
多任务、多场景等需求

汇报提纲

- I. 研究背景与意义
- II. 无线尺度规律：从规模扩展到能力演进**
- III. 高效模型架构：从多维表征到轻量推理
- IV. 可信智能表征：从物理先验到泛化决策
- V. 总结与展望**

2.1 大语言模型的基础理论

大语言模型架构



核心思想： 在海量**文本数据**上预训练的Transformer架构，掌握通用的**语言理解**、**逻辑推理**和**内容生成**

应用到**无线通信领域**

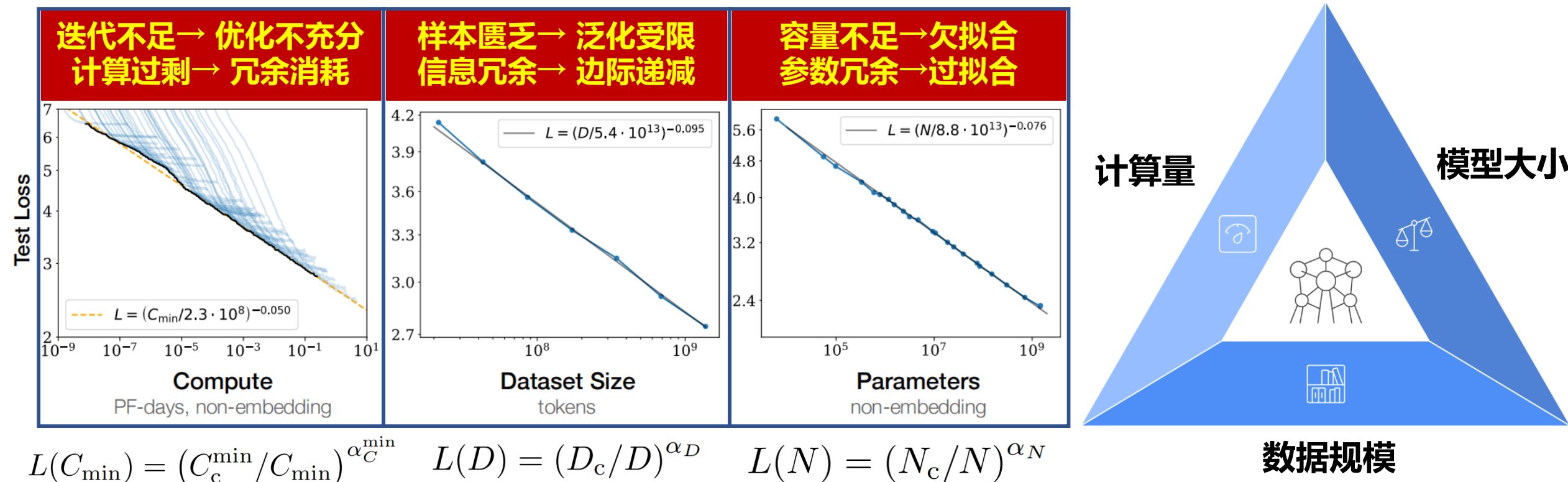
存在模态鸿沟，**无法**直接处理通信领域的数值数据

集中式大语言模型

- 克服模态鸿沟，构建一个能**直接**处理通信领域数据的专用模型
- 无缝集成到一个**端到端**、**面向物理层任务**的架构中

2.2 资源与效率的挑战 — 尺度定律

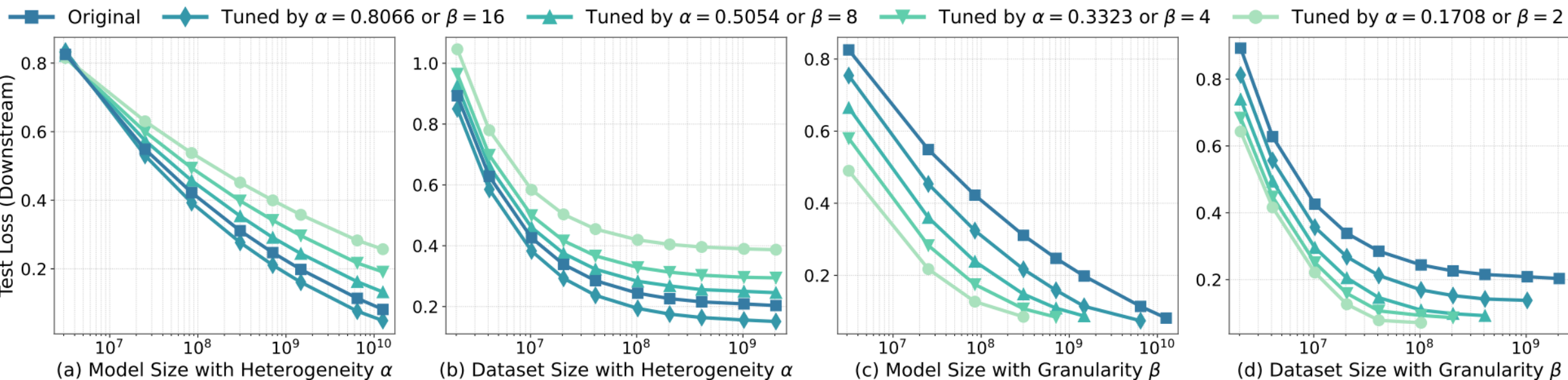
- 虽然集中式大语言模型功能强大，但其性能高度依赖**计算量**、**模型大小**和**数据集**三大维度
- 盲目、随意（**试错**）的部署会导致**资源利用率低**，需要引入**Scaling Law**实现**高效部署**



[2] J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei, “Scaling laws for neural language models,” arXiv:2001.08361, 2020.

2.3 无线领域的特殊性：传统尺度定律为何失效？

- **信道异构性**违背了传统Scaling Law的**同质性假设**，无法用单一常数来刻画复杂环境下的**泛化规律**
- **离散化颗粒度**直接决定了输入信道的**信息保真度**，而这一关键**物理特性**在传统Scaling Law所忽略



无法将传统LLM领域的Scaling Law直接应用于无线通信
无线信道的物理特性破坏了传统Scaling Law幂律公式的普适性

2.4 无线尺度定律的组成：信道异构性与离散化颗粒度

信道异构性 α

$$\mathbf{h}(\alpha) = \sqrt{\alpha}\mathbf{h}_{\text{LoS}} + \sqrt{1-\alpha}\mathbf{h}_{\text{NLoS}}$$

$$\mathcal{L}_{\min}(\alpha) \propto (1-\alpha)$$

离散化颗粒度 β

$$\mathcal{L}(\alpha, \beta) \approx \underbrace{\alpha\beta^{-\gamma_1}}_{\approx 0} + (1-\alpha)\beta^{\gamma_2}$$

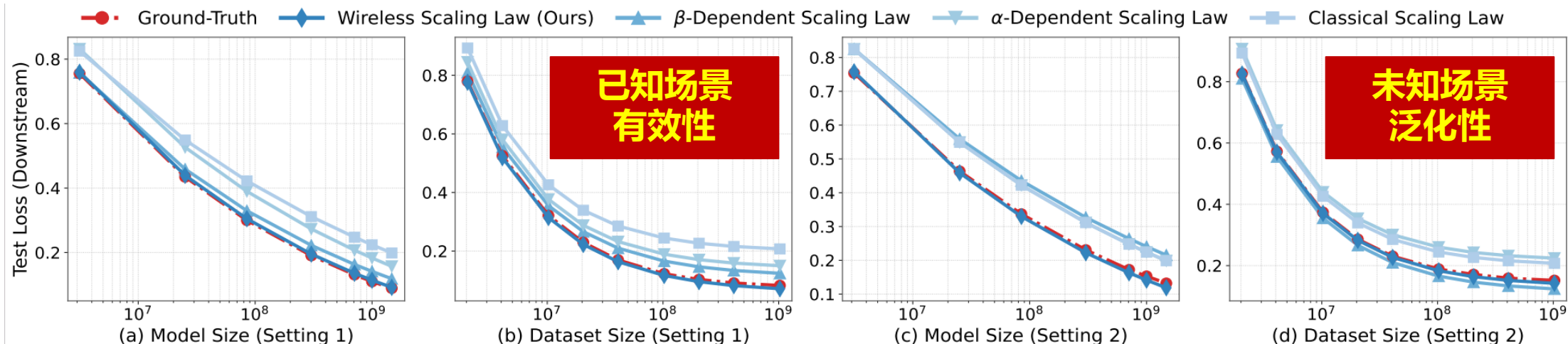
Scaling Paradigms	Scaling with Model Size (N) and Dataset Size (D)
Classical Scaling Law	Model: $\mathcal{L}(N) = -0.241 + (\frac{4901100}{N})^{0.144}$ Dataset: $\mathcal{L}(D) = 0.198 + (\frac{1211190}{D})^{0.692}$
α -Dependent Scaling Law	Model: $\mathcal{L}(N, \alpha) = -0.022\alpha - 0.225 + (\frac{75008\alpha+4853924}{N})^{0.105\alpha+0.076}$ Dataset: $\mathcal{L}(D, \alpha) = -0.359\alpha + 0.429 + (\frac{74246\alpha+1164312}{D})^{-0.100\alpha+0.758}$
β -Dependent Scaling Law	Model: $\mathcal{L}(N, \beta) = -1.311\beta^{0.047} + 1.303 + (\frac{73030\beta^{1.191}+365297}{N})^{0.460\beta^{-0.245}-0.051}$ Dataset: $\mathcal{L}(D, \beta) = 0.226\beta^{0.258} - 0.350 + (\frac{36493\beta^{0.361}+1083745}{D})^{1.003\beta^{0.049}-0.493}$
Wireless Scaling Law (Ours)	Model: $\mathcal{L}(N, \alpha, \beta) = 0.012\alpha - 1.662\beta^{0.036} + 1.641 + \frac{(208181\alpha+75190\beta^{1.185}+216948)^{e_1}}{N^{(e_1:0.102\alpha-0.518\beta^{-0.203}-0.179)}}$ Dataset: $\mathcal{L}(D, \alpha, \beta) = -0.344\alpha + 0.277\beta^{0.227} - 0.185 + \frac{(203835\alpha+17430\beta^{0.601}+964826)^{e_2}}{D^{(e_2:-0.098\alpha+1.737\beta^{0.036}-1.166)}}$

无线尺度定律, Wireless Scaling Law: $\mathcal{L}(N, D, \alpha, \beta) = L_{\min}(\alpha, \beta) + \frac{A(\alpha, \beta)}{N^{\alpha_n(\alpha, \beta)}} + \frac{B(\alpha, \beta)}{D^{\alpha_d(\alpha, \beta)}}$

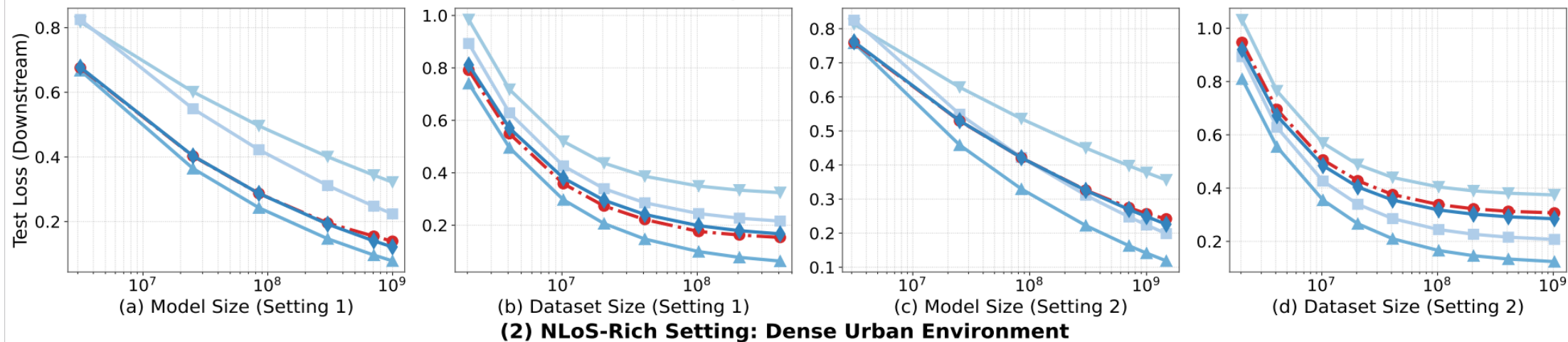
[3] Ziheng Liu, Jiayi Zhang, et.al, “Scaling Law for Large Wireless Models,” *Proc. AAAI*, 2026.

2.5 无线尺度定律的仿真分析

Wireless Scaling Law与传统Scaling Law的性能对比

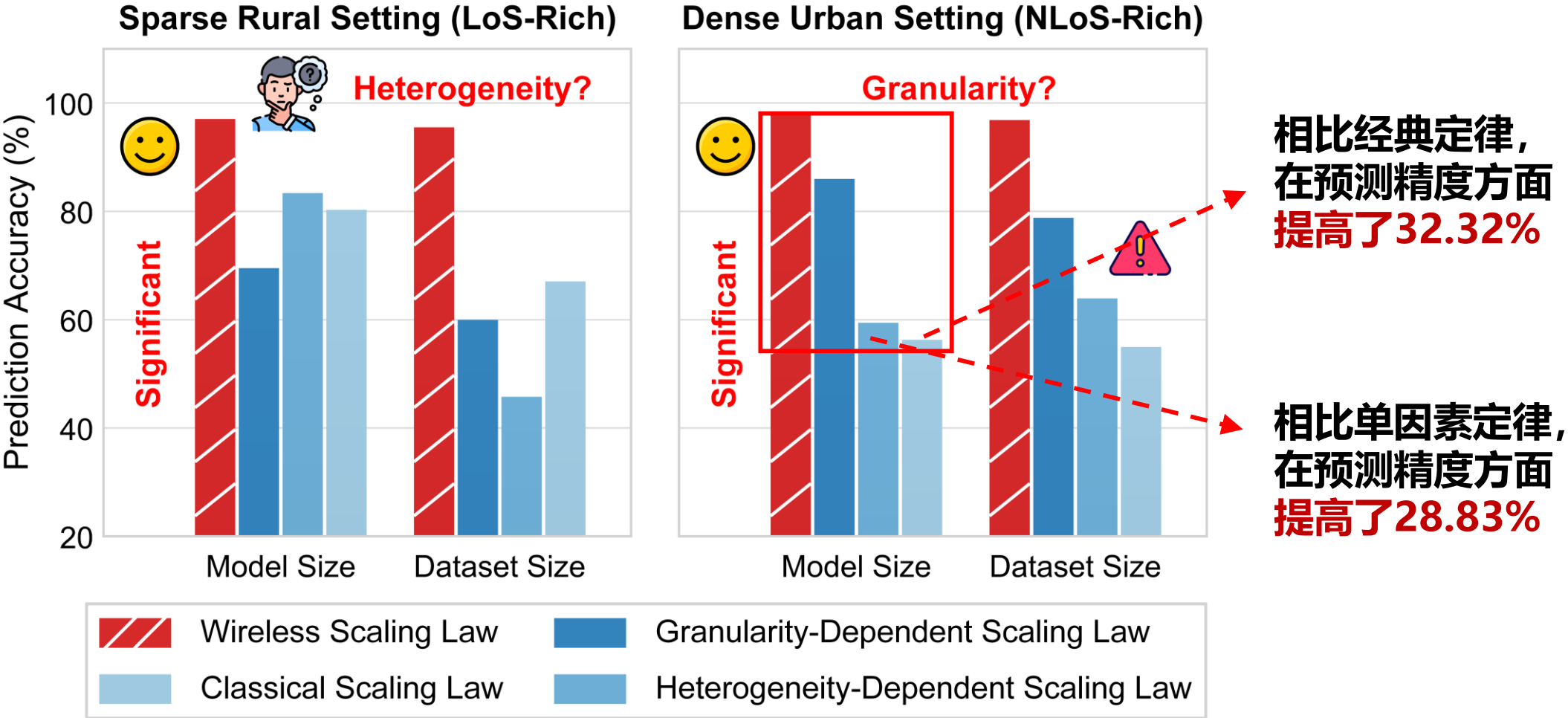


(1) LoS-Rich Setting: Sparse Rural Environment

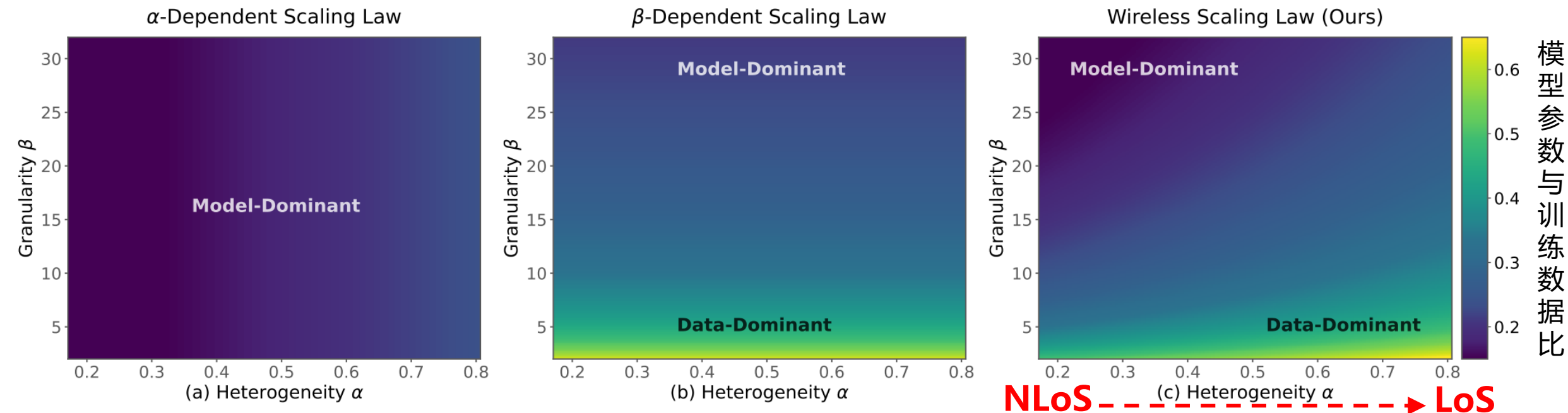


2.5 无线尺度定律的仿真分析

Wireless Scaling Law与传统Scaling Law的性能对比



2.6 无线尺度定律的启发：如何高效训练无线基座模型？



为设计功能强大且资源高效的无线基座模型规划出一条清晰可靠的路线图

- 首先，**计算资源的最优扩展策略**并非由一个**固定的“模型-数据”比例**来决定，而是一个随着数据**异构性**和**颗粒度动态**变化的函数。这意味着**不存在**一个**普适的“黄金比例”**
- 其次，这种依赖关系对数据的**“颗粒度”**尤其敏感，仅仅通过提升现有数据的**空间分辨率**，就能极大地释放模型性能，而无需额外增加数据量

汇报提纲

- I. 研究背景与意义
- II. 无线尺度规律：从规模扩展到能力演进
- III. 高效模型架构：从多维表征到轻量推理**
- IV. 可信智能表征：从物理先验到泛化决策
- V. 总结与展望**

3.1 高效模型架构的必要性

传统基座模型的困境

高开销: 巨大的模型参数和计算量，需要昂贵的云端算力，产生高能耗与数据传输开销

终端→基站→核心网→云端数据中心→计算→返回



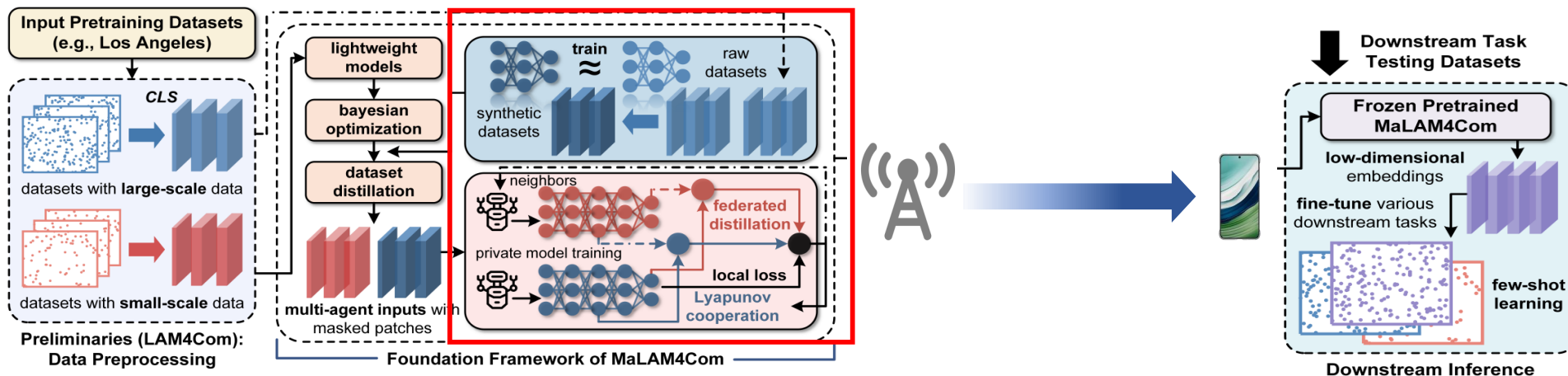
>100 ms

无线通信的苛刻要求

低开销: 无线资源调度、信道估计、波束赋形等任务需要在毫秒内完成决策



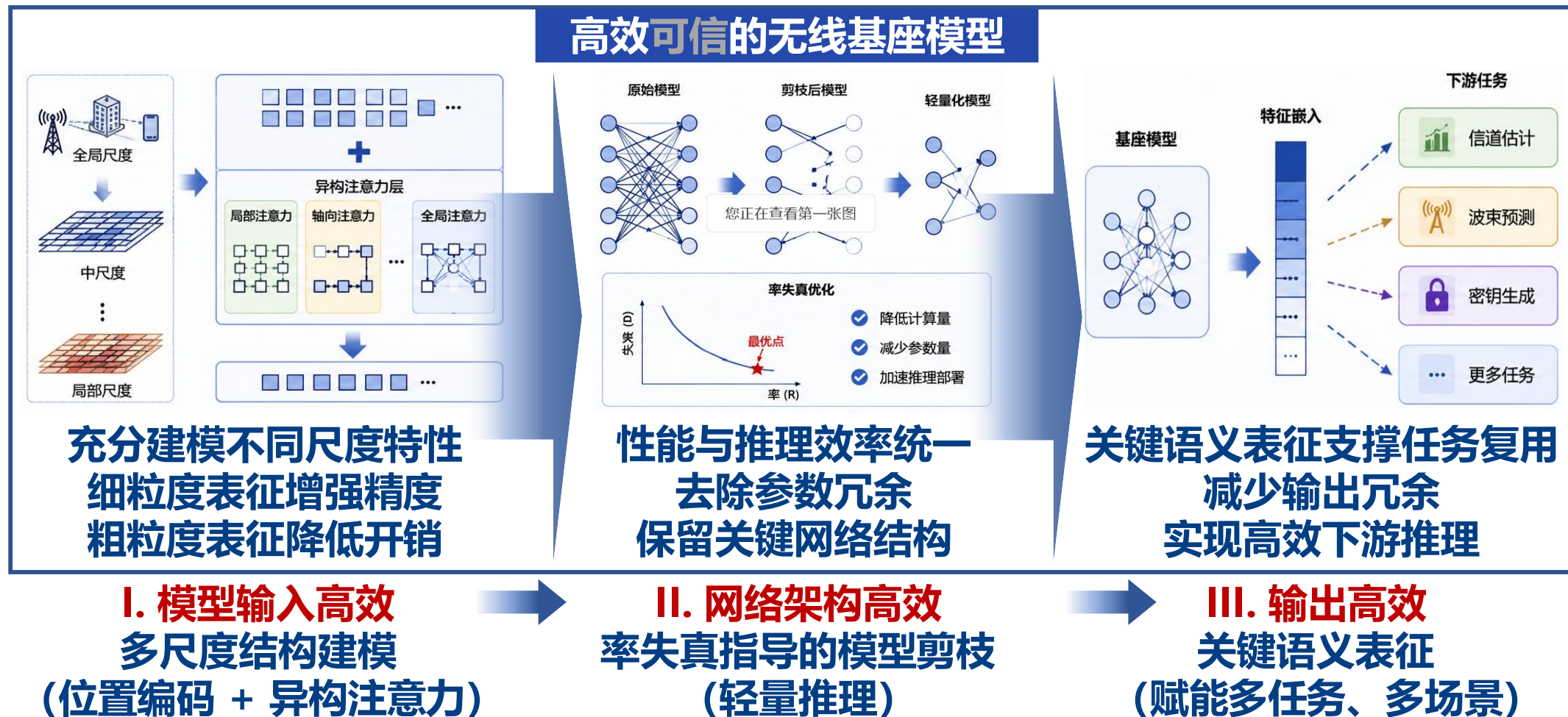
<1 ms



目标：实现毫秒级原生推理

3.1 高效模型架构的必要性

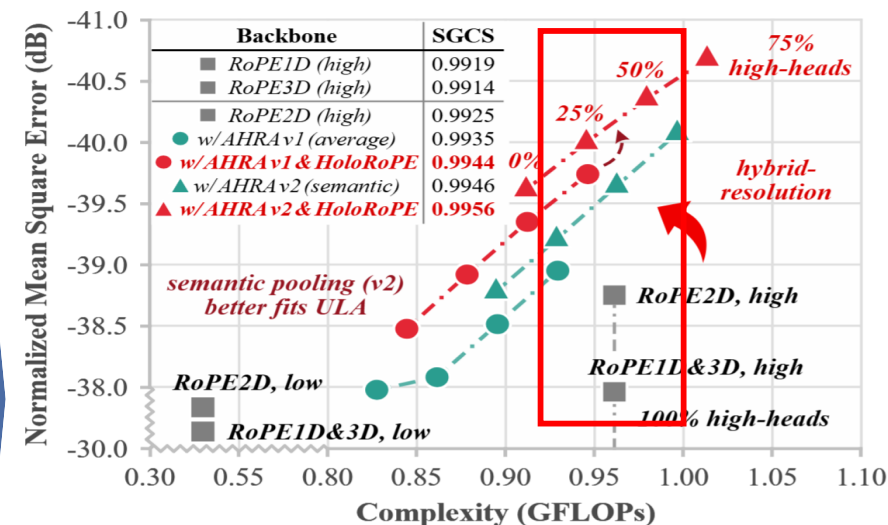
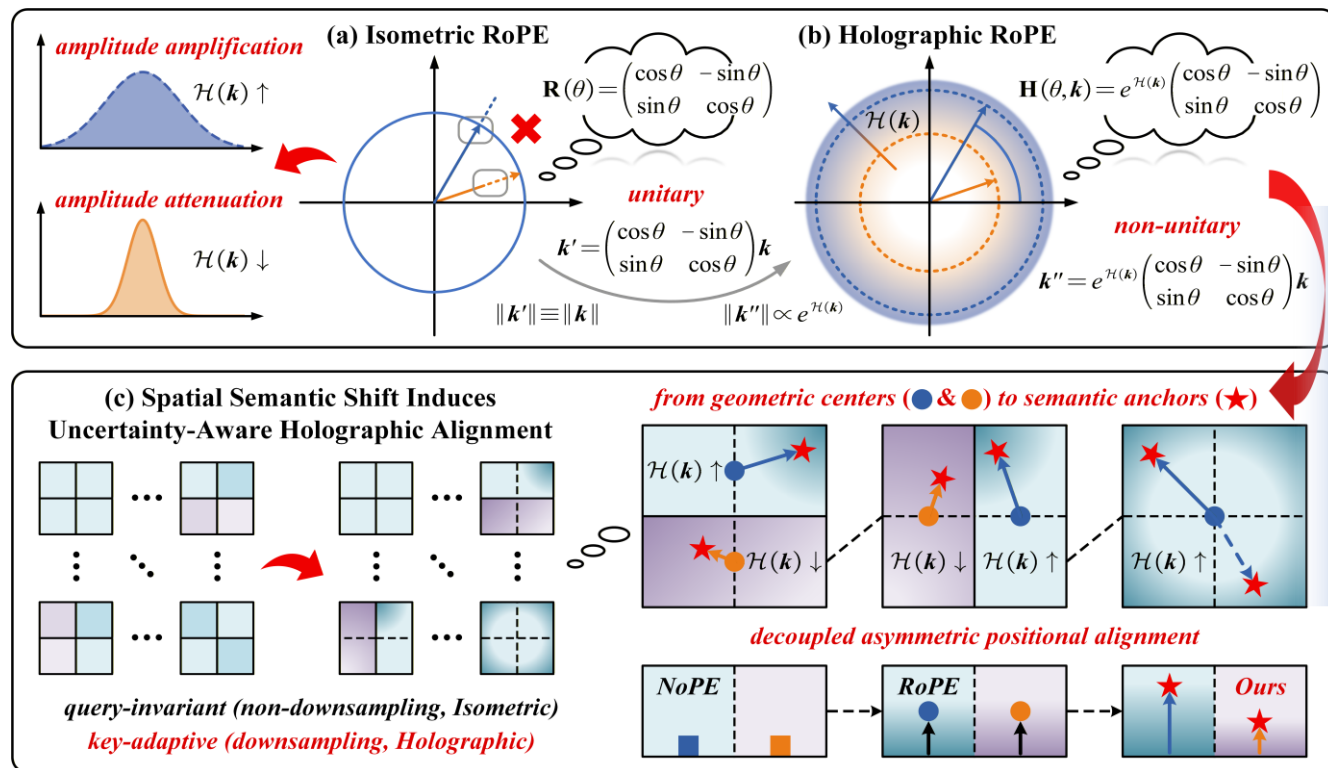
无线基座模型正推动无线智能从**任务专用**迈向**通用赋能**，而高效架构是实现大规模部署与应用的关键
因此，需要从**模型输入**、**网络架构**和**模型输出**三个层面协同设计高效模型架构



3.2 模型高效输入：结构感知的全息位置编码机制

核心
问题

如何实现**细节保真**、**开销降低**与**跨分辨率对齐**的统一？



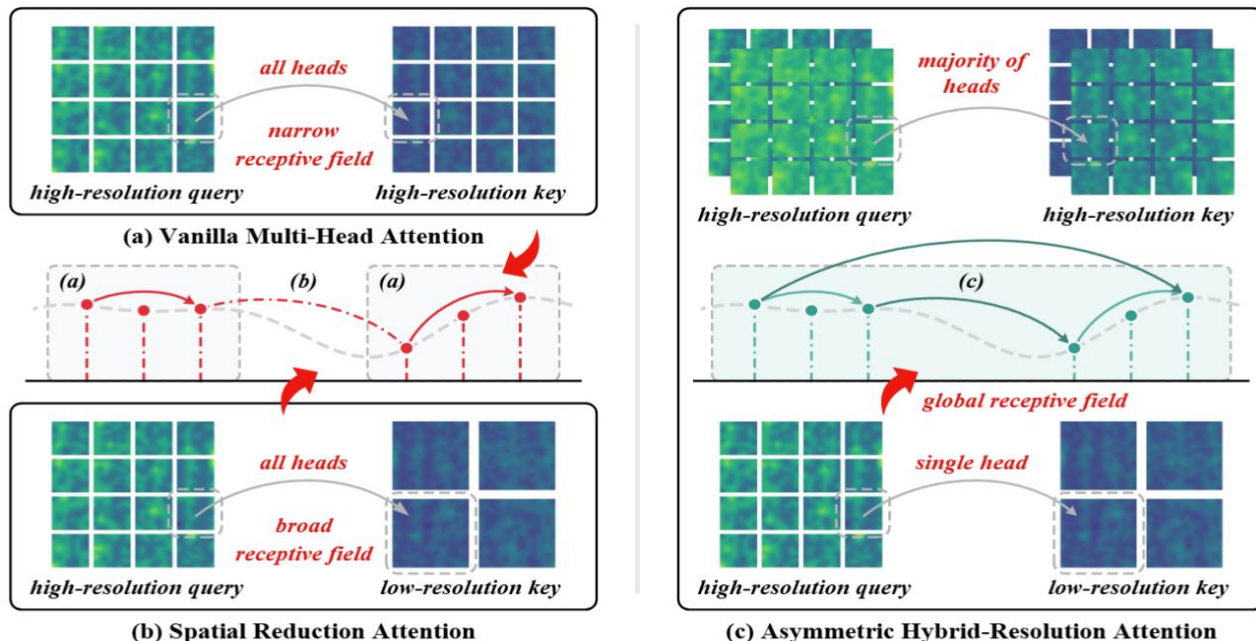
实现低复杂度下
高精度建模能力

全息位置编码赋能多尺度高效架构，支持基座模型轻量化部署

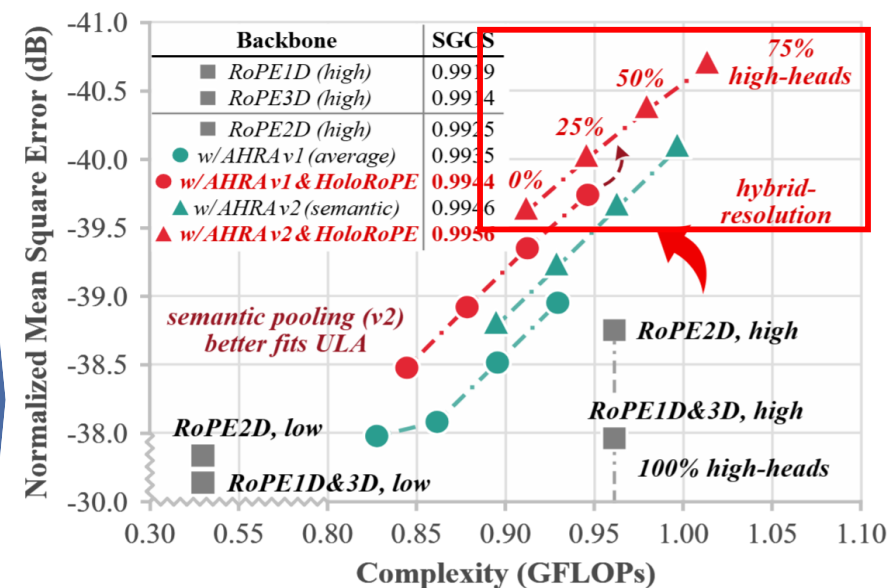
3.2 模型高效输入：结构感知的全息位置编码机制

核心
问题

如何在降低输入复杂度的同时，保持多尺度结构信息？



结构感知位置编码增强多尺度输入表达，
实现高效建模与轻量推理

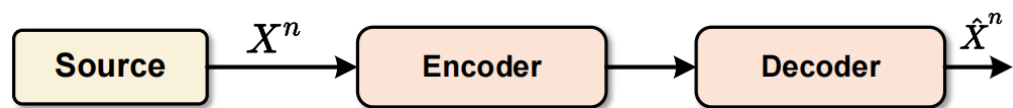


实现精度-复杂度
最优平衡

[4] Ziheng Liu, Jiayi Zhang, Bokai Xu, Chen Zhang, et.al, “Beyond Isometry: Holographic RoPE for Resolution-Agnostic Large Wireless Models,” under review, 2026.

3.3 网络架构高效：率失真理论

有损信源编解码



(a) Lossy Source Coding.

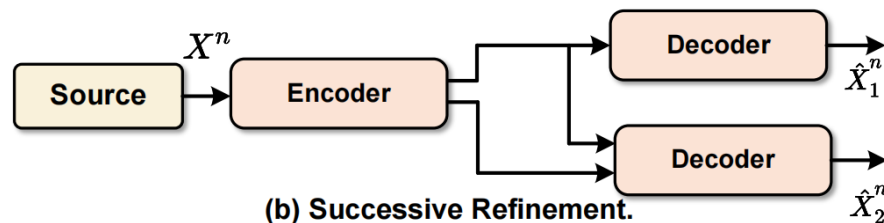
将原始模型压缩到一个目标大小

更加高效



低复杂度

逐次精炼编解码



(b) Successive Refinement.

编码一次，解码多次。
从统一的编码流中逐步解码出不同保真度的版本

- **信源**：原始的、未经剪枝的模型权重 X
- **码率**：剪枝后模型 $\hat{X}$ 的**有效参数量**，码率越低，部署的开销越低，**推理速度**越快
- **失真**：剪枝后模型性能的损失 D

模型剪枝即是对模型权重进行有损压缩

特定失真水平下模型压缩理论极限

压缩后的**模型**与**原始权重**之间的**差异刻画**：

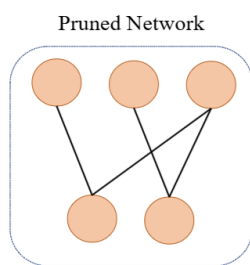
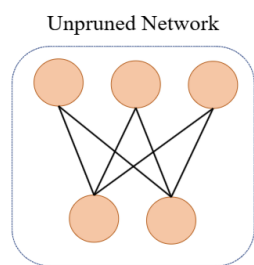
$$\lim_{n \rightarrow \infty} \mathbb{E}[d(X^n, g_n(f_n(X^n)))] \leq D.$$

率失真函数^[1]定义：

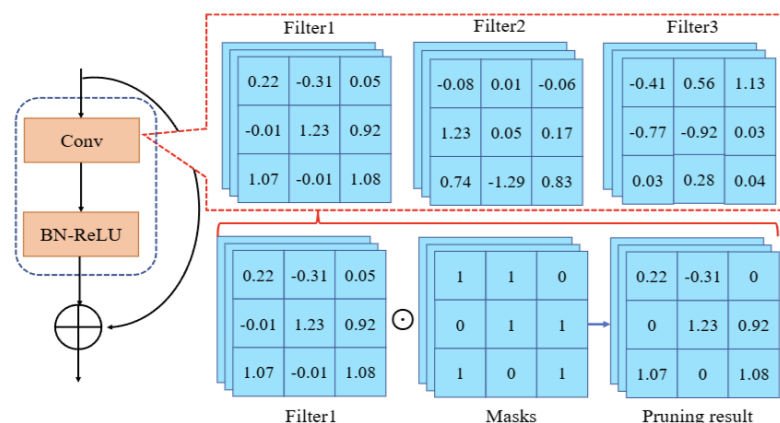
$$\mathcal{R}(D) = \min_{\mathcal{P}(\hat{X}|X): \mathbb{E}[d(x, \hat{x})] \leq D} \mathcal{I}(X; \hat{X}).$$

3.3 网络架构高效：冗余感知的模型剪枝机制

- **理论指导**：借鉴**率失真理论**进行剪枝
- **核心思想**：在可接受的失真范围内，找到**最低的码率**
- **关键挑战**：将率失真理论引入剪枝，建立**数学框架**评估
- **主要问题**：剪掉哪个网络模块，**性能损失最小**，但对模型大小和速度的**提升**最大？

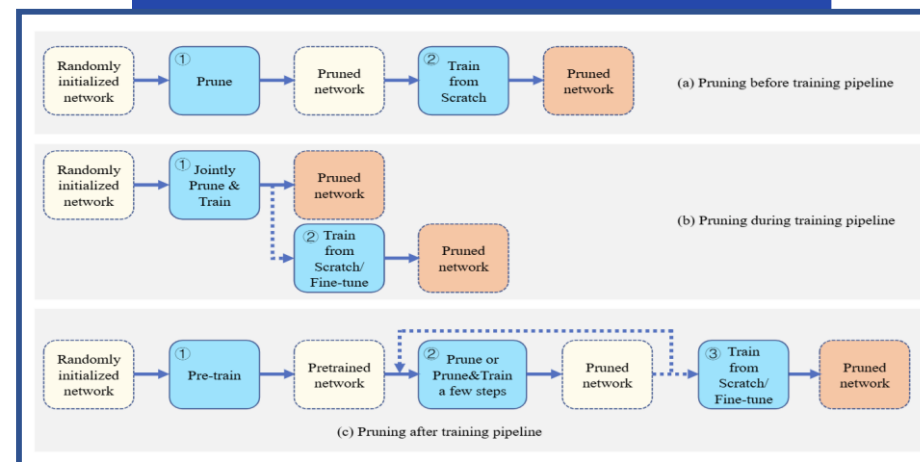


从神经元和连接的角度来看



从权重和掩码的角度来看

不同剪枝方法的比较



训练时剪枝 vs 训练后剪枝

实施训练后剪枝，可以剪枝任何现有模型，而无需增加训练的复杂性，或者重新训练模型

3.3 网络架构高效：模型剪枝的数学模型

- **传统剪枝**：从一个完整的基座模型开始，剪枝掉不重要的部分
- **所提方案**：受**率失真理论**启发，模型从**最稀疏的“骨架”模型**，逐步恢复成在目标压缩率下**性能最优的精简模型**

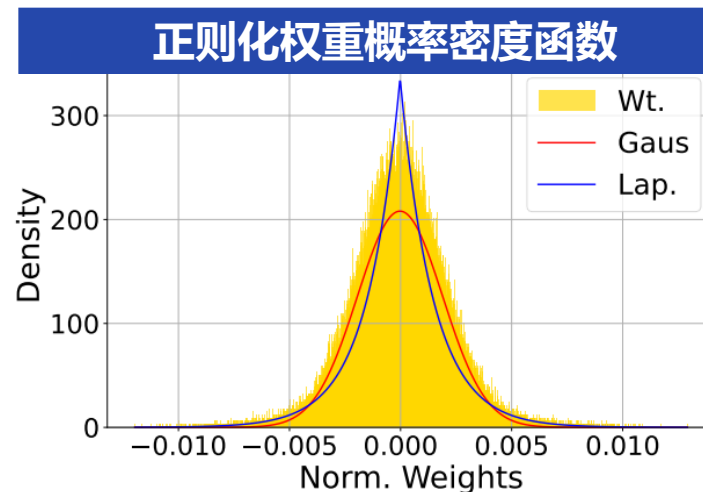
神经网络的数学刻画：

$$\mathcal{F}(\hat{\mathbf{X}}; \mathbf{g}) = \mathbf{g}^{(\mathcal{L})} \sigma \left(\mathbf{g}^{(\mathcal{L}-1)} \sigma \left(\dots \mathbf{g}^{(2)} \sigma \left(\mathbf{g}^{(1)} \hat{\mathbf{X}} \right) \right) \right)$$

模型输出上界：

$$\sup_{\|\hat{\mathbf{X}}\|_1 \leq 1} \|\mathcal{F}(\hat{\mathbf{X}}, \mathbf{g}) - \mathcal{F}(\hat{\mathbf{X}}, \hat{\mathbf{g}})\|_1 \leq \sum_{l=1}^{\mathcal{L}} (\|\mathbf{g}^{(l)} - \hat{\mathbf{g}}^{(l)}\|_1 \prod_{k \neq l} \|\mathbf{g}^{(k)}\|_1).$$

构建**编码器**和**解码器**



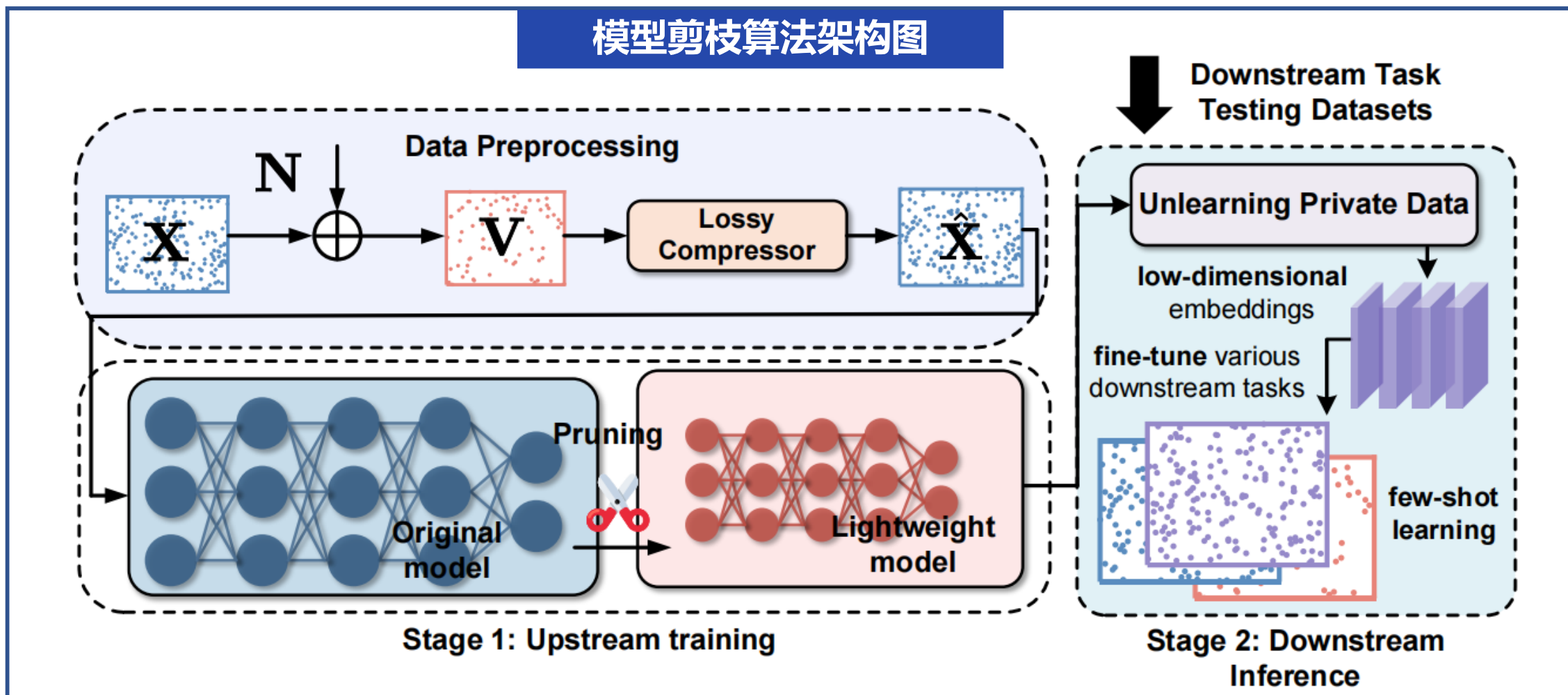
通过拉普拉斯分布对权重进行近似

根据信源权重**拉普拉斯分布**和**失真度量**，得到**率失真函数**

$$\mathcal{R}(D) = \begin{cases} -\ln(\lambda D), & 0 \leq D \leq 1/\lambda \\ 0, & D > 1/\lambda. \end{cases}$$

$$f_{\mathbf{Q}}(q) = \lambda^2 D^2 \delta(q) + (1 - \lambda^2 D^2) \frac{\lambda}{2} e^{-\lambda|q|},$$

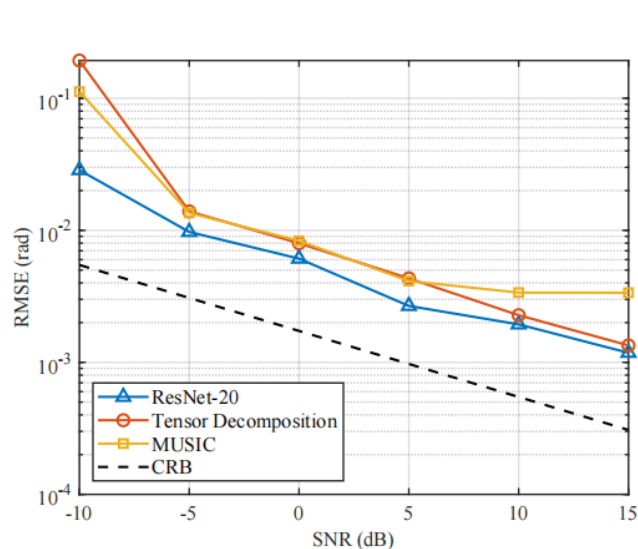
3.3 网络架构高效：模型剪枝的算法设计



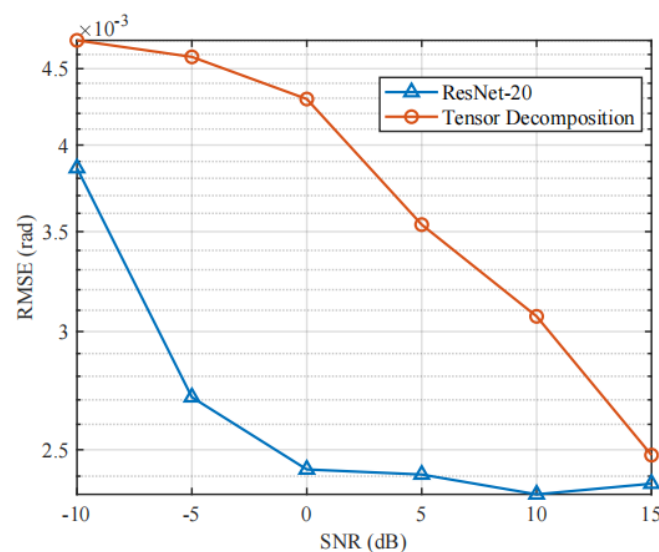
[5] Bokai Xu, **Jiayi Zhang**, Ziheng Liu, Chen Zhang, Derrick Wing Kwan Ng, and Bo Ai, “Towards Device-Side ISAC Model: A Secure and Lightweight Paradigm,” *submitted to IEEE TCOM*, 2026.

3.3 网络架构高效：模型剪枝的仿真结果与分析

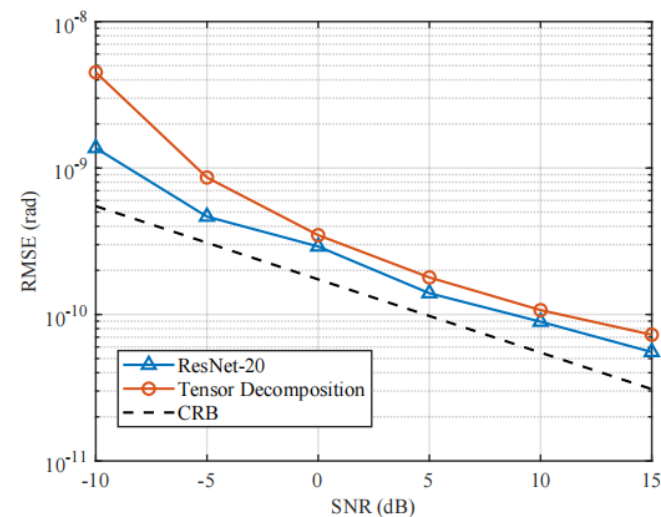
所提感知方案与传统信号处理方案对比



(a) AoA (AoD) estimation.



(b) Velocity estimation.

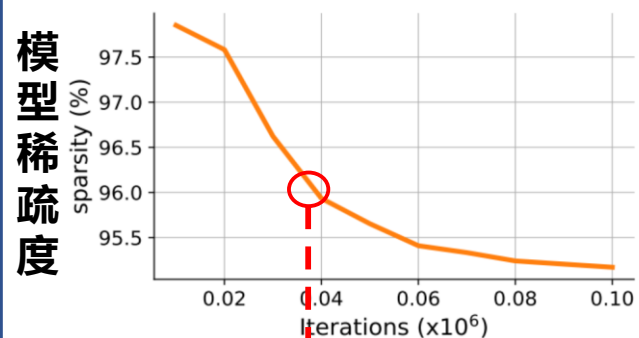


(c) Range estimation.

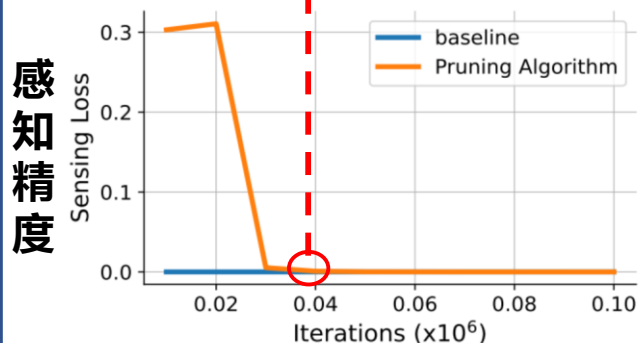
所提出方案优于传统信号处理方案（张量分解， MUSIC算法）

3.3 网络架构高效：模型剪枝的仿真结果与分析

感知稀疏性和精度



(a) Model sparsity.



(b) Model sensing loss.

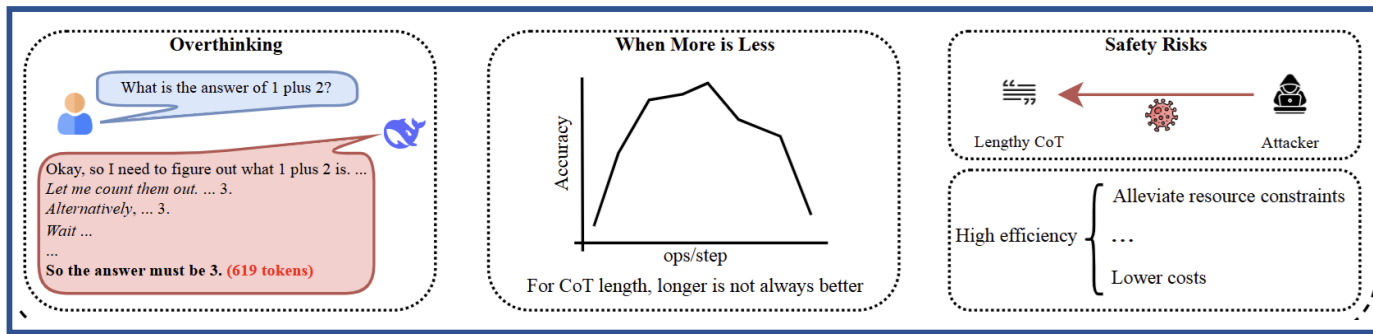
感知精度随剪枝率变化

TABLE III: Sensing Loss of Selected Methods for Summarized Pruning Ratios.

Method	Pruning Ratio										
	10.00%	19.00%	27.10%	40.95%	52.17%	61.25%	68.61%	79.41%	89.05%	96.23%	99.48%
Global	0.000184	0.000184	0.000184	0.000184	0.000187	0.000207	0.00028	0.002066	0.040002	0.21300	0.413103
LAMP	0.000307	0.000307	0.000307	0.00031	0.000322	0.000386	0.000479	0.001586	0.02428	0.1230	0.386148
Erk	0.000189	0.000189	0.000189	0.00019	0.000193	0.000213	0.000283	0.001157	0.016769	0.1123	0.324865
Our proposed	0.0000551	0.0000551	0.0000551	0.0000551	0.0000551	0.0000551	0.0000551	0.0000551	0.0000551	0.00388074	0.302918

所提出的剪枝方案能够保持感知性能不变的情况下，
降低模型复杂度96%

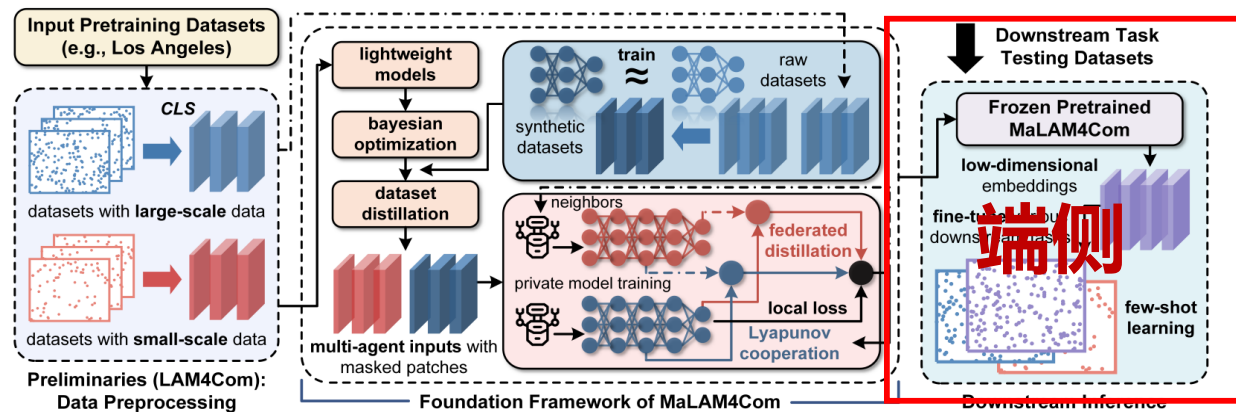
3.4 模型输出高效：高效模型架构的理论瓶颈



高效推理的目标： 在**有限**内存、算力、功耗下，实现**最佳性能**，并降低运行成本

大模型高效推理瓶颈

推理实时性： 动态调整天线波束、分配网络资源，这些决策必须在毫秒内完成，否则通信质量会急剧下降



基座模型推理较慢的原因

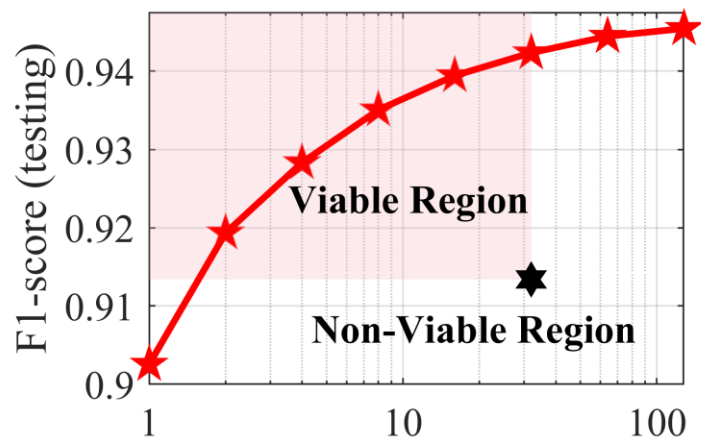
- 网络模型复杂
- 资源严重受限
- Auto-regression decoding机制

基座模型推理解决方案

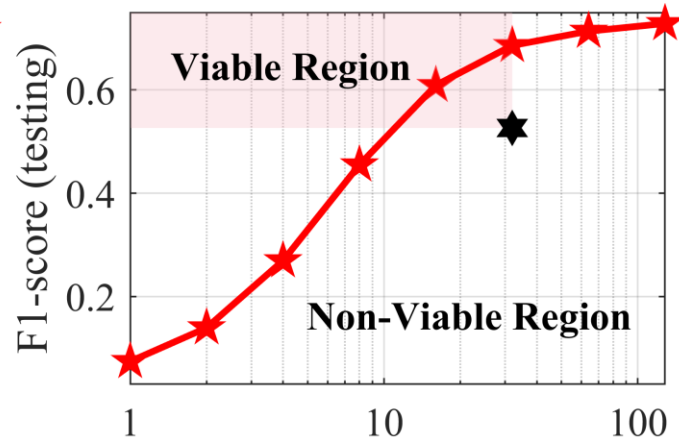
- 模型压缩
- 改变Transformer架构
- 底层算子优化
- **分布式推理架构**

3.4 模型输出高效：关键语义表征支撑任务复用

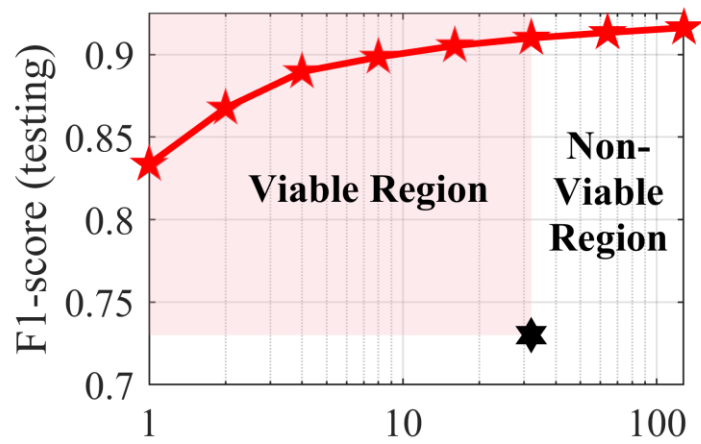
准确率



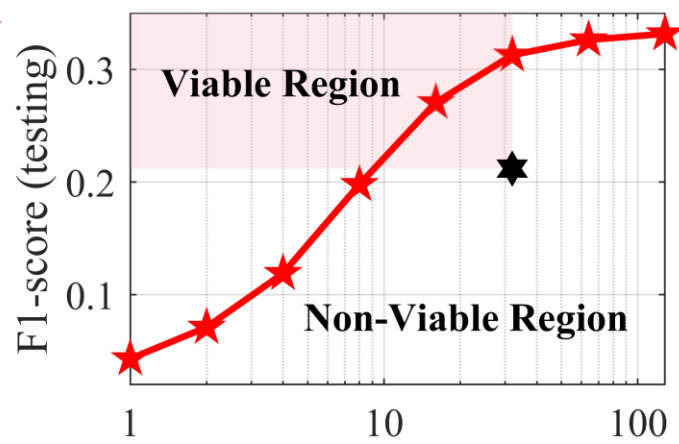
(a) Size of embeddings (5%)



(b) Size of embeddings (5%)



(c) Size of embeddings (0.5%)



(d) Size of embeddings (0.5%)

仿真结果分析

- 将基座模型的低维表征直接用于下游任务，可降低开销**96.88%**
- 随着用的信息越多，基座模型推理开销降低**越小**

表征长度和推理开销成正比

- **表征长度** $p=2D/s$
- **最低推理开销** $= 1/p$

D: 信道总共的元素
p: 元素token划分数
s: 每个token的长度

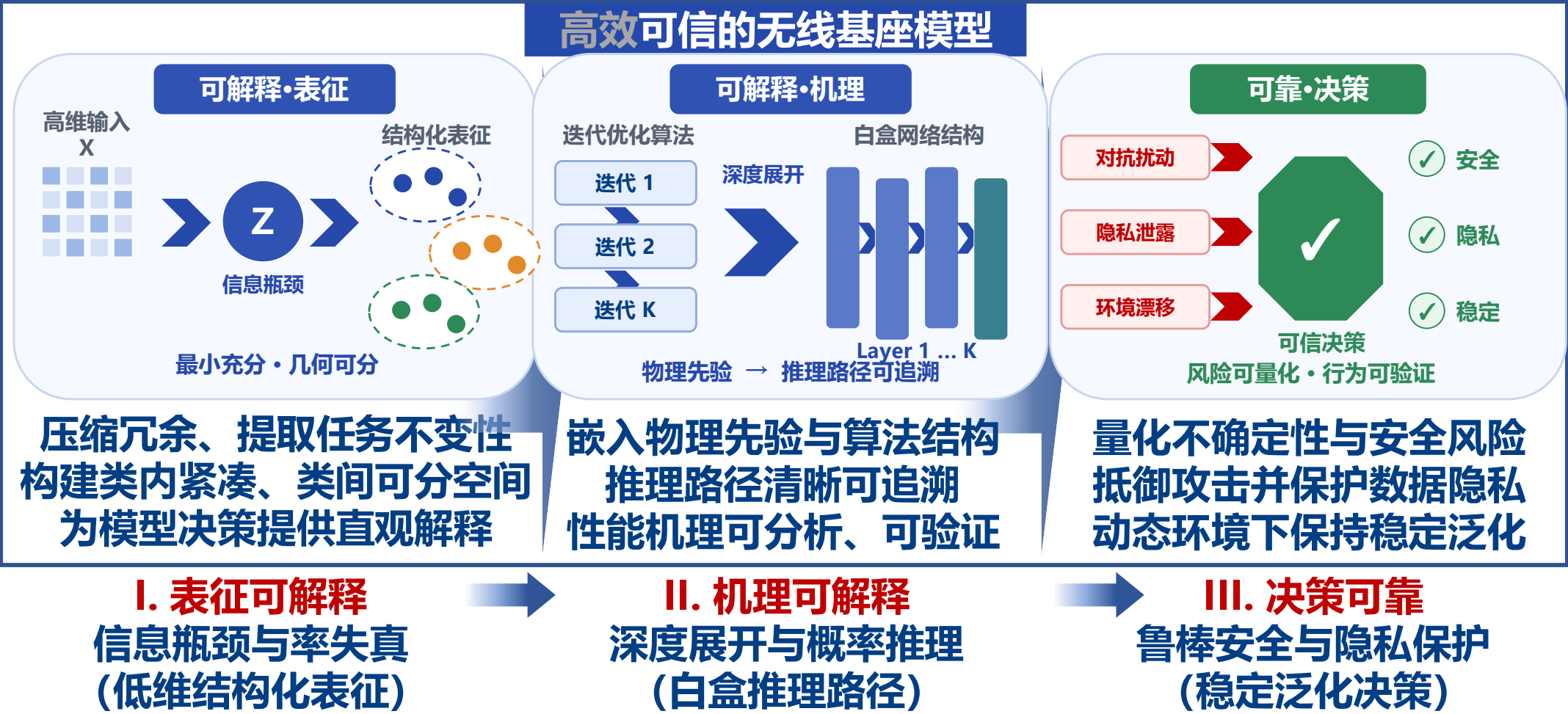
[1] Ziheng Liu, **Jiayi Zhang**, Yiyang Zhu, Enyu Shi, Bokai Xu, Dusit Niyato, Shi Jin, and Bo Ai, “MaLAM4Com: Multi-Agent Cooperative Large AI Models for Wireless Communications,” *IEEE JSAC*, vol. 44, pp. 2889-2903, 2026.

汇报提纲

- I. 研究背景与意义
- II. 无线尺度规律：从规模扩展到能力演进
- III. 高效模型架构：从多维表征到轻量推理
- IV. 可信智能表征：从物理先验到泛化决策**
- V. 总结与展望

4.1 可信模型架构的必要性

无线基座模型正从性能可用迈向可信可用，而黑盒决策、环境扰动与隐私风险制约其规模化部署
因此，需要从表征可解释、机理可解释和决策可靠三个层面协同设计可信模型架构



4.1 可信模型架构的必要性

传统AI (黑箱模型) 的局限性

- **缺乏可解释性:** 学习过程不透明, 优化机理难解释
- **计算复杂度高:** 模型推理开销大, 实时业务难满足
- **泛化能力有限:** 信道环境适应差, 鲁棒性能难保障



破解模型“黑盒”难题

可靠

高效

泛化

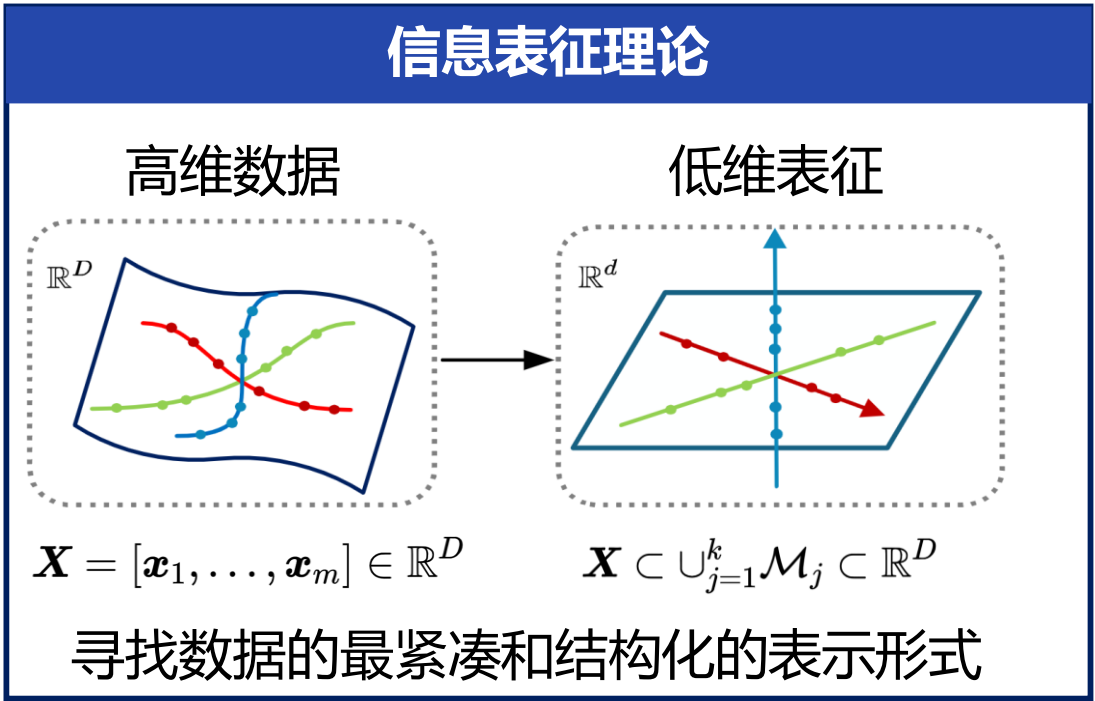
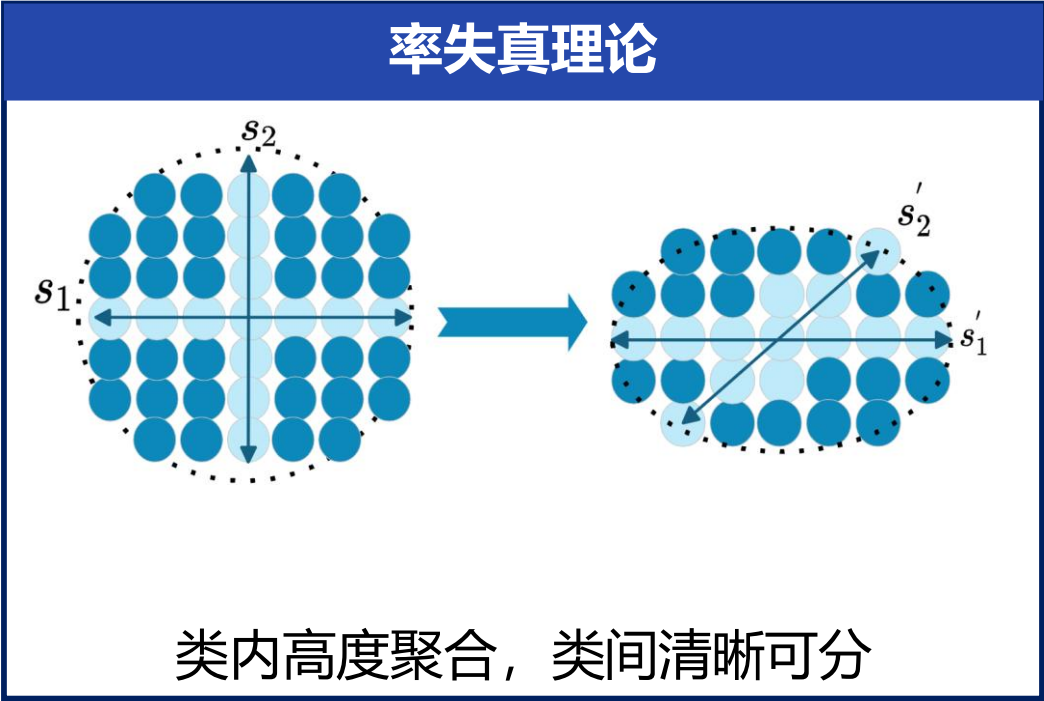


可解释AI (白盒模型) 的优势

- **可解释性:**
揭示内部优化机理, 决策过程清晰可验证
融入通信物理先验, 动态环境性能高鲁棒
- **高效性:**
平衡性能计算开销, 实时业务需求可满足

[6] Jiayao Yang, **Jiayi Zhang**, Bokai Xu, Jiakang Zheng, Zhilong Liu, Ziheng Liu, Dusit Niyato, M érouane Debbah, Zhu Han, and Bo Ai, “White-Box AI for Wireless Communications: Principles, Architectures, and Applications,” *IEEE Network*, 2026.

4.2 表征可解释：信息瓶颈与率失真



- | | | |
|----|--------------|---------------------------|
| 效率 | ➤ 最小化信息编码冗余 | ➤ 从信息论层面降低基座模型的计算复杂度与存储开销 |
| 泛化 | ➤ 提取任务内在不变性 | ➤ 提升模型于未知环境下的泛化能力与鲁棒性 |
| 可信 | ➤ 构建几何可分特征空间 | ➤ 为模型决策提供直观的几何可解释性 |

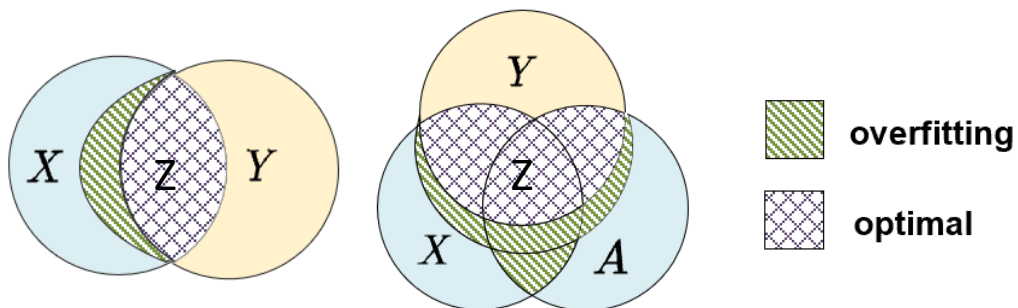
4.2 表征可解释：信息瓶颈与率失真

信息瓶颈理论

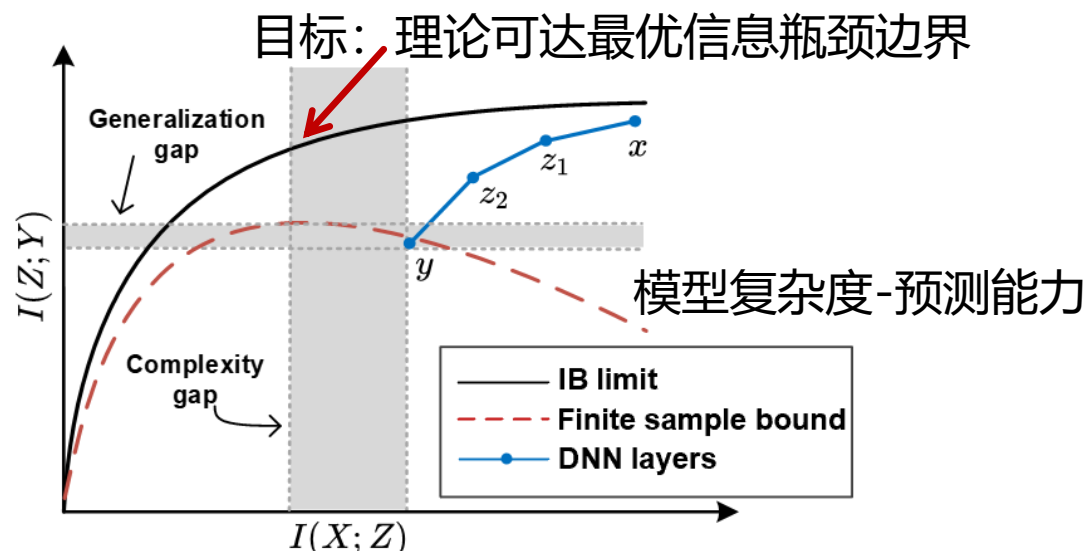
- 提供一种数学框架来**优化隐藏表示 Z**

$$\min_Z \text{IB}(\mathbf{X}, \mathbf{Y}; \mathbf{Z}) = -I(\mathbf{Z}; \mathbf{Y}) + \beta I(\mathbf{X}; \mathbf{Z})$$

- $I(\mathbf{Z}; \mathbf{Y})$: Z 中包含关于目标 Y 的信息
- $I(\mathbf{X}; \mathbf{Z})$: Z 保留关于输入 X 的信息
- 参数 β 权衡项 (调节压缩 vs 表达能力)



信息平面图



定量地度量和解释一个模型的泛化能力与结构效率

- **高效性**: 学习最小化表征, 压缩模型开销
- **鲁棒性**: 学习充分性表征, 增强抗干扰能力

4.2 表征可解释：基于边图信息瓶颈的可信表征学习

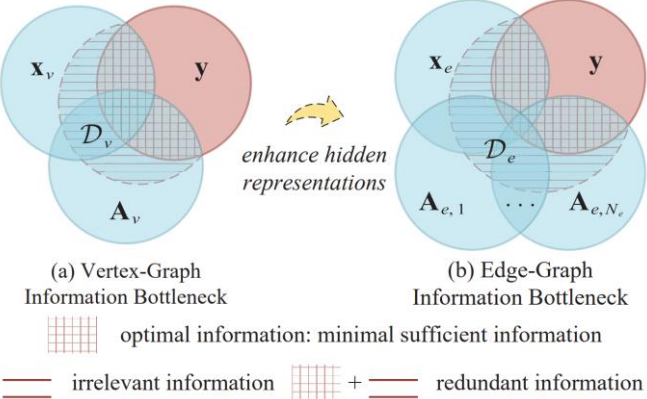
理论指导

对于图结构数据

什么是“好”的隐藏表示？

表征学习: 图信息瓶颈 (GIB) 原理

$$\min_{\mathbf{z}^L} \text{IB}(\mathbf{x}, \mathbf{y}; \mathbf{z}^L) \triangleq -I(\mathbf{y}; \mathbf{z}^L) + \beta I(\mathbf{x}; \mathbf{z}^L)$$



所提方案

➤ 顶点图瓶颈 (VGIB) 局限:

信道拓扑压缩到“用户”节点
造成信道结构信息损失

$$\min_{\mathbb{P}(\mathbf{z}_v^L | \mathcal{D}_v) \in \Omega_v} \text{VGIB}(\mathcal{D}_v, \mathbf{y}; \mathbf{z}_v^L) \triangleq -I(\mathbf{y}; \mathbf{z}_v^L) + \beta I(\mathcal{D}_v; \mathbf{z}_v^L)$$

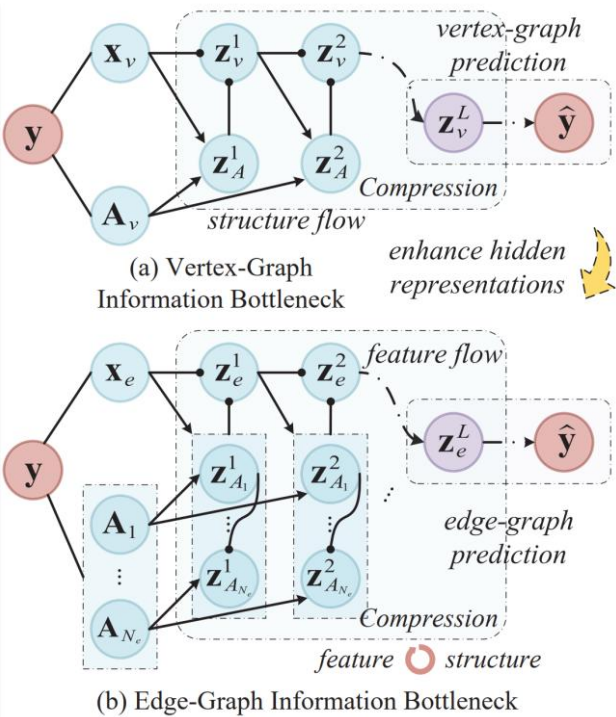
➤ 边图信息瓶颈 (EGIB):

保留信道的拓扑结构

解决了VGIB的信息损失问题

$$\min_{\mathbb{P}(\mathbf{z}_e^L | \mathcal{D}_e) \in \Omega_e} \text{EGIB}(\mathcal{D}_e, \mathbf{y}; \mathbf{z}_e^L) \triangleq -I(\mathbf{y}; \mathbf{z}_e^L) + \beta I(\mathcal{D}_e; \mathbf{z}_e^L)$$

模型架构

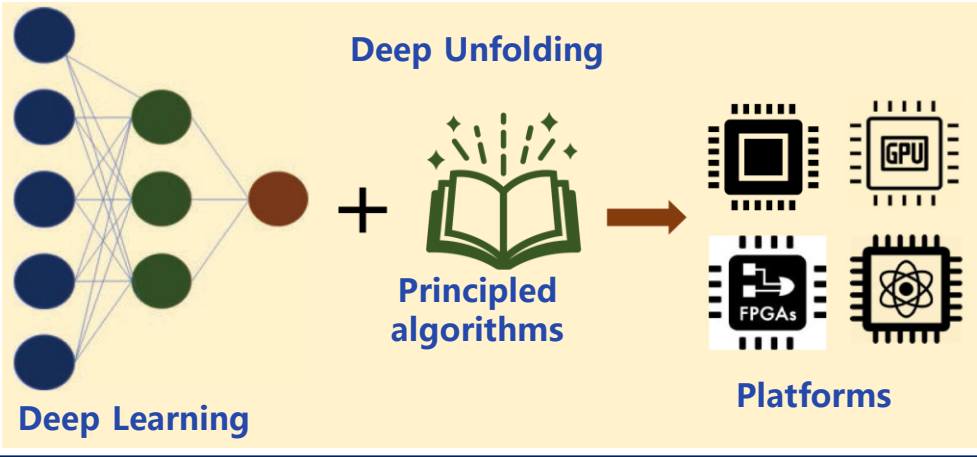


[7] Ziheng Liu, **Jiayi Zhang**, Yiyang Zhu, Enyu Shi, and Bo Ai, “Robust Multidimensional Graph Neural Networks for Signal Processing in Wireless Communications With Edge-Graph Information Bottleneck,” *IEEE TSP*, vol. 73, pp. 2688-2703, 2025.

4.3 机理可解释：深度展开与概率推理

理论优势

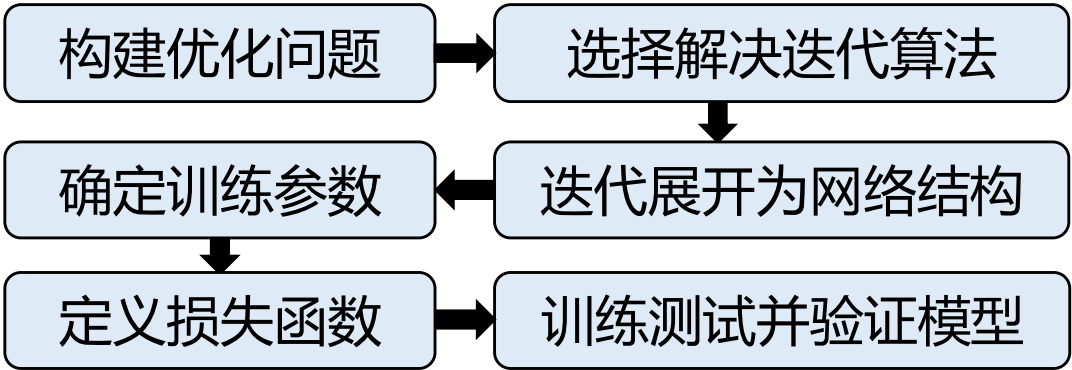
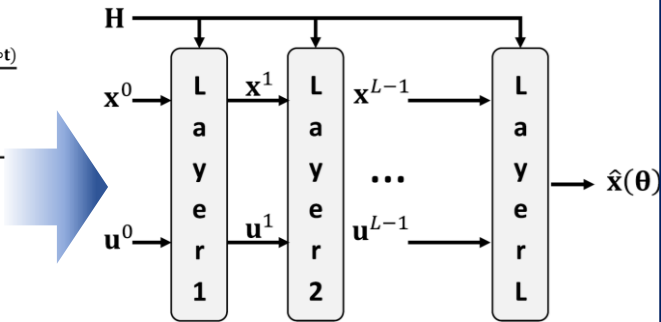
- **可解释性**：架构源于算法机理，白盒建模
- **高效性**：天然的轻量化结构，高效适配
- **可靠性**：继承算法理论保障，性能稳定
- **易训练性**：先验注入学习结构，快速收敛



理论原理

Algorithm Input: $\mathbf{x}^0, \mathbf{u}^0, \mathbf{H}$
Given: $\mathbf{D} = (\text{diag}(\mathbf{w}^0 \circ \mathbf{p}) + \mathbf{H}^T \mathbf{H})^{-1}$
 $\mathbf{p} = [\|\mathbf{h}_1\|_2^2, \|\mathbf{h}_2\|_2^2, \dots, \|\mathbf{h}_N\|_2^2]^T, \Gamma_{\beta, \mathbf{t}}(\mathbf{x}) = -\beta + \frac{\rho(\mathbf{x} + \beta \circ \mathbf{t})}{t} - \frac{\rho(\mathbf{x} - \beta \circ \mathbf{t})}{t}$
Learnable parameters: $\theta = \{\{\mathbf{w}^l, \mathbf{t}_l\}_{l=1}^L, \{\beta_l\}_{l=1}^L\}$
Output: $\hat{\mathbf{x}}(\theta)$

for $l = 1, 2, \dots, L$:

$$\mathbf{z}^l = \Gamma_{\beta_l, \mathbf{t}_{l-1}}(\mathbf{x}^{l-1} - \mathbf{u}^{l-1})$$
$$\mathbf{u}^l = \mathbf{z}^l - \mathbf{x}^{l-1} + \mathbf{u}^{l-1}$$
$$\mathbf{x}^l = \mathbf{D}(\mathbf{H}^T \mathbf{y} + (\mathbf{w}^l \circ \mathbf{p}) \circ (\mathbf{z}^l + \mathbf{u}^l))$$
$$\hat{\mathbf{x}}(\theta) = \Gamma_{1, \mathbf{t}_L}(\mathbf{x}^L)$$


4.3 机理可解释：深度展开与概率推理

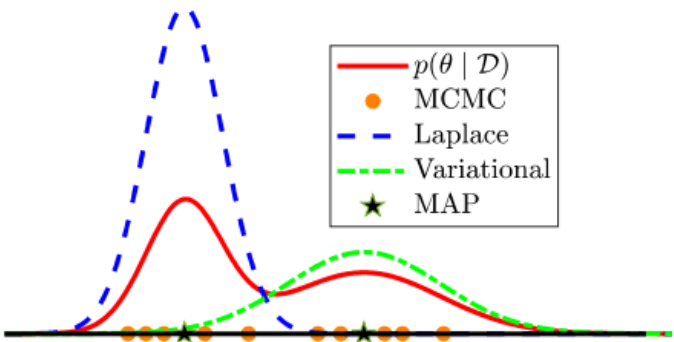
概率统计理论指导

用**概率模型**替代黑盒的 “**不可控经验拟合**”

贝叶斯理论

变分推断

- 推理过程由概率公式驱动
- 具备内在的理论可解释性与稳定性



概率统计理论优势

- **安全性**：贝叶斯驱动，量化不确定性
输出关键参数后验分布，避免点估计过自信
- **高效性**：消息传递，实现结构化推理
利用图拓扑迭代信息，高效近似联合分布
- **可解释性**：白盒结构，保证路径可追溯
更新规则与信息流明确，具备高结构可解释

[8] Bokai Xu, **Jiayi Zhang**, Qingfeng Lin, Huahua Xiao, Yik-Chung Wu, and Bo Ai, “Deep Unfolding Beamforming and Power Control Designs for Multi-Port Matching Networks,” *IEEE TWC*, vol. 24, no. 2, pp. 1401-1414, 2025.

4.3 机理可解释：深度展开的模型架构

算法对比

- 传统迭代WMMSE波束成形算法
 - 收敛慢、计算复杂度高、信令开销大

Algorithm 1 WMMSE Beamforming Algorithm

Input: CSI between AP m and UE k $\{\mathbf{H}_{m,k}\}$;

- 1: Initialize $\{\mathbf{F}_{m,k}\}$;
- 2: Update auxiliary variables $\{\mathbf{U}_{kc}^*\}$ via Eq. (16);
- 3: Update auxiliary variables $\{\mathbf{W}_{kc}^*\}$ via Eq. (17);
- 4: Update beamforming variables $\{\mathbf{F}_{m,k}^*\}$ via Eq. (18);
- 5: **Until** Converge

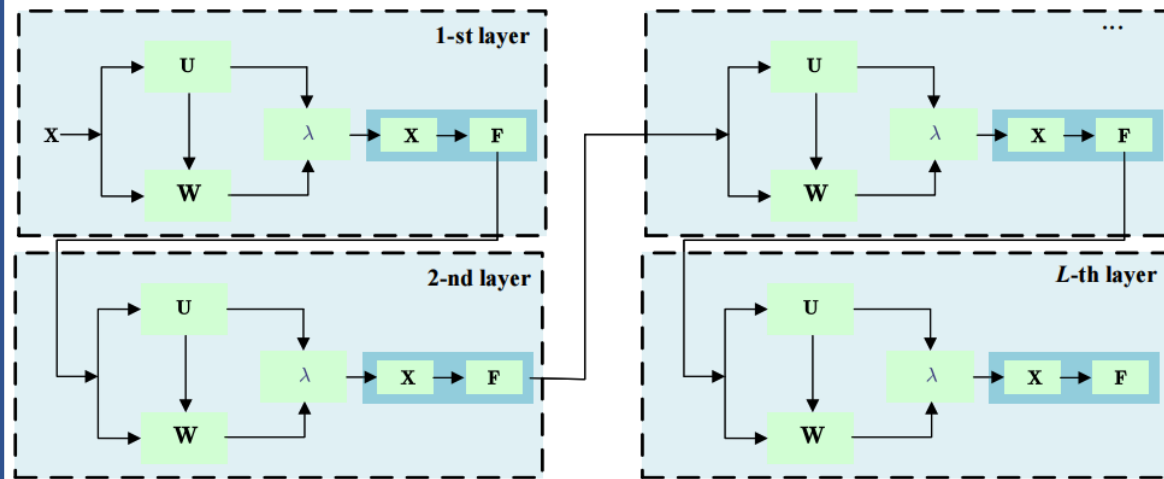
Output: CPU transmit $\{\mathbf{F}_{m,k}^*, k \in \mathcal{K}_m\}$ to each AP m .

$$\mathbf{U}_{kc} = (\mathbf{Q}_{kc} + \mathbf{H}_{kc} \mathbf{F}_{kc} \mathbf{F}_{kc}^H \mathbf{H}_{kc}^H)^{-1} \mathbf{H}_{kc} \mathbf{F}_{kc}, \forall k, \forall c.$$

$$\mathbf{W}_{kc} = (\mathbf{I}_{d_k^c} - \mathbf{U}_{kc}^H \mathbf{H}_{kc} \mathbf{F}_{kc})^{-1}, \forall k, \forall c.$$

$$\mathbf{F}_{mk} = (\mathbf{\Omega}_{mk}^c + \lambda_m \mathbf{I}_{N_r})^{-1} \alpha_k (\mathbf{H}_{mk}^H \mathbf{U}_{kc}^m \mathbf{W}_{kc}^m + \mathbf{\Xi}_{mk})$$

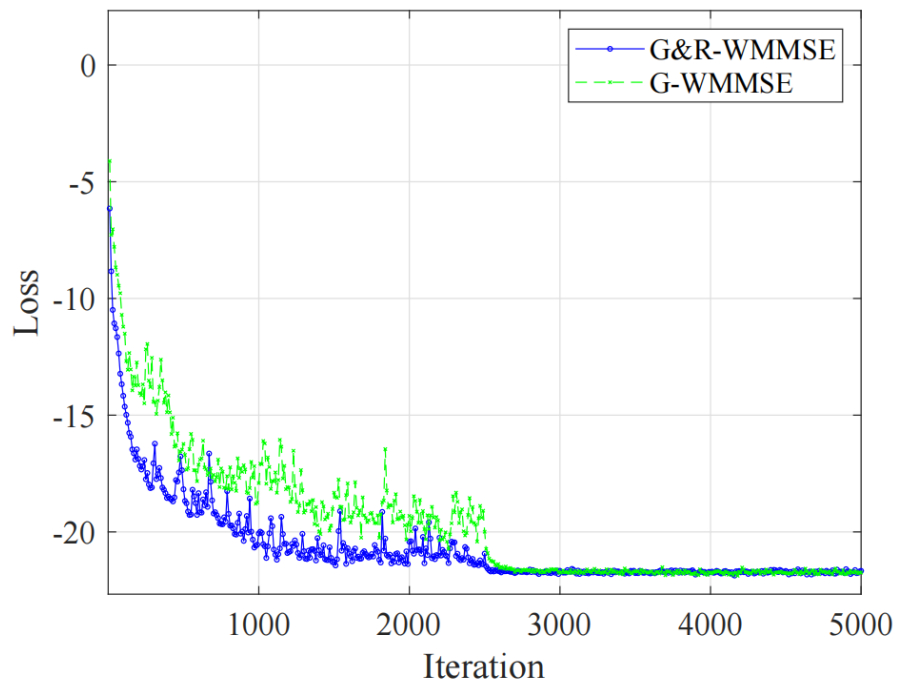
深度展开



- 算法展开：迭代过程 ➡ 网络层级结构
- 物理嵌入：矩阵更新 ➡ 图消息传递
- 端到端优化：固定参数 ➡ 网络可训练权重

4.3 机理可解释：深度展开的仿真结果

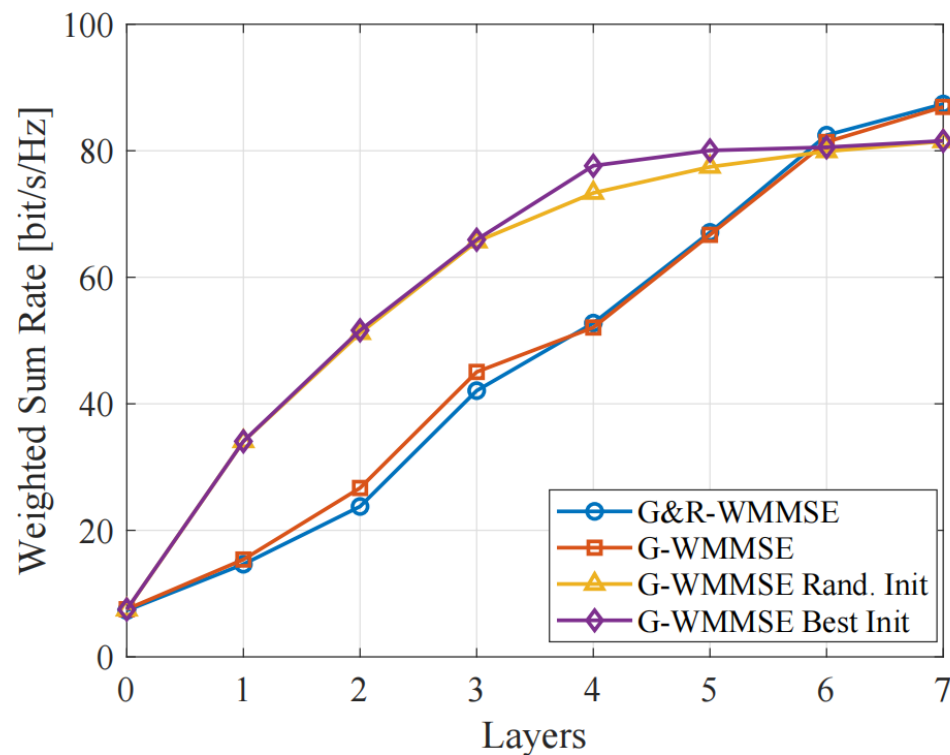
收敛性分析



➤ 损失值快速下降，更低损失值，高效收敛

G&R-WMMSE 通过深度展开技术，
将R-WMMSE映射为图卷积网络层级结构

性能对比



G&R-WMMSE算法
实现了更高的下行加权和速率性能

4.4 决策可靠：鲁棒安全与隐私保护

数据被对抗攻击

核心： 恶意构造无线信号扰动，欺骗并操纵部署在基站或终端的AI模型



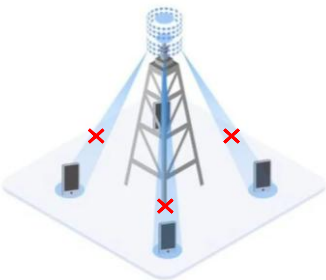
敏感数据需要遗忘

核心： 响应用户的被遗忘权，从无线基座模型中彻底、永久地抹除其数据组成



核心矛盾

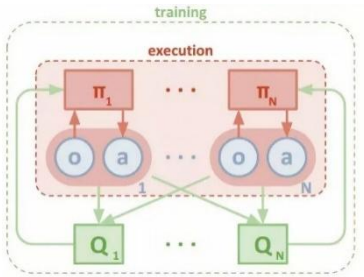
人工智能模型对海量数据的内在依赖性
VS
个人隐私保护的强约束性之间



波束赋形失败



网络瘫痪

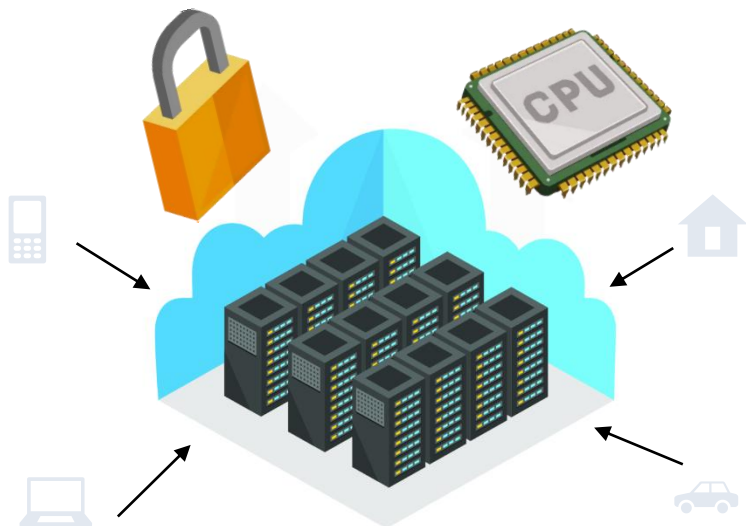


重训模型

成本高昂

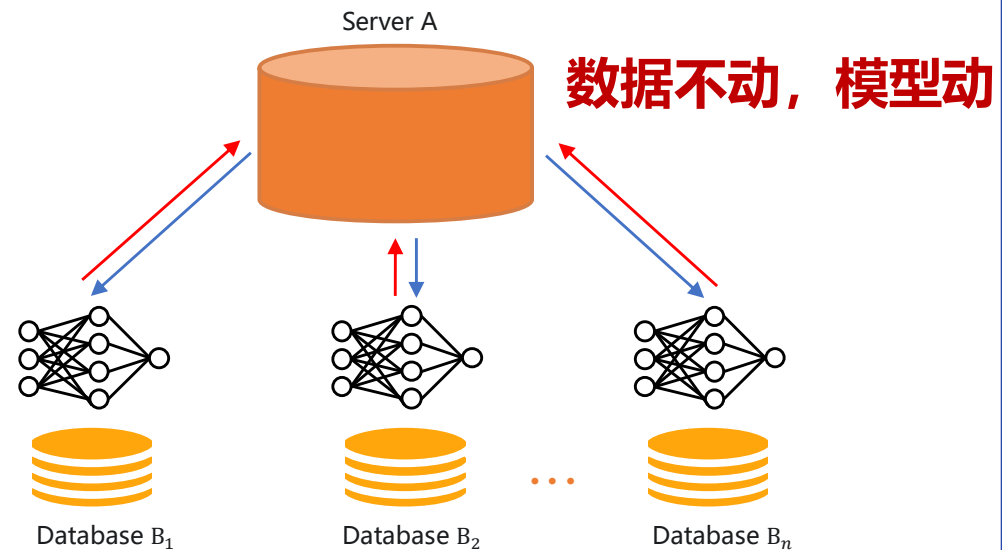
4.4 决策可靠：鲁棒安全与隐私保护

集中式模型训练



数据收集->隐私噪声注入->模型训练->安全保障

分布式模型训练

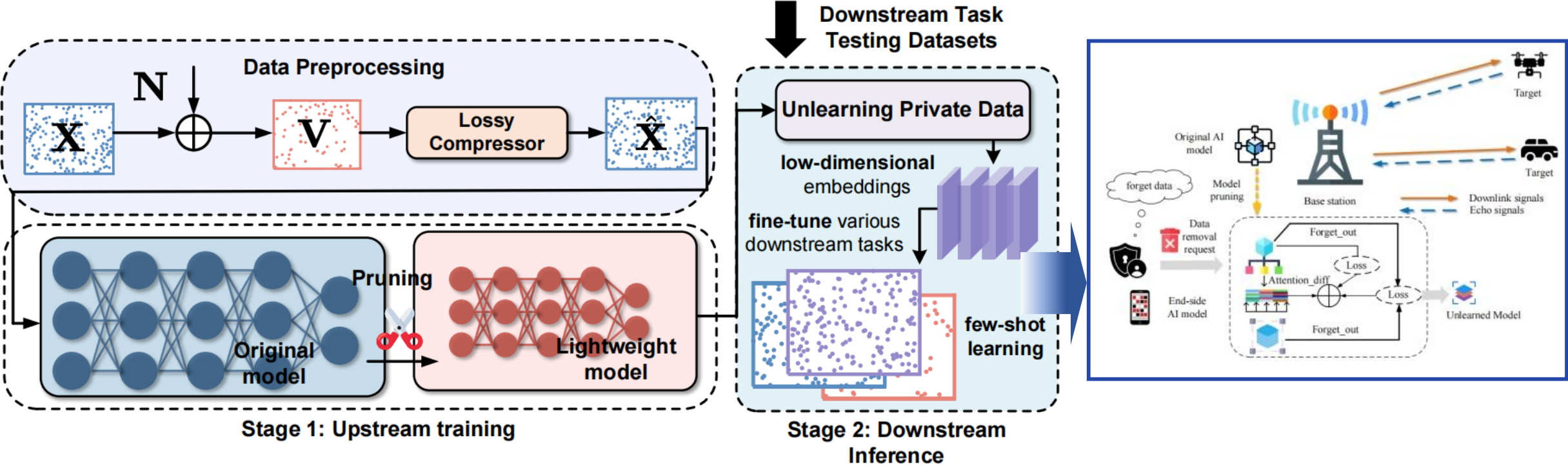


模型分发->本地训练->参数上传->安全聚合

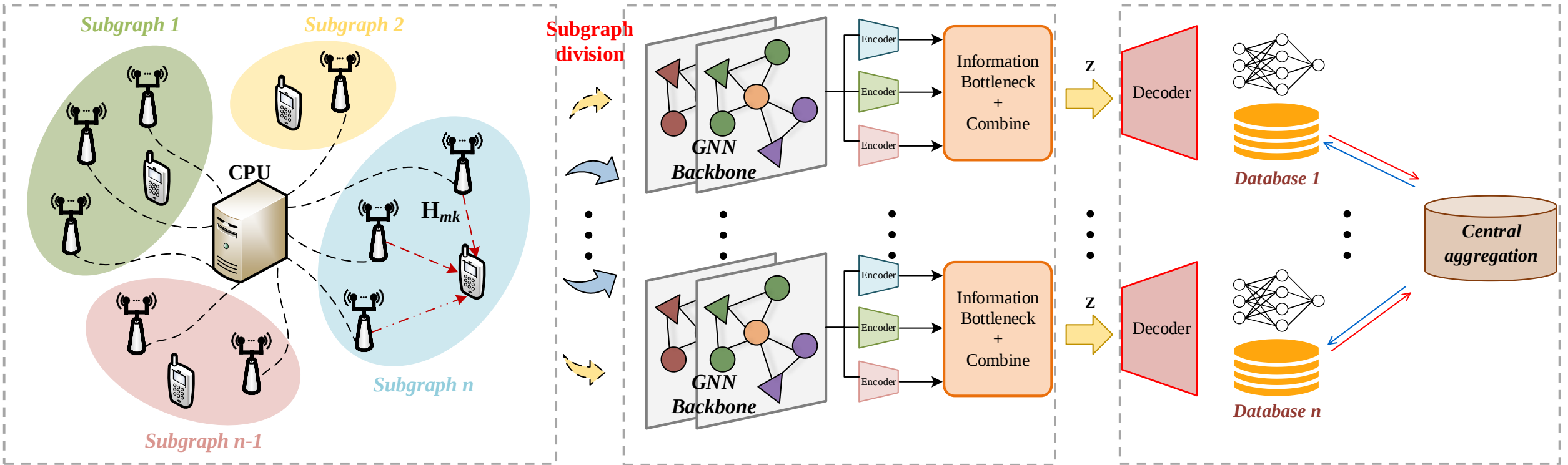
4.4 决策可靠：数据加噪

数据预先处理

在数据进入训练流程前，于服务器端进行清洗，切断数据与个人身份的直接关联



4.4 决策可靠：联邦无线网络



子图划分

$$\arg \min_{C_1, \dots, C_n} \sum_{i=1}^n \sum_{l \in C_i} \| \mathbf{h}_l - \mu_i \|^2$$

结构化特征压缩

$$\min_{\theta, \phi} -I(Z; Y) + \beta I(Z; X)$$

联邦聚合隐私保护

$$\mathbf{P} = \sigma \left(\sum_{i=1}^n \alpha_i \mathcal{F}_{\theta}(\mathbf{z}_i, \mathbf{M}_i) \right)$$

4.4 决策可靠：机器遗忘的仿真与分析



TABLE III: Comparison of Different Unlearning Methods on Various Metrics for Different Forget Sets. ($Err_{\mathcal{D}_r}$: Prediction error on retain set, $Err_{\mathcal{D}_f}$: Prediction error on forget set, Att_{prob} : Membership inference attack probability on forget set, $\mathcal{W}_{dist}$: Wasserstein distance between the unlearned and retrained model predictions on the forget set, AIN: Anamnesis Index.)

Forget Set	Metric	Original	Retrained	FineTune	Gaussian Amn	Blindspot
Angle range: $[0^\circ, 60^\circ]$	$Err_{\mathcal{D}_r} \downarrow$	0.0000590	0.0000178	0.0002302	0.0002962	0.000204913
	$Err_{\mathcal{D}_f} \uparrow$	0.000306	1.5549217	2.4666	1.53608477	1.52521724626
	$Att_{prob} \downarrow$	0.4565	0.5305	0.6675	0.6685	0.65
	$\mathcal{W}_{dist} \downarrow$	1.21707	-	0.31934215	0.07611	0.03601
	AIN $\uparrow$	-	-	0.3703703	0.63333333	0.91803
Velocity range: $[0, 30\text{m/s}]$	$Err_{\mathcal{D}_r} \downarrow$	0.000538320	0.0001256	0.000373286	0.00012542	0.000103156
	$Err_{\mathcal{D}_f} \uparrow$	0.000528796	0.00126219	0.00058530	0.0009765	0.0013157
	$Att_{prob} \downarrow$	0.043	0.498	0.265	0.5115	0.6335
	$\mathcal{W}_{dist} \downarrow$	0.0171629	-	0.0138161	0.00447313	0.0031209
	AIN $\uparrow$	-	-	0.5	0.5	0.5

无记忆指数

预测误差

所提方案能够实现对用户数据的有效遗忘

4.4 决策可靠：联邦无线网络的分层协同计算的仿真

子图划分

基于信道增益，将大规模网络动态分解为多个子图

结构化特征压缩

GNN提取信道特征，IB将高维特征压缩为低维z向量

联邦聚合

各子图传递z向量，上传轻量级的模型进行协同聚合

主要复杂度计算

集中式全局最优： $O(L \cdot K^2 \cdot M + K^3)$



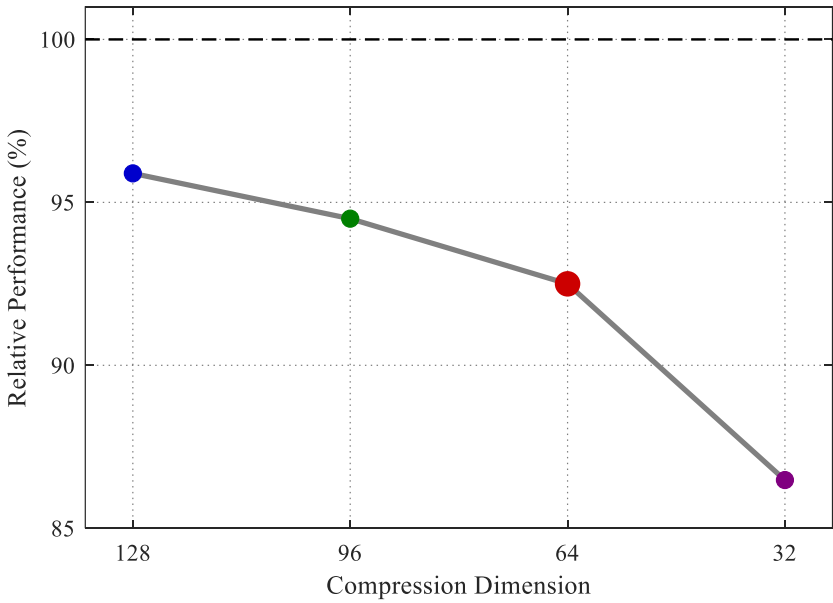
n =子图数, K_s =子图内UE数,
 z =压缩维度(128, 96, 64, 32)

联邦压缩隐私： $O(n \cdot K_s^2) + O(K \cdot z)$

通信开销

$O(L \cdot K \cdot M) \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow O(n \cdot z)$

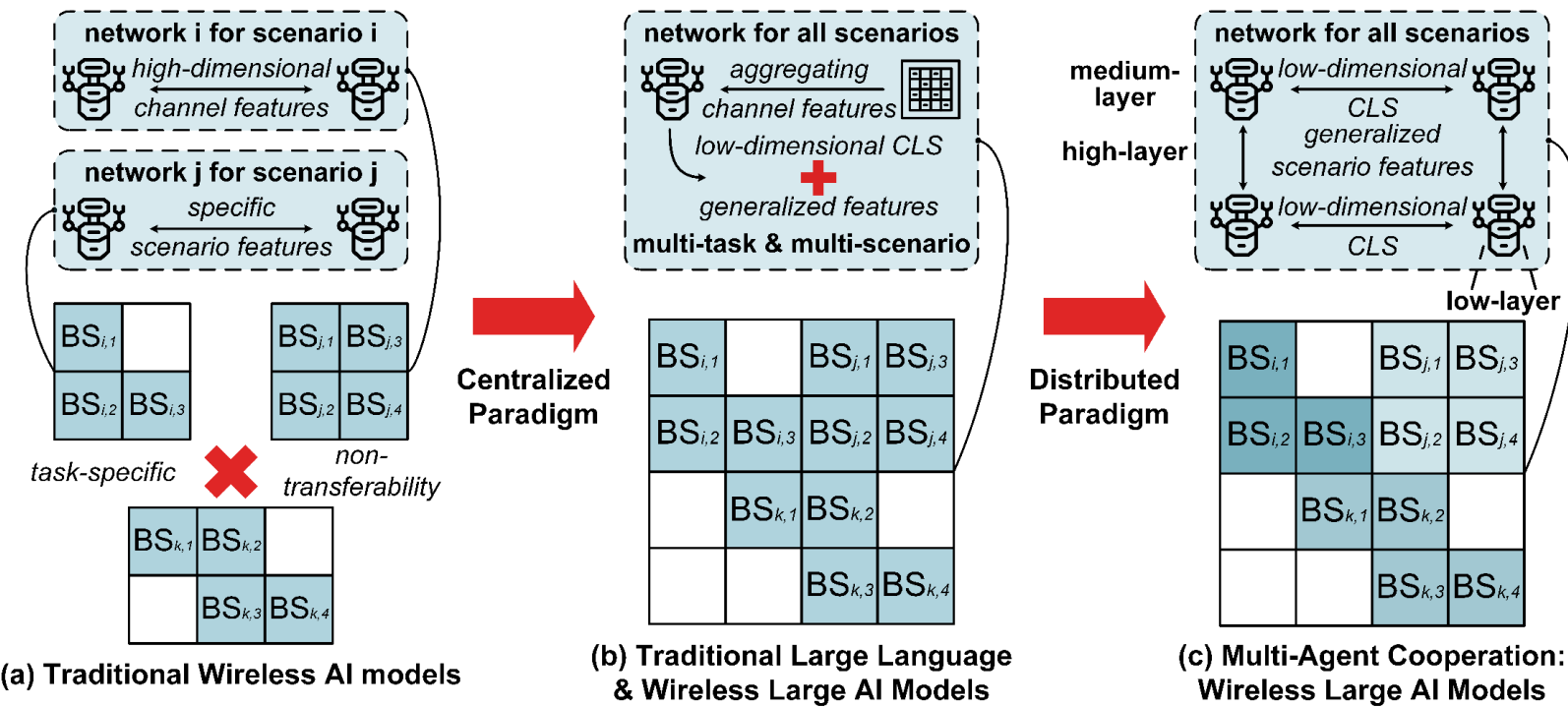
压缩维度



4.5 决策可靠应用：分布式协同基座模型

集中式基座模型的挑战：

- 资源开销巨大：终端设备难以承载**海量算力和参数**的WLAMs
- 训练效率低下：单一模型训练周期长，难以实时适应**动态**网络



分布式协同：

- **多智能体(Multi-Agent)：**将网络中的每个设备视为一个智能体，每个智能体拥有一个**轻量化的本地模型**
- **协同机制：**通过高效的信息交互共享知识，实现“1+1>2”的**群体智能**
- ✓ **优势：**显著提升**可扩展性、训练效率**，并保护**数据隐私**

4.5 决策可靠应用：分布式协同基座模型

TABLE I: Foundation Frameworks of Different AI Models for Wireless Communication Systems.

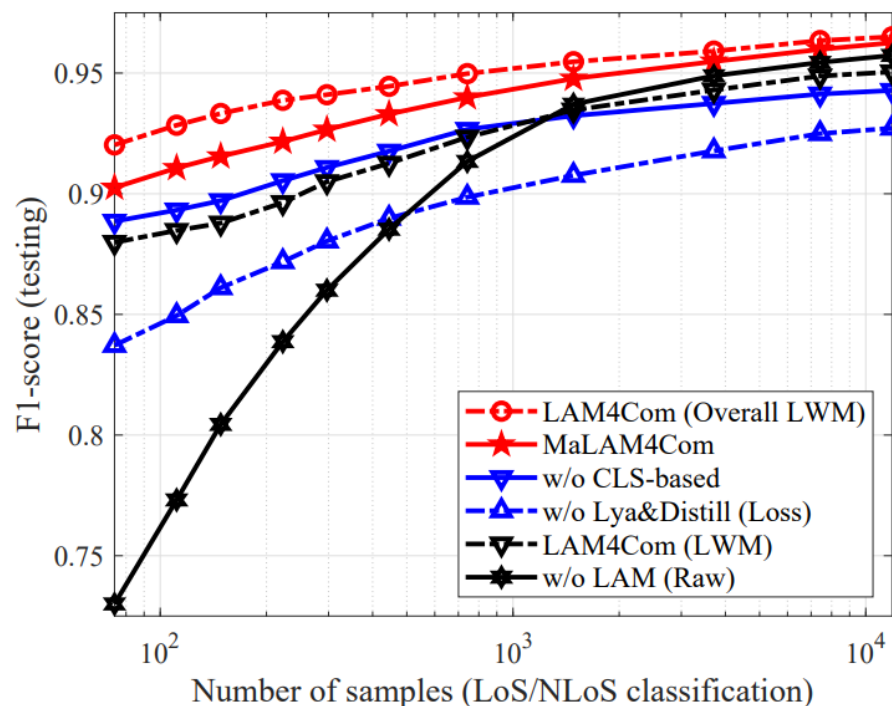
Reference	Applicable Object	Agent Configuration	Model Framework	Distributed Deployment	Generalized Tasks & Scenarios
J. Shao et al. (2024) [19]	Texts	Single-Agent	Large Language Models	—	✓
F. Jiang et al. (2024) [27], [28]	Images/Texts	Multi-Agent	Large Language Models	Partially	✓
M. Xu et al. (2025) [29]	Texts	Multi-Agent	Large Language Models	Partially	✓
Z. Liu et al. (2025) [7]	Channels	Multi-Agent	Wireless AI Models	✓	—
B. Liu et al. (2024) [20]	Channels	Single-Agent	Large Language Models	—	Partially
T. Zheng et al. (2024) [22]	Channels	Single-Agent	Large Language Models	—	✓
Z. Chen et al. (2024) [23]	Channels	Single-Agent	Wireless Large AI Models	—	✓
S. Alikhani et al. (2024) [24]	Channels	Single-Agent	Wireless Large AI Models	—	✓
T. Yang et al. (2025) [25]	Channels	Single-Agent	Wireless Large AI Models	—	✓
Z. Yang et al. (2024) [30]	Channels	Multi-Agent	Wireless Large AI Models	Partially	✓
This paper	Channels	Multi-Agent	Agent Cooperation: Wireless Large AI Models	✓	✓

- **联邦蒸馏：**
促进不同智能体之间的**知识高效迁移**
- **Lyapunov协同：**
确保系统在动态演化中保持**行为一致性**
- **实现稳定的分布式优化**



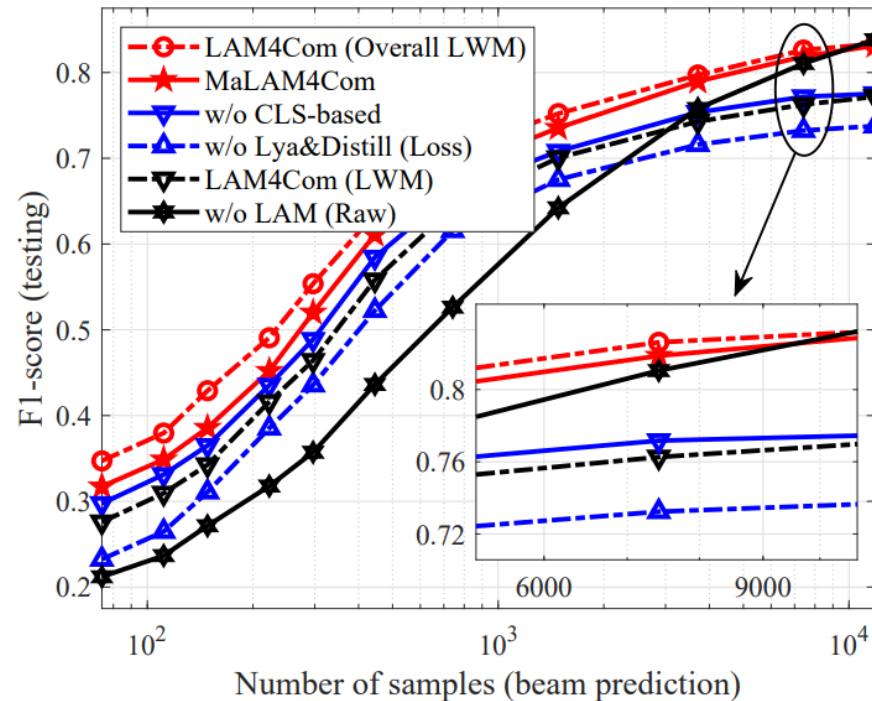
通过**多技术协同策略**，旨在构建一个**高效、可靠、安全且真正智能的AI原生无线通信系统**

4.5 决策可靠应用：分布式协同基座模型



(a) LoS/NLoS分类

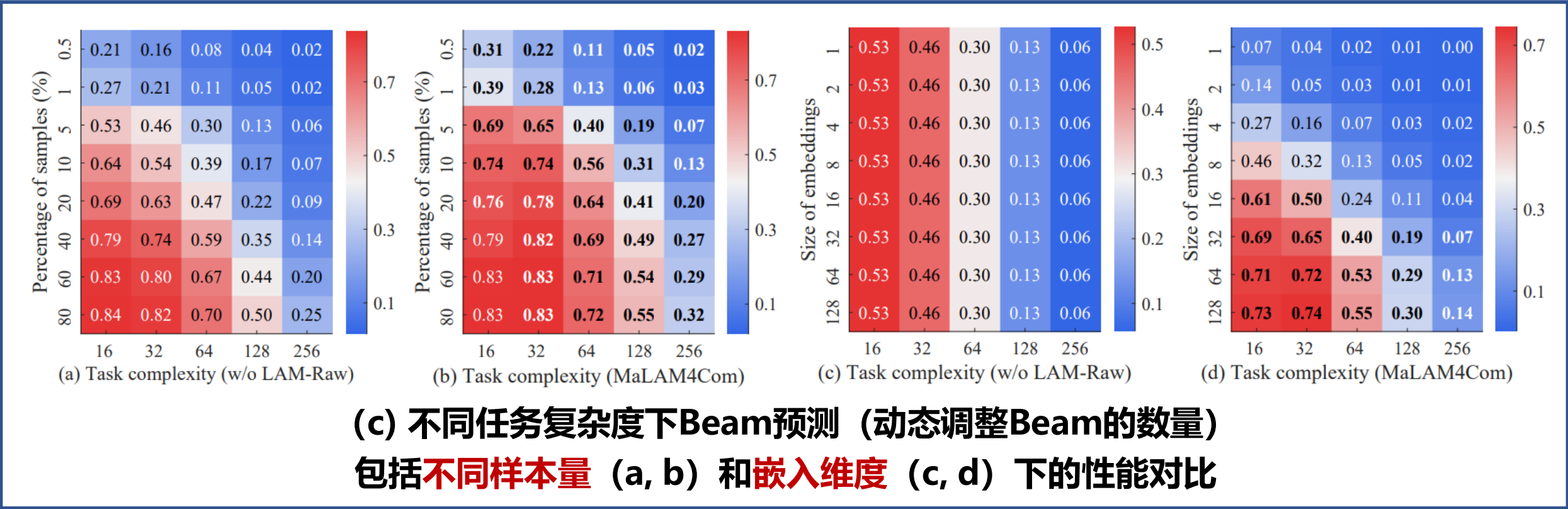
离散选择问题



(b) Beam预测

得益于强大的先验知识（自监督预训练），分布式协同基座模型在仅需少量样本的情况下，便能在多种下游无线任务上展现出卓越的泛化能力，使其能够快速适应动态的无线环境

4.5 决策可靠应用： 分布式协同基座模型

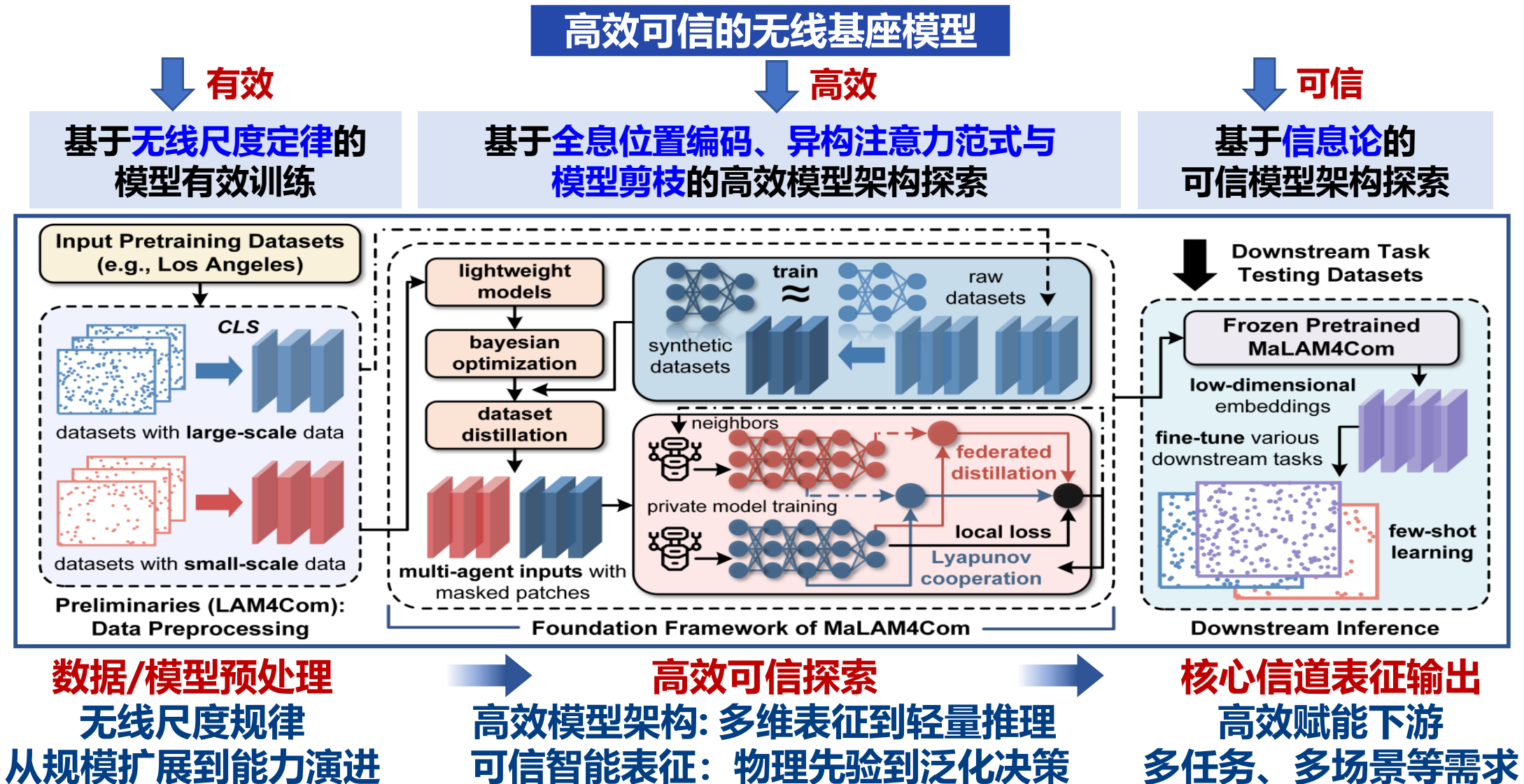


随着任务复杂度 (Beam数量从16增加到256), 分布式协同基座模型相对基线能够有效应对高复杂度的任务, 验证了其具备出色的少样本学习能力, 能从极少量数据中提取有效信息

汇报提纲

- I. 研究背景与意义
- II. 无线尺度规律：从规模扩展到能力演进
- III. 高效模型架构：从多维表征到轻量推理
- IV. 可信智能表征：从物理先验到泛化决策
- V. 总结与展望

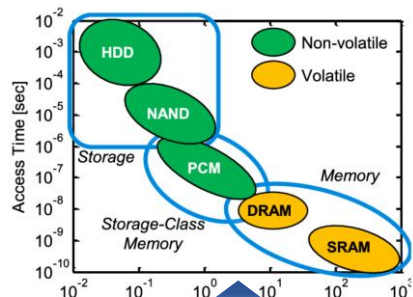
5.1 总结



5.2 展望与未来演进

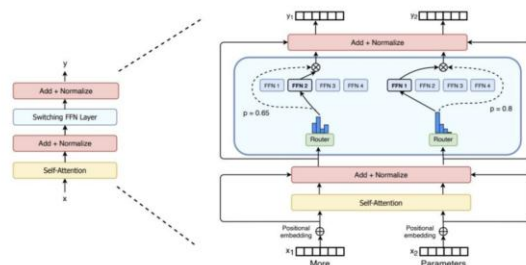
解决多模态下表征问题

增强模型表征泛化性
适配不同输入源与多样化任务



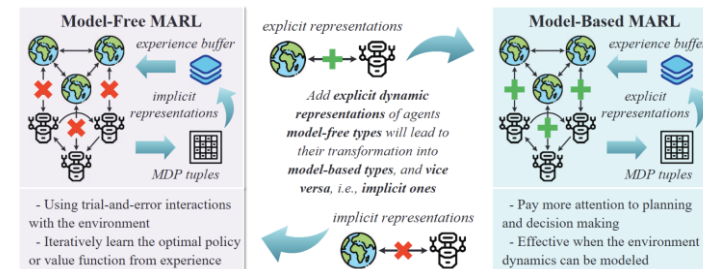
设计低开销的新型基座架构

设计高效低开销新型架构
全息位置编码与非对称注意力

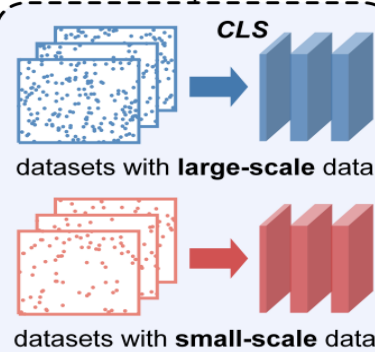


构建可验证的可信决策架构

量化不确定性与安全风险
支撑鲁棒泛化与隐私保护



Input Pretraining Datasets
(e.g., Los Angeles)



Preliminaries (LAM4Com):
Data Preprocessing

lightweight models

bayesian optimization

dataset distillation

multi-agent inputs with masked patches

neighbors

private model training

local loss

Lyapunov cooperation

train

synthetic datasets

raw datasets

federated distillation

local loss

Lyapunov cooperation

Lyapunov cooperation

Lyapunov cooperation

Downstream Task
Testing Datasets

Frozen Pretrained
MaLAM4Com

low-dimensional
embeddings

fine-tune various
downstream tasks

few-shot
learning

few-shot
learning

few-shot
learning

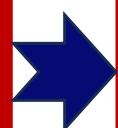
few-shot
learning

5.2 展望与未来演进：无线物理智能，通用模型向世界模型演进

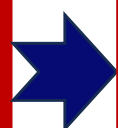
核心
问题

如何将无线通信从AI应用场景提升为物理智能问题？

全模态无线基座模型
多源融合感知



物理约束表征学习
电磁传播、空间几何



物理驱动世界模型
动态预测推演



面向无线物理智能，实现从任务识别到物理世界理解与推演

5.2 展望与未来演进：面向可信智能表征的具体路径

核心问题

如何学习**可解释、可泛化、可验证**的无线智能表征？



具体演进路径



从**黑盒特征学习**演进到物理先验、信息瓶颈与安全验证协同的**可信无线基座模型**

5.3 代表性工作

1. Ziheng Liu, **Jiayi Zhang**, Yiyang Zhu, Enyu Shi, Bokai Xu, Dusit Niyato, Shi Jin, and Bo Ai, “MaLAM4Com: Multi-Agent Cooperative Large AI Models for Wireless Communications,” *IEEE JSAC*, 2026.
2. Ziheng Liu, **Jiayi Zhang**, Haoyu Wang, Bokai Xu, Chen Zhang, Yiyang Zhu, and Enyu Shi, “Scaling Law for Large Wireless Models,” *Proc. AAAI*, 2026.
3. Jiayao Yang, **Jiayi Zhang**, Bokai Xu, Jiakang Zheng, Zhilong Liu, Ziheng Liu, Dusit Niyato, M érouane Debbah, Zhu Han, and Bo Ai, “White-Box AI for Wireless Communications: Principles, Architectures, and Applications,” *IEEE Network*, 2026.
4. Zhilong Liu, **Jiayi Zhang**, Bokai Xu, Derrick Wing Kwan Ng, Arumugam Nallanathan, and Bo Ai, “GCN-based Low-Complexity Downlink Beamforming for Cell-Free Massive MIMO Systems with Partially Coherent Joint Transmission,” *IEEE TWC*, 2025.
5. Bokai Xu, **Jiayi Zhang**, Qingfeng Lin, Huahua Xiao, Yik-Chung Wu, and Bo Ai, “Deep Unfolding Beamforming and Power Control Designs for Multi-Port Matching Networks,” *IEEE TWC*, vol. 24, no. 2, pp. 1401-1414, 2025.
6. Ziheng Liu, **Jiayi Zhang**, Yiyang Zhu, Enyu Shi, and Bo Ai, “Robust Multidimensional Graph Neural Networks for Signal Processing in Wireless Communications With Edge-Graph Information Bottleneck,” *IEEE TSP*, vol. 73, pp. 2688-2703, 2025.
7. **Jiayi Zhang**, Jing Zhang, Derrick Wing Kwan Ng, and Bo Ai, “Federated Learning-Based Cell-Free Massive MIMO System for Privacy-Preserving,” *IEEE TWC*, vol. 22, no. 7, pp. 4449-4460, Jul. 2023.



谢谢!



Homepage: <http://faculty.bjtu.edu.cn/8946/>

Github: <https://github.com/BJTU-MIMO>