Generative Decoding of Compressed CSI for MIMO Precoding Design

Hao Luo[*], Saeed R. Khosravirad[†], and Ahmed Alkhateeb[*]

[*] School of Electrical, Computer, and Energy Engineering, Arizona State University
[†] Nokia Bell Laboratories, New Providence, NJ, USA

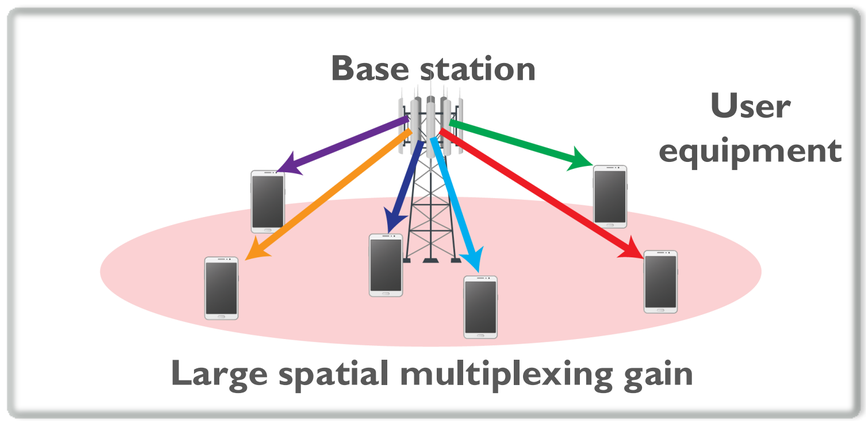
IEEE International Conference on Communications (ICC), 2026

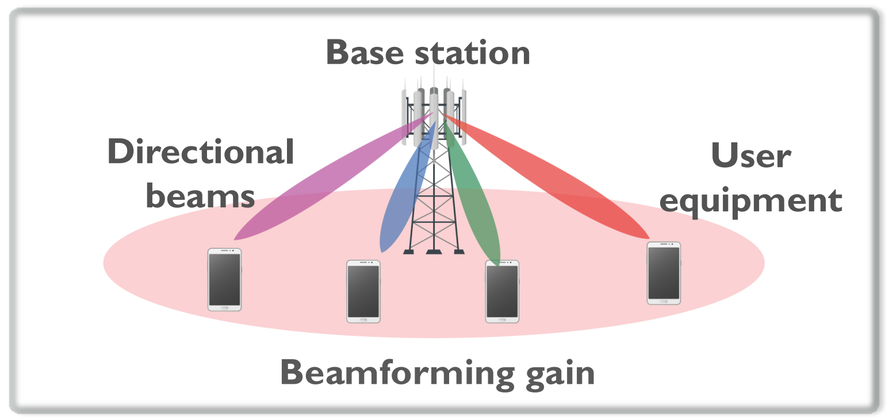
Wireless Intelligence Lab

IRA A. FULTON SCHOOLS OF engineering
ARIZONA STATE UNIVERSITY

Massive MIMO CSI feedback

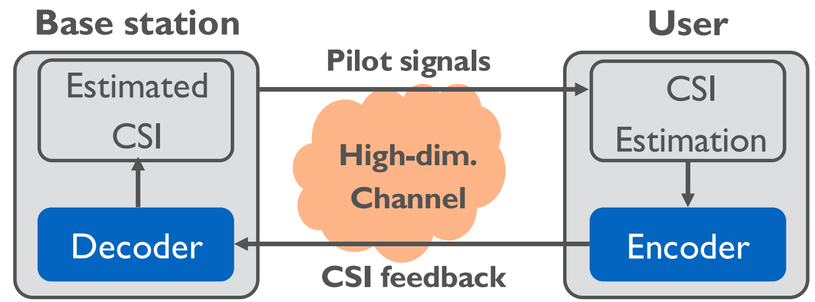
- Massive MIMO enhances spectral and energy efficiency





- However, the CSI acquisition overhead grows with the number of antenna

- ML encoder-decoder approaches face three key challenges
  - Interoperability: joint BS-UE training
  - Backward compatibility: requires significant changes to 3GPP standards
  - Data collection overhead: large real-world datasets are costly to gather



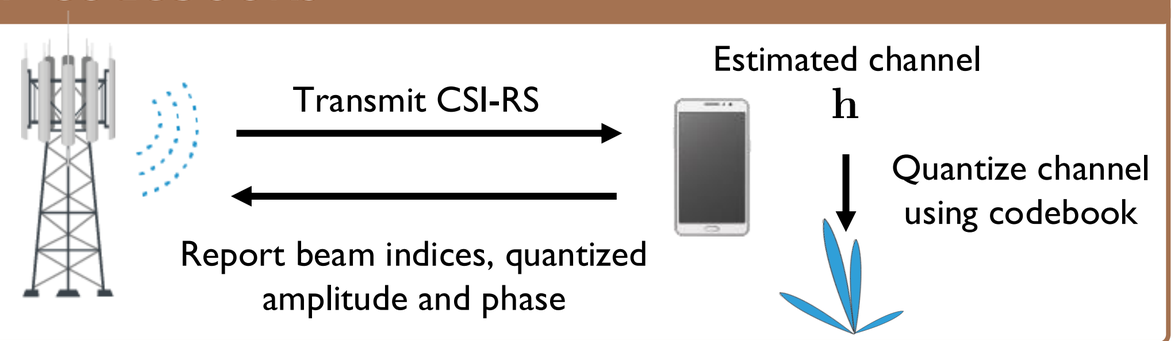
Can we improve the CSI quality at the BS without changing the UE-side encoder?

How can we reduce the dataset collection overhead for ML training?

Prior work

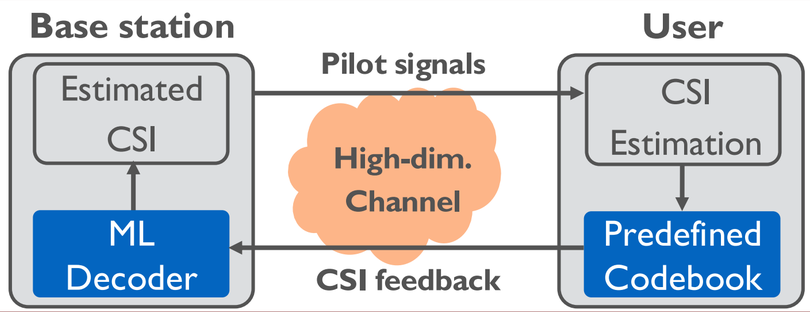
3GPP codebooks

- DFT beam grids
- Simple and standardized
- Lack site-specific env. knowledge



Existing decoder-only approaches [Guo'22, Chen'24]

- UE uses pre-defined codebook
- Preserves interoperability
- Data collection still a challenge



Our proposal

Keep 3GPP encoder at UE (standardized)

Propose a generative ML decoder with goal-oriented loss function

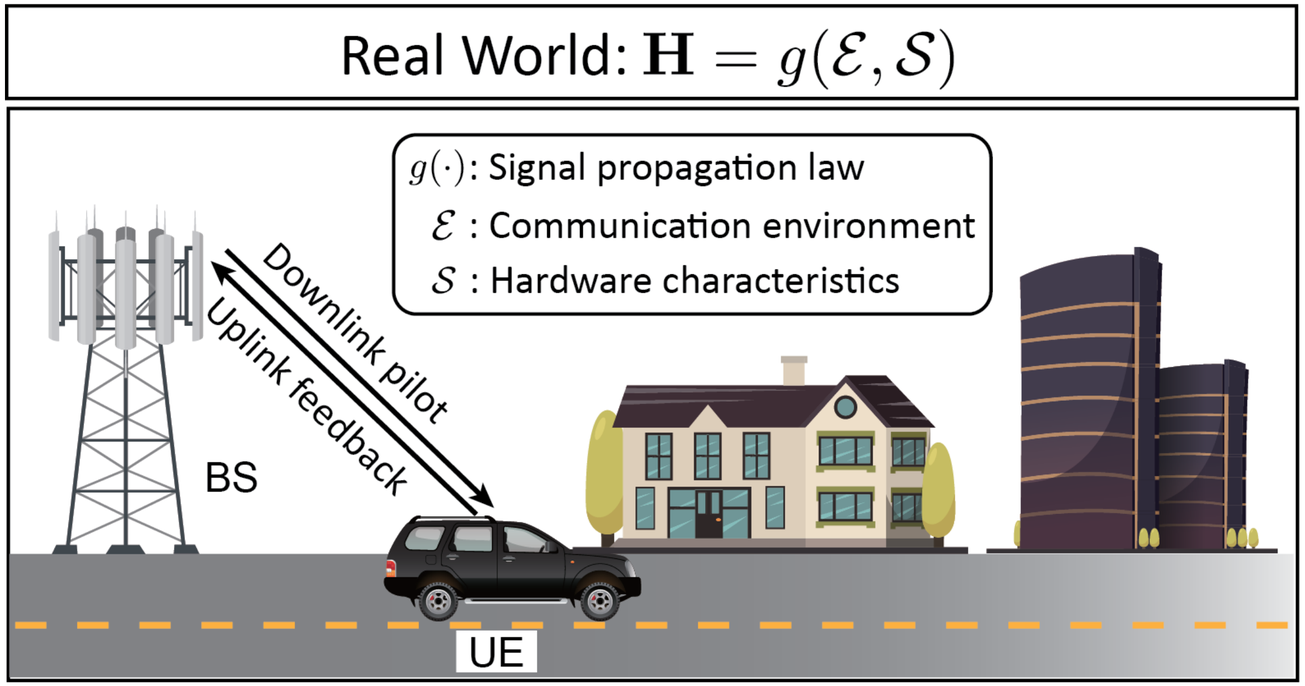
Use digital twin to generate synthetic CSI to reduce data collection overhead

[Guo'22] J J. Guo et al., "Environment Knowledge-Aided Massive MIMO Feedback Codebook Enhancement Using Artificial Intelligence," IEEE Trans. Commun.
[Chen'24] W. Chen et al., "CSI-PPPNet: A One-Sided One-for-All Deep Learning Framework for Massive MIMO CSI Feedback," IEEE Trans. Wirel. Commun.

System model

- Base station (BS)
  - N_t antennas
  - OFDM with K subcarriers
- Downlink signal model

$$y_{k,u} = \mathbf{h}_{k,u}^{\mathrm{H}} \mathbf{f}_{k,u} s_{k,u} + \sum_{v \neq u} \mathbf{h}_{k,u}^{\mathrm{H}} \mathbf{f}_{k,v} s_{k,v} + n_{k,u}$$



- Wideband geometric channel model

$$\mathbf{h}_{d,u} = \sum_{l=1}^{L} \alpha_{l,u}\, p(d T_{\mathrm{S}} - \tau_{l,u}) \mathbf{a}(\phi_{l,u}^{\mathrm{az}}, \phi_{l,u}^{\mathrm{el}})$$

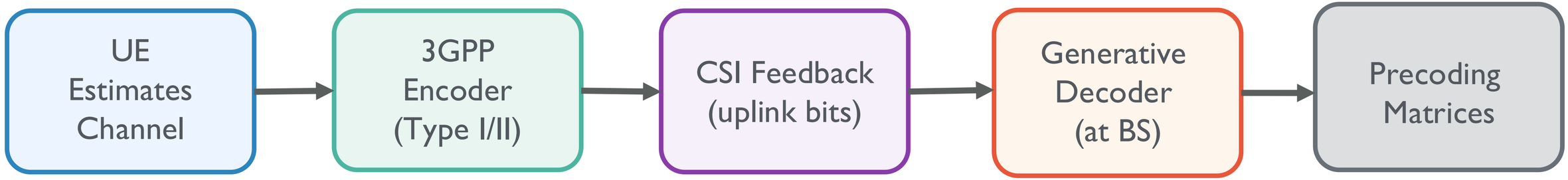
$$\mathbf{h}_{k,u} = \sum_{d=0}^{D-1} \mathbf{h}_{d,u} \exp\left(-j \frac{2\pi k}{K} d\right)$$

- Sum rate

$$R(\{\mathbf{H}_u\}_{u=1}^{U}, \{\mathbf{F}_k\}_{k=1}^{K}) = \frac{1}{K} \sum_{k=1}^{K} \sum_{u=1}^{U} \log_2 \left(1 + \frac{\left|\mathbf{h}_{k,u}^{\mathrm{H}}\, \mathbf{f}_{k,u}\right|^2}{\sum_{v \neq u} \left|\mathbf{h}_{k,u}^{\mathrm{H}}\, \mathbf{f}_{k,v}\right|^2 + \sigma^2}\right)$$

Problem formulation

- CSI feedback pipeline



- Objective: find optimal decoder parameters

$$\theta^\star = \arg\max_{\theta} \mathbb{E}_{\{\mathbf{H}_u\}_{u=1}^{U} \sim \mathcal{H}} \left[R(\{\mathbf{H}_u\}_{u=1}^{U}, \{\widehat{\mathbf{F}}_k\}_{k=1}^{K}) \right]$$

Learnable weights of decoder

Generated precoder

- To resolve the key challenges
  - ML encoder-decoder CSI compression is not interoperable and backward-compatible
  - Standard Type-I/II codebooks lack site-specific knowledge → suboptimal precoders

Decoder-only: UE uses 3GPP standard codebook – interoperable & backward-compatible

Goal-oriented loss: optimize achievable rate

3GPP Type-I and Type-II codebooks

- 3GPP Type-II codebook
  - Finer spatial resolution than Type-I
  - Capture multi-path characteristics

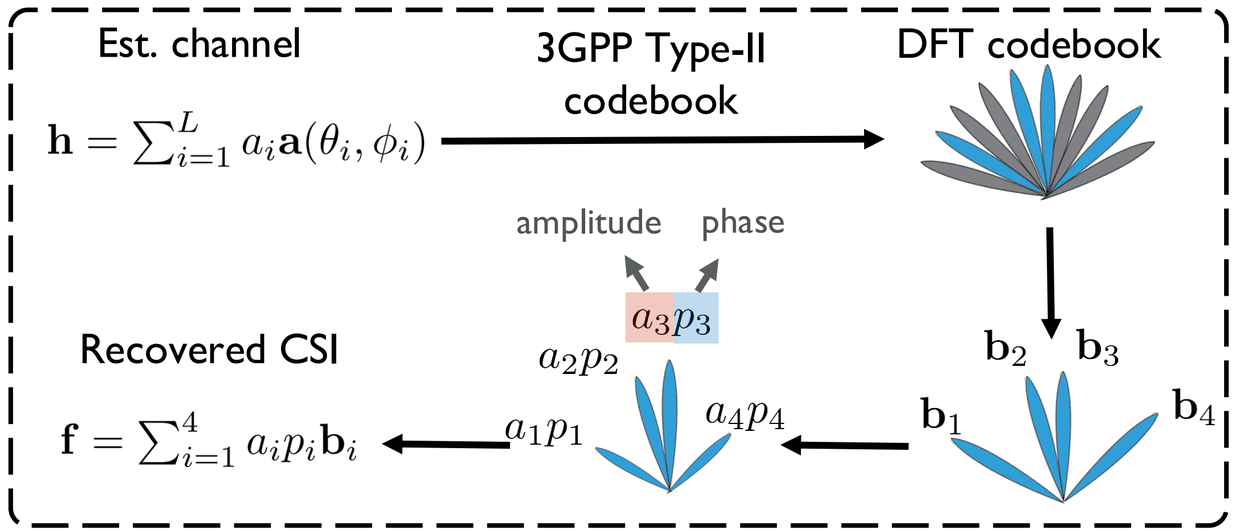
Quantized wideband amplitudes

Normalization factors

$$\mathbf{W}_u = \mathbf{W}_{1,u}\,\mathbf{W}_{\mathrm{C1},u}\,\mathbf{W}_{\mathrm{C2},u}\,\mathbf{\Lambda}_u$$

Wideband DFT beams

Quantized subband phases



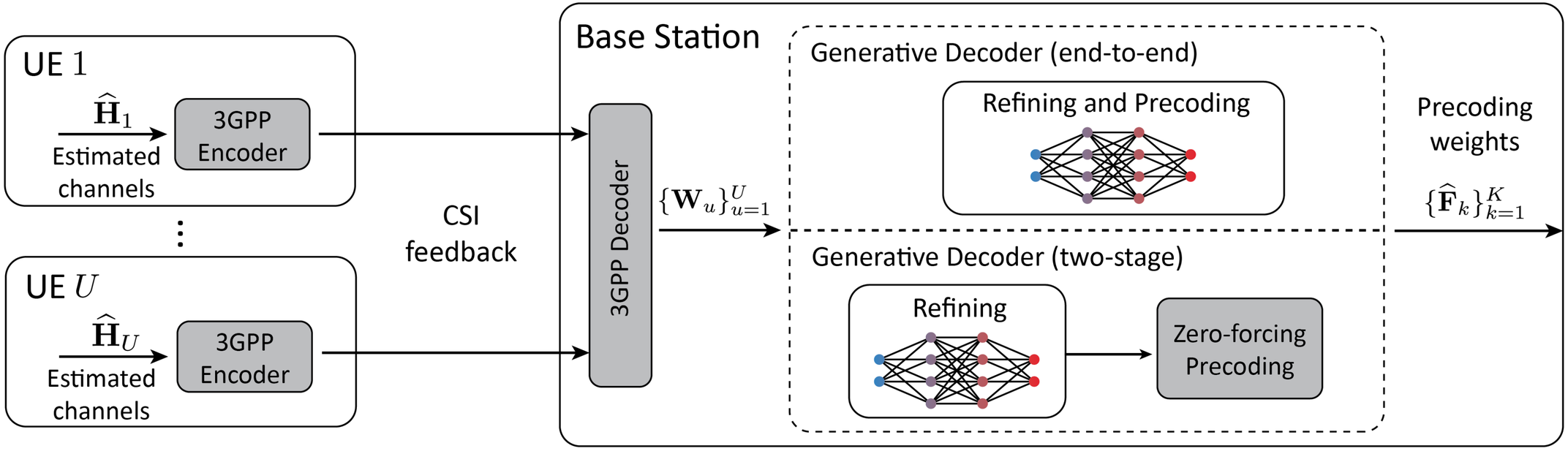
- Why use standardized codebooks?
  - UE implementation unchanged – full interoperability across vendors
  - BS only needs to improve its decoding – no standard revision required
  - Feedback bits are well-defined: easy to simulate and deploy

Our decoder works with both Type-I and Type-II – no UE changes needed

Key idea

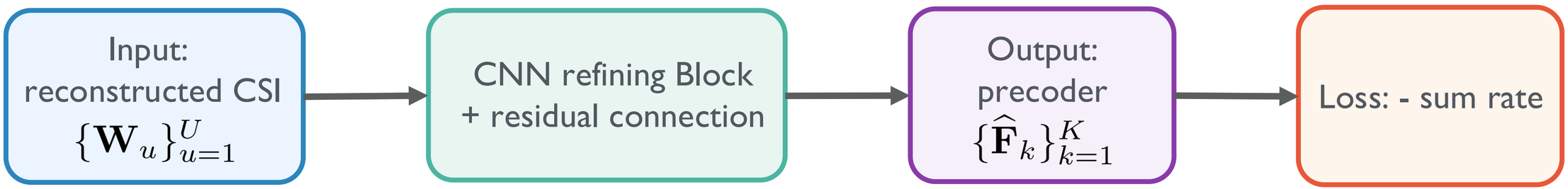
Interoperability	Backward compatible	Reduced data overhead
• BS-only deployment • No joint training with UE vendors • UE uses standard 3GPP codebook	• Fits within 3GPP spec • Precoder design is a vendor-side implementation detail	• Site-specific digital twin • Synthetic CSI for offline training • No massive real-world collection

- Proposed pipeline



Digital twin generates synthetic CSI for offline training → deploy in real-world system

Generative decoder – end-to-end approach



- Design philosophy
  - Decoder received compressed CSI from all users
  - Jointly refines and produces precoding matrices
  - Single model learns both CSI refinement and interference management

- Training – goal-oriented loss

$$\mathcal{L}_{\mathrm{e2e}}(\theta) = \frac{1}{N_{\mathrm{D}}} \sum_{n=1}^{N_{\mathrm{D}}} -R(\{\mathbf{H}_{n,u}\}_{u=1}^{U}, \{\widehat{\mathbf{F}}_{n,k}\}_{k=1}^{K})$$

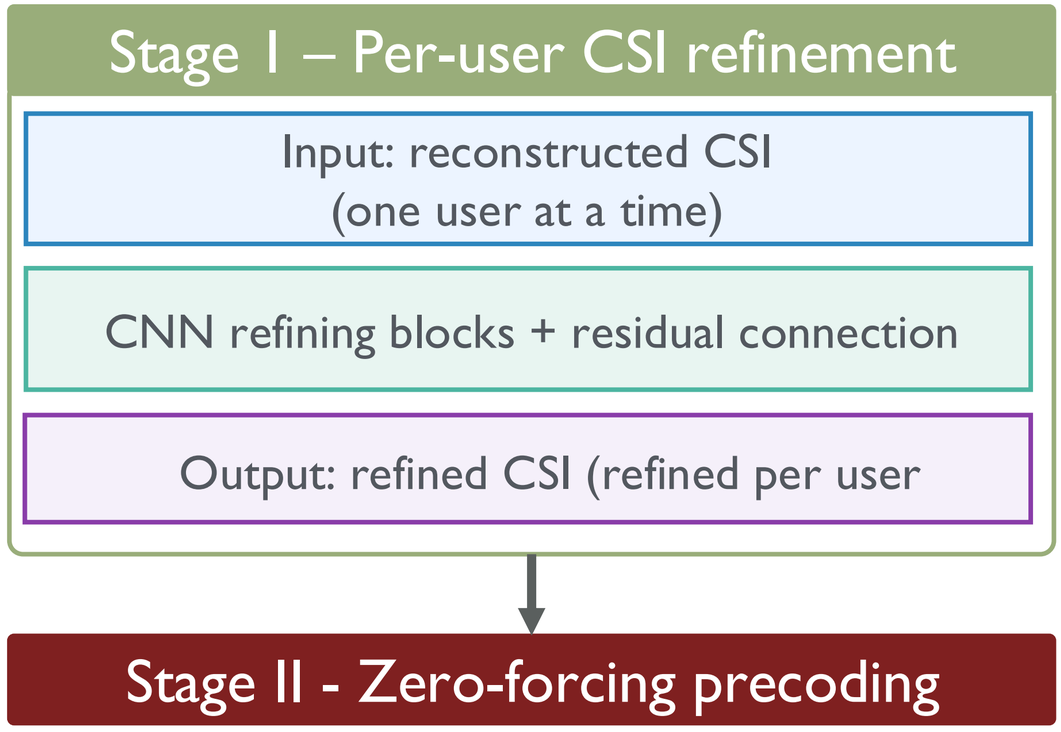
End-to-end: joint user refinement + precoding

Generative decoder – two-stage approach

- Design philosophy
  - Decouple the problem: first refine each user's CSI individually, then precode
  - Leverage refine CSI for zero-forcing (ZF) precoder computation
- Stage 1 – Per-user CSI refinement
  - Input: single user's compressed CSI
  - Decoder refines to each user
  - Loss:

$$\mathcal{L}_{\mathrm{ts,rate}}(\theta) = \frac{1}{N_{\mathrm{D}}} \sum_{n=1}^{N_{\mathrm{D}}} -R_{\mathrm{SU}}(\mathbf{H}_n, \widehat{\mathbf{W}}_u)$$

- Stage 2 – Zero-forcing precoding
  - Stack refined CSI per subcarrier k
  - Apply zero-forcing to stacked refined CSI



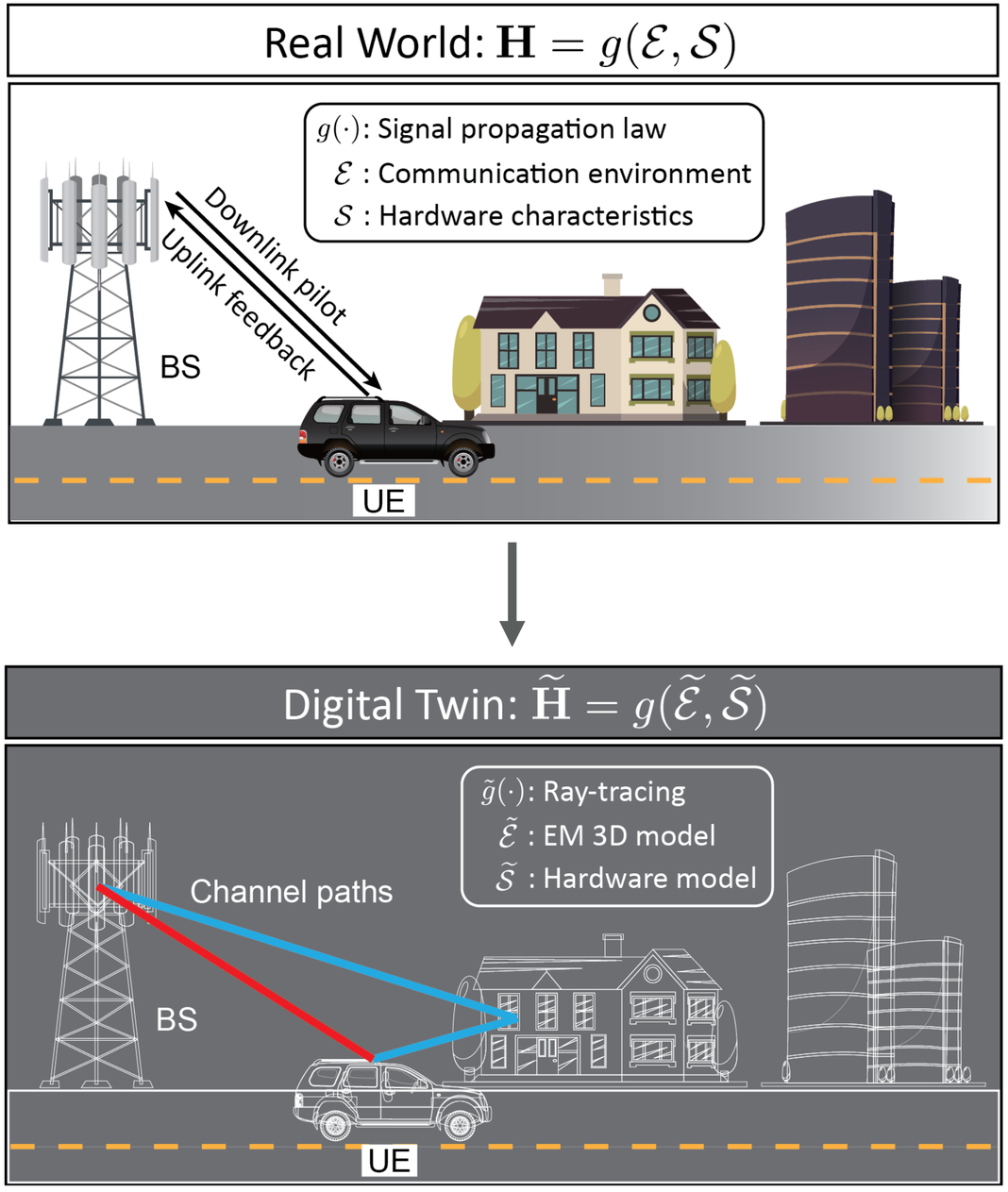
Two-stage: per-user refinement + zero-forcing precoding

Digital twin aided generative decoder

- Real-world channel is determined by three factors
  - Communication environment $\mathcal{E}$
  - Signal propagation law g
  - Hardware characteristics $\mathcal{S}$

- Digital twin approximates each factor
  - EM 3D model $\widetilde{\mathcal{E}}$
  - Ray tracing $\tilde{g}$
  - Hardware model $\widetilde{\mathcal{S}}$

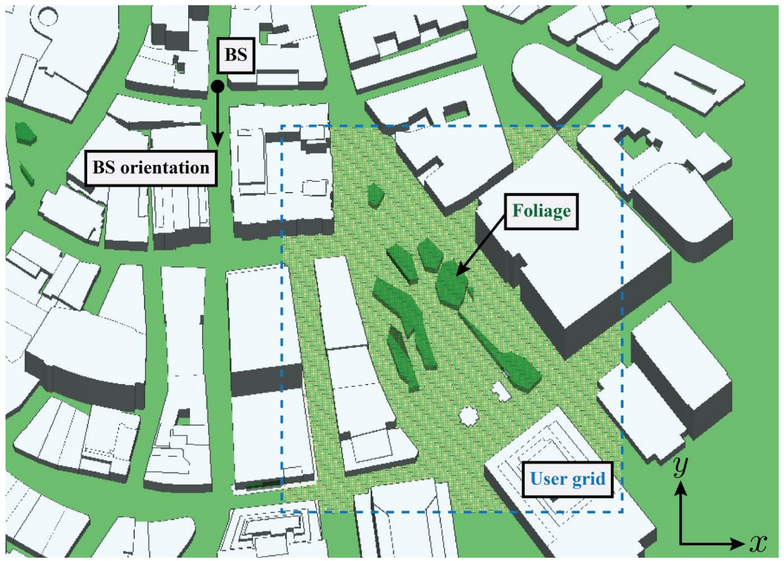
- Training and deployment workflow
  - Generate large synthetic CSI dataset from DT
  - Train the decoder offline on synthetic data
  - Deploy trained decoder in real-world system



Digital twins reduce data collection overhead while providing site-specific knowledge

Simulation setup

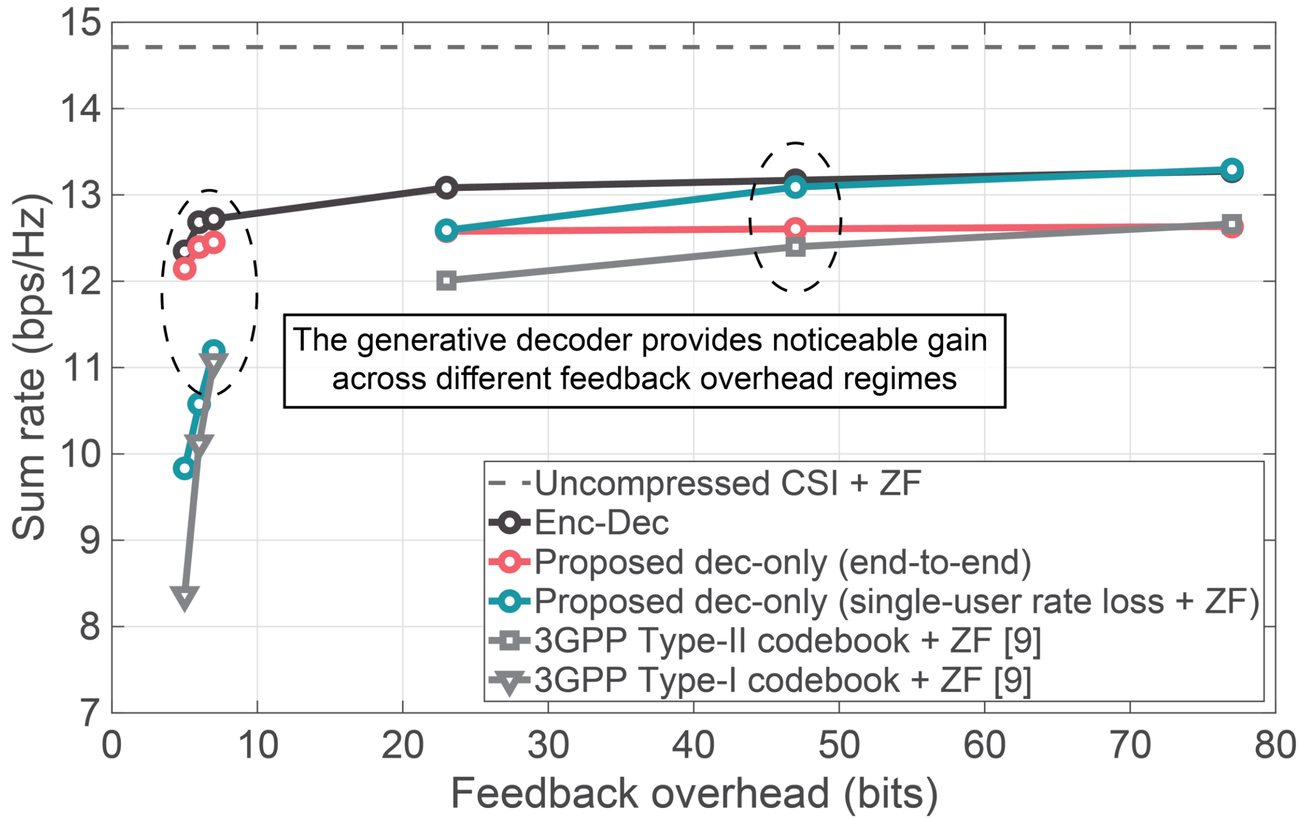
Boston Scenario



Parameter	Value
BS antennas	32 (ULA)
UE antennas	1
Carrier frequency	3.5 GHz
Subcarriers	288
Subcarrier spacing	30 kHz
BS EIRP	30 dBm
UE noise figure	7 dB
Number of users	2

- Target (real-world) scenario
  - Based on downtown Boston – including buildings and foliage
  - BS with 32 uniform linear array (ULA) at 15 m height
  - UE grid: 200m x 230m (mainly NLoS)
- Dataset generation
  - Wireless Insite ray-tracing simulator
  - DeepMIMO channel generator
  - 41,120 data points; 80% train / 20% test split
  - Consider two users as a data point for studying precoding
- Benchmarks
  - 3GPP Type-I/II codebooks
  - Reconstruction loss
  - Encoder-decoder (w/ straight-through estimator)
  - Generic dataset (20 ray-tracing scenarios, WINNER II)

Can the generative decoder effectively refine codebook-based CSI?

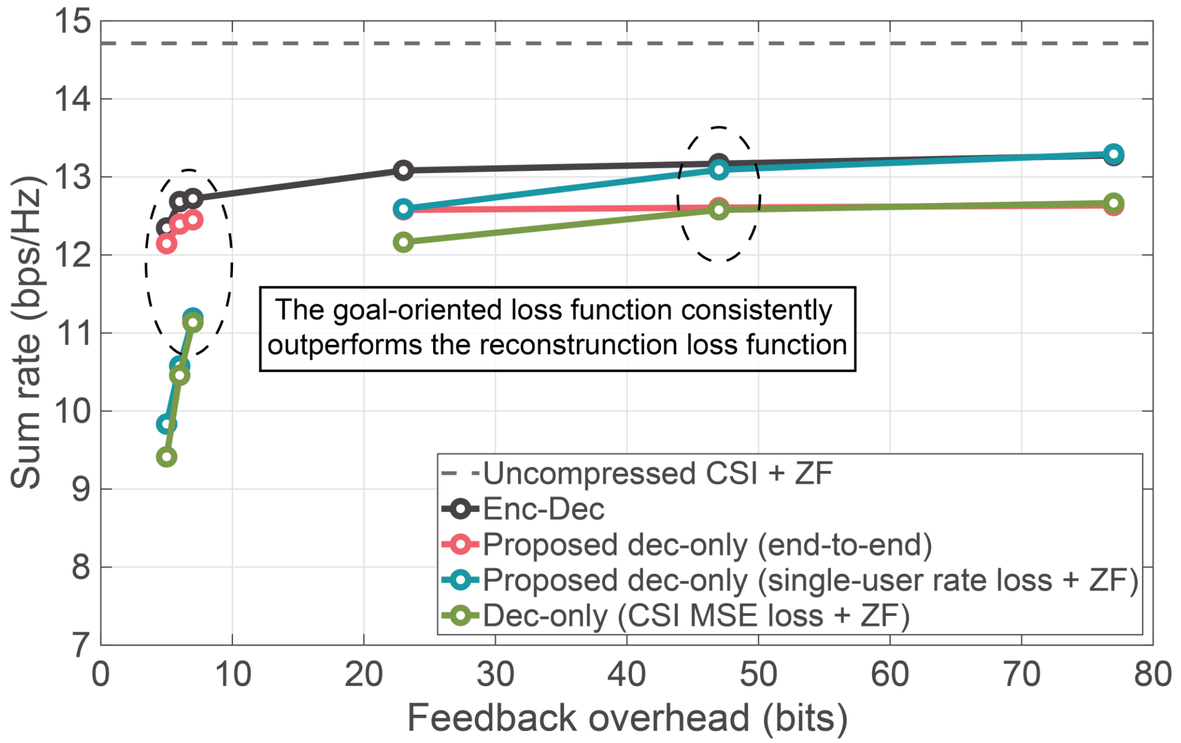


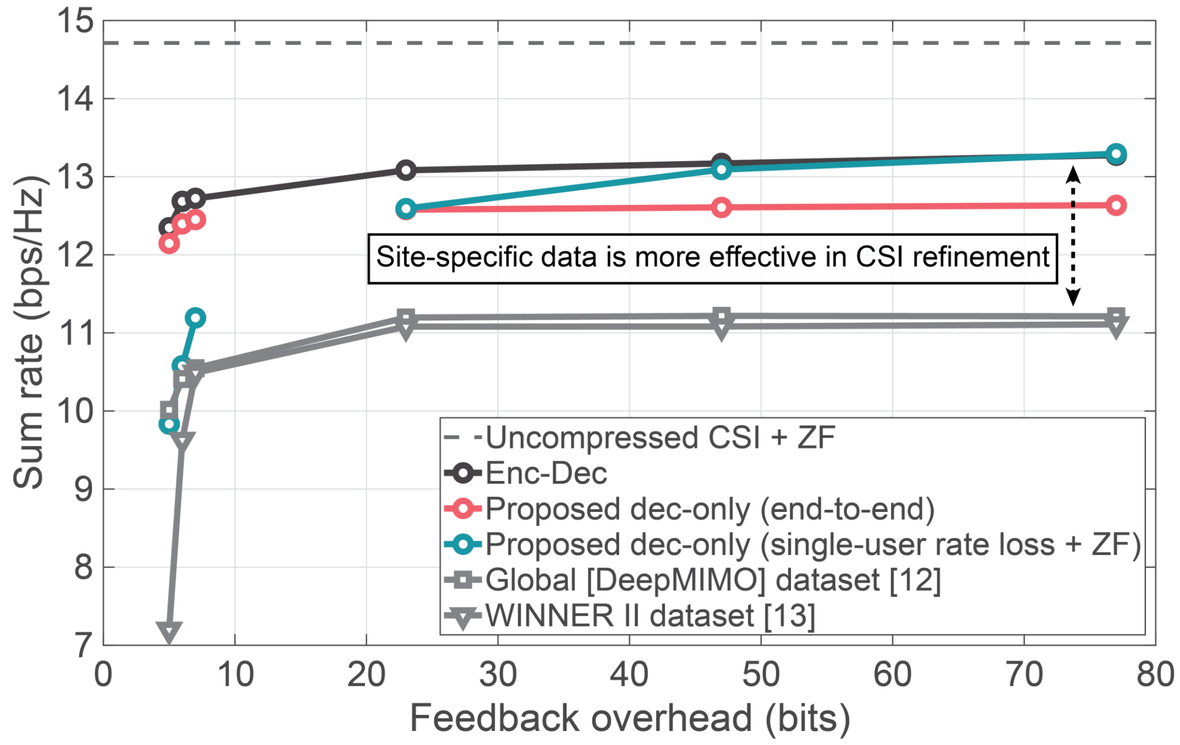
The generative decoder consistently outperforms 3GPP codebooks

Low feedback (Type-I): the end-to-end approach is superior

High feedback (Type-II): the two-stage method excels

What are the advantages of goal-oriented loss and site-specific dataset?





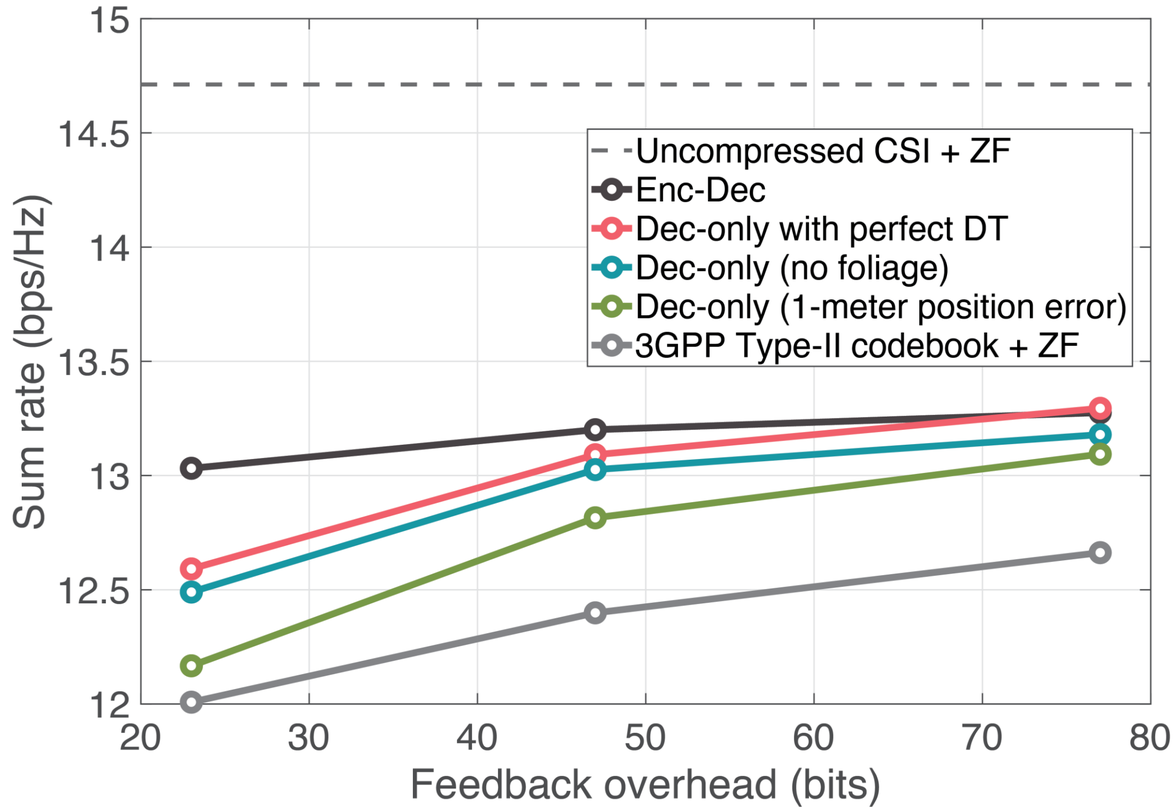
Goal-oriented loss outperforms reconstruction loss

Models trained on generic statistical or global datasets fail to generalize

Site-specific datasets are critical, providing the environmental knowledge for refinement

How does digital twin imperfection affect the performance?

- Digital twin scenarios
  - Foliage removed from the 3D scene
  - Building positions perturbed by random offset
- Observations
  - Performance degrades relative the perfect DT
  - Noticeably outperforms standard Type-II CB



The approach is robust to moderate levels of digital twin accuracy

Fine-tuning the model with a small amount of real-world data offers a future path

Conclusion and future work

- We propose a decoder-only generative CSI refinement solution for massive MIMO
  - UE uses 3GPP standardized codebook – ensures interoperability and backward compatibility
  - BS-side generative decoder refines compressed CSI for precoder design
- Two training approaches with goal-oriented loss
  - End-to-end: jointly refine and precode – excels at low feedback (Type-I)
  - Two-stage: refine per-user then zero-forcing – excels at high feedback (Type-II)
  - Goal-oriented (rate) loss consistently outperforms reconstruction (MSE) loss
- Digital twin data reduces real-world collection overhead
  - Synthetic data from site-specific DT significantly outperforms generic datasets
  - DT imperfections (missing foliage, position errors) cause manageable performance loss
- Future work
  - Scalability to more users – Extend the generative decoder to larger multi-user scenarios
  - Real-world calibration – Fine-tune the digital twin trained decoder with a small number of measurements
  - Tighter digital twin integration – Explore joint optimization of the digital twin modeling and the decoder

Q&A

Thank you!

Email: h.luo@asu.edu