AI/ML based CSI Compression

概述

超大规模 MIMO 系统是 5G 的关键技术。BS 发送 CSI-RS，UE 测量并反馈 CSI，用于 BS 侧确定预编码矩阵。超大规模 MIMO 系统在 FDD 系统以及 TDD 系统的部分场景均需要采用 UE 反馈的 CSI 用于下行预编码。UE 反馈测量的 CSI 的开销随着天线数、带宽增加而线性增加。考虑到上行资源受限，CSI 反馈的精度受限，影响预编码设计。

3GPP Rel-18 开始讨论 AI/ML 应用于空口，其中就包括了 CSI 压缩用例。AI 技术应用于 5G 早已开始。例如，5G 定义网络数据分析功能（Network Data Analytics Function, NWDAF）用于分析和处理其他网络功能单元的数据，应用于自动化运维、网络切片、网络节能等。

AI/ML CSI 压缩采用基于自编码器（auto-encoder）的双侧模型，对 UE 侧测量的 CSI 压缩，并将压缩的 CSI 用于 NW 侧的 CSI 重建。相比传统 eType II 码本，其反馈开销可以下降约 50%。然而，相比单侧模型的波束管理、定位，性能增益评估更快完成，采用双侧模型的 CSI 压缩，经过 2 个版本的迭代才进入 WI 阶段。

Rel-18 CSI 压缩研究的用例为空频域 CSI 压缩，包括预处理、后处理、量化/去量化阶段。信令框架以传统的 CSI 反馈为基线。讨论定义了 CSI 压缩的推理过程、性能监测方式、数据收集方式、协作训练类型，并给出了性能评估的结果。然而，在 Rel-18 评估时 CSI 压缩带来的性能增益不明显，需要较高的模型负载度，且 NW 和 UE 协同复杂。

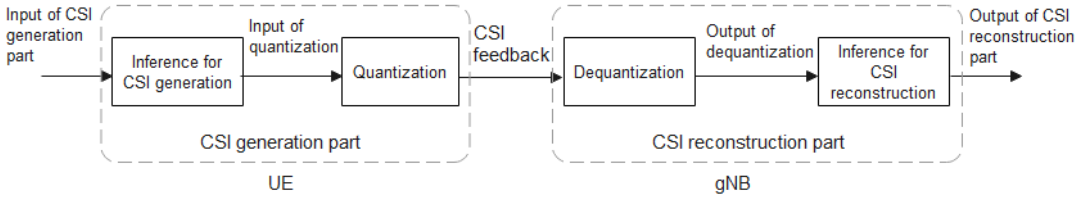
Rel-19 CSI 压缩研究扩展了 Rel-18 CSI 压缩研究的用例到空时频域 CSI 压缩，以进一步提升性能和复杂度的权衡。同时对 Rel-18 未收敛的协作训练进一步讨论，确定了 3 种可用的双侧模型协作训练方式。

Rel-20 CSI 压缩研究进入 WI。采纳空频域 CSI 压缩作为基线，标准化推理、数据收集、性能监测、配对、协作训练。Rel-19 讨论的其他用例并未涵盖在 Rel-20 的研究范围。当前，Rel-20 的研究工作才刚展开。

本文将详细分析 Rel-18 和 Rel-19 CSI 压缩的研究进展，并洞察 CSI 压缩的 Rel-20 的研究方向和 6G 的可能研究方向。

推理

CSI 压缩用例采用双侧模型。在 UE 侧部署 CSI 生成模型（或者编码器）、在 NW 侧部署 CSI 重建模型（或者解码器）。在 UE 侧将输入的 CSI 压缩和量化，在 NW 侧恢复和重建 CSI。CSI 压缩用例的推理流程如下框图：



其中，框图的各标注的解释如下：

- Input of CSI generation part: UE 侧模型的输入，也叫目标 CSI，为信道矩阵、预编码矩阵中的一种。
- Inference for CSI generation: 自编码器模型的编码器部分（或者 CSI 生成模型），用于将

输入压缩为隐向量。

- Quantization: 将隐向量的标量元素从连续值量化为离散值，获得 CSI feedback。支持的量化方式包括标量量化和向量量化。
- Dequantization: 去量化 CSI feedback，将离散值连续化为 Output of dequantization。去量化和量化规则应为 NW 和 UE 侧已知。
- Inference for CSI reconstruction: 自编码器的解码器部分（或者 CSI 重建模型），用于将 Output of dequantization 转化为重建的 CSI。
- Output of CSI reconstruction part: 即重建的 CSI 或者 recovery CSI。

模型输入输出

模型的输入和输出为信道矩阵或预编码矩阵。模型输入由 UE 测量 CSI-RS 获得，也称为目标 CSI。模型输出称为重建 CSI。在无重建损失下，重建 CSI 接近目标 CSI。目标 CSI 可以为原始信道矩阵时，可以为空频域或角度时延域的信道矩阵。为预编码矩阵时，可以为信道矩阵奇异值分解或者特征分解后得到，或者为类似于 eType II 码本的预编码矩阵。在 Rel-19 和 Rel-20 进一步讨论目标 CSI 的类型和格式。

量化和去量化

3GPP 讨论中比较了规定和未规定量化情况下的性能。规定量化方式可以获得比非量化方式更好的训练效果。其次，量化输出隐信息有利于降低 CSI 反馈开销。相比直接反馈浮点数的隐信息，反馈量化比特有利于降低反馈开销。最后，标准化量化方式有利于 UE 和 NW 侧确定负载大小和解码负载。

为了让 UE 和 NW 侧知道各自采用的量化方式，讨论确定的量化方式同步方式包括：

- VQ：同步 VQ 码本的格式和大小、CSI 生成模型输出的分段方式。
  - VQ 可以将隐向量的部分或所有元素映射到 VQ 码字。
  - 一个隐向量对应至少一个 VQ 分段，每个 VQ 分段对应多个元素，采用 VQ 码本映射到相应码字。
  - VQ 需要规定 VQ 分段的方式、分段所采用的 VQ 码本。
- SQ: 同步 SQ 码本的格式和大小。
  - SQ 可以采用均匀获非均匀量化。
  - SQ 码本需要规定量化比特和（隐向量的）浮点数的映射关系。

量化方式的同步方式包括：

- 模型配对
- 标准化量化码本的同步

性能监测

在 Rel-18 确定了性能监测的框架，包括支持 NW 侧性能监测、UE 侧性能检测，支持基于最终 KPI 和中间 KPI 的性能监测，定义性能监测的指标。在 Rel-19 细化了性能监测的 NW 侧和 UE 侧监测方向。

性能监测框架

类似于其他 AI/ML 空口用例，CSI 压缩性能监测可以划分为：

- NW 侧性能监测、UE 侧性能检测；
- 基于最终 KPI 的性能检测、基于中间 KPI 的性能监测。

采用的 KPI 包括：

- 中间 KPI（如 SGCS 或 NMSE）；
- 最终 KPI（如 BLER、吞吐量、ACK/NACK）。

其中，基于中间 KPI 的性能监测包括：

- NW 侧基于实测的 target CSI 和 UE 侧报告的关联的 CSI 报告计算 KPI。
- UE 侧基于 CSI 重建模型获得的输出。该输出可以是 NW 侧发送给 UE 的，也可以是在 UE 侧部署的 CSI 重建模型获得的。

性能监测方向

相比 R18，R19 进一步细化性能检测，包括性能恶化诊断（也就通常说的性能监测）以及性能恶化根源的诊断。性能恶化诊断方面，在 R18 基础上细化 NW 侧和 UE 性能监测的用例。性能恶化根源的诊断，是 R19 新引入的内容。对于跨厂商协作训练选项 1,3,4,5，双侧模型允许的离线工程使得性能恶化根源种类更多。

基于 R18 的讨论，性能监测可以分为基于最终 KPI 的监测和中间 KPI 的监测。由于基于最终 KPI 的监测不直接反映模型的性能，R19 的关注主要在基于中间 KPI 的性能监测。

对于 NW 侧性能检测，细化如下两种方式：

- Case 1：基于 UE 报告的 target CSI。Target CSI 的格式可能为传统 eType II 码本或者类似 eType II 的高精度码本。NW 基于 target CSI 以及 UE 报告的关联 CSI feedback 获得的重建 CSI，计算中间 KPI，如 SGCS。TR 中观测到，采用高精度的 target CSI，如采用新参数的 R16 eType II 码本，对于提升中间 KPI 的精度有益。考虑到 UE 未部署 CSI 重建模型，在 NW 侧基于高精度 target CSI 计算中间 KPI 更为准确。
- 基于 UE 发送的 SRS 的监测。

对于 UE 侧性能检测，细化如下方式：

- 基于 UE 侧部署的 CSI 重建模型的输出。这个 CSI 重建模型的可以是 NW 侧实际部署的 CSI 重建模型，或者是 NW 提供的一个参考模型，又或者是 UE 侧部署的代理模型。采用 UE 侧部署的 CSI 重建模型存在泛化性、额外的 LCM 开销的问题。
- 在没有重建 target CSI 的情况下直接估计中间 KPI。例如，采用 UE 侧部署的估计器。

类似于部署 CSI 重建模型，如何评估和监测估计器的性能是需要解决的问题。

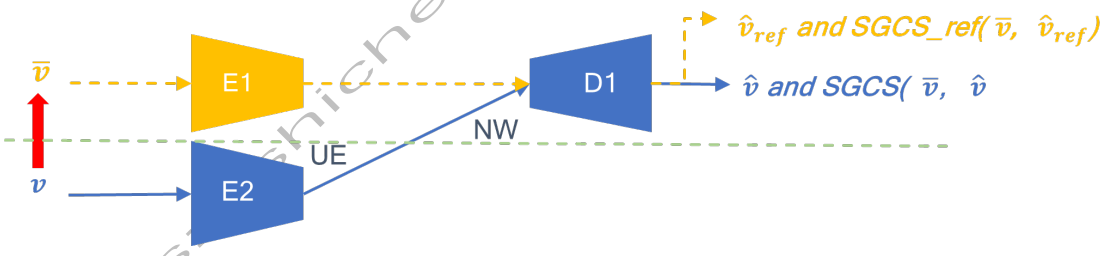
- 基于 NW 拿到重建 CSI 后构造的预编码 RS（包括 CSI-RS、DMRS）。例如，计算重建 CSI 构造的预编码 RS 的信道质量和 target CSI 构造的预编码 RS 的信道质量的差异。该方式存在反馈 CSI 到接收预编码后的 RS 的时延，对信道老化敏感。
- 基于 NW 发送的 CSI 重建模型的输出，该输出可以采用 eType II 码本或类似于 eType II 码本的高精度码本。在报告方式上，该输出可以采用 RRC 层传输，单次传输多个重建 CSI。需要考虑如何配对重建 CSI 和 UE 侧的 CSI feedback。另外，NW 发送重建 CSI，可能泄露 NW 隐私，为一些公司反对。

此外，在评估 NW 侧和 UE 侧性能监测时，需要考虑：

- 监测精度：监测精度如果具备泛化性，也应该考虑泛化性的影响。
- 监测开销、时延、复杂度以及精度的权衡：尤其是监测报告包括至少一个推理时机的监测结果。
- UE 报告的监测指标的可测试性，使得 NW 侧可以评估结果的有效性。

误差根源诊断是用于判别双侧模型性能恶化的起因。这个过程可以在性能监测过程中或性能监测后进行。对于跨厂商协作训练的方向 A 和方向 C，由于双侧模型在部署过程中和部署后基于各侧条件获取或微调模型，导致模型不适配。例如，NW 侧首先训练获得 CSI 生成模型 E1，CSI 重建模型 D1。NW 将 CSI 生成模型 E1 或训练数据集共享给 UE（选项 3a-1），用于 UE 侧生成 CSI 生成模型 E2，如前所述。

当执行 NW 侧性能监测时，NW 收集 target CSI 以及关联推理的 CSI feedback，以计算中间 KPI，如 SGCS。实际上，由于 NW 侧和 UE 侧推理实际采用的 E2 和 D1 不完全匹配，使得需要判定性能下降到底是哪一侧模型出了问题。基于 SGCS 的差异可以判别性能下降，这个监测流程如下图所示。



有三类因素可能导致性能恶化，包括：

- 数据漂移：如实测数据和训练数据集对应的信道条件不同，仅基于训练数据集不能反映实测信道。
- UE 侧模型问题：在离线工程中，UE 侧生成的 CSI 生成模型和 NW 侧的编码器不完全配对。
- NW 侧模型问题：NW 训练的双侧模型训练不加且在实际部署中推理性能差。又如，NW 侧模型在部署后基于新数据微调，可能使得该模型不匹配 UE 侧部署的 CSI 生成模型。

对于 NW 侧模型，因为部署有和 D1 配对的 E1。并且还能获得 UE 反馈的 E2。由此可以计算 2 个 SGCS 及其差异。由此，可以判定 NW 侧训练问题、数据漂移问题，如果既不是 NW 侧训练问题，又不是数据漂移问题，就可以认定为 UE 侧训练问题。同样，在 UE 侧也可以定义类似的机制，具体规则可以参考 R1-2500058 的描述。

跨厂商协作训练

在 Rel-18 初步确定了双侧训练类型优先级，确定了在 UE 侧和 NW 侧分别训练模型的相应部分，解决双侧模型训练的隐私性和互操作性问题。但胃就具体的协作训练双侧模型的数据集、模型的交互达成一致。在 Rel-19 进一步讨论跨厂商协作的方向，并收敛出 3 个可行的训练方式。

协作训练类型

Rel-18 确定了 CSI 压缩的协作训练类型包括：

- Type 1：在单个实体联合（jointly）训练双侧模型，如 UE 或 NW 侧；
- Type 2：在 UE 和 NW 侧联合（jointly）训练双侧模型，如通过在线训练。
- Type 3：在 UE 和 NW 侧分别双侧模型相应部分，如 UE 侧训练编码器模型，NW 侧训练解码器模型。可以是 UE 侧首先开启模型训练，也可以是 NW 侧首先开启模型训练。

Type 1 和 Type 2 采用联合训练，意味着双侧模型需在同一个循环内完成前向传播和后向传播。Type 1 是在单个节点，Type 2 需要 UE 侧和 NW 侧的梯度交互。在 Rel-18 的讨论中，Type 2 优先级低，因为梯度交互导致传输开销大、时延要求高。

各训练方式的优缺点总结如下：

	Type 1	Type 2	Type 3
隐私性	N	Y	Y
泛化性	不适合另一侧特有的条件	--	Y
可更新性（允许单侧模型更新）	N	N	Y

在协作训练方式上，采用 Type 3 时，UE 和 NW 采用序列训练（sequential training）方式时需要交互数据集或其他信息。数据集的交互或者共享可以通过空口信令或者非空口信令的方式实现。此时，需要研究 NW 共享数据集/其他信息由 UE 或 NW 发起、数据类型/格式、量化和去量化的信息的协议影响。这一部分内容在 R19 得到进一步研究。

协作训练方向

在 R19 的讨论中跨厂商协作需考虑降低复杂度以及提升性能。首先，会议讨论了 5 个可选项目，而后收敛并确定 3 个跨厂商协作训练方向。R20 的后续标准研究就基于这 3 个跨厂商协作训练方向进行。

在 R19 RAN1#116 次会议的讨论中，起初就确定了跨厂商协作的 5 个选项，规范 CSI 压缩用例跨厂商协作需标准化的内容。这 5 个选项的模型设计灵活性和互操作性的复杂度不同，总的来说不能同时做到最优，必须要在模型设计的灵活性和双侧模型用例的互操作性（interoperability）之间取舍 [R1-2400166]。5 个选项的内容如下：

- 选项 1：完全标准化的参考模型，参考模型包括模型的结构以及模型的参数。
- 选项 2：标准化数据集。

- 选项 3：标准化参考模型结构。在 NW 侧和 UE 侧之间交互模型参数。
- 选项 4：标准化数据或数据集格式。在 NW 侧和 UE 侧之间交互数据集。
- 选项 5：标准化模型格式。在 NW 侧和 UE 侧之间交互参考模型。

这 5 个选项内容并非完全独立，而是相互重叠。在提出这几个选项后的几次会，确定了选项的优先级，选项 3~5 的内容得到进一步细化，而选项 2 则作为低优先级。基于 RAN1#116bis 会议结论可知，从 RAN1 角度，选项 1 和选项 2 存在 2 个问题：首先可扩展性最差，难以适应复杂的部署环境，其次，制定统一的模型或数据集，涉及的标准工作量大。选项 1 最后成为独立的跨厂商协作方向，即方向 C。

在 3GPP 规范中，通常规定发送端信号和信道，而把接收端交给实现，以灵活处理多用户、干扰抑制和信道估计。就 CSI 压缩用例而言，规定 UE 侧的行为，允许 NW 侧的灵活处理就可以保持互操作性[R1-2400166]。选项 3~5 细化内容如下：

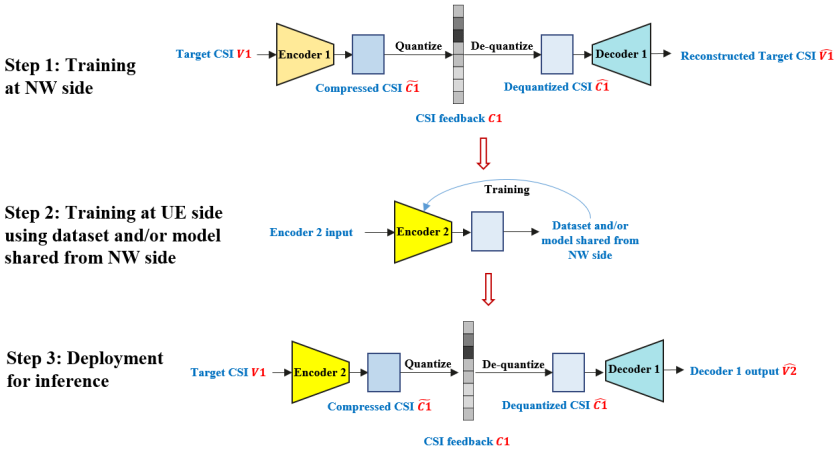
- 选项 3a/5a：NW 侧将模型或模型参数发送给 UE 侧，可以发送 CSI 生成模型、CSI 重建模型、双侧模型中的一种。此外，还可以发送附加信息，如性能目标、数据集或用于数据集收集的信息。
  - 选项 3a-1/5a-1：从 NW 侧交互给 UE 侧的内容是 CSI 生成模型。对选项 3a 为模型参数，对选项 5a 为模型。
  - 选项 3a-2/5a-2：从 NW 侧交互给 UE 侧的内容是 CSI 重建模型。对选项 3a 为模型参数，对选项 5a 为模型。
  - 选项 3a-3/5a-3：从 NW 侧交互给 UE 侧的内容是双侧模型。对选项 3a 为模型参数，对选项 5a 为模型。
- 选项 3b/5b：NW 侧将模型或模型参数发送给 UE 侧。UE 侧得到后不经过离线工程而直接用于推理。对于选项 3b，交互的内容为 CSI 生成模型的模型参数。对于选项 5b，交互的内容为 CSI 重建模型。
- 选项 4：NW 侧将数据集发送给 UE 侧。UE 接收到数据集后，UE 侧（在 UE 侧的 OTT 服务器）可以离线工程。此外，还可以发送附加信息，如性能目标帮组 UE 侧离线工程或提供性能指导（guidance）。子选项包括：
  - 选项 4-1：数据集包括{target CSI 和 CSI feedback}。
  - 选项 4-2：数据集包括{ CSI feedback 和重建 CSI}。
  - 选项 4-3：数据集包括{ CSI feedback、CSI feedback 和重建 CSI}。

基于以上 5 个选项，在 RAN1#118 次会，跨厂商协作按照 NW 共享给 UE 的内容归类为以下 3 个方向：

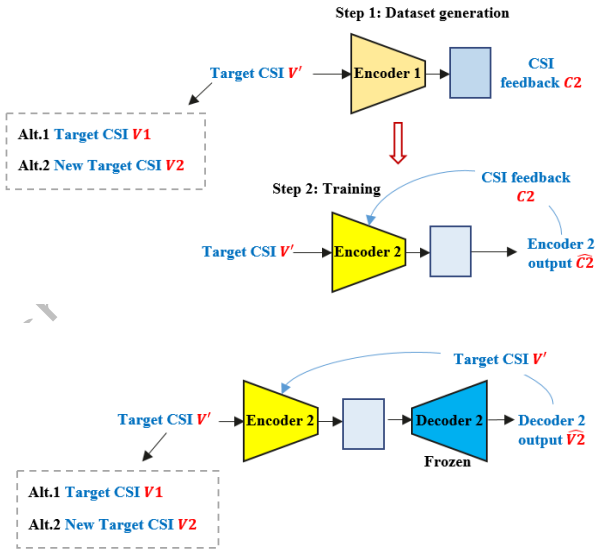
- 方向 A：NW 侧发起的跨厂商协作。NW 发送的内容用于 UE 侧离线工程 CSI 生成模型。方向 A 包括了选项 3a/5a 以及选项 4。
  - 更进一步的，会议还讨论各选项可能的训练 CSI 生成模型的步骤、性能监测的中间 KPI。
  - 在 RAN1#118bis 确定了方向 A 的选项 3a-1/4-1 训练方式。实际上，在 FL summary 中把选项 5a-1 和 3a-1 一起描述，未做区分，而最终的结论中移除了选项 5a-1。
    - 步骤 1：NW 侧训练基于训练数据集获得双侧模型 E1 和 D1。
    - 步骤 2：UE 侧基于 NW 发送的数据集或 CSI 生成模型的参数直接或间接的获取 E1。具体的训练方式可以参考 R1-2408840 的描述。
      - 对于选项 3a-1：NW 发送 E1 的模型参数给 UE。
        - UE 有 2 种实现方式获得 CSI 生成模型 E2。方式 1： UE 直接训练 E2。由于选项 3a-1，UE 侧训练仅只有模型参数，必须借助 UE 侧收集数据训练得到 D2，因此该方式可能导致 D2 和 D1 存在数据不匹

配。方式 2：UE 先基于 UE 侧收集的数据（此外，NW 可能交互 target CSI 给 UE）以及 E1 训练得到 D2，然后基于 D2 训练得到 E2。

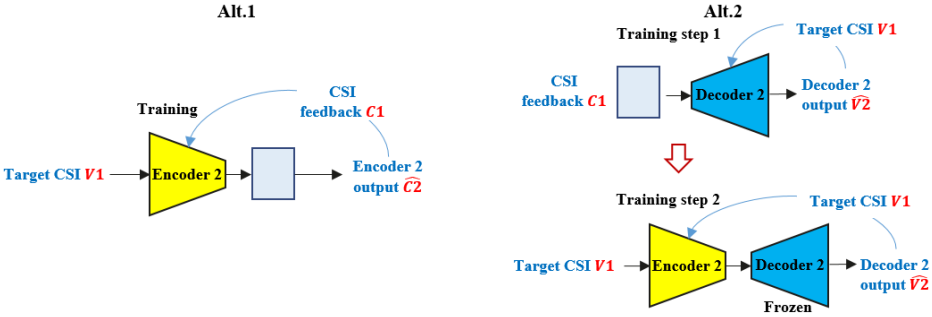
- 对应选项 4-1：NW 发送{target CSI, CSI feedback}数据集给 UE。
  - UE 有 2 种获得 CSI 生成模型 E2 的方式。方式 1： UE 直接训练 E2。相比选项 3a-1，选项 4-1 可能存在 UE 侧训练 D2 不准确的问题。方式 2：UE 首先基于训练数据集获得 D2，然后基于 D2 训练 E2。



方向 A 的训练方式，参考 R1-2407656



选项 3a-1 的训练过程，参考 R1-2407656



选项 4-1，步骤 2 的训练过程，参考 R1-2407656

- 方向 B：NW 侧训练模型。NW 发送的内容用于 UE 侧直接用于推理。因此，方向 B 包括了选项 3b。由于选项 5b 没有得到后续的研究，因此不包括在方向 B 中。
- 方向 C：NW 侧和 UE 侧均可以在未通知另一侧的情况下完成离线工程。包括选项 1。
  - 进一步的，会议还讨论并确定了标准化可扩展的双侧模型结构。该模型结构适用于用例 0，采用 Transformer 架构，模型输入为端口子带域表示的 token 和 feature。

对于每个方向，R19 都细致的研究了待解决问题，涉及到可扩展性、性能、开销、数据不匹配、隐私泄露问题的评估和解决方案。这里不做赘述。这些方向中涉及到多个选项，经过筛选，RAN1 于 RAN1#121 次会收敛结论如下：

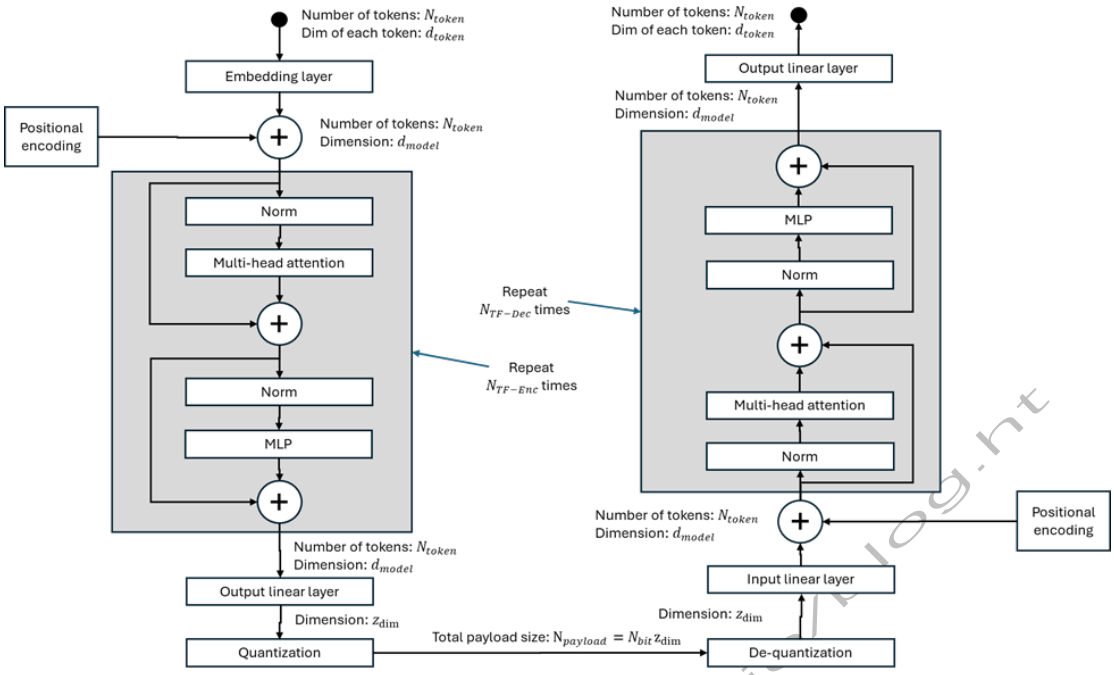
- 方向 A 和方向 C 都是可行的跨厂商协作训练方式。方向 B 得到的研究很少，只有零星的厂商关注，未声明为可行的跨厂商协作训练方式。
- 方向 A 支持选项 3a-1 和 4-1。选项 3a-1 中，在进行参考的编码器模型参数交互时，可以有也可以没有{target CSI}。RAN1 认为，可以通过定义标准化的模型参数、数据集共享的信令，用于降低跨厂商协作复杂度。这些标准化信令可以是通过 OTA 也可以是非 OTA 的信令传输。其他选项不支持的原因是重建模型的参数、重建 CSI 的分发会泄露 NW 隐私。
- 方向 C 可以作为最低性能要求。方向 C 采用的模型结构和参数为采用人造数据训练获得。RAN1 获得的可扩展的双侧模型结构是可行的，但如何支持互操作性则是 RAN4 需要研究的内容。
- 选项 3a-1 和方向 C 都需要规定参考模型结构和参数。选项 3a-1 规定编码器的参考模型结构和参数，方向 C 规定参考双侧模型的结构和参数。选项 3a-1 采用的参考模型结构可以和方向 C 的参考模型结构一致，便于降低标准化复杂度。选项 3a-1 采用的编码器模型结构需要基于人造数据确定。

模型的可用性和泛化性

在 AI/ML 通用框架的讨论中初步讨论了模型泛化性影响。即，支持场景/配置特定的模型可以获得更好的性能，但模型的泛化性较差。而泛化性好的模型，则在特定场景/配置难以取得最优的性能。通过模型切换、模型更新、模型泛化可以提升在特定场景的性能。

对于方向 A 选项 3a-1 和方向 C，都依赖于标准化的可扩展模型结构。可扩展是指对于不同数量的 Tx 端口数、CSI feedback 负载大小、带宽、时隙时，模型具备可用性（feasibility）。对于方向 A 选项 3a-1，关注可扩展模型结构的规范，而方向 C 则关注可扩展的模型结构和模型参数的规范。在 R19 讨论了多个子用例，在可扩展模型结构中也研究了这些子用例的影响，但明确的是，这些子用例都会基于用例 0 扩展。因此，定义用例 0 的可扩展模型结构成为基础问题。

最终会议确定的可扩展模型结构如下所示：



在 RAN1 的假设中，该模型结果的输入为预编码矩阵，可以为空频域、角度时延域的表征，如 eType II 码本的 W2. 该模型的核心为 transformer。同时，考虑了量化和解量化的影响。另外，RAN1 假设该模型用于单侧的推理。扩展到多层时，采用相同的模型结构。具体的参数解释参考 TR 38.843 的表格 5.1.6.

支持层数>1

如前所述的模型推理为对单个层的 CSI 采用一个双侧模型压缩和解压缩。当扩展到多层 CSI 的压缩时，模型应该如何部署？讨论确定的几种可能的模型部署方式包括：

- Option 1-1 (rank specific)：给定秩，可以确定一个应采用的双侧 CSI 压缩模型。不同秩所对应的 CSI 压缩模型不同。
- Option 1-2 (rank common): 不同秩采用相同双侧 CSI 压缩模型。
- Option 2-1 (layer specific and rank specific): 不同秩用不同的 CSI 压缩模型，不同层用不同 CSI 压缩模型。不同秩、层的组合，其采用的 CSI 压缩模型不同。
- Option 2-2 (layer-specific and rank common): 不同秩采用相同的 CSI 压缩模型，不同层用不同 CSI 压缩模型。不同层，其采用的 CSI 压缩模型不同。
- Option 3 (layer-common and rank common)：不同秩、层的组合，其采用的 CSI 压缩模型相同。

选项 3 最简便，但未考虑不同层、不同秩目标 CSI 的差异。选项 2-2 或取得最有性能，但训练复杂。Rel-18 未就选项达成一致。具体的讨论在 Rel-20 标准化阶段或取得结论。

数据收集

考虑到双侧模型的训练采用协作训练类型 3，在 NW 侧和 UE 侧均可以收集数据用于

各自的模型训练。

为了 NW 侧模型训练，需要定义 NW 侧数据收集的机制。NW 侧数据收集的内容为目标 CSI 以及辅助信息。讨论明确如下协议影响：

- CSI-RS/SRS 测量增强以实现高精度测量。
- 数据的类型和格式的规定：类型包括信道矩阵和或预编码矩阵（即目标 CSI 的类型）。格式为数据的量化方式，包括标量量化和向量量化。
  - 标量量化是指将目标 CSI 的元素逐个量化，如量化实部和虚部或者量化幅度和相位。
  - 向量量化是指基于 Rel-16 eType II 或 Rel-18 用于 PMI 预测的 eType II 码本。一种方式是直接复用现有码本，另一种方式是增强 eType II 码本的参数组合。
  - 各公司给出开销和训练性能。可知，采用标量量化方式反馈开销大，计算开销相对较低。现有的 eType II 码本的 PC6 和 PC8 均可取得较好的性能，同时反馈开销最低。增强 eType II，如采用更大非零系数比例、空频域基向量数，复杂度相比 Float32 标量量化降低(94%~97.5%)，性能有一定提升(相比 eType II PC6.8, 0.7%~5.4% SGCS gain)。是否需要增强相关测量和报告配置是需要进一步研究的问题，如是否支持 L3 信令报告？
- 辅助信息：时间戳、小区 ID、数据类别 ID（用于区分场景、配置、站点、数据质量）。怎么定义数据质量是需要进一步研究的问题。

为了 UE 侧模型训练，需要定义 UE 侧数据收集。UE 侧数据收集的协议影响如下：

- CSI-RS 测量增强以实现高精度测量。
- 辅助信息：数据类别 ID，用于区分数据收集的场景、配置、站点等信息。该信息不应该暴露 NW 侧的隐私。
- UE 侧数据收集的发起，可以是 UE 侧请求或者 NW 配置中的一种。

用例

在 Rel-18 研究结束后，CSI 压缩用例并没有进入 WI 阶段。原因如下：

TR 38.843—Rel-18 Version:

At least the following aspects are the reasons for the lack of RAN1 consensus on the recommendation of CSI compression for normative work:

- Trade-off between performance and complexity/overhead.
- Issues related to inter-vendor training collaboration.

Rel-18 评估空频域 CSI 压缩的性能，认为双侧 CSI 压缩性能增益受限，只有不到 10% 的 SGCS 增益。不同 CSI 反馈开销，吞吐量提升不到 10%。单纯压缩空频域 CSI 带来的提升受限。且模型复杂度高，普遍在 1M~17M 参数个数，10M~800M FLOPs。

另一方面的争议在于跨厂商协作训练的讨论未完全细化，争议很多。

TR 38.843—Rel-18 Version:

The pros and cons are analysed for each training collaboration types, and each training collaboration type has its own benefits and limitations in different aspects. The study has investigated the feasibility of the studied training collaboration types and necessity of corresponding potential RAN1 specification impact. However, not all aspects have been

concluded.

这两个问题为 R19 CSI 压缩研究的主要问题。了解决性能和增益/开销的权衡，R19 将扩展空频域 CSI 压缩到空时频域 CSI 压缩。跨厂商协作的遗留问题继续在 R19 讨论。

其中一个重要方面为子用例的扩展。为了实现性能和复杂度的权衡，R19 扩展空频域 CSI 压缩到空时频域 CSI 压缩。在 RAN1#116 次会议，达成如下扩展用例的共识。

Table 5.1-3: Additional sub-use cases for CSI feedback using two-sided model

Case	Target CSI slot(s)	Whether the UE uses past CSI information	Whether the NW uses past CSI information
0	Present slot	No	No
1	Present slot	Yes	No
2	Present slot	Yes	Yes
3	Future slot(s)	Yes	No
4	Future slot(s)	Yes	Yes
5	Present slot	No	Yes

名称解释：

- Target CSI slot(s)：报告的 CSI feedback 对应的时隙；
- Present slot：用于生成 CSI 报告的最近时刻的 CSI-RS 测量时隙。CSI feedback 是压缩的当前时隙的结果。
- Future slot：当前时隙之后的时隙。未来时隙可能包含当前时隙。CSI feedback 是预测的未来时隙的结果。
- Past CSI information：UE 和 NW 侧模型可以使用了历史 CSI 信息用于模型推理。

各子用例的说明如下：

- Case 0: 为 R18 空频域 CSI 压缩用例。
- Case 1: 为 R18 空频域 CSI 压缩用例。区别于 Case 0，UE 侧可以应用历史 CSI 信息用于 CSI 压缩，NW 侧恢复 CSI 并没有用到历史 CSI 信息。
- Case 2：UE 和 NW 侧都利用历史 CSI 信息用于压缩当前时隙的 CSI。例如，UE 侧编码器输入为当前时隙测量的 target CSI 和历史 CSI，UE 侧编码器反馈压缩的 CSI。在 NW 侧的模型输入为当前时隙反馈的 CSI 和历史 CSI。由于 CSI 压缩利用了数据的时间相关性，压缩程度可以更高，反馈开销更低。

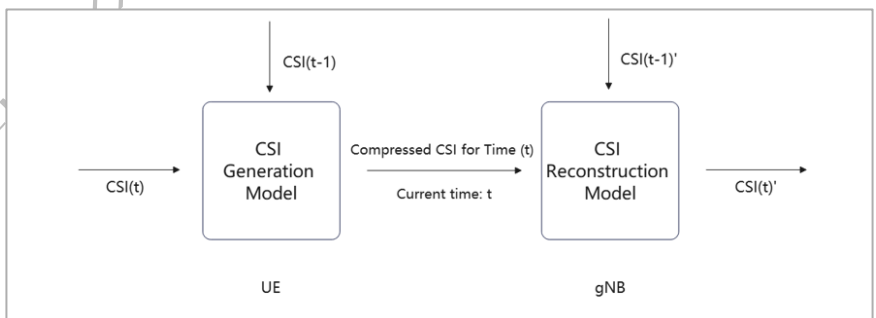


Figure 1 The procedure of leveraging past CSI instances to assist the current CSI compression

From R1-2400265

- Case 3：UE 基于历史 CSI 预测未来时刻的 CSI，并将压缩后的 CSI 发送给 NW。NW 基于压缩的 CSI 直接恢复预测的 CSI。例如，预测和压缩模型为 2 个独立模型，先基于历史 CSI 预测得到未来 CSI，然后将未来 CSI 压缩后反馈给 NW，NW 侧只部署 CSI 重建模型恢复未来 CSI。

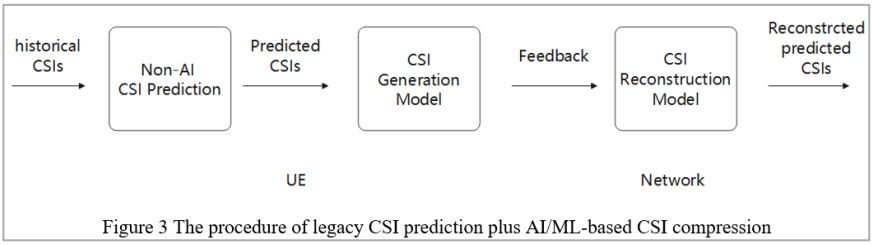


Figure 3 The procedure of legacy CSI prediction plus AI/ML-based CSI compression

From R1-2400265

- Case 4：UE 和 NW 都基于历史 CSI 预测未来 CSI。UE 侧编码器输入为过去 CSI，输出为压缩的未来 CSI。NW 侧恢复未来 CSI 时还用到了过去 CSI。和 Case 2 类似，区别是输出为未来 CSI，可以实现 CSI 压缩/预测。

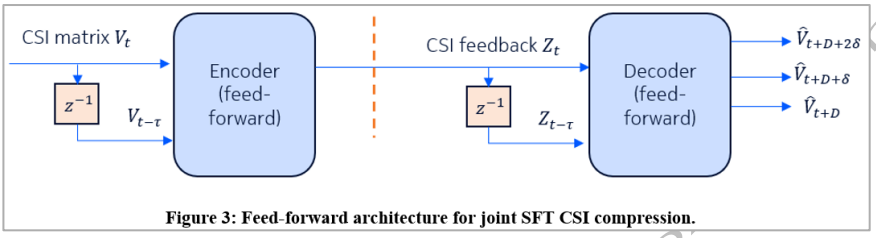


Figure 3: Feed-forward architecture for joint SFT CSI compression.

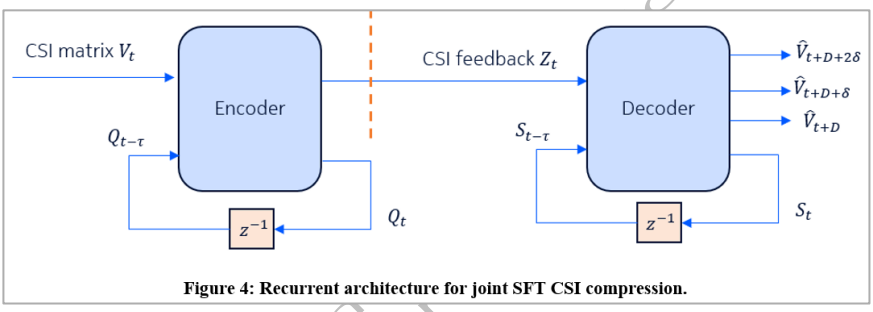


Figure 4: Recurrent architecture for joint SFT CSI compression.

From R1-2400796

- Case 5：凑数的，为富士通提出。NW 侧 CSI 重建模型可以基于历史 CSI 提升恢复性能。

在 RAN1#121 次会议，绝大多数赞同以 Case 0 作为规范性工作的基线，同时 Case 2,3 有一定性能和复杂度的权衡，但协议影响相对更大。未有公司支持 Case 1,4,5。最终，Case 0,2,3 的优先级高于 Case 1,4,5。

CSI 报告配置

主要讨论集中在 Rel-18，Rel-19 几乎没有进展。而后则进入 Rel-20 的讨论中。

CSI 压缩报告的内容除了 CSI feedback，还应以传统的 CSI 报告机制作为基线。报告内容还可以包括 RI, CQI 信息。

CQI 的确定方式可能采用 CSI 重建模型，也可以不采用 CSI 重建模型。方式包括：

- Option 1: 不采用重建模型的输出计算 CQI，于 RAN1 可能无标准影响。
  - Option 1a: 基于实测的 target CSI 计算。
  - Option 1b: 基于实测的 target CSI 计算，并做调整。
  - Option 1c: 基于传统码本计算。
- Option 2:采用重建模型的输出计算 CQI。
  - Option 2a: UE 侧部署 CSI 重建模型，然后计算获得。UE 侧的重建模型可能不同于

NW 侧的重建模型。导致 UE 侧重建模型性能监测的需求，以及可能的 NW 侧隐私泄露。

- Option 2b: 采用 2 部分计算方法。第一步是 NW 发送经过 CSI feedback 生成的预编码，然后 UE 基于接受的预编码后的 CSI-RS 测量获得 CQI。时延过大，信道可能已经变化，

CSI 压缩推理报告配置也在 R18 讨论过，已经明确 CSI 压缩推理报告会以传统码本报告为基线。首先，CSI 报告可以划分为 CSI part 1（包括第一个码字的 CQI，RI 以及 part 2 的信息）和 CSI part 2（包含 CSI 生成模型的输出）两个部分。其次，为确定 CSI 报告内容，CSI 报告配置中可能也包含 CSI-RS 资源、码本子集约束、秩约束等。CSI 映射和省略优先级的规则可能需要重新定义。报告配置也会包括其他和负载相关的信息。

考虑到 CSI 压缩可能同时占用 APU 和 CPU，CSI 处理过程也可能发送变化，如 PU 的占用、PU 的时间线等。

Rel-20 展望

在 RAN#108 次会议，AI 空口第二阶段的 WI 立项成功。主要讨论 CSI 压缩的标准化。WI 将 CSI 压缩分为 3 个议程，分别是推理、除推理外的其他部分以及跨厂商协作训练，均由 RAN1 主导。数据收集、LCM 由 RAN2 主导，互操作性和 RRM 要求由 RAN4 主导。

当前，Rel-20 已启动研究工作，有相应的研究点值得深入讨论。

在推理部分，需要确定目标 CSI 的类型、格式，该类型和格式可能通用于推理、监测、数据收集、数据集共享，为重中之重。其次是推理报告的配置和上报，当前确定在现有的 CSI 报告框架中增强，如负载的确定、报告格式等。

在其他部分，需要确定 NW 侧数据收集的格式和信令。此外，性能监测的讨论需要收敛。NW 侧数据收集的格式可以复用到监测和数据集共享，相对标准工作量更小。UE 侧性能监测涉及到代理模型或者 NW 侧发送高精度的重建 CSI，其标准化难度大。

在跨厂商协作训练部分，需要规定数据集、模型参数共享的格式以及辅助信息。其相应的容器则由高层定义，或采用对协议透明的方式发送。数据集、模型参数格式和内容需要进一步研究。

面向 6G，CSI 压缩的讨论主要是扩展 Rel-20 未涉及的其他子用例或者训练方式。在 6G 的第一个版本，如何引入 CSI 压缩，是否需要增强是值得研究的点。