

模型监控方法、装置、系统、存储介质及程序产品

本申请要求于 2023 年 07 月 14 日提交中国国家知识产权局、申请号为 202310870686.9、发明名称为“模型监控方法、装置、系统、存储介质及程序产品”的中国专利申请的优先权，其全部内容通过引用结合在本申请中。

技术领域

本申请涉及通信技术领域，尤其涉及一种模型监控方法、装置、系统、存储介质及程序产品。

背景技术

模型监控指的是监控人工智能 (artificial intelligence, AI) 模型的性能。判断 AI 模型是否在正常工作，如果 AI 模型性能较差，则需要切换至非 AI 模式，或者更换 AI 模型，或者更新 AI 模型等。模型监控可以是监控 AI 模型输出的准确性（也可称为中间关键绩效指标 (key performance indicator, KPI)）。监控 AI 模型输出的准确性是通过对比 AI 模型的输出与对应的真值 (ground-truth) 之间的差异，判断 AI 模型的性能是否满足要求。

AI 模型可以用于信息的预测。例如，AI 模型的输出是 $t+5$ 时刻的预测信息，则通常使用 $t+5$ 时刻的真值信息与 $t+5$ 时刻的预测信息进行对比，如果中间 KPI 满足预设要求，则说明该 AI 模型性能较好；反之，则说明该 AI 模型性能较差。而考虑到 AI 模型用在实际系统中，预测的信息用于基站调度下行传输，但是基站调度的下行传输并不一定正好在预测信息所对应的时刻。例如，AI 模型预测的信息为 $t+5$ 时刻的信息，即如果参考信号 (reference signal, RS) 的周期为 5ms，则该 AI 模型仅能预测 5ms 格点上的信息，但基站有可能在任意时刻调度下行传输，如在 $t+3$ 时刻调度下行传输，则仅监控预测信息在其对应的时刻上的性能并不能反应出使用该预测信息进行下行传输时的实际性能。

如何灵活地监控预测模型的性能，使得预测性能能够反应出实际的系统性能，目前没有相应的解决方案。

发明内容

本申请提供一种模型监控方法、装置、系统、存储介质及程序产品，以期支持灵活地监控预测模型的性能，使得预测性能能够反应出实际的系统性能。

第一方面，提供了一种模型监控方法，所述方法包括：获取第一信息，所述第一信息包括一个或多个真值信息分别对应的一个或多个第一索引；以及获取所述一个或多个第一索引分别对应的一个或多个真值信息；其中，所述一个或多个真值信息用于获取人工智能 AI 模型的性能的监控结果，所述 AI 模型的输出为预测信息。

在该方面中，终端设备通过获取一个或多个真值信息的一个或多个第一索引，可以获取特定的时间单元对应的真值信息，从而可以灵活地监控预测模型的性能。

第二方面，提供了一种模型监控方法，所述方法包括：获取第一信息，所述第一信息包括用于测量真值信息的一个或多个参考信号的时域资源位置；以及根据所述第一信息，确定一个或多个真值信息；其中，所述一个或多个真值信息用于获取所述 AI 模型的性能的监控结果，所述 AI 模型的输出为预测信息，所述一个或多个真值信息中的至少一个真值信息的时域资源位置与所述预测信息的时域资源位置不同。

在该方面中，终端设备通过获取用于测量真值信息的一个或多个参考信号的时域资源位置，可以获取任一时域资源位置对应的真值信息，从而可以灵活地监控预测模型的性能。

在一种可能的实现中，所述方法还包括：发送所述一个或多个真值信息。

在该实现中，通过向接入网设备发送一个或多个真值信息，使得接入网设备可以根据接收到的一个或多个真值信息，获取实际的系统的模型监控结果，使得可以灵活地监控预测模型的性能，预测性能能够反应出实际的系统性能。

在另一种可能的实现中，所述方法还包括：获取一个或多个预测信息，所述一个或多个预测信息为人工智能 AI 模型的输出；以及发送所述一个或多个预测信息。

在该实现中，通过向接入网设备发送一个或多个真值信息，使得接入网设备可以根据接收到的一个或

多个真值信息和一个或多个预测信息，获取实际的系统的模型监控结果，使得可以灵活地监控预测模型的性能，预测性能能够反应出实际的系统性能。

在又一种可能的实现中，所述方法还包括：获取一个或多个预测信息，所述一个或多个预测信息为人工智能 AI 模型的输出；根据所述一个或多个真值信息和所述一个或多个预测信息，获取所述 AI 模型的一个或多个监控结果。

在该实现中，终端设备存储有 AI 模型，可以预测信息，且终端设备还可以获取监控结果，使得接入网设备可以获取实际的系统的模型监控结果，使得可以灵活地监控预测模型的性能，预测性能能够反应出实际的系统性能。

在又一种可能的实现中，所述获取一个或多个预测信息，包括：接收第二信息，所述第二信息包括所述一个或多个预测信息分别对应的一个或多个第二索引，或所述第二信息包括所述一个或多个预测信息中的每个预测信息相对于参考资源对应的时间单元的相对时间单元值；以及根据所述第二信息，获取一个或多个预测信息。

在该实现中，终端设备可以根据接入网设备的指示获取相应的预测信息。

在又一种可能的实现中，所述一个或多个第一索引、所述一个或多个第二索引是基于参考资源对应的时间单元的。

在又一种可能的实现中，所述参考资源对应的时间单元为以下任意一个：上报监控结果报告的资源所在的第一时间单元；或调度所述模型监控的下行控制信令所在的第二时间单元；或基于所述第一时间单元或所述第二时间单元确定的第三时间单元；或预定义/预配置的第四时间单元。

在又一种可能的实现中，发送监控结果报告；其中，所述监控结果报告包括所述一个或多个监控结果；或所述监控结果报告包括所述一个或多个监控结果中大于或等于第一阈值的至少一个监控结果；或所述监控结果报告包括所述一个或多个监控结果中小于或等于第二阈值的至少一个监控结果；或所述监控结果报告包括所述一个或多个监控结果中的最优监控结果对应的预测信息的索引和真值信息的索引；或所述监控结果报告包括所述一个或多个监控结果中的最差监控结果对应的预测信息的索引和真值信息的索引；或所述监控结果报告包括多个监控结果中的第一监控结果的绝对值、以及所述多个监控结果中除所述第一监控结果之外的其它监控结果相对于所述第一监控结果的相对值。

在上述实现方式中，终端设备可以向接入网设备上报获取的所有监控结果，使得接入网设备可以获取完整的监控结果；

在某些场景下，接入网设备可以仅知道是否存在监控结果表明模型性能较好，无需获取所有的监控结果。因此，该监控结果报告包括上述一个或多个监控结果中大于或等于第一阈值的至少一个监控结果。该第一阈值可以是接入网设备配置或预定义的。该实现方式可以节省信令开销。

在某些场景下，接入网设备可以仅知道是否存在监控结果表明模型性能较差，无需获取所有的监控结果。因此，该监控结果报告可以包括上述一个或多个监控结果中小于或等于第二阈值的至少一个监控结果。该第二阈值可以是接入网设备配置或预定义的。该第二阈值小于上述第一阈值。该实现方式可以节省信令开销。

在某些场景下，接入网设备可以仅知道是否存在监控结果表明模型性能较好，无需获取所有的监控结果。因此，该监控结果报告可以包括上述一个或多个监控结果中的最优监控结果对应的预测信息的索引和真值信息的索引。接入网设备可以预先接收到了一个或多个预测信息、一个或多个真值信息，则可以根据最优监控结果对应的预测信息的索引和真值信息的索引，自身获得该最优监控结果。该实现方式可以节省信令开销，且对终端设备的计算能力要求较低。

在某些场景下，接入网设备可以仅知道是否存在监控结果表明模型性能较差，无需获取所有的监控结果。因此，该监控结果报告可以包括上述一个或多个监控结果中的最差监控结果对应的预测信息的索引和真值信息的索引。接入网设备可以预先接收到了一个或多个预测信息、一个或多个真值信息，则可以根据最差监控结果对应的预测信息的索引和真值信息的索引，自身获得该最差监控结果。该实现方式可以节省信令开销，且对终端设备的计算能力要求较低。

在某些场景下，终端设备可以采用差分上报方式，该监控结果报告包括多个监控结果中的第一监控结果的绝对值、以及多个监控结果中除第一监控结果之外的其它监控结果相对于第一监控结果的相对值。该实现方式可以节省信令开销。

在又一种可能的实现中，所述第一监控结果是基于在获取所述多个监控结果时，使用的所述真值信息的顺序和/或使用的所述预测信息的顺序确定的。

在该实现方式中,可以根据终端设备生成监控结果时使用真值信息和/或预测信息的顺序,确定一个或多个监控结果的生成顺序,从而可以根据该生成顺序确定第一监控结果。

在又一种可能的实现中,终端设备可以根据接入网设备的指示上报部分监控结果。例如,可以通过位图(bitmap)的方式指示终端设备上报上述所有监控结果中的部分监控结果。

在又一种可能的实现中,所述监控结果报告和用于发送参考信号测量结果的信息反馈在同一个信息报告中上报,或所述监控结果报告和用于发送参考信号测量结果的信息反馈在不同的信息报告中上报。

第三方面,提供了一种模型监控方法,所述方法包括:发送第一信息,所述第一信息包括一个或多个真值信息分别对应的一个或多个第一索引;以及接收所述一个或多个第一索引分别对应的一个或多个真值信息;其中,所述一个或多个真值信息用于获取人工智能 AI 模型的性能的监控结果,所述 AI 模型的输出为预测信息。

在该方面中,接入网设备通过向终端设备指示一个或多个真值信息的一个或多个第一索引,使得终端设备可以获取特定的时间单元对应的真值信息,从而可以灵活地监控预测模型的性能。

第四方面,提供了一种模型监控方法,所述方法包括:发送第一信息,所述第一信息包括用于测量真值信息的一个或多个参考信号的时域资源位置;以及根据所述第一信息,接收一个或多个真值信息;其中,所述一个或多个真值信息用于获取所述 AI 模型的性能的监控结果,所述 AI 模型的输出为预测信息,所述一个或多个真值信息中的至少一个真值信息的时域资源位置与所述预测信息的时域资源位置不同。

在该方面中,接入网设备通过向终端设备指示用于测量真值信息的一个或多个参考信号的时域资源位置,使得终端设备可以获取任一时域资源位置对应的真值信息,从而可以灵活地监控预测模型的性能。

在一种可能的实现中,所述方法还包括:获取一个或多个预测信息;以及根据所述一个或多个真值信息和所述一个或多个预测信息,获取所述 AI 模型的性能的一个或多个监控结果。

在另一种可能的实现中,所述方法还包括:接收监控结果报告,所述监控结果报告包括一个或多个监控结果,所述一个或多个监控结果为根据所述一个或多个真值信息和一个或多个预测信息获取的。

在又一种可能的实现中,所述方法包括:发送第二信息,所述第二信息包括所述一个或多个预测信息分别对应的一个或多个第二索引,或所述第二信息包括所述一个或多个预测信息中的每个预测信息相对于参考资源对应的时间单元的相对时间单元值;以及接收所述一个或多个预测信息。

在又一种可能的实现中,所述一个或多个第一索引、所述一个或多个第二索引是基于参考资源对应的时间单元的。

在又一种可能的实现中,所述参考资源对应的时间单元为以下任意一个:上报监控结果报告的资源所在的第一时间单元;或调度所述模型监控的下行控制信令所在的第二时间单元;或基于所述第一时间单元或所述第二时间单元确定的第三时间单元;或预定义/预配置的第四时间单元。

在又一种可能的实现中,所述监控结果报告包括所述一个或多个监控结果;或所述监控结果报告包括所述一个或多个监控结果中大于或等于第一阈值的至少一个监控结果;或所述监控结果报告包括所述一个或多个监控结果中小于或等于第二阈值的至少一个监控结果;或所述监控结果报告包括所述一个或多个监控结果中的最优监控结果对应的预测信息的索引和真值信息的索引;或所述监控结果报告包括所述一个或多个监控结果中的最差监控结果对应的预测信息的索引和真值信息的索引;或所述监控结果报告包括多个监控结果中的第一监控结果的绝对值、以及所述多个监控结果中除所述第一监控结果之外的其它监控结果相对于所述第一监控结果的相对值。

在又一种可能的实现中,所述第一监控结果是基于在获取所述多个监控结果时,使用的所述真值信息的顺序和/或使用的所述预测信息的顺序确定的。

在又一种可能的实现中,接入网设备也可以指示终端设备上报部分监控结果。例如,可以通过位图的方式指示终端设备上报上述所有监控结果中的部分监控结果。

在又一种可能的实现中,所述监控结果报告和用于发送参考信号测量结果的信息反馈在同一个信息报告中上报,或所述监控结果报告和用于发送参考信号测量结果的信息反馈在不同的信息报告中上报。

在上述第一方面至第四方面中的任一方面中,所述预测信息是预测信道状态信息(channel state information, CSI),所述真值信息是真值 CSI;或所述预测信息是预测波束的信息,所述真值信息是真值波束的信息。

其中,预测波束的信息可以是预测的最优波束或预测的最优波束的参考信号接收功率(reference signal receiving power, RSRP)中的一项或多项。真值波束的信息可以是实际的最优波束或实际的最优波束的 RSRP 中的一项或多项,换句话说,所述真值波束信息包括以下至少一种:真值最优波束,真值参考信号

接收功率。

上述第一方面、第二方面的方法可以由终端设备执行,也可以由应用于终端设备的模块(例如处理器、芯片、或芯片系统等)执行,还可以由能实现全部或部分终端设备功能的逻辑节点、逻辑模块或软件实现。

上述第三方面、第四方面的方法可以由接入网设备执行,也可以由应用于接入网设备的模块(例如处理器、芯片、或芯片系统等)执行,还可以由能实现全部或部分接入网设备功能的逻辑节点、逻辑模块或软件实现。

第五方面,提供了一种模型监控装置,用于实现上述第一方面或第一方面的任意一种实现中的模型监控方法;或者,用于实现上述第二方面或第二方面的任意一种实现中的模型监控方法。该装置可以是终端设备,也可以是应用于终端设备的模块(例如处理器、芯片、或芯片系统等),还可以是能实现全部或部分终端设备功能的逻辑节点、逻辑模块或软件。一种实现中,该模型监控装置可以包括发送单元、接收单元,还可以包括处理单元。发送单元和接收单元可以是独立的,也可以是结合在一起的(可以称为“收发单元”)。

第六方面,提供了一种模型监控装置,用于实现上述第三方面或第三方面的任意一种实现中的模型监控方法;或者,用于实现上述第四方面或第四方面的任意一种实现中的模型监控方法。该装置可以是接入网设备,也可以是应用于接入网设备的模块(例如处理器、芯片、或芯片系统等),还可以是能实现全部或部分接入网设备功能的逻辑节点、逻辑模块或软件。一种实现中,该模型监控装置可以包括发送单元、接收单元,还可以包括处理单元。发送单元和接收单元可以是独立的,也可以是结合在一起的(可以称为“收发单元”)。

在一种可能的实现方式中,上述第五方面至第六方面中的模型监控装置包括用于分别执行上述第一方面至第四方面中的任一方面或任意一种实现中的方法的模块。

在另一种可能的实现方式中,上述第五方面至第六方面中的模型监控装置包括与存储器耦合的处理器;所述处理器被配置为实现所述装置执行上述模型监控方法中相应的功能。存储器用于与处理器耦合,其保存所述装置必要的程序(指令)和/或数据。可选的,所述模型监控装置还可以包括通信接口用于实现所述装置与其他网元之间的通信。可选的,该存储器可以位于该模型监控装置内部,也可以位于该模型监控装置外部。

在又一种可能的实现方式中,上述第五方面至第六方面中的模型监控装置包括处理器和收发装置,所述处理器与所述收发装置耦合,所述处理器用于执行计算机程序或指令,以控制所述收发装置进行信息的接收和发送;当所述处理器执行所述计算机程序或指令时,所述处理器还用于通过逻辑电路或执行代码指令实现上述方法。其中,所述收发装置可以为收发器、收发电路或输入输出接口,用于接收来自所述模型监控装置之外的其它模型监控装置的信号并传输至所述处理器或将来自所述处理器的信号发送给所述模型监控装置之外的其它模型监控装置。当所述模型监控装置为芯片时,所述收发装置为收发电路或输入输出接口。

当上述第五方面至第六方面中的模型监控装置为芯片时,发送单元可以是输出单元,比如输出电路或者通信接口;接收单元可以是输入单元,比如输入电路或者通信接口。当所述模型监控装置为终端时,发送单元可以是发射器或发射机;接收单元可以是接收器或接收机。

第七方面,提供了一种计算机可读存储介质,所述计算机可读存储介质中存储有计算机程序或指令,当所述计算机程序或指令被执行时,实现上述各方面所述的方法。

第八方面,提供了一种包含指令的计算机程序产品,当该指令在模型监控装置上运行时,使得模型监控装置执行上述各方面所述的方法。

第九方面,提供了一种模型监控系统,该模型监控系统包括第五方面所述的模型监控装置和第六方面所述的模型监控装置。

附图说明

图1为本申请实施例提供的无线通信系统的一简化示意图;

图2A为神经元结构的一种示意图;

图2B为神经网络的一种示意图;

图2C为一种AI应用框架;

图3为一种引入了AI网元的通信系统的示意图;

图4A~图4D为本申请实施例的网络架构的示意图;

图5为本申请实施例提供的一种模型监控方法的流程示意图;

图 6 为本申请实施例提供的另一种模型监控方法的流程示意图；
图 7 为本申请实施例示例的一个基于 AI 模型预测 CSI 的示意图；
图 8 为本申请实施例提供的又一种模型监控方法的流程示意图；
图 9 为本申请实施例示例的参考资源的示意图；
图 10 为本申请实施例提供的另一种模型监控方法的流程示意图；
图 11 为本申请实施例提供的又一种模型监控方法的流程示意图；
图 12 为本申请实施例提供的又一种模型监控方法的流程示意图；
图 13 为本申请实施例提供的又一种模型监控方法的流程示意图；
图 14 为本申请实施例示例的监控结果报告上报示意图；
图 15 为本申请实施例提供的又一种模型监控方法的流程示意图；
图 16 为本申请实施例提供的又一种模型监控方法的流程示意图；
图 17 为本申请实施例提供的又一种模型监控方法的流程示意图；
图 18 为本申请实施例提供的一种模型监控装置的结构示意图；
图 19 为本申请实施例提供的另一种模型监控装置的结构示意图。

具体实施方式

下面结合本申请实施例中的附图对本申请实施例进行描述。

为了使本申请的目的、技术方案和优点更加清楚，下面将结合附图对本申请作进一步地详细描述。

本申请如下涉及的至少一个（项），指示一个（项）或多个（项）。多个（项），是指两个（项）或两个（项）以上。“和/或”，描述关联对象的关联关系，表示可以存在三种关系，例如，A 和/或 B，可以表示：单独存在 A，同时存在 A 和 B，单独存在 B 这三种情况。字符“/”一般表示前后关联对象是一种“或”的关系。另外，应当理解，尽管在本申请中可能采用术语第一、第二等来描述各对象，但这些对象不应限于这些术语。这些术语仅用来将各对象彼此区分开。

本申请如下描述中所提到的术语“包括”和“具有”以及它们的任何变形，意图在于覆盖不排除的包含。例如包含了一系列步骤或单元的过程、方法、系统、产品或设备，没有限定于已列出的步骤或单元，而是可选地还包括其他没有列出的步骤或单元，或可选地还包括对于这些过程、方法、产品或设备固有的其它步骤或单元。需要说明的是，本申请中，“示例性的”或者“例如”等词用于表示作例子、例证或说明。本申请中被描述为“示例性的”或者“例如”的任何方法或设计方案不应被解释为比其它方法或设计方案更优选或更具优势。确切而言，使用“示例性的”或者“例如”等词旨在以具体方式呈现相关概念。

本申请提供的技术可以应用于各种通信系统，例如，该通信系统可以是第四代（4th generation, 4G）通信系统（例如长期演进(long term evolution, LTE)系统）、第五代（5th generation, 5G）通信系统、全球互联微波接入(worldwide interoperability for microwave access, WiMAX)或者无线局域网(wireless local area network, WLAN)系统、或者多种系统的融合系统，或者是未来的通信系统，例如 6G 通信系统等。其中，5G 通信系统还可以称为新无线（new radio, NR）系统。

通信系统中的一个网元可以向另一个网元发送信号或从另一个网元接收信号。其中信号可以包括信息、信令或者数据等。其中，网元也可以被替换为实体、网络实体、设备、终端设备、通信模块、节点、通信节点等等，本申请中以网元为例进行描述。例如，通信系统可以包括至少一个终端设备和至少一个接入网设备。接入网设备可以向终端设备发送下行信号，和/或终端设备可以向接入网设备发送上行信号。此外，可以理解的是，若通信系统中包括多个终端设备，多个终端设备之间也可以互发信号，即信号的发送网元和信号的接收网元均可以是终端设备。

本申请实施例提供的模型监控方法可以应用于 5G、6G、卫星通信等无线通信系统中。参见图 1，图 1 为本申请实施例提供的无线通信系统的一简化示意图。如图 1 所示，该无线通信系统包括无线接入网 100。无线接入网 100 可以是下一代（例如 6G 或更高版本）无线接入网，或传统（例如 5G、4G）无线接入网。一个或多个终端设备（120a-120j，统称为 120）可以相互连接，或连接到无线接入网 100 中的一个或多个网络设备（110a、110b，统称为 110）。可选的，图 1 只是示意图，该无线通信系统中还可以包括其它设备，如还可以包括核心网设备、无线中继设备和/或无线回传设备等，在图 1 中未画出。

可选的，在实际应用中，该无线通信系统可以同时包括多个网络设备（也称为接入网设备），也可以同时包括多个终端设备。一个网络设备可以同时服务于一个或多个终端设备。一个终端设备也可以同时接入一个或多个网络设备。本申请实施例对该无线通信系统中包括的终端设备和网络设备的数量不做限定。

其中，网络设备可以是网络侧的一种用于发射或接收信号的实体。网络设备可以为终端设备通过无线

方式接入到该无线通信系统中的接入设备，如网络设备可以是基站。基站可以广义的覆盖如下中的各种名称，或与如下名称进行替换，比如：RAN 节点、节点 B (NodeB)、演进型基站 (evolved NodeB, eNB)、下一代基站 (next generation NodeB, gNB)、开放无线接入网 (open radio access network, O-RAN) 中的接入网设备、中继站、接入点、传输点(transmitting and receiving point, TRP)、发射点(transmitting point, TP)、主站 MeNB、辅站 SeNB、多制式无线 (multi-standard radio, MSR) 节点、家庭基站、网络控制器、接入节点、无线节点、接入点 (access point, AP)、传输节点、收发节点、基带单元 (building baseband unit, BBU)、射频拉远单元 (remote radio unit, RRU)、有源天线单元 (active antenna unit, AAU)、射频头 (remote radio head, RRH)、集中式单元 (centralized unit, CU)、分布式单元 (distributed unit, DU)、无线单元 (radio unit, RU)、集中式单元控制面 (CU control plane, CU-CP) 节点、集中式单元用户面 (CU user plane, CU-UP) 节点、定位节点等。基站可以是宏基站、微基站、中继节点、施主节点或类似物，或其组合。网络设备还可以指用于设置于前述设备或装置内的通信模块、调制解调器或芯片。网络设备还可以是移动交换中心以及设备到设备 (device-to-device, D2D)、车辆外联 (vehicle-to-everything, V2X)、机器到机器

(machine-to-machine, M2M) 通信中承担基站功能的设备、6G 网络中的网络侧设备、未来的通信系统中承担基站功能的设备等。网络设备可以支持相同或不同接入技术的网络。本申请的实施例对网络设备所采用的具体技术和具体设备形态不做限定。

网络设备可以是固定的，也可以是移动的。例如，基站 110a、110b 是静止的，并负责来自终端设备 120 的一个或多个小区中的无线传输和接收。图 1 中示出的直升机或无人机 120i 可以被配置成充当移动基站，并且一个或多个小区可以根据移动基站 120i 的位置移动。在其他示例中，直升机或无人机 (120i) 可以被配置成用作与基站 110b 通信的终端设备。

本申请中，用于实现如上接入网络功能的通信装置可以是接入网设备，也可以是具有接入网络的部分功能的网络设备，也可以是能够支持实现接入网络功能的装置，例如芯片系统、硬件电路、软件模块、或硬件电路加软件模块，该装置可以被安装在接入网设备中或者和接入网设备匹配使用。本申请的方法中，以用于实现接入网设备功能的通信装置是接入网设备为例进行描述。

终端设备可以是用户侧的一种用于接收或发射信号的实体，如手机。终端设备可以用于连接人、物和机器。终端设备可通过网络设备与一个或多个核心网进行通信。终端设备包括具有无线连接功能的手持式设备、连接到无线调制解调器的其他处理设备或车载设备等。终端设备可以是便携式、袖珍式、手持式、计算机内置的或者车载的移动装置。终端设备 120 可以广泛应用于各种场景，例如蜂窝通信、D2D、V2X、端到端 (point-to-point, P2P)、机器到机器 (machine-to-machine, M2M)、机器类型通信 (machine type communication, MTC)、物联网 (internet of things, IoT)、虚拟现实 (virtual reality, VR)、增强现实 (augmented reality, AR)、工业控制、自动驾驶、远程医疗、智能电网、智能家具、智能办公、智能穿戴、智能交通、智慧城市、无人机、机器人、遥感、被动传感、定位、导航与跟踪、自主交付与移动等。终端设备 120 的一些举例为：3GPP 标准的用户设备 (user equipment, UE)、固定设备、移动设备、手持设备、可穿戴设备、蜂窝电话、智能电话、会话发起协议 (session initiated protocol, SIP) 电话、笔记本电脑、个人计算机、智能书、车辆、卫星、全球定位系统 (global positioning system, GPS) 设备、目标跟踪设备、无人机、直升机、飞行器、船只、遥控设备、智能家居设备、工业设备、个人通信业务 (personal communication service, PCS) 电话、无线本地环路 (wireless local loop, WLL) 站、个人数字助理 (personal digital assistant, PDA)、无线网络摄像头、平板电脑、掌上电脑、移动互联网设备 (mobile internet device, MID)、可穿戴设备如智能手表、VR 设备、AR 设备、工业控制 (industrial control) 中的无线终端、车联网系统中的终端、无人驾驶 (self driving) 中的无线终端、智能电网 (smart grid) 中的无线终端、运输安全 (transportation safety) 中的无线终端、智慧城市 (smart city) 中的无线终端如智能加油器，高铁上的终端设备以及智慧家庭 (smart home) 中的无线终端，如智能音响、智能咖啡机、智能打印机等。终端设备 120 可以为以上各种场景中的无线设备或用于设置于无线设备的装置，例如，上述设备中的通信模块、调制解调器或芯片等。终端设备也可以称为终端、终端设备、用户设备 (user equipment, UE)、移动台 (mobile station, MS)、移动终端 (mobile terminal, MT) 等。终端设备还可以是未来的无线通信系统中的终端设备。终端设备可以用于专用网设备或者通用设备中。本申请的实施例对终端设备所采用的具体技术和具体设备形态不做限定。

可选的，终端设备可以用于充当基站。例如，UE 可以充当调度实体，其在 V2X、D2D 或 P2P 等中的 UE 之间提供侧行链路信号。如图 1 所示，蜂窝电话 120a 和汽车 120b 利用侧行链路信号彼此通信。蜂窝电话 120a 和智能家居设备 120e 之间通信，而无需通过基站 110b 中继通信信号。

本申请中，用于实现终端设备功能的通信装置可以是终端设备，也可以是具有以上终端设备的部分功

能的终端设备,也可以是能够支持实现以上终端设备的功能的装置,例如芯片系统,该装置可以被安装在终端设备中或者和终端设备匹配使用。本申请中,芯片系统可以由芯片构成,也可以包括芯片和其他分立器件。本申请提供的技术方案中,以通信装置是终端设备或UE为例进行描述。

可选的,无线通信系统通常由小区组成,基站提供小区的管理,基站向小区中多个移动台(mobile station, MS)提供通信服务。其中基站包含基带单元(baseband unit, BBU)和远端射频单元(remote radio unit, RRU)。BBU和RRU可以放置在不同的地方,例如:RRU拉远,放置于高话务量的区域,BBU放置于中心机房。BBU和RRU也可以放置在同一机房。BBU和RRU也可以为一个机架下的不同部件。可选的,一个小区可以对应于一个载波或成员载波。

在一些部署中,本申请实施例所提及的网络设备可以为包括CU、或DU、或包括CU和DU的设备、或者控制面CU节点(中央单元控制面(central unit-control plane, CU-CP))和用户面CU节点(中央单元用户面(central unit-user plane, CU-UP))以及DU节点的设备。例如,网络设备可以包括gNB-CU-CP、gNB-CU-UP和gNB-DU。

在一些部署中,由多个RAN节点协作协助终端实现无线接入,不同RAN节点分别实现基站的部分功能。例如,RAN节点可以是CU, DU, CU-CP, CU-UP,或者RU等。CU和DU可以是单独设置,或者也可以包括在同一个网元中,例如BBU中。RU可以包括在射频设备或者射频单元中,例如包括在RRU、AAU或RRH中。

RAN节点可以支持一种或多种类型的前传接口,不同前传接口,分别对应具有不同功能的DU和RU。若DU和RU之间的前传接口为通用公共无线电接口(common public radio interface, CPRI),DU被配置用于实现基带功能中的一项或多项,RU被配置用于实现射频功能中的一项或多项。若DU和RU之间的前传接口为另一种接口,其相对于CPRI,将下行和/或上行的部分基带功能,比如,针对下行,预编码(precoding),数字波束赋形(beamforming, BF),或快速傅立叶反变换(inverse fast Fourier transform, IFFT)/添加循环前缀(cyclic prefix, CP)中的一项或多项,从DU中移至RU中实现,针对上行,数字波束赋形(beamforming, BF),或快速傅立叶变换(fast Fourier transform, FFT)/去除循环前缀(cyclic prefix, CP)中的一项或多项,从DU中移至RU中实现。在一种可能的实现方式中,该接口可以为增强型通用公共无线电接口(enhanced common public radio interface, eCPRI)。在eCPRI架构下,DU和RU之间的切分方式不同,对应不同类型(category, Cat)的eCPRI,比如eCPRI Cat A, B, C, D, E, F。

以eCPRI Cat A为例,对于下行传输,以层映射为切分,DU被配置用于实现层映射及之前的一项或多项功能(即编码、速率匹配,加扰,调制,层映射中的一项或多项),而层映射之后的其他功能(例如,RE映射,数字波束赋形(beamforming, BF),或快速傅立叶反变换(inverse fast Fourier transform, IFFT)/添加循环前缀(cyclic prefix, CP)中的一项或多项)移至RU中实现。对于上行传输,以解RE映射为切分,DU被配置用于实现解映射及之前的一项或多项功能(即解码,解速率匹配,解扰,解调,离散傅里叶逆变换(inverse discrete Fourier transform, IDFT),信道均衡,解RE映射中的一项或多项功能),而解映射之后的其他功能(例如,数字BF或快速傅里叶变换(fast Fourier transform, FFT)/去CP中的一项或多项)移至RU中实现。可以理解的是,关于各种类型的eCPRI所对应的DU和RU的功能描述,可以参考eCPRI协议,在此不予赘述。

一种可能的设计中,BBU中用于实现基带功能的处理单元称为基带高层(base band high, BBH)单元,RRU/AAU/RRH中用于实现基带功能的处理单元称为基带低层(base band low, BBL)单元。

在不同系统中,CU(或CU-CP和CU-UP)、DU或RU也可以有不同的名称,但是本领域的技术人员可以理解其含义。例如,在ORAN系统中,CU也可以称为O-CU(开放式CU),DU也可以称为O-DU,CU-CP也可以称为O-CU-CP,CU-UP也可以称为O-CU-UP,RU也可以称为O-RU。本申请中的CU(或CU-CP、CU-UP)、DU和RU中的任一单元,可以通过软件模块、硬件模块、或者软件模块与硬件模块结合来实现。

本申请实施例中,用于实现网络设备的功能的装置可以是网络设备;也可以是能够支持网络设备实现该功能的装置,例如芯片系统、硬件电路、软件模块、或硬件电路加软件模块。该装置可以被安装在网络设备中或者和网络设备匹配使用。在本申请实施例中仅以用于实现网络设备的功能的装置为网络设备为例进行说明,不对本申请实施例的方案构成限定。

可以理解的是,本申请可以应用在接入网设备和终端设备之间。

接入网设备和终端设备之间的协议层结构

接入网设备和终端设备之间的通信遵循一定的协议层结构。该协议层结构可以包括控制面协议层结构

和用户面协议层结构。例如，控制面协议层结构可以包括无线资源控制（radio resource control, RRC）层、分组数据汇聚层协议（packet data convergence protocol, PDCP）层、无线链路控制（radio link control, RLC）层、媒体接入控制（medium access control, MAC）层和物理层等协议层的功能。例如，用户面协议层结构可以包括 PDCP 层、RLC 层、MAC 层和物理层等协议层的功能，在一种可能的实现中，PDCP 层之上还可以包括业务数据适配协议（service data adaptation protocol, SDAP）层。

可选的，接入网设备和终端设备之间的协议层结构还可以包括人工智能（artificial intelligence, AI）层，用于传输 AI 功能相关的数据。

以接入网设备和终端设备之间的数据传输为例，数据传输需要经过用户面协议层，比如经过 SDAP 层、PDCP 层、RLC 层、MAC 层、物理层。其中，SDAP 层、PDCP 层、RLC 层、MAC 层和物理层也可以统称为接入层。根据数据的传输方向分为发送或接收，上述每层又分为发送部分和接收部分。以下行数据传输为例，PDCP 层自上层取得数据后，将数据传送到 RLC 层与 MAC 层，再由 MAC 层生成传输块，然后通过物理层进行无线传输。数据在各个层中进行相对应的封装。例如，某一层从该层的上层收到的数据视为该层的服务数据单元（service data unit, SDU），经过该层封装后成为协议数据单元（protocol data unit, PDU），再传递给下一个层。

示例性的，终端设备还可以具有应用层和非接入层。其中，应用层可以用于向终端设备中所安装的应用程序提供服务，比如，终端设备接收到的下行数据可以由物理层依次传输到应用层，进而由应用层提供给应用程序；又比如，应用层可以获取应用程序产生的数据，并将数据依次传输到物理层，发送给其它通信装置。非接入层可以用于转发用户数据，比如将从应用层接收到的上行数据转发给 SDAP 层，或者将从 SDAP 层接收到的下行数据转发给应用层。

为了在无线网络中支持 AI 技术，网络中还可能引入 AI 节点。

可选地，AI 节点可以部署于该通信系统中的如下位置中的一项或多项：接入网设备、终端设备、或核心网设备等，或者，AI 节点也可单独部署，例如，部署于上述任一项设备之外的位置，比如，过顶（over the top, OTT）系统的主机或云端服务器中。AI 节点可以与通信系统中的其它设备通信，其它设备例如可以以为以下中的一项或多项：网络设备，终端设备，或，核心网的网元等。

可以理解，本申请对于 AI 节点的数量不予限制。例如，当有多个 AI 节点时，多个 AI 节点可以基于功能进行划分，如不同的 AI 节点负责不同的功能。

还可以理解，AI 节点可以是各自独立的设备，也可以集成于同一设备中实现不同的功能，或者可以是硬件设备中的网络元件，也可以是在专用硬件上运行的软件功能，或者是平台（例如，云平台）上实例化的虚拟化功能，本申请对于上述 AI 节点的具体形态不作限定。

AI 节点可以为 AI 网元或 AI 模块。

这些网元节点，例如核心网设备、接入网节点（RAN 节点）、终端或 OAM 中的一个或多个设备中设置有一个或多个 AI 模块。所述接入网节点可以作为单独的 RAN 节点，也可以包括多个 RAN 节点，例如，包括 CU 和 DU。所述 CU 和/或 DU 也可以设置有一个或多个 AI 模块。可选的，CU 还可以被拆分为 CU-CP 和 CU-UP。CU-CP 和/或 CU-UP 中设置有一个或多个 AI 模型。

所述 AI 模块用以实现相应的 AI 功能。不同网元中部署的 AI 模块可以相同或不同。AI 模块的模型根据不同的参数配置，AI 模块可以实现不同的功能。AI 模块的模型可以是基于以下一项或多项参数配置的：结构参数（例如神经网络层数、神经网络宽度、层间的连接关系、神经元的权值、神经元的激活函数、或激活函数中的偏置中的至少一项）、输入参数（例如输入参数的类型和/或输入参数的维度）、或输出参数（例如输出参数的类型和/或输出参数的维度）。其中，激活函数中的偏置还可以称为神经网络的偏置。

一个 AI 模块可以具有一个或多个模型。一个模型可以推理得到一个输出，该输出包括一个参数或者多个参数。不同模型的学习过程、训练过程、或推理过程可以部署在不同的节点或设备中，或者可以部署在相同的节点或设备中。

通信系统中包括 RAN 智能控制器（RAN intelligent controller, RIC）。例如，该 RIC 可以是上述 AI 模块，用于实现 AI 相关的功能。该 RIC 包括近实时 RIC（near-real time RIC, near-RT RIC）和非实时 RIC（non-real time RIC, Non-RT RIC）。其中，非实时 RIC 主要处理非实时的信息，比如，对时延不敏感的数据，该数据的时延可以为秒级。实时 RIC 主要处理近实时的信息，比如，对时延相对敏感的数据，该数据的时延为数十毫秒级。

近实时 RIC 用于进行模型训练和推理。例如，用于训练 AI 模型，利用该 AI 模型进行推理。近实时 RIC 可以从 RAN 节点（例如 CU、CU-CP、CU-UP、DU 和/或 RU）和/或终端获得网络侧和/或终端侧的信

息。该信息可以作为训练数据或者推理数据。可选的，近实时 RIC 可以将推理结果递交给 RAN 节点和/或终端。可选的，CU 和 DU 之间，和/或 DU 和 RU 之间可以交互推理结果。例如近实时 RIC 将推理结果递交给 DU，DU 将其发给 RU。

非实时 RIC 也用于进行模型训练和推理。例如，用于训练 AI 模型，利用该模型进行推理。非实时 RIC 可以从 RAN 节点（例如 CU、CU-CP、CU-UP、DU 和/或 RU）和/或终端获得网络侧和/或终端侧的信息。该信息可以作为训练数据或者推理数据，推理结果可以被递交给 RAN 节点和/或终端。可选的，CU 和 DU 之间，和/或 DU 和 RU 之间可以交互推理结果，例如非实时 RIC 将推理结果递交给 DU，由 DU 将其发给 RU。

近实时 RIC、非实时 RIC 也可以分别作为一个网元单独设置。可选的，近实时 RIC、非实时 RIC 也可以作为其他设备的一部分，例如，近实时 RIC 设置在 RAN 节点中（例如，CU，DU 中），而非实时 RIC 设置在 OAM 中、云服务器中、核心网设备、或者其他网络设备中。

示例性地，近实时 RIC、非实时 RIC 在网络架构中的设置可以如图 4A~图 4D 所示：

如图 4A 中的 (a) 所示，第一种可能的实现中，接入网设备中包括近实时 RIC 模块，用于进行模型学习和/或推理。

如图 4A 中的 (b) 所示，第二种可能的实现中，在通信系统中，接入网设备之外可以包括非实时 RIC，可选的，该非实时 RIC 可以位于 OAM 中或核心网设备中。

如图 4A 中的 (c) 所示，第三种可能的实现中，接入网设备中包括近实时 RIC，接入网设备之外还包括非实时 RIC。可选的，非实时 RIC 可以位于 OAM 中或者核心网设备中。

相对图 4A 中的 (c)，图 4B 中将 CU 分离为了 CU-CP 和 CU-UP。近实时 RIC 和非实时 RIC 的设置于图 4A 中的 (c) 相同。

如图 4C 所示，可选的，接入网设备中包括一个或多个 AI 实体，该 AI 实体的功能类似上述近实时 RIC。可选的，OAM 中包括一个或多个 AI 实体，该 AI 实体的功能类似上述非实时 RIC。可选的，核心网设备中包括一个或多个 AI 实体，该 AI 实体的功能类似上述非实时 RIC。当 OAM 和核心网设备中都包括 AI 实体时，他们各自的 AI 实体所训练得到的模型不同，和/或用于进行推理的模型不同。本申请中，模型不同可以包括以下至少一项不同：模型的结构参数（例如模型的层数、和/或权值等）、模型的输入参数、或模型的输出参数。

相对图 4C，图 4D 中的接入网设备分离为 CU 和 DU。可选的，CU 中可以包括 AI 实体，该 AI 实体的功能类似上述近实时 RIC。可选的，DU 中可以包括 AI 实体，该 AI 实体的功能类似上述近实时 RIC。当 CU 和 DU 中都包括 AI 实体时，他们各自的 AI 实体所训练得到的模型不同，和/或用于进行推理的模型不同。可选的，还可以进一步将图 4D 中的 CU 拆分为 CU-CP 和 CU-UP。可选的，CU-CP 中可以部署有一个或多个 AI 模型。和/或，CU-UP 中可以部署有一个或多个 AI 模型。可选的，图 4C 或图 4D 中，接入网设备的 OAM 和核心网设备的 OAM 可以分开独立部署。

应理解，图 1 所示的通信系统中各个设备的数量、类型仅作为示意，本申请并不限于此，实际应用中还可以在通信系统中还可以包括更多的终端设备、更多的接入网设备，还可以包括其它网元，例如可以包括核心网设备，和/或用于实现人工智能功能的网元。

可以理解的是，终端设备、接入网设备、核心网设备、或用于实现人工智能功能的网元中的一项或多项所实现的全部或部分功能均可以进行虚拟化，也即，通过专有处理器或通用处理器中的一项或多项和相应的软件模块来实现。其中，终端设备和接入网设备因涉及空口传输的接口，该接口的收发功能可由硬件来实现。核心网设备，如操作维护管理（operation administration and maintenance, OAM）网元，均可虚拟化。可选的，虚拟化后的终端设备、接入网设备、核心网设备、或用于实现人工智能功能的网元中的一项或多项功能可以由云端设备来实现，比如过顶(over the top, OTT)系统中的云端设备来实现。

为便于理解，下面先对本申请涉及的 AI 技术进行介绍。可以理解的是，该介绍并不作为对本申请的限定。

(1) AI 模型

AI 指由人制造出来的机器所表现出来的智能。通常人工智能是指通过普通计算机程序来呈现人类智能的技术。人工智能可以定义为模仿人类、且具有与人类思维相关的认知功能的机器或计算机，如学习和解决问题。人工智能能够从过去的经验中学习，做出合理的决策，并快速回应。人工智能的目标是通过构建具有象征意义的推理或推理的计算机程序来理解智能。

机器学习是实现人工智能的一个途径，即以机器学习为手段解决人工智能中的问题。机器学习理论主

要是设计和分析一些让计算机可以自动“学习”的算法。机器学习算法是一类从数据中自动分析获得规律，并利用规律对未知数据进行预测的算法。因为学习算法中涉及了大量的统计学理论，机器学习与推断统计学联系尤为密切，也被称为统计学习理论。

AI 模型为能实现 AI 功能的算法或者计算机程序，是 AI 技术功能的具体实现，AI 模型表征了模型的输入和输出之间的映射关系。AI 模型的类型可以是神经网络、线性回归模型、决策树模型、支持向量机 (support vector machine, SVM)、贝叶斯网络、Q 学习模型或者其他机器学习 (machine learning, ML) 模型。

(2) 神经网络

神经网络是 AI 或机器学习技术的一种具体实现形式。根据通用近似定理，神经网络在理论上可以逼近任意连续函数，从而使得神经网络具备学习任意映射的能力。因此神经网络可以对复杂的高维度问题进行准确地抽象建模。

神经网络的思想来源于大脑组织的神经元结构。例如，每个神经元都对其输入值进行加权求和运算，通过一个激活函数输出运算结果。如图 2A 所示，为神经元结构的一种示意图。假设神经元的输入为 $x = [x_0, x_1, \dots, x_n]$ ，与各个输入对应的权值分别为 $w = [w, w_1, \dots, w_n]$ ，其中， w_i 作为 x_i 的权值，用于对 x_i 进行加权。根据权值对输入值进行加权求和的偏置例如为 b 。激活函数的形式可以有多种，假设一个神经元的激活函数为： $y = f(z) = \max(0, z)$ ，则该神经元的输出为：

$y = f(\sum_{i=0}^n w_i * x_i + b) = \max(0, \sum_{i=0}^n w_i * x_i + b)$ 。再例如，一个神经元的激活函数为： $y = f(z) = z$ ，则该神经元的输出为：

$y = f(\sum_{i=0}^n w_i * x_i + b) = \sum_{i=0}^n w_i * x_i + b$ 。其中， b 、 w_i 、 x_i 可以是小数、整数（例如 0、正整数或负整数）、或复数等各种可能的取值。神经网络中不同神经元的激活函数可以相同或不同。

神经网络一般包括多个层，每层可包括一个或多个神经元。通过增加神经网络的深度和/或宽度，能够提高该神经网络的表达能力，为复杂系统提供更强大的信息提取和抽象建模能力。其中，神经网络的深度可以是指神经网络包括的层数，其中每层包括的神经元个数可以称为该层的宽度。在一种实现方式中，神经网络包括输入层和输出层。神经网络的输入层将接收到的输入信息经过神经元处理，将处理结果传递给输出层，由输出层得到神经网络的输出结果。在另一种实现方式中，神经网络包括输入层、隐藏层和输出层，可参考图 2B 的神经网络的示意图。神经网络的输入层将接收到的输入信息经过神经元处理，将处理结果传递给中间的隐藏层，隐藏层对接收的处理结果进行计算，得到计算结果，隐藏层将计算结果传递给输出层或者相邻的隐藏层，最终由输出层得到神经网络的输出结果。其中，一个神经网络可以包括一个隐藏层，或者包括多个依次连接的隐藏层，不予限制。

以 AI 模型的类型为神经网络为例，本申请涉及的 AI 模型可以为深度神经网络 (deep neural network, DNN)。根据网络的构建方式，DNN 可以包括前馈神经网络 (feedforward neural network, FNN)、卷积神经网络 (convolutional neural networks, CNN) 和递归神经网络 (recurrent neural network, RNN)。

(3) 训练数据集和推理数据

训练数据集用于 AI 模型的训练，训练数据集可以包括 AI 模型的输入，或者包括 AI 模型的输入和目标输出。其中，训练数据集包括一个或多个训练数据，训练数据可以是输入至 AI 模型的训练样本，也可以是 AI 模型的目标输出。其中，目标输出也可以被称为标签或者标签样本。训练数据集是机器学习重要的部分之一，模型训练本质上就是从训练数据中学习它的某些特征，使得 AI 模型的输出尽可能接近目标输出，如 AI 模型的输出与目标输出之间的差异尽可能地小。训练数据集的构成与选取，在一定程度上可以决定训练出来的 AI 模型的性能。

另外，在 AI 模型（如神经网络）的训练过程中，可以定义损失函数。损失函数描述了 AI 模型的输出值与目标输出值之间的差距或差异。本申请并不限制损失函数的具体形式。AI 模型的训练过程就是通过调整 AI 模型的模型参数，使得损失函数的取值小于门限，或者使得损失函数的取值满足目标需求的过程。例如，AI 模型为神经网络，调整神经网络的模型参数包括调整如下参数中的至少一种：神经网络的层数、宽度、神经元的权值、或神经元的激活函数中的参数。

推理数据可以作为已训练好的 AI 模型的输入，用于 AI 模型的推理。在模型推理过程中，将推理数据输入 AI 模型，可以得到对应的输出即为推理结果。

(4) AI 模型的设计

AI 模型的设计主要包括数据收集环节（例如收集训练数据和/或推理数据）、模型训练环节以及模型推理环节。进一步地还可以包括推理结果应用环节。参见图 2C，示意了一种 AI 应用框架。在前述数据收集环节中，数据源 (data source) 用于提供训练数据集和推理数据。在模型训练环节中，通过对数据源提供

的训练数据 (training data) 进行分析或训练, 得到 AI 模型。其中, AI 模型表征了模型的输入和输出之间的映射关系。通过模型训练节点学习得到 AI 模型, 相当于利用训练数据学习得到模型的输入和输出之间的映射关系。在模型推理环节中, 使用经由模型训练环节训练后的 AI 模型, 基于数据源提供的推理数据进行推理, 得到推理结果。该环节还可以理解为: 将推理数据输入到 AI 模型, 通过 AI 模型得到输出, 该输出即为推理结果。该推理结果可以指示: 由执行对象使用 (执行) 的配置参数、和/或由执行对象执行的操作。在推理结果应用环节中进行推理结果的发布, 例如推理结果可以由执行 (actor) 实体统一规划, 例如执行实体可以发送推理结果给一个或多个执行对象 (例如, 核心网设备、接入网设备、或终端设备等) 去执行。又如执行实体还可以反馈模型的性能给数据源, 便于后续实施模型的更新训练。

可以理解的是, 在通信系统中可以包括具备人工智能功能的网元。上述 AI 模型设计相关的环节可以由一个或多个具备人工智能功能的网元执行。一种可能的设计中, 可以在通信系统中已有网元内配置 AI 功能 (如 AI 模块或者 AI 实体) 来实现 AI 相关的操作, 例如 AI 模型的训练和/或推理。例如该已有网元可以是接入网设备 (如 gNB)、终端设备、核心网设备、或网管等。其中, 网管可以根据运营商网络运营的实际需要, 将网络的管理工作划分为 3 类: 操作 (Operation)、管理 (Administration)、维护 (Maintenance)。网管又可以称为操作维护管理 (operation administration and maintenance, OAM) 网元, 简称 OAM。操作主要完成日常网络和业务进行的分析、预测、规划和配置工作; 维护主要是对网络及其业务的测试和故障管理等进行的日常操作活动, 网管可以检测网络运行状态、优化网络连接和性能, 提升网络运行稳定性, 降低网络维护成本。或者另一种可能的设计中, 也可以在通信系统中引入独立的网元来执行 AI 相关的操作, 如训练 AI 模型。该独立的网元可以称为 AI 网元或者 AI 节点等, 本申请对此名称不进行限制。该 AI 网元可以和通信系统中的接入网设备之间直接连接, 也可以通过第三方网元和接入网设备实现间接连接。其中, 第三方网元可以是认证管理功能 (authentication management function, AMF) 网元、用户面功能 (user plane function, UPF) 网元等核心网网元、OAM、云服务器或者其他网元, 不予限制。示例性的, 参见图 3, 该通信系统包括接入网设备 110、终端设备 120、130, 还在上述图 1 所示的通信系统中引入了 AI 网元 140。

本申请中, 一个模型可以推理得到一个参数, 或者推理得到多个参数。不同模型的训练过程可以部署在不同的设备或节点中, 也可以部署在相同的设备或节点中。不同模型的推理过程可以部署在不同的设备或节点中, 也可以部署在相同的设备或节点中。

其中, 模型参数可以包括如下的一种或多种模型的结构参数 (例如模型的层数、和/或权值等)、模型的输入参数 (如输入维度、输入端口数)、或模型的输出参数 (如输出维度、输出端口数)。可以理解, 输入维度可以指的是一个输入数据的大小, 例如输入数据为一个序列时, 该序列对应的输入维度可以指示该序列的长度。输入端口数可以指的是输入数据的数量。类似地, 输出维度可以指的是一个输出数据的大小, 例如输出数据为一个序列时, 该序列对应的输出维度可以指示该序列的长度。输出端口数可以指的是输出数据的数量。针对背景技术提出的有可能基于预测模型得到的预测性能不能反应出实际的系统性能, 本申请提供一种模型监控方案, 终端设备通过获取一个或多个真值信息的一个或多个第一索引, 可以获取特定的时间单元, 如所需的时间单元, 对应的真值信息, 从而可以灵活地监控预测模型的性能。

本申请中“向...(例如终端设备)发送信息”或者附图中的相关示意可以理解为该信息的目的端是终端设备。可以包括直接或间接的向终端设备发送信息。“从...(例如终端设备)接收信息”或者“接收来自...(例如终端设备)的信息”, 或者附图中的相关示意可以理解为该信息的源端是终端设备, 可以包括直接或间接的从终端设备接收信息。信息在信息发送的源端和目的端之间可能会被进行必要的处理, 例如格式变化等, 但目的端可以理解来自源端的有效信息。本申请中类似的表述可以做类似的理解, 在此不再赘述。

下面详细描述本申请实施例提供的模型监控方法。可以理解的, 本申请中是以接入网设备和终端设备作为该交互示意的执行主体为例进行示意的, 但本申请并不限制交互示意的执行主体。例如, 本申请提供的方法中的接入网设备也可以是应用于接入网设备的芯片、芯片系统、或处理器, 还可以是能够实现全部或部分接入网设备的逻辑节点、逻辑模块或软件; 本申请提供的方法中的终端设备也可以是应用于终端设备的芯片、芯片系统、或处理器, 还可以是能够实现全部或部分终端设备功能的逻辑节点、逻辑模块或软件。

如图 5 所示, 为本申请实施例提供的一种模型监控方法的流程示意图。示例性地, 该方法可以包括以下步骤:

S501.终端设备获取第一信息。

本申请实施例涉及的模型监控指的是监控 AI 模型的性能, 判断 AI 模型是否在正常工作, 如果 AI 模型性能较差, 则需要切换至非 AI 模式, 或者更换 AI 模型, 或者更新 AI 模型等。模型监控可以是监控 AI

模型输出的准确性（也可称为中间 KPI）。监控 AI 模型输出的准确性是通过对比 AI 模型的输出与对应的标签或者真值之间的差异，判断 AI 模型的性能是否满足要求。中间 KPI 通常包括广义余弦相似度

（generalized cosine similarity, GCS）、平方广义余弦相似度（square generalized cosine similarity, SGCS）、最小均方误差（mean square error, MSE）、归一化均方误差（normalized mean square error, NMSE）等。

计算 AI 模型的中间 KPI 需要两种信息，一个是待监控的信息，如 AI 模型输出的预测信息，一个是标签信息或真值信息，即 $KPI(H, \hat{H})$ ，其中 H 为真值信息， $\hat{H}$ 为预测信息。真值信息指的是该信息可以被认为是准确的，例如终端设备测量的信息相对于预测的信息可以被认为是真值信息。 $KPI(H, \hat{H})$ 一般用于表示 H 和 $\hat{H}$ 之间的差异性度量或相似性度量。

本实施例通过真值信息的索引指示终端设备获取哪个真值信息。

终端设备获取第一信息。其中，该第一信息包括一个或多个真值信息分别对应的一个或多个第一索引。示例性地，终端设备可以获取接入网设备发送的上述第一信息，也可以获取自身预先存储的上述第一信息。

S502.终端设备获取一个或多个第一索引分别对应的一个或多个真值信息。

终端设备获取一个或多个真值信息分别对应的一个或多个第一索引后，测量一个或多个索引分别对应的时隙上的参考信号，得到一个或多个第一索引分别对应的一个或多个真值信息。

其中，该一个或多个真值信息用于获取 AI 模型的性能的监控结果。该 AI 模型的输出为预测信息。

相对于现有技术只能周期性地获取格点上模型监控结果，本实施例终端设备通过获取一个或多个真值信息的一个或多个第一索引，可以获取特定的时间单元对应的真值信息。

根据本申请实施例提供的一种模型监控方法，终端设备通过获取一个或多个真值信息分别对应的一个或多个第一索引，可以获取特定的时间单元对应的真值信息，从而可以灵活地监控预测模型的性能。

如图 6 所示，为本申请实施例提供的另一种模型监控方法的流程示意图。与图 5 所示实施例不同的是，该实施例不是直接通过真值信息的索引指示终端设备获取哪个真值信息，而是根据用于测量真值信息的参考信号的时域资源位置来确定真值信息。示例性地，该方法可以包括以下步骤：

S601.终端设备获取第一信息。

用于测量真值信息（即用于模型监控）的参考信号可以与用于模型推断（作为模型的输入，用于预测信息）的参考信号是同一组参考信号，即属于同一个参考信号配置；用于测量真值信息的参考信号可以与用于模型推断的参考信号也可以是不同组的参考信号，即为模型监控专门配置一个参考信号。

本实施例中，用于测量真值信息的参考信号与用于模型推断的参考信号是不同组的参考信号，终端设备可以根据用于测量真值信息的参考信号确定真值信息。因此，终端设备可以获取第一信息，该第一信息包括用于测量真值信息的一个或多个参考信号的时域资源位置。

示例性地，终端设备可以获取接入网设备发送的上述第一信息，也可以获取自身预先存储的上述第一信息。

S602.终端设备根据第一信息，确定一个或多个真值信息。

终端设备获取到第一信息后，可以根据该第一信息，确定一个或多个真值信息。

在一个实现方式中，终端设备默认可以测量第一信息携带的一个或多个参考信号的时域资源位置上的一个或多个参考信号，得到一个或多个真值信息。

在另一个实现方式中，若接入网设备通过第一信息配置了用于测量真值信息的一个或多个参考信号的时域资源位置，则终端设备默认可以测量第一信息携带的一个或多个参考信号的时域资源位置上的一个或多个参考信号，得到一个或多个真值信息。

在又一个实现方式中，终端设备还可以获取用于模型推断的一个或多个参考信号的时域资源位置。终端设备默认可以测量第一信息携带的一个或多个参考信号的时域资源位置上的一个或多个参考信号，以及测量用于模型推断的一个或多个参考信号的时域资源位置上的一个或多个参考信号，得到一个或多个真值信息。

在又一个实现方式中，接入网设备也可以指示终端设备测量第一信息携带的一个或多个参考信号的时域资源位置上的一个或多个参考信号，获得一个或多个真值信息，或者指示终端设备测量用于模型推断的一个或多个参考信号的时域资源位置上的一个或多个参考信号，获得一个或多个真值信息。

其中，该一个或多个真值信息用于获取 AI 模型的性能的监控结果。该 AI 模型的输出为预测信息。

可以看出，本实施例中，获得的一个或多个真值信息中的至少一个真值信息的时域资源位置与预测信息的时域资源位置不同。

根据本申请实施例提供的一种模型监控方法，终端设备通过获取用于测量真值信息的一个或多个参考

信号的时域资源位置，可以获取特定的时域资源位置对应的真值信息，从而可以灵活地监控预测模型的性能。

上述方案可以用于 CSI 预测模型的监控。其中，在通信系统（例如，LTE 通信系统或 NR 通信系统等）中，网络设备需要基于 CSI 决定调度终端设备的下行数据信道的资源、MCS 以及预编码等配置。可以理解，CSI 属于一种信道信息，是一种能够反映信道特征、信道质量的信息。

CSI 测量指的是接收端根据发送端发送的参考信号求解信道信息，即利用信道估计方法估计出信道信息。示例性地，参考信号可以包括信道状态信息参考信号（channel state information reference signal, CSI-RS）、同步信号/广播信道块（synchronizing signal/ physical broadcast channel block, SSB）、信道探测参考信号（sounding reference signal, SRS）或解调参考信号（demodulation reference signal, DMRS）等中的一项或多项。CSI-RS、SSB 以及 DMRS 等可以用于测量下行 CSI。SRS 和 DMRS 等可以用于测量上行 CSI。

以 FDD 通信场景为例，在 FDD 通信场景中，由于上下行信道不具备互易性或者说无法保证上下行信道的互易性，网络设备通常会向终端设备下行参考信号，终端设备根据接收到的下行参考信号进行信道测量、干扰测量估计下行 CSI。终端设备根据协议预定义的方式或网络设备配置的方式生成 CSI 报告，并反馈给网络设备，以使其获取下行 CSI。

示例性地，CSI 可以包括以下至少一项：信道质量指示（channel quality indication, CQI）、预编码矩阵指示（precoding matrix indicator, PMI）、秩指示（rank indicator, RI）、CSI-RS 资源指示（CSI-RS resource indicator, CRI）、层指示（layer indicator, LI）、参考信号接收功率（reference signal receiving power, RSRP）或信号与干扰加噪声比（signal to interference plus noise ratio, SINR）等。信号与干扰加噪声比也可以称为信干噪比。

其中，RI 用于指示终端设备建议的下行传输的层数，CQI 用于指示终端设备判断的当前信道条件所能支持的调制编码方式，PMI 用于指示终端设备建议的预编码。PMI 所指示的预编码的层数与 RI 对应。

应理解，上述 CSI 报告所指示的 RI、CQI 和 PMI 等仅为终端设备的建议值，网络设备可以按照该 CSI 报告所指示的信息中的部分或全部进行下行传输。或者，网络设备也可以不按照该 CSI 报告所指示的信息进行下行传输。

基于 AI 的 CSI 预测指的是将若干个历史/当前 CSI 信息输入到 AI 预测模型，输出预测的未来 CSI。如图 7 所示，为本申请实施例例的一个基于 AI 模型预测 CSI 的示意图，可以将 t-5 和 t 这 2 个历史/当前 CSI 信息输入 AI 模型中，AI 模型可以输出未来 t+5 时刻的 CSI 信息。预测 CSI 可以是以下任意一种形式：信道响应、信道矩阵、信道特征矩阵、预编码矩阵。其中，信道响应、信道矩阵是信道本身；信道特征矩阵、预编码矩阵是对信道提取出的特征构成的矩阵。

计算 AI 模型的中间 KPI 需要两种 CSI 信息，一个是待监控的 CSI，如 AI 模型输出的预测 CSI，一个是标签 CSI 或真值 CSI，即 $KPI(H, \hat{H})$ ，其中 H 为真值 CSI， $\hat{H}$ 为预测 CSI。真值 CSI 指的是该 CSI 可以被认为是准确的，例如终端设备测量的 CSI 相对于预测的 CSI 可以被认为是真值 CSI。 $KPI(H, \hat{H})$ 一般用于表示 H 和 $\hat{H}$ 之间的差异性度量或相似性度量。

如前所述，例如 KPI 可以 GCS、SGCS、MSE 或 NMSE。

KPI 为 GCS 时，则 $KPI(H, \hat{H}) = GCS(H, \hat{H}) = \frac{1}{N} \sum_{i=1}^N \frac{\|\hat{w}_i^H w_i\|}{\|\hat{w}_i\| \|w_i\|}$ ， $\hat{w}_i^H$ 是预测的每个资源单元的 CSI， w_i 为每个资源单元的真值 CSI。

KPI 为 SGCS 时，则 $KPI(H, \hat{H}) = SGCS(H, \hat{H}) = \frac{1}{N} \sum_{i=1}^N \left(\frac{\|\hat{w}_i^H w_i\|}{\|\hat{w}_i\| \|w_i\|} \right)^2$ ， $\hat{w}_i^H$ 是预测的每个资源单元的 CSI， w_i 为每个资源单元的真值 CSI。

KPI 为 MSE 时，则 $KPI(H, \hat{H}) = MSE(H, \hat{H}) = \frac{1}{N} \sum_{i=1}^N \|w_i - \hat{w}_i\|^2$ ， $\hat{w}_i^H$ 是预测的每个资源单元的 CSI， w_i 为每个资源单元的真值 CSI。

KPI 为 NMSE 时，则 $KPI(H, \hat{H}) = NMSE(H, \hat{H}) = \frac{1}{N} \sum_{i=1}^N \frac{\|w_i - \hat{w}_i\|^2}{\|w_i\|^2}$ ， $\hat{w}_i^H$ 是预测的每个资源单元的 CSI， w_i 为每个资源单元的真值 CSI。

针对终端设备是否存储有 AI 模型，分别结合不同的实施例进行描述如何对 CSI 预测模型进行监控：

如图 8 所示，为本申请实施例提供的又一种模型监控方法的流程示意图。在该实施例中，终端设备未存储有 AI 模型。示例性地，该方法可以包括以下步骤：

S801.终端设备获取第一信息。

本实施例通过真值 CSI 的索引指示终端设备获取哪个真值 CSI。

终端设备获取第一信息。其中,该第一信息包括一个或多个真值 CSI 分别对应的一个或多个第一索引。示例性地,终端设备可以获取接入网设备发送的上述第一信息,也可以获取自身预先存储的上述第一信息。

在一个实现方式中,终端设备可以通过参考资源的时刻确定真值 CSI 的索引。其中,参考资源的时刻可以是协议定义的或接入网设备配置的,参考资源在时域上是一个时间单元(时间单元包括但不限于以下任意一种:符号、时隙、微时隙、子帧、无线帧,本申请实施例以时隙为例进行描述)。如图 9 所示,示例了参考资源,该参考资源对应的时隙为 t 时隙。以参考资源所在的时隙为参考,例如,参考资源所在时隙之前的第 n 个真值 CSI 的索引为 $-n$,参考资源所在时隙及其之后的第 n 个真值 CSI 的索引为 $n-1$ 。又例如,参考资源所在时隙及其之前的第 n 个真值 CSI 的索引为 $-n+1$,参考资源所在时隙之后的第 n 个真值 CSI 的索引为 n 。又例如,参考资源所在时隙的之前的第 K 个真值 CSI 的索引为 0,之后的真值 CSI 的索引依次增加, K 为预定义的或接入网设备配置的参数。又例如,参考资源所在时隙之后的第 K 个真值 CSI 的索引为 0,之前的真值 CSI 的索引依次增加。

上述第一信息中包括的一个或多个真值 CSI 分别对应的一个或多个第一索引为相对于参考资源所在时隙的索引。例如在图 9 所示的例子中,参考资源所在时隙为 t 时隙,则 $t+2$ 时隙的真值 CSI 的索引为 0, $t+4$ 时隙的真值 CSI 的索引为 1。

接入网设备向终端设备发送信道状态信息-参考信号(channel state information-reference signal, CSI-RS),终端设备接收 CSI-RS 并测量,得到真值 CSI。用于测量真值 CSI(即用于模型监控)的 CSI-RS 可以与用于模型推断(作为模型的输入,用于预测 CSI)的 CSI-RS 是同一组 CSI-RS,即属于同一个 CSI-RS 配置;用于测量真值 CSI 的 CSI-RS 可以与用于模型推断的 CSI-RS 也可以是不同组的 CSI-RS,即为模型监控专门配置一个 CSI-RS。

则在另一个实现方式中,真值 CSI 的索引也可以是用于测量真值 CSI 的 CSI-RS 的索引。仍参考图 9,还示例了用于测量真值 CSI 的 CSI-RS 的时域位置,用于模型推断的 CSI-RS 为周期为 5ms 的 CSI-RS(例如 $t-5$ 、 t 、 $t+5$ 时隙对应的 CSI-RS);用于模型监控的 CSI-RS 为非周期的或半静态的 CSI-RS,为 $t+2$ 、 $t+4$ 时隙对应的 CSI-RS。 $t+2$ 时隙的真值 CSI 的索引为 0, $t+4$ 时隙的真值 CSI 的索引为 1, $t+5$ 时隙的真值 CSI 的索引为 2。

在又一个实现方式中,用于测量真值 CSI 的 CSI-RS 可以仅包括用于监控的 CSI-RS,例如在图 9 所示的例子中,CSI 参考资源在 t 时隙, $t+2$ 时刻的真值 CSI 的索引为 0, $t+4$ 时刻的真值 CSI 的索引为 1。

在又一个实现方式中,用于测量真值 CSI 的 CSI-RS 也可以包括用于监控的 CSI-RS 和预测 CSI 所在时刻的 CSI-RS,例如在图 9 所示的例子中, $t+2$ 时刻的真值 CSI 的索引为 0, $t+4$ 时刻的真值 CSI 的索引为 1, $t+5$ 时刻的真值 CSI 的索引为 2。

S802.终端设备获取一个或多个第一索引分别对应的一个或多个真值 CSI。

终端设备获取一个或多个真值 CSI 分别对应的一个或多个第一索引后,测量一个或多个索引分别对应的时隙上的 CSI-RS,得到一个或多个第一索引分别对应的一个或多个真值 CSI。

其中,该一个或多个真值 CSI 用于获取 AI 模型的性能的监控结果。该 AI 模型的输出为预测 CSI。

相对于现有技术只能周期性地获取格点上模型监控结果,本实施例终端设备通过获取一个或多个真值 CSI 的一个或多个第一索引,可以获取特定的时间单元,如所需的时间单元,对应的真值 CSI。

进一步地,该方法还可以包括以下步骤(图中以虚线表示):

S803.终端设备向接入网设备发送一个或多个真值 CSI。相应地,接入网设备接收该一个或多个真值 CSI。

本实施例中,终端设备未存储 AI 模型,不能进行 CSI 预测。因此,终端设备获取一个或多个真值 CSI 后,向接入网设备发送该一个或多个真值 CSI。

S804.接入网设备获取一个或多个预测 CSI。

接入网设备存储有 AI 模型,可以利用接收到的上述一个或多个真值 CSI 作为模型的输入信息,输出一个或多个预测 CSI。

例如,接入网设备接收到了 $t+2$ 和 $t+4$ 时隙的真值 CSI,接入网设备可以将 $t+2$ 和 $t+4$ 时隙的真值 CSI 作为输入信息,输出 $t+5$ 时隙的预测 CSI。

S805.接入网设备根据一个或多个真值 CSI 和一个或多个预测 CSI,获取 AI 模型的性能的一个或多个监控结果。

AI 模型的性能的监控结果为 $KPI(H, \hat{H})$,其中 H 为真值 CSI, $\hat{H}$ 为预测 CSI。则针对每个真值 CSI,接入网设备可以根据该真值 CSI 和一个或多个预测 CSI,获得一个或多个监控结果;或者也可以针对每个预

测 CSI, 接入网设备可以根据一个或多个真值 CSI, 获得一个或多个监控结果。

进一步地, 为了对比预测 CSI 和真值 CSI 谁更准确, 接入网设备也可以将 T1 时刻的真值 CSI 用作 T2 时刻的预测 CSI。例如假设 CSI-RS 周期为 5ms, 下行传输在 t+3 时刻, 如果不是在监控阶段, 则 t+3 时刻没有 CSI-RS, 接入网设备可以将 t 时刻的真值 CSI 用作 t+3 时刻的预测 CSI, 得到 t+3 时刻的模型监控结果: $KPI(H_{t+3}, H_t)$ 。

则针对背景技术提出的问题, 由于可以获取 t+3 时刻的真值 CSI, 则可以获取 t+3 时刻的模型性能, 使得模型预测出的预测 CSI 能够反应出实际的系统性能, 从而可以将该预测 CSI 用于通信。

根据本申请实施例提供的一种模型监控方法, 终端设备通过获取一个或多个真值 CSI 分别对应的一个或多个第一索引, 可以获取特定的时间单元对应的真值 CSI, 从而可以灵活地监控预测模型的性能。

如图 10 所示, 为本申请实施例提供的另一种模型监控方法的流程示意图。在该实施例中, 终端设备未存储有 AI 模型。与图 8 所示实施例不同的是, 该实施例不是直接通过真值 CSI 的索引指示终端设备获取哪个真值 CSI, 而是根据用于测量真值 CSI 的 CSI-RS 的时域资源位置来确定真值 CSI。示例性地, 该方法可以包括以下步骤:

S1001. 终端设备获取第一信息。

用于测量真值 CSI (即用于模型监控) 的 CSI-RS 可以与用于模型推断 (作为模型的输入, 用于预测 CSI) 的 CSI-RS 是同一组 CSI-RS, 即属于同一个 CSI-RS 配置; 用于测量真值 CSI 的 CSI-RS 可以与用于模型推断的 CSI-RS 也可以是不同组的 CSI-RS, 即为模型监控专门配置一个 CSI-RS。

仍参考图 9, 还示例了用于测量真值 CSI 的 CSI-RS 的时域位置, 用于模型推断的 CSI-RS 为周期为 5ms 的 CSI-RS (例如 t-5、t、t+5 时隙对应的 CSI-RS); 用于模型监控的 CSI-RS 为非周期的或半静态的 CSI-RS, 为 t+2、t+4 时隙对应的 CSI-RS, 即用于测量真值 CSI 的 CSI-RS 与用于模型推断的 CSI-RS 是不同组的 CSI-RS。

本实施例中, 用于测量真值 CSI 的 CSI-RS 与用于模型推断的 CSI-RS 是不同组的 CSI-RS, 终端设备可以根据用于测量真值 CSI 的 CSI-RS 确定真值 CSI。因此, 终端设备可以获取第一信息, 该第一信息包括用于测量真值 CSI 的一个或多个 CSI-RS 的时域资源位置。

示例性地, 终端设备可以获取接入网设备发送的上述第一信息, 也可以获取自身预先存储的上述第一信息。

S1002. 终端设备根据第一信息, 确定一个或多个真值 CSI。

终端设备获取到第一信息后, 可以根据该第一信息, 确定一个或多个真值 CSI。

在一个实现方式中, 终端设备默认可以测量第一信息携带的一个或多个 CSI-RS 的时域资源位置上的一个或多个 CSI-RS, 得到一个或多个真值 CSI。例如, 在图 9 的示例中, 终端设备可以测量 t+2 和 t+4 时刻的 CSI-RS, 分别得到对应的真值 CSI。

在另一个实现方式中, 若接入网设备通过第一信息配置了用于测量真值 CSI 的一个或多个 CSI-RS 的时域资源位置, 则终端设备默认可以测量第一信息携带的一个或多个 CSI-RS 的时域资源位置上的一个或多个 CSI-RS, 得到一个或多个真值 CSI。例如, 在图 9 的示例中, 终端设备可以测量 t+2 和 t+4 时刻的 CSI-RS, 分别得到对应的真值 CSI。

在又一个实现方式中, 终端设备还可以获取用于模型推断的一个或多个 CSI-RS 的时域资源位置。终端设备默认可以测量第一信息携带的一个或多个 CSI-RS 的时域资源位置上的一个或多个 CSI-RS, 以及测量用于模型推断的一个或多个 CSI-RS 的时域资源位置上的一个或多个 CSI-RS, 得到一个或多个真值 CSI。例如, 在图 9 的示例中, 终端设备可以测量 t+2、t+4 和 t+5 时刻的 CSI-RS, 分别得到对应的真值 CSI。

在又一个实现方式中, 接入网设备也可以指示终端设备测量第一信息携带的一个或多个 CSI-RS 的时域资源位置上的一个或多个 CSI-RS, 获得一个或多个真值 CSI, 或者指示终端设备测量用于模型推断的一个或多个 CSI-RS 的时域资源位置上的一个或多个 CSI-RS, 获得一个或多个真值 CSI。

其中, 该一个或多个真值 CSI 用于获取 AI 模型的性能的监控结果。该 AI 模型的输出为预测 CSI。

可以看出, 本实施例中, 通过上述实现方式获得的一个或多个真值 CSI 中的至少一个真值 CSI 的时域资源位置与预测 CSI 的时域资源位置不同。

进一步地, 该方法还可以包括以下步骤 (图中以虚线表示):

S1003. 终端设备向接入网设备发送一个或多个真值 CSI。相应地, 接入网设备接收该一个或多个真值 CSI。

该步骤的具体实现可参考图 8 所示实施例的步骤 S803, 在此不再赘述。

S1004.接入网设备获取一个或多个预测 CSI。

其中，该一个或多个预测 CSI 为上述 AI 模型的输出。

该步骤的具体实现可参考图 8 所示实施例的步骤 S804，在此不再赘述。

S1005.接入网设备根据接收到的一个或多个真值 CSI 和一个或多个预测 CSI，获取 AI 模型的性能的一个或多个监控结果。

该步骤的具体实现可参考图 8 所示实施例的步骤 S805，在此不再赘述。

根据本申请实施例提供的一种模型监控方法，终端设备通过获取用于测量真值 CSI 的一个或多个 CSI-RS 的时域资源位置，可以获取特定的时域资源位置对应的真值 CSI，从而可以灵活地监控预测模型的性能。

如图 11 所示，为本申请实施例提供的又一种模型监控方法的流程示意图。与图 8、图 10 所示实施例不同的是，在该实施例中，终端设备存储有 AI 模型，可以预测 CSI。示例性地，该方法可以包括以下步骤：

S1101.终端设备获取第一信息。

其中，该第一信息包括一个或多个真值 CSI 的一个或多个第一索引。

该步骤的具体实现可参考图 8 所示实施例的步骤 S801，在此不再赘述。

S1102.终端设备获取一个或多个第一索引分别对应的一个或多个真值 CSI。

该步骤的具体实现可参考图 8 所示实施例的步骤 S802，在此不再赘述。

S1103.终端设备获取一个或多个预测 CSI。

本实施例不限定上述步骤 S1101、S1102 和 S1103 的顺序，可以先执行步骤 S1101、S1102，再执行步骤 S1103；也可以同时执行上述步骤 S1101、S1102 和 S1103；或者也可以先执行步骤 S1103，再执行步骤 S1101、S1102。

本实施例中，接入网设备可以指示终端设备获取哪些时刻的预测 CSI，终端设备根据接入网设备的指示，获取一个或多个预测 CSI。

具体地，接入网设备可以向终端设备发送第二信息，终端设备根据该第二信息，获取一个或多个预测 CSI。

在一个实现方式中，该第二信息包括一个或多个预测 CSI 的一个或多个第二索引。该一个或多个第二索引是基于参考资源对应的时间单元的。其中，参考资源的时刻是协议定义的，参考资源在时域上是一个时间单元（时间单元包括但不限于以下任意一种：符号、时隙、微时隙、子帧、无线帧，本申请实施例以时隙为例进行描述）。

该参考资源对应的时间单元为以下任意一个：

上报监控结果报告的资源所在的第一时间单元；或

调度模型监控的下行控制信令（downlink control information, DCI）所在的第二时间单元；或

基于第一时间单元或第二时间单元确定的第三时间单元；或

预定义/预配置的第四时间单元。

在本申请的一些实施例中，终端设备获取到预测 CSI 后，可以生成 AI 模型的监控结果报告，并在第一时间单元向接入网设备上报该监控结果报告。则参考资源对应的时间单元可以为上报监控结果报告的资源所在的第一时间单元。例如，该参考资源对应的时隙为用于上报该监控结果报告的资源所在的时隙之前或之后的第 A 个 CSI-RS 所在的时隙，其中，A 为接入网设备配置的或预定义的参数。例如，A=1。

在本申请的一些实施例中，接入网设备可以调度终端设备进行模型监控，具体地，接入网设备在第二时间单元向终端设备发送 DCI，该 DCI 用于指示终端设备进行模型监控。则参考资源对应的时间单元可以为调度模型监控的 DCI 所在的第二时间单元。

该参考资源对应的时间单元还可以是基于第一时间单元或第二时间单元确定的第三时间单元。例如，T 为上述第一时间单元或第二时间单元，第三时间单元为 $T + \Delta t$ 。其中， Δt 可以是接入网设备配置的或预定义的参数。

在确定了参考资源对应的时间单元后，本实现方式中，一个或多个预测 CSI 的一个或多个第二索引是基于上述参考资源对应的时间单元的。例如，参考资源所在时隙之前的第 n 个预测 CSI 的索引为 -n，参考资源所在时隙及其之后的第 n 个预测 CSI 的索引为 n-1。又例如，参考资源所在时隙及其之前的第 n 个预测 CSI 的索引为 -n+1，参考资源所在时隙之后的第 n 个预测 CSI 的索引为 n。又例如，参考资源所在时隙之前的第 K 个预测 CSI 的索引为 0，之后的预测 CSI 的索引依次增加，K 为预定义的或接入网设备配置的参数。又例如，CSI 参考资源所在时隙之后的第 K 个预测 CSI 的索引为 0，之前的预测 CSI 的索引依次增

加。

在另一个实现方式中，该第二信息包括一个或多个预测 CSI 中的每个预测 CSI 相对于参考资源对应的时间单元的相对时间单元值。即也可以通过预测 CSI 相对于参考资源对应的时间单元的相对时间单元值来指示预测 CSI。例如，0 表示参考资源所在时隙，-n 表示参考资源所在时隙之前的第 n 个时隙，n 表示 CSI 参考资源所在时隙之后的第 n 个时隙。

终端设备在接收到上述第二信息后，可以自身 AI 模型的实现，生成一个或多个预测 CSI。例如，如图 9 所示，参考资源所在时隙为 t，该第二信息包括的第二索引为 1，而用于模型推断的 CSI-RS 为周期为 5ms 的 CSI-RS，因此，该第二信息用于指示获取 t+5 时隙的预测 CSI。终端设备可以根据自身 AI 模型的实现，将 t-5 和 t 时隙的实际测量得到的 CSI 作为模型的输入信息，得到 t+5 时隙的预测 CSI。

进一步地，该方法还可以包括以下步骤（图中以虚线表示）：

S1104.终端设备向接入网设备发送一个或多个真值 CSI 以及一个或多个预测 CSI。相应地，接入网设备接收该一个或多个真值 CSI 以及一个或多个预测 CSI。

终端设备在获取了一个或多个真值 CSI 以及一个或多个预测 CSI，可以向接入网设备发送该一个或多个真值 CSI 以及一个或多个预测 CSI。

S1105.接入网设备根据一个或多个真值 CSI 和一个或多个预测 CSI，获取 AI 模型的性能的一个或多个监控结果。

该步骤的具体实现可参考图 8 所示实施例的步骤 S805，在此不再赘述。

根据本申请实施例提供的一种模型监控方法，终端设备通过获取一个或多个真值 CSI 分别对应的一个或多个第一索引，可以获得特定的时间单元对应的真值 CSI，以及可以根据接入网设备的指示获取相应的预测 CSI，从而使得接入网设备可以获得实际的系统的模型监控结果，使得可以灵活地监控预测模型的性能，预测性能能够反应出实际的系统性能。

如图 12 所示，为本申请实施例提供的又一种模型监控方法的流程示意图。在该实施例中，终端设备存储有 AI 模型，可以预测 CSI。与图 11 所示实施例不同的是，该实施例根据用于测量真值 CSI 的 CSI-RS 的时域资源位置确定真值 CSI。示例性地，该方法可以包括以下步骤：

S1201.终端设备获取第一信息。

其中，该第一信息包括用于测量真值 CSI 的一个或多个 CSI-RS 的时域资源位置。

该步骤的具体实现可参考图 10 所示实施例的步骤 S1001，在此不再赘述。

S1202.终端设备根据第一信息，确定一个或多个真值 CSI。

该步骤的具体实现可参考图 10 所示实施例的步骤 S1002，在此不再赘述。

S1203.终端设备获取一个或多个预测 CSI。

该步骤的具体实现可参考图 11 所示实施例的步骤 S1103，在此不再赘述。

S1204.终端设备向接入网设备发送一个或多个真值 CSI 以及一个或多个预测 CSI。相应地，接入网设备接收该一个或多个真值 CSI 以及一个或多个预测 CSI。

该步骤的具体实现可参考图 11 所示实施例的步骤 S1104，在此不再赘述。

本实施例不限定上述步骤 S1201、S1202 和 S1203 的顺序，可以先执行步骤 S1201、S1202，再执行步骤 S1203；也可以同时执行上述步骤 S1201、S1202 和 S1203；或者也可以先执行步骤 S1203，再执行步骤 S1201、S1202。

S1205.接入网设备根据一个或多个真值 CSI 和一个或多个预测 CSI，获取 AI 模型的性能的一个或多个监控结果。

该步骤的具体实现可参考图 8 所示实施例的步骤 S805，在此不再赘述。

根据本申请实施例提供的一种模型监控方法，终端设备通过获取用于测量真值 CSI 的一个或多个 CSI-RS 的时域资源位置，可以获得特定的时域资源位置对应的真值 CSI，以及可以根据接入网设备的指示获取相应的预测 CSI，从而使得接入网设备可以获得实际的系统的模型监控结果，使得可以灵活地监控预测模型的性能，预测性能能够反应出实际的系统性能。

如图 13 所示，为本申请实施例提供的又一种模型监控方法的流程示意图。与前述实施例不同的是，在该实施例中，终端设备存储有 AI 模型，可以预测 CSI，且终端设备还可以获取监控结果。示例性地，该方法可以包括以下步骤：

S1301.终端设备获取第一信息。

其中，该第一信息包括一个或多个真值 CSI 的一个或多个第一索引。

该步骤的具体实现可参考图 8 所示实施例的步骤 S801, 在此不再赘述。

S1302. 终端设备获取一个或多个第一索引分别对应的一个或多个真值 CSI。

该步骤的具体实现可参考图 8 所示实施例的步骤 S802, 在此不再赘述。

S1303. 终端设备获取一个或多个预测 CSI。

该步骤的具体实现可参考图 11 所示实施例的步骤 S1103, 在此不再赘述。

本实施例不限定上述步骤 S1301、S1302 和 S1303 的顺序, 可以先执行步骤 S1301、S1302, 再执行步骤 S1303; 也可以同时执行上述步骤 S1301、S1302 和 S1303; 或者也可以先执行步骤 S1303, 再执行步骤 S1301、S1302。

S1304. 终端设备根据一个或多个真值 CSI 和一个或多个预测 CSI, 获取 AI 模型的性能的一个或多个监控结果。

终端设备获取了一个或多个真值 CSI 和一个或多个预测 CSI 后, 可以针对每个真值 CSI, 根据该真值 CSI 和一个或多个预测 CSI, 获得一个或多个监控结果; 或者也可以针对每个预测 CSI, 根据一个或多个真值 CSI, 获得一个或多个监控结果。

例如, 终端设备获取了 2 个预测 CSI, 分别是 $\hat{H}_t$ 和 $\hat{H}_{t+5}$, 以及 2 个真值 CSI, 分别是 H_{t+2} 和 H_{t+4} , 则终端设备可以计算得到 4 个 KPI, 分别是 $KPI(H_{t+2}, \hat{H}_t)$, $KPI(H_{t+2}, \hat{H}_{t+5})$, $KPI(H_{t+4}, \hat{H}_t)$ 和 $KPI(H_{t+4}, \hat{H}_{t+5})$ 。

进一步地, 接入网设备也可以指示每个监控结果是通过哪个预测 CSI 和哪个真值 CSI 计算得到。

S1305. 终端设备向接入网设备发送监控结果报告。

终端设备在获取了上述一个或多个监控结果后, 可以向网络设备发送监控结果报告。

终端设备发送监控结果报告的实现方式包括但不限于:

一个实现方式为, 该监控结果报告包括上述一个或多个监控结果。在该实现方式中, 终端设备可以向接入网设备上报获取的所有监控结果, 使得接入网设备可以获取完整的监控结果。例如, 在上述示例中, 终端设备可以上报上述 4 个 KPI。

具体上报时, 终端设备可以按照预设的顺序将上述一个或多个监控结果上报给接入网设备。例如, 如果监控结果是 KPI 结果, 对于多个预测 CSI 和/或多个真值 CSI, 可以按照接入网设备指示时的顺序, 或按照它们在时间上的顺序, 对于预测 CSI 和真值 CSI 之间, 可以先按照预测 CSI 的顺序再按照真值 CSI 的顺序, 或先按照真值 CSI 的顺序再按照预测 CSI 的顺序上报。以 4 个 KPI 为 $KPI(H_{t+2}, \hat{H}_t)$, $KPI(H_{t+2}, \hat{H}_{t+5})$, $KPI(H_{t+4}, \hat{H}_t)$ 和 $KPI(H_{t+4}, \hat{H}_{t+5})$ 为例, 按照时间从前到后的顺序, 以及先按真值 CSI 的顺序再按待监控 CSI 的顺序, 4 个 KPI 的顺序为 $KPI(H_{t+2}, \hat{H}_t)$, $KPI(H_{t+4}, \hat{H}_t)$, $KPI(H_{t+2}, \hat{H}_{t+5})$, $KPI(H_{t+4}, \hat{H}_{t+5})$ 。

另一个实现方式为, 在某些场景下, 接入网设备可以仅知道是否存在监控结果表明模型性能较好, 无需获取所有的监控结果。因此, 该监控结果报告包括上述一个或多个监控结果中大于或等于第一阈值的至少一个监控结果。该第一阈值可以是接入网设备配置或预定义的。例如, 在上述示例中, 终端设备可以上报上述 4 个 KPI 中大于或等于第一阈值的一个或多个 KPI。该实现方式可以节省信令开销。

又一个实现方式为, 在某些场景下, 接入网设备可以仅知道是否存在监控结果表明模型性能较差, 无需获取所有的监控结果。因此, 该监控结果报告可以包括上述一个或多个监控结果中小于或等于第二阈值的至少一个监控结果。该第二阈值可以是接入网设备配置或预定义的。该第二阈值小于上述第一阈值。例如, 在上述示例中, 终端设备可以上报上述 4 个 KPI 中小于或等于第二阈值的一个或多个 KPI。该实现方式可以节省信令开销。

又一个实现方式为, 在某些场景下, 接入网设备可以仅知道是否存在监控结果表明模型性能较好, 无需获取所有的监控结果。因此, 该监控结果报告可以包括上述一个或多个监控结果中的最优监控结果对应的预测 CSI 的索引和真值 CSI 的索引。接入网设备可以预先接收到了一个或多个预测 CSI、一个或多个真值 CSI, 则可以根据最优监控结果对应的预测 CSI 的索引和真值 CSI 的索引, 自身获得该最优监控结果。该实现方式可以节省信令开销, 且对终端设备的计算能力要求较低。

可替换地, 如前所述, 接入网设备通过第一消息指示了用于计算监控结果的一个或多个预测 CSI 的索引, 可以用与接入网设备相同的指示方法, 在该监控结果报告可以包括上述一个或多个监控结果中的最优监控结果对应的预测 CSI 在接入网设备指示的一个或多个预测 CSI 的顺序中的序号; 以及如前所述, 接入网设备还指示了用于计算监控结果的一个或多个真值 CSI 的索引, 可以用与接入网设备相同的指示方法, 在该监控结果报告可以包括上述一个或多个监控结果中的最优监控结果对应的真值 CSI 在接入网设备指示的一个或多个真值 CSI 的顺序中的序号。

又一个实现方式为, 在某些场景下, 接入网设备可以仅知道是否存在监控结果表明模型性能较差, 无

需获取所有的监控结果。因此，该监控结果报告可以包括上述一个或多个监控结果中的最差监控结果对应的预测 CSI 的索引和真值 CSI 的索引。接入网设备可以预先接收到了一个或多个预测 CSI、一个或多个真值 CSI，则可以根据最差监控结果对应的预测 CSI 的索引和真值 CSI 的索引，自身获得该最差监控结果。该实现方式可以节省信令开销，且对终端设备的计算能力要求较低。

可替换地，如前所述，接入网设备通过第一消息指示了用于计算监控结果的一个或多个预测 CSI 的索引，可以用与接入网设备相同的指示方法，在该监控结果报告可以包括上述一个或多个监控结果中的最差监控结果对应的预测 CSI 在接入网设备指示的一个或多个预测 CSI 的顺序中的序号；以及如前所述，接入网设备还指示了用于计算监控结果的一个或多个真值 CSI 的索引，可以用与接入网设备相同的指示方法，在该监控结果报告可以包括上述一个或多个监控结果中的最差监控结果对应的真值 CSI 在接入网设备指示的一个或多个真值 CSI 的顺序中的序号。

又一个实现方式为，可以采用差分上报方式，该监控结果报告包括多个监控结果中的第一监控结果的绝对值、以及多个监控结果中除第一监控结果之外的其它监控结果相对于第一监控结果的相对值。以 4 个待上报 KPI 为 $KPI(H_{t+2}, \hat{H}_t)$, $KPI(H_{t+2}, \hat{H}_{t+5})$, $KPI(H_{t+4}, \hat{H}_t)$ 和 $KPI(H_{t+4}, \hat{H}_{t+5})$ 为例，上报量为 $KPI(H_{t+2}, \hat{H}_t)$, $\Delta KPI1$, $\Delta KPI2$, $\Delta KPI3$ ，其中 $\Delta KPI1 = KPI(H_{t+2}, \hat{H}_{t+5}) - KPI(H_{t+2}, \hat{H}_t)$, $\Delta KPI2 = KPI(H_{t+4}, \hat{H}_t) - KPI(H_{t+2}, \hat{H}_t)$, $\Delta KPI3 = KPI(H_{t+4}, \hat{H}_{t+5}) - KPI(H_{t+2}, \hat{H}_t)$ 。该实现方式可以节省信令开销。

进一步地，上述第一监控结果可以是基于在获取多个监控结果时，使用的真值 CSI 的顺序和/或使用的预测 CSI 的顺序确定的。如前所述，终端设备可以针对每个真值 CSI，根据该真值 CSI 和一个或多个预测 CSI，获得一个或多个监控结果，而对于每个真值 CSI，接入网设备指示了用于计算监控结果的一个或多个真值 CSI 的索引；或者也可以针对每个预测 CSI，根据一个或多个真值 CSI，获得一个或多个监控结果，而对于每个预测 CSI，接入网设备通过第一消息指示了用于计算监控结果的一个或多个预测 CSI 的索引。因此，可以根据终端设备生成监控结果时使用真值 CSI 和/或预测 CSI 的顺序，确定一个或多个监控结果的生成顺序，从而可以根据该生成顺序确定第一监控结果。例如，第一监控结果可以是顺序生成的一个或多个监控结果中的第一个监控结果或最后一个监控结果。

又一个实现方式为，接入网设备也可以指示终端设备上报部分监控结果。例如，可以通过位图的方式指示终端设备上报上述 4 个 KPI 结果中的部分 KPI 结果。

进一步地，上述监控结果报告可以和 CSI 反馈在同一个 CSI 报告中上报，也可以和 CSI 反馈在不同的 CSI 报告中上报。其中，CSI 反馈指的是用于上报 CSI-RS 的测量结果。

如果监控结果可以和 CSI 反馈在同一个 CSI 报告中上报，则在哪个用于 CSI 反馈的 CSI 报告中上报可以是接入网设备配置的，也可以是预定义的。例如，预定义在最后一个用于测量真值 CSI 的 CSI-RS 后的第一个 CSI 报告中上报，或者在最后一个用于测量真值 CSI 的 CSI-RS 后 T1 时间之后的第一个 CSI 报告中上报。其中，T1 为预留给终端设备计算 KPI 的时间，可以是接入网设备配置的或预定义的，T1 可以与终端设备的处理能力有关。需要说明的是，如果监控结果报告可以和 CSI 反馈在同一个 CSI 报告中上报，该 CSI 报告中上报的预测 CSI 和监控结果报告对应的待监控的预测 CSI 不是同一个 CSI，如图 14 所示，为本申请实施例示例的监控结果报告上报示意图，在 t+7 时刻上报 CSI 报告，该 t+7 时刻的 CSI 报告中上报的预测 CSI 是 t+10 时刻的预测 CSI，是由 t+5 时刻及其之前的真值 CSI 预测的；而该 CSI 报告中监控结果报告对应的待监控的预测 CSI 是 t+5 时刻的预测 CSI，是由 t 时刻及其之前的真值 CSI 预测的。

如果监控结果报告独立于 CSI 反馈上报，则监控结果报告的上报资源为接入网设备配置的。且终端设备不期望该上报资源早于该监控结果报告对应的最后一个用于测量真值 CSI 的 CSI-RS，或者终端设备不期望该上报资源早于该监控结果对应的最后一个用于测量真值 CSI 的 CSI-RS 后 T2 时间。T2 为预留给终端设备计算 KPI 的时间，可以是接入设备配置的或预定义的，T2 可以与终端设备的处理能力有关。

根据本申请实施例提供的一种模型监控方法，终端设备通过获取一个或多个真值 CSI 分别对应的一个或多个第一索引，可以获取特定的时间单元对应的真值 CSI，以及可以根据接入网设备的指示获取相应的预测 CSI，并根据获取的真值 CSI 和预测 CSI，获得模型监控结果，从而使得接入网设备可以获取实际的系统的模型监控结果，使得可以灵活地监控预测模型的性能，预测性能能够反应出实际的系统性能。

如图 15 所示，为本申请实施例提供的又一种模型监控方法的流程示意图。与图 8、图 10 所示的实施例不同的是，在该实施例中，终端设备存储有 AI 模型，可以预测 CSI，且终端设备还可以获取监控结果。与图 13 所示实施例不同的是，该实施例根据用于测量真值 CSI 的 CSI-RS 的时域资源位置确定真值 CSI。示例性地，该方法可以包括以下步骤：

S1501.终端设备获取第一信息。

其中，该第一信息包括用于测量真值 CSI 的一个或多个 CSI-RS 的时域资源位置。

该步骤的具体实现可参考图 10 所示实施例的步骤 S1001，在此不再赘述。

S1502.终端设备根据第一信息，确定一个或多个真值 CSI。

该步骤的具体实现可参考图 10 所示实施例的步骤 S1002，在此不再赘述。

S1503.终端设备获取一个或多个预测 CSI。

该步骤的具体实现可参考图 11 所示实施例的步骤 S1103，在此不再赘述。

S1504.终端设备根据一个或多个真值 CSI 和一个或多个预测 CSI，获取 AI 模型的性能的一个或多个监控结果。

该步骤的具体实现可参考图 13 所示实施例的步骤 S1304，在此不再赘述。

S1505.终端设备向接入网设备发送监控结果报告。

该步骤的具体实现可参考图 13 所示实施例的步骤 S1305，在此不再赘述。

根据本申请实施例提供的一种模型监控方法，终端设备通过获取用于测量真值 CSI 的一个或多个 CSI-RS 的时域资源位置，可以获取特定的时域资源位置对应的真值 CSI，以及可以根据接入网设备的指示获取相应的预测 CSI，并根据获取的真值 CSI 和预测 CSI，获得模型监控结果，从而使得接入网设备可以获得实际的系统的模型监控结果，使得可以灵活地监控预测模型的性能，预测性能能够反应出实际的系统性能。

上述实施例描述的方案也可以用于波束预测场景：

波束预测是通过历史/当前的波束测量结果预测未来的最优波束。

波束预测的 KPI 可以是 Top-1 准确性、带 ndB 边际的 Top1 准确性、Top-K 准确性、参考信号接收功率（reference signal receiving power, RSRP）差异等。

其中，Top-1 准确性指的是预测的最优波束是否实际的最优波束，或多次监控中预测的最优波束是实际的最优波束的比例；

带 ndB 边际的 Top1 准确性指的是预测的最优波束的 RSRP 与实际最优波束 RSRP 的差值是否小于 ndB，或多次监控中预测的最优波束的 RSRP 与实际最优波束 RSRP 的差值小于 ndB 的比例；

Top-K 准确性指的是预测出的最优波束是否在实际最优的 K 个波束中，或多次监控中预测出的最优波束在实际最优的 K 个波束中的比例，或实际的最优波束是否在预测的 K 个最优波束中，或多次监控中实际的最优波束在预测的 K 个最优波束中的比例；

RSRP 差异指的是预测的最优波束的 RSRP 和实际最优波束的 RSRP 的差异，或预测的最优波束的 RSRP 与该波束的实际 RSRP 的差异。

如图 16 所示，为本申请实施例提供的又一种模型监控方法的流程示意图。示例性地，该方法可以包括以下步骤：

S1601.终端设备获取第一信息。

与 CSI 预测类似，计算波束预测的 KPI 也需要两种信息，一种是预测信息，例如预测的最优波束的索引或 RSRP 中的一项或多项，另一种是真值信息，例如实际的最优波束的索引或 RSRP 中的一项或多项。接入网设备也可以指示终端设备使用哪个预测信息和真值信息计算 KPI。

本实施例通过真值波束的信息的索引指示终端设备获取哪个真值波束的信息。

终端设备获取第一信息。其中，该第一信息包括一个或多个真值波束的信息分别对应的一个或多个第一索引。示例性地，终端设备可以获取接入网设备发送的上述第一信息，也可以获取自身预先存储的上述第一信息。

该步骤的具体实现可参考图 8 所示实施例中 CSI 预测中获取第一信息的具体实现，在此不再赘述。

S1602.终端设备获取一个或多个第一索引分别对应的一个或多个真值波束的信息。

终端设备获取一个或多个真值波束的信息分别对应的一个或多个第一索引后，测量一个或多个索引分别对应的参考信号，如 CSI-RS，得到一个或多个第一索引分别对应的一个或多个真值波束的信息。

其中，该一个或多个真值波束的信息用于获取 AI 模型的性能的监控结果。该 AI 模型的输出为预测波束的信息。

相对于现有技术只能周期性地获取格点上模型监控结果，本实施例终端设备通过获取一个或多个真值波束的信息的一个或多个第一索引，可以获取特定的时间单元，如所需的时间单元，对应的真值波束的信息。

进一步地，终端设备可以向接入网设备发送一个或多个真值波束的信息，接入网设备获取一个或多个

预测波束的信息，并根据一个或多个真值波束的信息和一个或多个预测波束的信息，获取 AI 模型的性能的一个或多个监控结果。

或者，终端设备还可以获取一个或多个预测波束的信息，并向接入网设备发送一个或多个真值波束的信息以及一个或多个预测波束的信息，以及接入网设备根据一个或多个真值波束的信息和一个或多个预测波束的信息，获取 AI 模型的性能的一个或多个监控结果。

根据本申请实施例提供的一种模型监控方法，终端设备通过获取一个或多个真值波束的信息分别对应的一个或多个第一索引，可以获取特定的时间单元对应的真值波束的信息，从而可以灵活地监控预测模型的性能。

如图 17 所示，为本申请实施例提供的又一种模型监控方法的流程图。与图 16 所示实施例不同的是，该实施例不是直接通过真值波束的信息的索引指示终端设备获取哪个真值波束的信息，而是根据用于测量真值波束的信息的参考信号的时域资源位置来确定真值波束的信息。示例性地，该方法可以包括以下步骤：

S1701.终端设备获取第一信息。

用于测量真值波束的信息（即用于模型监控）的参考信号可以与用于模型推断（作为模型的输入，用于预测波束的信息）的参考信号是同一组参考信号，即属于同一个参考信号配置；用于测量真值波束的信息的参考信号可以与用于模型推断的参考信号也可以是不同组的参考信号，即为模型监控专门配置一个参考信号。

本实施例中，用于测量真值波束的信息的参考信号与用于模型推断的参考信号是不同组的参考信号，终端设备可以根据用于测量真值波束的信息的参考信号确定真值波束的信息。因此，终端设备可以获取第一信息，该第一信息包括用于测量真值波束的信息的一个或多个参考信号的时域资源位置。

示例性地，终端设备可以获取接入网设备发送的上述第一信息，也可以获取自身预先存储的上述第一信息。

该步骤的具体实现可参考图 10 所示实施例中 CSI 预测中获取第一信息的具体实现，在此不再赘述。

S1702.终端设备根据第一信息，确定一个或多个真值波束的信息。

终端设备获取到第一信息后，可以根据该第一信息，确定一个或多个真值波束的信息。

在一个实现方式中，终端设备默认可以测量第一信息携带的一个或多个参考信号的时域资源位置上的一个或多个参考信号，得到一个或多个真值波束的信息。

在另一个实现方式中，若接入网设备通过第一信息配置了用于测量真值波束的信息的一个或多个参考信号的时域资源位置，则终端设备默认可以测量第一信息携带的一个或多个参考信号的时域资源位置上的一个或多个参考信号，得到一个或多个真值波束的信息。

在又一个实现方式中，终端设备还可以获取用于模型推断的一个或多个参考信号的时域资源位置。终端设备默认可以测量第一信息携带的一个或多个参考信号的时域资源位置上的一个或多个参考信号，以及测量用于模型推断的一个或多个参考信号的时域资源位置上的一个或多个参考信号，得到一个或多个真值波束的信息。

在又一个实现方式中，接入网设备也可以指示终端设备测量第一信息携带的一个或多个参考信号的时域资源位置上的一个或多个参考信号，获得一个或多个真值波束的信息，或者指示终端设备测量用于模型推断的一个或多个参考信号的时域资源位置上的一个或多个参考信号，获得一个或多个真值波束的信息。

其中，该一个或多个真值波束的信息用于获取 AI 模型的性能的监控结果。该 AI 模型的输出为预测波束的信息。

可以看出，本实施例中，获得的一个或多个真值波束的信息中的至少一个真值波束的信息的时域资源位置与预测波束的信息的时域资源位置不同。

进一步地，终端设备可以向接入网设备发送一个或多个真值波束的信息，接入网设备获取一个或多个预测波束的信息，并根据一个或多个真值波束的信息和一个或多个预测波束的信息，获取 AI 模型的性能的一个或多个监控结果。

或者，终端设备还可以获取一个或多个预测波束的信息，并向接入网设备发送一个或多个真值波束的信息以及一个或多个预测波束的信息，以及接入网设备根据一个或多个真值波束的信息和一个或多个预测波束的信息，获取 AI 模型的性能的一个或多个监控结果。

根据本申请实施例提供的一种模型监控方法，终端设备通过获取用于测量真值波束的信息的一个或多个参考信号的时域资源位置，可以获取特定的时域资源位置对应的真值波束的信息，从而可以灵活地监控

预测模型的性能。

可以理解的是，以上各个实施例中，由接入网设备实现的方法和/或步骤，也可以由可用于接入网设备的部件（例如芯片或者电路）实现；由终端设备实现的方法和/或步骤，也可以由可用于终端设备的部件（例如芯片或者电路）实现。

上述主要从各个网元之间交互的角度对本申请实施例提供的方案进行了介绍。相应地，本申请实施例还提供了模型监控装置，该模型监控装置用于实现上述各种方法。该模型监控装置可以为上述方法实施例中的接入网设备，或者为可用于接入网设备的部件；或者，该模型监控装置可以为上述方法实施例中的终端设备，或者为可用于终端设备的部件。可以理解的是，该模型监控装置为了实现上述功能，其包含了执行各个功能相应的硬件结构和/或软件模块。本领域技术人员应该很容易意识到，结合本文中所公开的实施例描述的各示例的单元及算法步骤，本申请能够以硬件或硬件和计算机软件的结合形式来实现。某个功能究竟以硬件还是计算机软件驱动硬件的方式来执行，取决于技术方案的特定应用和设计约束条件。专业技术人员可以对每个特定的应用来使用不同方法来实现所描述的功能，但是这种实现不应认为超出本申请的范围。

本申请实施例可以根据上述方法实施例中对模型监控装置进行功能模块的划分，例如，可以对应各个功能划分各个功能模块，也可以将两个或两个以上的功能集成在一个处理单元中。上述集成的模块既可以采用硬件的形式实现，也可以采用软件功能模块的形式实现。需要说明的是，本申请实施例中对模块的划分是示意性的，仅仅为一种逻辑功能划分，实际实现时可以有另外的划分方式。

基于上述模型监控方法的同一构思，本申请还提供了如下模型监控装置：

如图 18 所示，为本申请实施例提供的一种模型监控装置的结构示意图，该模型监控装置 1800 包括收发单元 1801 和处理单元 1802；其中：

该模型监控装置用于实现上述方法实施例中终端设备的功能时，处理单元 1802 用于执行如图 6 所示实施例中 S601 和 S602 的操作；或者，处理单元 1802 用于执行如图 7 所示实施例中 S701 和 S702 的操作；收发单元 1801 用于执行如图 8 所示实施例中 S803 中终端设备的操作，以及处理单元 1802 用于执行如图 8 所示实施例中 S801 和 S802 的操作；或者，收发单元 1801 用于执行如图 10 所示实施例中 S1003 中终端设备的操作，以及处理单元 1802 用于执行如图 10 所示实施例中 S1001 和 S1002 的操作；或者，收发单元 1801 用于执行如图 11 所示实施例中 S1104 中终端设备的操作，以及处理单元 1802 用于执行如图 11 所示实施例中的步骤 S1101~S1103；或者，收发单元 1801 用于执行如图 12 所示实施例中 S1204 中终端设备的操作，以及处理单元 1802 用于执行如图 12 所示实施例中的步骤 S1201~S1203；或者，收发单元 1801 用于执行如图 13 所示实施例中 S1305 中终端设备的操作，以及处理单元 1802 用于执行如图 13 所示实施例中的步骤 S1301~S1304；或者，收发单元 1801 用于执行如图 15 所示实施例中 S1505 中终端设备的操作，以及处理单元 1802 用于执行如图 15 所示实施例中的步骤 S1501~S1504；或者，处理单元 1802 用于执行如图 16 所示实施例中 S1601 和 S1602 的操作；或者，处理单元 1802 用于执行如图 17 所示实施例中 S1701 和 S1702 的操作。

该模型监控装置用于实现上述方法实施例中接入网设备的功能时，收发单元 1801 用于执行如图 8 所示实施例中 S803 中接入网设备的操作，以及处理单元 1802 用于执行如图 8 所示实施例中 S804 和 S605 的操作；或者，收发单元 1801 用于执行如图 10 所示实施例中 S1003 中接入网设备的操作，以及处理单元 1802 用于执行如图 10 所示实施例中 S1004 和 S1005 的操作；或者，收发单元 1801 用于执行如图 11 所示实施例中 S1104 中接入网设备的操作，以及处理单元 1802 用于执行如图 11 所示实施例中的步骤 S1105；或者，收发单元 1801 用于执行如图 12 所示实施例中 S1204 中接入网设备的操作，以及处理单元 1802 用于执行如图 12 所示实施例中的步骤 S1205；或者，收发单元 1801 用于执行如图 13 所示实施例中 S1305 中接入网设备的操作；或者，收发单元 1801 用于执行如图 15 所示实施例中 S1505 中接入网设备的操作。

有关上述收发单元 1801 和处理单元 1802 的具体实现可参考上述方法实施例中的描述。

如图 19 所示，为本申请实施例提供的另一种模型监控装置的结构示意图，该模型监控装置 1900 包括一个或多个处理器 1901（图中示例了一个处理器）。可选的，该模型监控装置 1900 还可以包括存储器 1903（图中以虚线表示）。该存储器 1903 用于存储处理器 1901 执行的指令，或存储处理器 1901 运行指令所需要的输入数据，或存储处理器 1901 运行指令后产生的数据。可选的，模型监控装置 1900 还可以包括接口电路 1902（图中以虚线表示），处理器 1901 和接口电路 1902 之间相互耦合。可以理解的是，接口电路 1902 可以为收发器或输入输出接口。其中，处理器 1901 用于实现上述图 18 所示实施例中处理单元 1802 的功能；以及接口电路 1902 用于实现上述图 18 所示实施例中收发单元 1801 的功能。

当上述模型监控装置为应用于接入网设备的芯片时，该芯片实现上述方法实施例中接入网设备的功能。该芯片从接入网设备中的其它模块（如射频模块或天线）接收信息，该信息是终端设备发送给接入网设备的；或者，该芯片向接入网设备中的其它模块（如射频模块或天线）发送信息，该信息是接入网设备发送给终端设备的。

当上述模型监控装置为应用于终端设备的芯片时，该芯片实现上述方法实施例中终端设备的功能。该芯片从终端设备中的其它模块（如射频模块或天线）接收信息，该信息是接入网设备发送给终端设备的；或者，该芯片向终端设备中的其它模块（如射频模块或天线）发送信息，该信息是终端设备发送给接入网设备的。

此外，需要说明的是，前述收发单元和/或处理单元可通过虚拟模块实现，例如处理单元可通过软件功能单元或虚拟装置实现，收发单元可以通过软件功能或虚拟装置实现。或者，处理单元或收发单元也可以通过实体装置实现，例如若该装置采用芯片/芯片电路实现，所述收发单元可以是输入输出电路和/或通信接口，执行输入操作（对应前述接收操作）、输出操作（对应前述发送操作）；处理单元为集成的处理器或者微处理器或者集成电路。

本申请中对模块的划分是示意性的，仅仅为一种逻辑功能划分，实际实现时可以有另外的划分方式，另外，在本申请各个示例中的各功能模块可以集成在一个处理器中，也可以是单独物理存在，也可以两个或两个以上模块集成在一个模块中。上述集成的模块既可以采用硬件的形式实现，也可以采用软件功能模块的形式实现。

可以理解的是，本申请的实施例中的处理器可以是中央处理单元（central processing unit, CPU），还可以是其它通用处理器、数字信号处理器（digital signal processor, DSP）、专用集成电路（application specific integrated circuit, ASIC）、现场可编程门阵列（field programmable gate array, FPGA）或者其它可编程逻辑器件、晶体管逻辑器件，硬件部件或者其任意组合。通用处理器可以是微处理器，也可以是任何常规的处理器。

本申请实施例还提供了一种计算机可读存储介质，该计算机可读存储介质中存储有计算机程序或指令，当计算机程序或指令被执行时，实现上述实施例中的方法。

本申请实施例还提供了一种包含指令的计算机程序产品，当该指令在计算机上运行时，使得计算机执行上述实施例中的方法。

本申请实施例还提供了一种模型监控系统，包括上述的模型监控装置。

本申请实施例还提供了一种电路，该电路与存储器耦合，该电路被用于执行上述实施例中的方法。该电路可包括芯片电路。

当上述模型监控装置为应用于基站的模块时，该基站模块实现上述方法实施例中基站的功能。该基站模块从基站中的其它模块（如射频模块或天线）接收信息，该信息是终端设备发送给基站的；或者，该基站模块向基站中的其它模块（如射频模块或天线）发送信息，该信息是基站发送给终端设备的。这里的基站模块可以是基站的基带芯片，也可以是 CU、DU 或其他模块，也可以是开放式无线接入网（open radio access network, O-RAN）架构下的装置，例如开放式 CU、开放式 DU 等装置。

需要说明的是，以上单元或单元的一个或多个可以软件、硬件或二者结合来实现。当以上任一单元或单元以软件实现的时候，所述软件以计算机程序指令的方式存在，并被存储在存储器中，处理器可以用于执行所述程序指令并实现以上方法流程。

在本申请中，处理器可以是通用处理器、数字信号处理器、专用集成电路、现场可编程门阵列或者其他可编程逻辑器件、分立门或者晶体管逻辑器件、分立硬件组件，或者，前述器件中的用于实现处理功能的全部或部分电路，可以实现或者执行本申请中的公开的各方法、步骤及逻辑框图。通用处理器可以是微处理器或者任何常规的处理器等。结合本申请所公开的方法的步骤可以直接体现为硬件处理器执行完成，或者用处理器中的硬件及软件模块组合执行完成。

当以上单元或单元以硬件实现的时候，该硬件可以是 CPU、微处理器、数字信号处理（digital signal processing, DSP）芯片、微控制单元（microcontroller unit, MCU）、人工智能处理器、ASIC、SoC、FPGA、PLD、专用数字电路、硬件加速器或非集成的分立器件中的任一个或任一组合，其可以运行必要的软件或不依赖于软件以执行以上方法流程。

可选的，本申请实施例还提供了一种芯片系统，包括：至少一个处理器和接口，该至少一个处理器通过接口与存储器耦合，当该至少一个处理器运行存储器中的计算机程序或指令时，使得该芯片系统执行上述任一方法实施例中的方法。可选的，该芯片系统可以由芯片构成，也可以包含芯片和其他分立器件，本

申请实施例对此不作具体限定。

本申请中的存储器还可以是电路或者其它任意能够实现存储功能的装置,用于存储程序指令和/或数据。存储器是能够用于携带或存储具有指令或数据结构形式的期望的程序代码并能够由计算机存取的任何其他介质,但不限于此。例如,存储器可以是非易失性存储器,比如数字通用光盘(digital versatile disc, DVD)、硬盘(hard disk drive, HDD)或固态硬盘(solid-state drive, SSD)等,还可以是易失性存储器(volatile memory),例如随机存取存储器(random-access memory, RAM)。

应理解,在本申请的描述中,除非另有说明,“/”表示前后关联的对象是一种“或”的关系,例如,A/B可以表示A或B;其中A,B可以是单数或者复数。并且,在本申请的描述中,除非另有说明,“多个”是指两个或两个以上。“以下至少一项(个)”或其类似表达,是指的这些项中的任意组合,包括单项(个)或复数项(个)的任意组合。例如,a,b,或c中的至少一项(个),可以表示:a,b,c,a-b,a-c,b-c,或a-b-c,其中a,b,c可以是单个,也可以是多个。另外,为了便于清楚描述本申请实施例的技术方案,在本申请的实施例中,采用了“第一”、“第二”等字样对功能和作用基本相同的相同项或相似项进行区分。本领域技术人员可以理解“第一”、“第二”等字样并不对数量和执行次序进行限定,并且“第一”、“第二”等字样也并不限定一定不同。同时,在本申请实施例中,“示例性的”或者“例如”等词用于表示作例子、例证或说明。本申请实施例中被描述为“示例性的”或者“例如”的任何实施例或设计方案不应被解释为比其它实施例或设计方案更优选或更具优势。确切而言,使用“示例性的”或者“例如”等词旨在以具体方式呈现相关概念,便于理解。

在上述实施例中,可以全部或部分地通过软件、硬件、固件或者其任意组合来实现。当使用软件程序实现时,可以全部或部分地以计算机程序产品的形式来实现。该计算机程序产品包括一个或多个计算机指令。在计算机上加载和执行计算机程序指令时,全部或部分地产生按照本申请实施例所述的流程或功能。所述计算机可以是通用计算机、专用计算机、计算机网络、或者其他可编程装置。所述计算机指令可以存储在计算机可读存储介质中,或者从一个计算机可读存储介质向另一个计算机可读存储介质传输,例如,所述计算机指令可以从一个网站站点、计算机、服务器或者数据中心通过有线(例如同轴电缆、光纤、数字用户线(digital subscriber line, DSL))或无线(例如红外、无线、微波等)方式向另一个网站站点、计算机、服务器或数据中心进行传输。

尽管在此结合各实施例对本申请进行了描述,然而,在实施所要求保护的本申请过程中,本领域技术人员通过查看所述附图、公开内容、以及所附权利要求书,可理解并实现所述公开实施例的其他变化。在权利要求中,“包括”(comprising)一词不排除其他组成部分或步骤,“一”或“一个”不排除多个的情况。单个处理器或其他单元可以实现权利要求中列举的若干项功能。相互不同的从属权利要求中记载了某些措施,但这并不表示这些措施不能组合起来产生良好的效果。

可以理解的是,在本申请的实施例中涉及的各种数字编号仅为描述方便进行的区分,并不用来限制本申请的实施例的范围。上述各过程的序号的大小并不意味着执行顺序的先后,各过程的执行顺序应以其功能和内在逻辑确定。

在上述实施例中,对各个实施例的描述都各有侧重,某个实施例中未详述的部分,可以参见其他实施例的相关描述。

本申请实施例装置中的部件可以根据实际需要进行合并、划分和删减。本领域的技术人员可以将本说明书中描述的不同实施例以及不同实施例的特征进行结合或组合。

在本申请中,在无逻辑矛盾的前提下,各示例之间可以相互引用,例如方法实施例之间的方法和/或术语可以相互引用,例如装置实施例之间的功能和/或术语可以相互引用,例如装置示例和方法示例之间的功能和/或术语可以相互引用。

权利要求书

- 1.一种模型监控方法，其特征在于，所述方法包括：
获取第一信息，所述第一信息包括一个或多个真值信息分别对应的一个或多个第一索引；
获取所述一个或多个第一索引分别对应的一个或多个真值信息；
其中，所述一个或多个真值信息用于获取人工智能 AI 模型的性能的监控结果，所述 AI 模型的输出为预测信息。
- 2.一种模型监控方法，其特征在于，所述方法包括：
获取第一信息，所述第一信息包括用于测量真值信息的一个或多个参考信号的时域资源位置；
根据所述第一信息，确定一个或多个真值信息；
其中，所述一个或多个真值信息用于获取所述 AI 模型的性能的监控结果，所述 AI 模型的输出为预测信息，所述一个或多个真值信息中的至少一个真值信息的时域资源位置与所述预测信息的时域资源位置不同。
- 3.根据权利要求 1 或 2 所述的方法，其特征在于，所述方法还包括：
发送所述一个或多个真值信息。
- 4.根据权利要求 1 或 2 所述的方法，其特征在于，所述方法还包括：
获取一个或多个预测信息，所述一个或多个预测信息为人工智能 AI 模型的输出；
发送所述一个或多个预测信息。
- 5.根据权利要求 1 或 2 所述的方法，其特征在于，所述方法还包括：
获取一个或多个预测信息，所述一个或多个预测信息为人工智能 AI 模型的输出；
根据所述一个或多个真值信息和所述一个或多个预测信息，获取所述 AI 模型的性能的一个或多个监控结果。
- 6.根据权利要求 4 或 5 所述的方法，其特征在于，所述获取一个或多个预测信息，包括：
接收第二信息，所述第二信息包括所述一个或多个预测信息分别对应的一个或多个第二索引，或所述第二信息包括所述一个或多个预测信息中的每个预测信息相对于参考资源对应的时间单元的相对时间单元值；
根据所述第二信息，获取一个或多个预测信息。
- 7.根据权利要求 6 所述的方法，其特征在于，所述一个或多个第一索引、所述一个或多个第二索引是基于参考资源对应的时间单元的。
- 8.根据权利要求 7 所述的方法，其特征在于，所述参考资源对应的时间单元为以下任意一个：
上报监控结果报告的资源所在的第一时间单元；或
调度所述模型监控的下行控制信令所在的第二时间单元；或
基于所述第一时间单元或所述第二时间单元确定的第三时间单元；或
预定义/预配置的第四时间单元。
- 9.根据权利要求 5 所述的方法，其特征在于，所述方法还包括：
发送监控结果报告；
其中，所述监控结果报告包括所述一个或多个监控结果；或
所述监控结果报告包括所述一个或多个监控结果中大于或等于第一阈值的至少一个监控结果；或
所述监控结果报告包括所述一个或多个监控结果中小于或等于第二阈值的至少一个监控结果；或
所述监控结果报告包括所述一个或多个监控结果中的最优监控结果对应的预测信息的索引和真值信息的索引；或

所述监控结果报告包括所述一个或多个监控结果中的最差监控结果对应的预测信息的索引和真值信息的索引；或

所述监控结果报告包括多个监控结果中的第一监控结果的绝对值、以及所述多个监控结果中除所述第一监控结果之外的其它监控结果相对于所述第一监控结果的相对值。

10.根据权利要求9所述的方法,其特征在于,所述第一监控结果是基于在获取所述多个监控结果时,使用的所述真值信息的顺序和/或使用的所述预测信息的顺序确定的。

11.如权利要求9或10所述的方法,其特征在于,所述监控结果报告包括所有监控结果中的部分监控结果。

12.如权利要求9-11中任一项所述的方法,其特征在于,所述监控结果报告和用于发送参考信号测量结果的信息反馈在同一个信息报告中上报,或所述监控结果报告和用于发送参考信号测量结果的信息反馈在不同的信息报告中上报。

13.一种模型监控方法,其特征在于,所述方法包括:

发送第一信息,所述第一信息包括一个或多个真值信息分别对应的一个或多个第一索引;

接收所述一个或多个第一索引分别对应的一个或多个真值信息;

其中,所述一个或多个真值信息用于获取人工智能AI模型的性能的监控结果,所述AI模型的输出为预测信息。

14.一种模型监控方法,其特征在于,所述方法包括:

发送第一信息,所述第一信息包括用于测量真值信息的一个或多个参考信号的时域资源位置;

根据所述第一信息,接收一个或多个真值信息;

其中,所述一个或多个真值信息用于获取所述AI模型的性能的监控结果,所述AI模型的输出为预测信息,所述一个或多个真值信息中的至少一个真值信息的时域资源位置与所述预测信息的时域资源位置不同。

15.根据权利要求13或14所述的方法,其特征在于,所述方法还包括:

获取一个或多个预测信息;

根据所述一个或多个真值信息和所述一个或多个预测信息,获取所述AI模型的性能的一个或多个监控结果。

16.根据权利要求13或14所述的方法,其特征在于,所述方法还包括:

接收监控结果报告,所述监控结果报告包括一个或多个监控结果,所述一个或多个监控结果为根据所述一个或多个真值信息和一个或多个预测信息获取的。

17.根据权利要求15所述的方法,其特征在于,所述方法包括:

发送第二信息,所述第二信息包括所述一个或多个预测信息分别对应的一个或多个第二索引,或所述第二信息包括所述一个或多个预测信息中的每个预测信息相对于参考资源对应的时间单元的相对时间单元值;

接收所述一个或多个预测信息。

18.根据权利要求17所述的方法,其特征在于,所述一个或多个第一索引、所述一个或多个第二索引是基于参考资源对应的时间单元的。

19.根据权利要求18所述的方法,其特征在于,所述参考资源对应的时间单元为以下任意一个:

上报监控结果报告的资源所在的第一时间单元;或

调度所述模型监控的下行控制信令所在的第二时间单元;或

基于所述第一时间单元或所述第二时间单元确定的第三时间单元；或
预定义/预配置的第四时间单元。

20.根据权利要求 16 所述的方法，其特征在于：

所述监控结果报告包括所述一个或多个监控结果；或

所述监控结果报告包括所述一个或多个监控结果中大于或等于第一阈值的至少一个监控结果；或

所述监控结果报告包括所述一个或多个监控结果中小于或等于第二阈值的至少一个监控结果；或

所述监控结果报告包括所述一个或多个监控结果中的最优监控结果对应的预测信息的索引和真值信息的索引；或

所述监控结果报告包括所述一个或多个监控结果中的最差监控结果对应的预测信息的索引和真值信息的索引；或

所述监控结果报告包括多个监控结果中的第一监控结果的绝对值、以及所述多个监控结果中除所述第一监控结果之外的其它监控结果相对于所述第一监控结果的相对值。

21.根据权利要求 20 所述的方法，其特征在于，所述第一监控结果是基于在获取所述多个监控结果时，使用的所述真值信息的顺序和/或使用的所述预测信息的顺序确定的。

22.如权利要求 20 或 21 所述的方法，其特征在于，所述监控结果报告包括所有监控结果中的部分监控结果。

23.如权利要求 20-22 中任一项所述的方法，其特征在于，所述监控结果报告和用于发送参考信号测量结果的信息反馈在同一个信息报告中上报，或所述监控结果报告和用于发送参考信号测量结果的信息反馈在不同的信息报告中上报。

24.如权利要求 1-23 中任一项所述的方法，其特征在于，所述预测信息是预测信道状态信息 CSI，所述真值信息是真值 CSI；或所述预测信息是预测波束的信息，所述真值信息是真值波束的信息。

25.一种模型监控装置，其特征在于，包括用于执行如权利要求 1-24 中任意一项所述的方法的模块。

26.一种模型监控装置，其特征在于，包括与存储器耦合的处理器，所述处理器被配置为实现所述装置执行如权利要求 1-24 中任一项所述的方法中相应的功能。

27.一种计算机可读存储介质，所述计算机可读存储介质中存储有计算机程序或指令，其特征在于，当所述计算机程序或指令被执行时，实现如权利要求 1-24 中任意一项所述的方法。

28.一种包含指令的计算机程序产品，其特征在于，当所述指令在模型监控装置上运行时，使得所述模型监控装置执行如权利要求 1-24 中任意一项所述的方法。