

虚拟角色的动作决策方法、装置、设备及存储介质

本申请要求于 2023 年 09 月 15 日提交，申请号为 202311195505.3、发明名称为“虚拟角色的动作决策方法、装置、设备及存储介质”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

本申请实施例涉及人工智能技术领域，特别涉及一种虚拟角色的动作决策方法、装置、设备及存储介质。

背景技术

现如今，在策划电子游戏过程中，通常通过使 NPC（Non-Player Character，非玩家角色）的行为逻辑与真人玩家的行为逻辑尽可能一致，以使得游戏中的 NPC 具有高拟人性。

相关技术中，通常使用行为树（behavior tree）结构来实现 NPC 的行为逻辑决策，行为树是用于实现 NPC 角色复杂行为的工具，在控制 NPC 执行动作时计算机设备从行为树的根节点开始按照执行的顺序遍历，直至达到终结状态。在遍历行为树的过程中，计算机设备会基于不同叶子节点返回的状态信息（成功、失败或运行中）基于设定的规则确定下一个执行的节点，从而实现对 NPC 的行为逻辑决策。

然而，行为树结构是一种基于规则的算法，在设定的规则不变的情况下，会导致 NPC 在场景中的表现过于单一，若需要使得 NPC 能够具有高拟人性，则需要预先设定复杂的规则，导致人力成本较高且对于不同地图场景的适配性较差。

发明内容

本申请实施例提供了一种虚拟角色的动作决策方法、装置、设备及存储介质。所述技术方案如下：

一方面，本申请实施例提供了一种虚拟角色的动作决策方法，所述方法由计算机设备执行，所述方法包括：

获取状态信息，所述状态信息用于表征第一虚拟角色所处对局的对局状态；

将所述状态信息输入动作决策模型，得到所述动作决策模型中 n 个动作输出头串行输出的 n 个目标子动作，其中，不同动作输出头对应不同动作类型，且 n 个所述动作输出头基于所述动作类型之间的依赖关系串行连接，所述依赖关系用于表征不同动作类型下子动作之间的依赖限制情况， n 为正整数；

控制所述第一虚拟角色执行由 n 个所述目标子动作构成的目标动作。

另一方面，本实施例提供了一种虚拟角色的动作决策装置，所述装置包括：

获取模块，用于获取状态信息，所述状态信息用于表征第一虚拟角色所处对局的对局状态；

决策模块，用于将所述状态信息输入动作决策模型，得到所述动作决策模型中 n 个动作输出头串行输出的 n 个目标子动作，其中，不同动作输出头对应不同动作类型，且 n 个所述动作输出头基于所述动作类型之间的依赖关系串行连接，所述依赖关系用于表征不同动作类型下子动作之间的依赖限制情况， n 为正整数；

控制模块，用于控制所述第一虚拟角色执行由 n 个所述目标子动作构成的目标动作。

另一方面，本申请实施例提供了一种计算机设备，所述计算机设备包括处理器和存储器，所述存储器中存储有至少一条计算机指令，所述至少一条计算机指令由所述处理器加载并执行以实现如上述方面所述的虚拟角色的动作决策方法。

另一方面，本申请实施例提供了一种计算机可读存储介质，所述可读存储介质中存储有至少一条计算机指令，所述至少一条计算机指令由处理器加载并执行以实现如上述方面所述

的虚拟角色的动作决策方法。

另一方面，本申请实施例提供了一种计算机程序产品，该计算机程序产品包括计算机指令，该计算机指令存储在计算机可读存储介质中。计算机设备的处理器从计算机可读存储介质读取该计算机指令，处理器执行该计算机指令，使得该计算机设备执行上述方面提供的虚拟角色的动作决策方法。

本申请实施例提供的技术方案带来的有益效果至少包括：

本申请实施例中，通过动作决策模型基于状态信息实现动作决策，从而能够根据当前状态，确定出合理的目标动作。其中，动作决策模型中包含 n 个动作输出头串行连接， n 个动作输出头分别用于输出不同的目标子动作， n 个动作输出头基于动作类型之间的依赖关系串行连接，能够在动作决策时考虑到不同动作之间的依赖关系，并在串行输出 n 个目标子动作的过程中，使得 n 个目标子动作在时序上前后因果关联，从而提高目标动作的合理性，提升了虚拟角色的拟人性。

附图说明

图 1 示出了本申请一个示例性实施例提供的计算机系统的结构框图；

图 2 示出了本申请一个示例性实施例提供的虚拟角色的动作决策方法的流程图；

图 3 示出了本申请一个示例性实施例提供的射线探测方案的示意图；

图 4 示出了本申请一个示例性实施例提供的二维深度图的示意图；

图 5 示出了本申请一个示例性实施例提供的用于提供虚拟环境的应用程序的用户界面示意图；

图 6 示出了本申请一个示例性实施例提供的确定目标动作的过程的流程图；

图 7 示出了本申请一个示例性实施例提供的 n 个动作输出头的自回归结构的示意图；

图 8 示出了相关技术中一种对并行动作输出头执行动作遮罩的示意图；

图 9 示出了本申请一个示例性实施例提供的移动方向与朝向方向叠加影响移动状态的示意图；

图 10 示出了本申请一个示例性实施例提供的可视情况感知的示意图；

图 11 示出了本申请一个示例性实施例提供的确定竖直方向有效瞄准范围的示意图；

图 12 示出了本申请一个示例性实施例提供的水平方向有效瞄准范围的示意图；

图 13 示出了本申请一个示例性实施例提供的动作决策模型的示意图；

图 14 示出了本申请一个示例性实施例提供的动作决策模型训练过程的流程图；

图 15 示出了本申请一个示例性实施例提供的训练过程中客户端与训练动作决策模型的服务器之间交互的示意图；

图 16 示出了本申请一个示例性实施例通过的应用以及训练过程中实现动作决策请求的决策模式示意图；

图 17 示出了本申请一个示例性实施例通过的应用过程中实现动作决策请求的决策模式示意图；

图 18 示出了本申请一个示例性实施例提供的虚拟角色的动作决策装置的结构示意图；

图 19 示出了本申请一个示例性实施例提供的计算机设备的结构示意图。

具体实施方式

为使本申请的目的、技术方案和优点更加清楚，下面将结合附图对本申请实施方式作进一步地详细描述。

首先，对本申请实施例中涉及的名词进行介绍。

智能体：能通过传感器感知环境和通过执行器作用于环境，对于每种可能的感知序列，智能体应当选择一个动作来执行，该动作能够在它具有指示和感知序列提供的证据条件下，使它的性能达到期望最大值。本申请实施例中，虚拟角色作为智能体感知虚拟环境，获取环

境感知信息，并执行动作决策模型确定的目标动作。

虚拟环境：是应用程序在计算机设备上运行时显示（或提供）的虚拟环境。该虚拟环境可以是对真实世界的仿真环境，也可以是半仿真半虚构的环境，还可以是纯虚构的环境。虚拟环境可以是二维虚拟环境、2.5 维虚拟环境和三维虚拟环境中的任意一种，本申请对此不加以限定。下述实施例以虚拟环境是三维虚拟环境来举例说明。

虚拟角色：是指在虚拟环境中的可活动对象。该可活动对象可以是虚拟角色、虚拟动物、动漫人物中的至少一种。可选地，当虚拟环境为三维虚拟环境时，虚拟角色可以是三维虚拟模型，每个虚拟角色在三维虚拟环境中具有自身的形状和体积，占据三维虚拟环境中的一部分空间。可选地，虚拟角色是基于三维人体骨骼技术构建的三维角色，该虚拟角色通过穿戴不同的皮肤来实现不同的外在形象。在一些实现方式中，虚拟角色也可以采用 2.5 维或 2 维模型来实现，本申请实施例对此不加以限定。

相关技术中，通过基于人工规则的层次行为树结构来实现虚拟角色的动作决策过程，从根节点开始进行搜索执行并返回“成功”、“失败”、“运行中”其中之一的执行结果，借此来控制虚拟角色的行为。

然而，行为树的规则一成不变，基于行为树执行虚拟角色的动作决策使得虚拟角色表现固定单一，难以实现对虚拟角色的行为逻辑的复杂控制，且在复杂场景中，需要行为树具有大量的规则，导致消耗大量的人工成本。

此外，在相关技术中，可能会采用监督学习的方式，通过模仿学习真实玩家控制虚拟角色产生的对照数据，训练出符合人类玩家行为的网络模型。

然而，由于地图场景复杂多样，并非所有地图场景均存在足够多的真实玩家，以供计算机设备数据采集，因此，采用监督学习的方式训练动作决策模型，浪费大量的时间及人力成本，且通过监督学习的方式训练的动作决策模型的决策策略过于拟合在训练数据的分布上，在实际部署时难以实现有效的动作决策。

因此，本申请实施例提供了一种虚拟角色的动作决策方法，通过获取状态信息，并根据强化训练完成的动作决策模型基于状态信息在 n 个动作输出头串行输出目标子动作，使得输出的目标子动作具有较高的合理性以及拟人性。

本申请中的计算机设备可以为台式计算机、膝上编写计算机、手机、平板电脑、台式电脑、电子书阅读器、MP3（Moving Picture Experts Group Audio Layer III，动态影像专家压缩标准音频层面 3）播放器、MP4（Moving Picture Experts Group Audio Layer IV，动态影像专家压缩标准音频层面 4）播放器、智能语音交互设备、智能家电以及车载终端等等。该计算机设备中安装和运行有支持虚拟环境的应用程序，比如支持三维虚拟环境的应用程序。该应用程序可以是虚拟现实应用程序、三维地图程序、TPS（Third-Person Shooter，第三人称射击）游戏、FPS（First-Person Shooter，第一人称射击）游戏、MOBA（Multiplayer Online Battle Arena，多人在线战术竞技）游戏中的任何一种。可选地，该应用程序可以是单机版的应用程序，比如单机版的 3D 游戏程序，也可以是网络联机版的应用程序。下述实施例是以在游戏中的应用来举例说明。

基于虚拟环境的游戏往往由一个或多个游戏世界的地图构成，游戏中的虚拟环境模拟现实世界的场景，目标虚拟角色在虚拟环境中可以进行行走、跑步、跳跃、射击、格斗、驾驶、攀爬、滑翔、切换使用虚拟道具、使用虚拟道具攻击其他虚拟角色等动作。

请参考图 1，其示出了本申请一个示例性实施例提供的计算机系统的结构框图。该计算机系统可以包括：终端 110、第一服务器 120 以及第二服务器 130。

在本申请实施例中，终端 110 包括但不限于手机、平板电脑、笔记本电脑、台式电脑、电子书阅读器、智能语音交互设备、智能家电以及车载终端等设备。终端 110 中运行有支持

虚拟环境的应用程序 111，该应用程序可以是 TPS（Third-Person Shooter，第三人称射击）游戏或者 FPS（First-Person Shooter，第一人称射击）游戏。当终端 110 运行应用程序 111 时，终端 110 的屏幕上显示应用程序 111 的用户界面。用户使用终端 110 控制位于虚拟环境中的虚拟角色进行活动，或者由终端控制位于虚拟环境中的 NPC（Non-Player Character，非玩家控制角色）进行活动。虚拟角色的活动包括但不限于：调整身体姿态、爬行、步行、奔跑、骑行、飞行、跳跃、驾驶、拾取、射击、攻击、投掷、释放技能中的至少一种。

第一服务器 120 包括一台服务器、多台服务器、云计算平台和虚拟化中心中的至少一种。本申请实施例中，第一服务器 120 用于为支持三维虚拟环境的应用程序提供后台服务，可以为 DS（Dedicated Server，专用服务器）。可选地，第一服务器 120 承担主要计算工作，终端承担次要计算工作；或者，第一服务器 120 承担次要计算工作，终端承担主要计算工作；或者，第一服务器 120 和终端之间采用分布式计算架构进行协同计算。

第二服务器 130 包括一台服务器、多台服务器、云计算平台和虚拟化中心中的至少一种。本申请实施例中，第二服务器 130 用于支持三维虚拟环境的应用程序提供虚拟角色的动作决策服务。在第二服务器 130 提供的动作决策服务开启时，在 ETCD（et-see-dee，分布式键值对存储）中注册自身地址。在具有虚拟角色的动作决策需求的情况下，第一服务器 120 向服务调度请求连接推理服务，并且获取服务调度返回的动作决策服务的地址。从而第一服务器 120 通过得到的地址向第二服务器 130 提供的动作决策服务发送动作决策请求，在动作决策服务基于训练后的动作决策模型确定目标动作后，向第一服务器 120 返回动作指令。

在训练动作决策模型的过程中，第二服务器 130 基于第一服务器 120 发送的状态信息，通过强化学习的方式训练动作决策模型。

本申请实施例可以应用在虚拟环境中虚拟角色的导航场景、虚拟动作完成设定任务的场景或者虚拟角色与其他虚拟角色的对局场景中，本实施例对此不进行限定。下面对本申请实施例在虚拟场景中的应用过程进行示意性的说明。

本申请实施例提供的虚拟角色的动作决策方法应用于游戏场景中，当游戏场景中要实现对第一虚拟角色的动作决策的情况下，通过终端 110 或者第一服务器 120 获取第一虚拟角色所处对局的状态信息，并通过第二服务器 130 训练完成的动作决策模型基于状态信息，得到 n 个动作输出头串行输出的 n 个目标子动作目标动作。而在训练过程中，获取样本状态信息，并将样本状态信息输入动作决策模块，以训练动作决策模型。

需要说明的是，本申请实施例还可以应用于云技术、人工智能、智慧交通、辅助驾驶等各种场景。上述实施环境中仅作为示意性的例子，并不对本申请实施例的应用场景构成限定。

为了方便表述，下述各个实施例以虚拟场景中虚拟角色的动作决策方法由计算机设备执行为例进行说明。

请参考图 2，其示出了本申请一个示例性实施例提供的虚拟角色的动作决策方法的流程图。本实施例以该方法由计算机设备执行为例进行说明，该方法包括如下步骤：

步骤 201，获取状态信息。

其中，状态信息用于表征第一虚拟角色所处对局的对局状态。

可选的，状态信息可以包括第一虚拟角色在对局中的角色状态信息、第一虚拟角色对所处虚拟环境的环境感知信息，还可以包括其他能够表现游戏场景对局状态的信息。其中，角色状态信息为一维向量，环境感知信息为二维图像。

其中，角色状态信息是指能够反映第一虚拟角色当前状态的信息，可以包括第一虚拟角色当前的属性信息、第一虚拟角色与其他虚拟角色交互产生的交互信息（用于表征第一虚拟角色在对局中与其他虚拟角色之间的交互状态）以及第一虚拟角色与环境交互的交互信息（用于表征第一虚拟角色在对局中与虚拟环境之间的交互状态）。

在游戏场景中，计算机设备可以以游戏帧为单位，在每个游戏帧运行过程中，获取当前状态信息，或者，计算机设备基于一定的周期间隔获取当前状态信息。其中，每秒内包含的

游戏帧数越多，则界面显示越平滑，游戏帧数可以为 36，则计算机设备每秒获取 36 帧状态信息。而在帧数较多的情况下，例如帧数为 72，对于计算机设备性能要求较高，则此时可以设定周期为 3，即每三个游戏帧获取一帧状态信息，每秒获取 24 帧状态信息。

在一种可能的实施方式中，角色状态信息为一维向量，即为向量特征，在游戏场景中角色状态信息可以从游戏引擎接口直接获取，此外一部分角色状态信息也可以通过射线探测的方式获取。

示意性的，请参考表 1，其示出了本申请一个示例性实施例提供的角色状态信息的分类。

表 1

<<属性信息>>	<<交互信息>>
<第一/第二虚拟角色共有>	<第一虚拟角色与第二虚拟角色之间>
血量	是否受到伤害
位置	是否视野内可见
朝向	是否躲避攻击
当前虚拟道具数	欧氏距离
阵营	寻路距离
是否进攻	基于第一虚拟角色朝向×距离相对坐标
是否静步	攻击距离
是否移动	<第一虚拟角色与掩体之间>
是否冲刺	是否在掩体中
是否切换道具	掩体是否安全
是否蹲伏	寻路距离
是否跳跃	欧氏距离
是否侧身	掩体相对坐标
是否趴下	<第一虚拟角色与障碍物之间>
是否架枪	是否存在射线遮挡
是否出局	相对距离

在一些实施例中，角色状态信息中包含多种能够描述当前局势状态的标量信息，可能为第一虚拟角色的状态信息以及第一虚拟角色与环境交互的信息。实质上，环境感知信息是第一虚拟角色对周围虚拟环境感知得到的，刻画了第一虚拟角色相对于虚拟环境的相对特征，因此，只需要在足够丰富的虚拟环境中训练动作决策模型，即可使得动作决策模型获取到的相对特征足够丰富多样，从而保证动作决策模型拥有对环境感知的泛化能力。

然而，角色状态信息中可能会包含一些与特定地图场景具有强关联的特异性特征，因此计算机设备会基于该特征进行进一步的泛化处理。

可选的，计算机设备通过泛化处理角色状态信息中包含的绝对状态信息，得到相对状态信息，相对状态信息是第一虚拟角色相对于第二虚拟角色或者障碍物的状态信息。

具体的，可以采用相对信息取代绝对信息的方式实现泛化处理，例如将第一虚拟角色、第二虚拟角色以及掩体点等绝对位置信息或绝对位置朝向，替换为第一虚拟角色与第二虚拟角色之间、第一虚拟角色与掩体点或障碍物之间的相对位置即位置朝向。

角色状态信息中包含第一虚拟角色与虚拟环境之间的交互信息，该交互信息可以通过射线探测的方式获取。

可选的，以第一虚拟角色所处位置为起始点，向四周发射环境感知射线。环境感知射线在碰撞到障碍物表面的情况下，会沿环境感知射线的法线方向反射。进而计算机设备可以获取环境感知射线的反射情况，从而确定第一虚拟角色与虚拟环境之间交互的角色状态信息。例如表 1 中第一虚拟角色与掩体之间的交互信息以及第一虚拟角色与障碍物之间的交互信息。

示意性的，图 3 示出了本申请一个示例性实施例提供的射线探测方案的示意图，计算机

设备以第一虚拟角色的腰部为起始点，向四周发射环境感知射线，在环境感知射线碰撞到虚拟环境中障碍物的情况下发生反射，从而根据环境感知射线的反射情况，得到第一虚拟角色与障碍物之间的交互信息。

可选的，为更加准确的获取第一虚拟角色与障碍物之间的交互信息，计算机设备以第一虚拟角色所处位置的至少两种高度为起始点，向四周发射环境感知射线。并且根据环境感知射线的反射情况生成第一虚拟角色与障碍物之间的交互信息。

其中，相同高度发射的不同环境感知射线位的发射方向于同一水平高度。例如，以第一虚拟角色所处位置脚部、腰部以及头部作为起始点，向四周发射环境感知射线（即环形射线），从而获取第一虚拟角色与障碍物之间的交互信息。

在一种可能的实施方式中，以不同高度为起始点发射的环境感知射线的数量相同，即每一层环形射线的数量相同。不同高度的环形射线的反射情况能够表征不同高度下虚拟环境中障碍物与第一虚拟角色的距离。每一层中的环形射线数量越多，则对于第一虚拟角色周围障碍物的感知越精细，对于计算机设备的性能需求越高。

可选的，对局中的环境感知信息为第一虚拟角色朝向方向的二维深度图，该二维深度图用于表征第一虚拟角色朝向方向上障碍物的深度情况。

在一种可能的实施方式中，计算机设备基于第一虚拟角色在虚拟环境中所处位置以及第一虚拟角色的朝向，通过射线探测获取环境感知信息。

具体的，以第一虚拟角色为起始点，向第一虚拟角色的朝向方向发射环境感知射线。在环境感知射线碰撞到障碍物表面的情况下，沿环境感知射线的法线方向反射，从而使得计算机设备能够根据环境感知射线的反射情况，生成环境感知信息。

示意性的，请参考图4，其示出了本申请一个示例性实施例提供的二维深度图的示意图，虚拟环境的示意图与二维深度图相对应，其中，每个方格中的数字代表图像中该区域的灰度值，灰度值范围0~255，白色为255，黑色为0，灰度值越小，该图像区域内的像素颜色越深。环境感知信息为第一虚拟角色朝向方向上的二维深度图，深度图中的像素颜色越深，则表明该像素对应的障碍物与第一虚拟角色的距离越近，深度图的像素颜色越浅，则表明该像素对应的障碍物与第一虚拟角色的距离越远。

可选的，二维深度图所表征虚拟环境的精细程度与分辨率呈正相关关系，二维深度图的采样分辨率越大，则二维深度图表征的第一虚拟角色前方障碍物的深度信息越精细。

步骤202，将状态信息输入动作决策模型，得到动作决策模型中n个动作输出头串行输出的n个目标子动作。

其中，不同动作输出头对应不同动作类型，且n个动作输出头基于动作类型之间的依赖关系串行连接，依赖关系用于表征不同动作类型下子动作之间的依赖限制情况，n为正整数。

在动作决策模型中，n个动作输出头属于动作决策模型的输出层，用于预测不同维度下的动作，从而输出不同维度的动作参数。其中，第一虚拟角色支持执行的动作维度可能包括移动状态维度、移动方向维度、朝向维度、进攻维度以及姿态维度等等。动作决策模型中，n个动作输出头分别对应n个动作维度，则n个动作输出头输出n个不同维度的动作参数，例如，在移动方向维度对应的动作输出头输出的动作参数为具体的移动方向，如90°方向；朝向维度对应的动作输出头输出的动作参数为能表征方向的转向角度，如-5°，表征向逆时针方向旋转5°；进攻维度对应的动作输出头输出的动作参数为进攻或者不进攻，等等。本实施例对于具体的动作维度划分以及不同动作维度对应的动作输出头输出的动作参数不进行限定。

可选的，计算机设备通过正交分解第一虚拟角色的可执行动作，得到n个动作类型，不同动作类型中包含至少两个可执行动作，不同动作类型下的子动作可独立控制，即可以控制第一虚拟角色仅执行一个动作类型中的子动作。

其中，正交分解用于将第一虚拟角色的可执行动作划分为多个离散化动作类型，正交分解后得到的n个动作类型互为正交动作类型，不同正交动作类型中的子动作之间不存在交集，

即不存在一个子动作属于两个动作类型的情况，不同正交动作类型支持同时执行，且同一正交动作类型中的不同子动作不支持同时执行。

在游戏场景中，存在控制第一虚拟角色的最基本、最细粒的原子动作，这类动作不可进一步进行拆分细化。在一个动作类型中的子动作即为原子动作，在同一动作类型中的不同子动作之间不能够同时执行。例如，第一虚拟角色可执行动作中包括匍匐前进这一个动作，则可将这个动作划为两个原子动作分别为匍匐以及前进，其中匍匐对应姿态动作类型，而前进对应移动方向动作类型。再例如，第一虚拟角色可执行动作中包括原地探头射击这一动作，则可将这个动作划分为停留在原地、探头以及射击这三个原子动作，其中停留在原地对应移动状态动作类型，探头对应探头动作类型，射击对应进攻动作类型。

示意性的，请参考图 5，其示出了本申请一个示例性提供的提供虚拟环境的应用程序的用户界面示意图。其中包含移动状态控件 501、进攻控件 502、姿态调整控件 503、水平转向控件以及竖直转向控件 504、左右探头控件 505 以及移动状态控件 506，不同控件可触发控制虚拟角色执行不同动作。

对上述图中示出的各个控件对应的动作正交分解，即可以得到多种动作类型。

在一些实施例中，不同动作类型之间可能存在依赖关系，例如，虚拟角色在执行攻击操作时才会站立或侧身，则第一虚拟角色的姿态动作决策需要依赖于进攻动作类型的动作决策。并且，依赖限制情况指示不同动作类型中的子动作之间可能存在依赖限制关系，例如，在第一虚拟角色执行冲刺动作时不支持同时执行进攻动作，则冲刺动作与进攻动作之间存在依赖限制关系。因此，为使动作决策模型在动作决策时能够考虑到不同动作之间的依赖关系，以及不同动作类型下不同子动作之间的依赖限制情况，计算机设备将动作决策层中的 n 个动作输出头基于动作类型之间的依赖关系串行连接。

可选的，在确定不同动作输出头的连接顺序时，计算机设备可以根据不同动作类型对于第一虚拟角色所要完成任务的关联程度，对动作输出头的连接顺序排序。例如，在第一虚拟角色的任务为击败敌人，则可以确定该任务与进攻动作类型的关联程度较高，则可以将其对应的动作输出头排列在靠前的位置；而该任务与侧身动作类型的关联程度显然较低，则可以将侧身动作类型对应的动作输出头排列在最后的位置。

步骤 203，控制第一虚拟角色执行由 n 个目标子动作构成的目标动作。

在一些实施例中，动作决策模型部署于第二服务器中，则在通过动作决策模型确定目标动作的情况下，第二服务器将动作执行指令发送至用于支持虚拟环境的应用程序提供后台服务的第一服务器，再由第一服务器将该动作执行指令发送至客户端以实现第一虚拟角色执行动作的控制。

在一种可能的实施方式中，计算机设备对不同动作类型中的各个子动作名称分别编码，得到多个候选动作标签，计算机设备基于动作决策模型中的动作决策层得到各动作类型下各子动作的概率分布，基于动作执行概率从子动作标签中确定目标子动作对应标签，在确定目标子动作标签的情况下，计算机设备对目标子动作标签进行解码，得到目标子动作，并控制第一虚拟角色执行由多个目标子动作构成的目标动作。

综上所述，本申请实施例中，通过动作决策模型基于状态信息实现动作决策，从而能够根据当前状态，确定出合理的目标动作。其中，动作决策模型中包含 n 个动作输出头串行连接， n 个动作输出头分别用于输出不同的目标子动作， n 个动作输出头基于动作类型之间的依赖关系串行连接，能够在动作决策时考虑到不同动作之间的依赖关系，并在串行输出 n 个目标子动作的过程中，使得 n 个目标子动作在时序上前后因果关联，从而提高目标动作的合理性，提升了虚拟角色的拟人性。

本申请实施例中，动作决策模型包括信息处理层、特征提取层以及动作决策层，动作决策层包括 n 个动作输出头。下面将通过一个示例性实施例对确定 n 个目标子动作的过程进行说明。

请参考图 6，其示出了本申请一个示例性实施例提供的确定目标动作的过程的流程图。

步骤 601，获取状态信息。

本行步骤的实施方式可以参照上述步骤 201，本实施例对此不做赘述。

步骤 602，将状态信息输入动作决策模型，通过信息处理层对状态信息编码，得到状态编码。

可选的，状态信息既包括第一虚拟角色在对局中的角色状态信息，还包括第一虚拟角色所处虚拟环境的环境感知信息。其中，角色状态信息用于表征第一虚拟角色在对局中与第二虚拟角色之间的交互状态，以及第一虚拟角色在对局中与虚拟环境之间的交互状态，角色状态信息为一维向量，环境感知信息为二维图像。在一些实施例中，计算机设备通过信息处理层采用不同编码方式对状态信息中包含的角色状态信息以及环境感知信息编码。

可选的，信息处理层中包括标量编码器以及图像编码器。计算机设备将状态信息输入动作决策模型，通过信息处理层中的标量编码器对角色状态信息编码，得到状态信息编码结果。再通过信息处理层中的图像编码器对环境感知信息编码，得到环境信息编码结果。最后通过信息处理层拼接状态信息编码结果以及环境信息编码结果，得到状态编码。

可选的，标量编码器中包含一维卷积核。计算机设备通过一维卷积核对角色状态信息中的属性信息以及交互信息执行一维卷积处理，得到状态信息编码结果。

可选的，图像编码器中包含二维卷积核。计算机设备通过二维卷积核对状态信息中的环境感知信息编码，得到环境信息编码结果。二维卷积核是一个二维矩阵，在通过二维卷积核对二维深度图编码的过程中，通过将该二维矩阵和输入的二维深度图逐元素相乘并求和，即完成对环境感知信息的特征编码。

在得到状态信息编码结果以及环境信息编码结果之后，计算机设备通过全连接层以及激活函数拼接状态信息编码结果以及感知信息编码结果，即可以得到状态编码。其中，全连接层用于将前层卷积计算得到的特征空间映射到样本标记空间（即将特征表示整合为一个值），能够减少特征位置对于分类结果的影响，提高动作决策模型的鲁棒性。激活函数用于加入非线性因素，从而提高动作决策模型的表达能力。

步骤 603，将状态编码输入特征提取层，通过特征提取层提取状态编码特征，得到融合特征。

在一种可能的实施方式中信息处理层中还包括 LSTM（Long Short Term Memory，长短期记忆）网络。计算机设备通过将状态编码输入 LSTM 网络，通过特征提取，以获取到融合特征。在特征提取的过程中，长短期记忆网络会保留具有更重要价值的特征，并将价值较低的特征遗忘掉，从而实现特征提取。

可选的，该 LSTM 网络可以通过其他卷积神经网络进行替换，本实施例对此不进行限定。

步骤 604，基于融合特征，通过动作决策层中的 n 个动作输出头确定 n 个目标子动作。

在一些实施例中，计算机设备将融合特征输入动作决策层，并且由于动作决策层中的 n 个动作输出头为串行连接，因此将融合特征输入动作决策层即为将融合特征首先输入动作决策层的第一动作输出头，通过动作决策层中第一动作输出头从第一动作类型中的子动作中确定第一目标子动作。

可选的，动作决策模型中还包括信息过滤层，信息过滤层与标量编码器的输出端相连，且信息过滤层与动作决策层的输入端相连。计算机设备在获取到角色状态信息的情况下，通过信息过滤层过滤角色状态信息，得到过滤后的角色状态信息，其中，过滤后的角色状态信息与第一动作输出头对应的第一动作类型具有相关关系。进而，计算机设备再将过滤后的角色状态信息输入第一动作输出头。并且，由于第一动作输出头与所要完成的任务关联程度最高，则将与第一动作类型相关度最高的角色状态信息与融合特征一同输入动作决策层，有利于增强第一动作输出头的合理决策能力。

示意性的，表 2 示出了本申请一个示例性实施例提供的正交分解后动作输出头与动作类型之间的对应关系。

表 2

动作输出头名称	动作维度 (包含子动作的数目)	动作类型
第一动作输出头	2	是否进攻
第二动作输出头	4	姿态
第三动作输出头	8	移动方向
第四动作输出头	8	水平朝向
第五动作输出头	8	竖直朝向
第六动作输出头	3	是否探头
第七动作输出头	4	移动姿态

在一种可能的实施方式中，由于不同动作输出头之间串行连接，因此，计算机设备在将融合特征输入动作决策层中的第一动作输出头，通过第一动作输出头从第一动作类型中确定第一目标子动作之后，还需要将已确定的第一目标子动作至第 $i-1$ 目标子动作分别对应的嵌入编码向量，以及融合特征输入动作决策层，通过动作决策层中第 i 动作输出头从第 i 动作类型中确定第 i 目标子动作。

其中，第一目标子动作至第 $i-1$ 目标子动作分别由第一动作输出头至第 $i-1$ 动作输出头确定， i 小于或等于 n ，且 i 大于 1。

可选的， n 个动作输出头通过自回归嵌入层连接。由于不同动作类型之间存在依赖关系，则在前序输出头确定目标子动作后，应当基于自回归嵌入层将已确定的至少一个目标子动作以及融合特征输入到后序的动作输出头中，以使得后序动作输出头根据依赖关系确定目标子动作。

其中，自回归是指在统计上处理时间顺序的方法，用同一变量之前各个时刻预测当前时刻该变量的表现，自回归具有时间序列的自相关性。嵌入 (embedding) 的本质是数据压缩，用较低维度的特征表述具有冗余信息的较高维度的特征。

请参考图 7，其示出了本申请一个示例性实施例提供的 n 个动作输出头的自回归结构的示意图。在一个动作输出头中包含全连接层、Logits 层、采样层以及嵌入层。Logits 层即为确定的动作类型中各子动作之间的执行概率分布，从而确定出目标子动作，在确定目标子动作后，通过嵌入层将以确定的子动作传输至下一串行连接的动作输出头，由该动作输出头融合前序已经确定的各个目标子动作，从而根据已确定的目标子动作确定新的目标子动作。此外，信息处理层也会将融合特征输入至每个动作输出头。

本申请实施例中，通过正交分解离散化动作控件以降低动作控件维度，并且基于动作决策模型中的 n 个动作输出头串行确定 n 个目标子动作，能够通过正交分解离散化动作控件，有效解耦结构化动作空间。并且串行动作输出头结合，能够提升动作决策模型的合理性，且在训练模型的过程中增加训练效率。

在一些实施例中，在通过 n 个动作输出头串行输出 n 个目标子动作的过程中，考虑到不同动作类型下的各个子动作之间存在依赖限制关系，因此为了屏蔽不同动作类型中不合理的子动作，或者与已确定目标子动作之间存在依赖限制关系的子动作，计算机设备还可以在动作决策模型中设置动作遮罩层，通过动作遮罩层遮罩不同动作类型中的子动作。

本质上，动作遮罩是用于屏蔽掉不同动作类型中不合理或者与已确定目标子动作之间存在依赖限制关系的子动作。通过动作遮罩，降低被遮罩的子动作在动作输出头得到的概率分布中的概率，从而在依照概率采样确定目标子动作时，被遮罩的子动作难以被确定为目标子动作。通过动作遮罩的方式，能够在动作决策过程中直接排除掉不必要探索的子动作分支，从而减小动作空间，在提升动作决策模型探索效率的同时保证动作决策模型的输出准确性。

其中，存在依赖限制关系的至少两个子动作属于不同的动作类型，且存在依赖限制关系的至少两个子动作不能同时执行。可选的，存在第一子动作与第二子动作之间存在依赖限制

关系, 则这两个子动作不能同时执行, 因此在已经确定第一子动作为目标子动作的情况下, 需要对第二子动作执行动作遮罩。可选的, 存在第二子动作结合第三子动作后, 与第四子动作之间存在依赖限制关系, 则第二子动作与第三子动作的结合, 和第四子动作不能同时执行, 因此在已经确定第二子动作以及第三子动作均为目标子动作的情况下, 需要对第四子动作执行动作遮罩。

相关技术中, 在将动作遮罩应用于并行动作输出头的过程中, 各个动作输出头并行输出子动作, 动作输出头之间并行连接, 每个动作输出头会输出不同的子动作指令, 对于被执行动作遮罩的动作输出头, 其输出动作被选择的概率趋近于 0。

请参考图 8, 其示出了相关技术中一种对并行动作输出头执行动作遮罩的示意图, 其中, 在动作遮罩时, 在网络原始 logits 动作输出层, 通过对不合理动作的 logits 加上一个绝对值很大的负数, 保证该动作对应的 logits 值比所有合理的动作的 logits 值小, 进而重塑了输出动作层的 logits 分布, 将不合理动作被采样的概率映射接近于零。图中 a_1 至 a_{n+1} 表征不同动作的 logits 分布, P_1 至 P_{n+1} 表征各动作对应采样概率。对应图中 $i=3$ 的情况下, 对 a_3 对应的动作进行动作遮罩, 加上一个绝对值很大的负数, 得到新的 logits 分布, 再次依照概率进行采样时, 则难以采样到 a_3 对应的动作。

可选的, 由于不同动作类型下的子动作存在依赖限制关系, 存在依赖限制关系的至少两个子动作之间不支持同时执行, 则在存在已确定的前序的目标子动作的情况下, 可能在后续动作类型中存在部分子动作与已确定的目标子动作之间具有依赖限制关系。并且, 由于正交分解后的动作空间维度依旧很高, 会导致丰富的拟人化原子动作空间组合爆炸, 因此, 可以根据不同子动作之间组合的合理性(即依赖限制关系)确定子动作, 提升动作组合的拟人性。此外, 在对动作决策模型训练的过程中, 也可以采用动作遮罩的方式减少动作决策层无效的决策策略探索, 加速训练效率的同时提升动作决策模型动作决策的拟人性。

动作遮罩可以用于强化学习模型(对应本申请实施例中的动作决策模型), 可以在大规模决策空间中, 屏蔽掉不合理或者无效的动作集, 从而减少无意义的动作探索, 能够使动作决策层更快收敛。

可选的, 动作决策模型中包含动作遮罩层, 动作遮罩层与动作决策层中 n 个动作输出头相连接, 用于对各个输出头动作决策过程中各个动作类型中的子动作执行动作遮罩。

在一种可能的实施方式中, 计算机设备基于已确定的第一目标子动作至第 $i-1$ 目标子动作, 以及不同动作类型下子动作之间的依赖限制情况所指示的依赖限制关系, 对第 i 动作类型中的子动作执行动作遮罩, 其中, 具有依赖限制关系的至少两个子动作不支持同时执行。从而计算机设备通过动作决策层中第 i 动作输出头, 从第 i 动作类型中未被遮罩的子动作中确定第 i 目标子动作。

可选的, 在动作遮罩之前, 计算机设备首先基于已确定的目标子动作, 在待决策动作类型中, 确定与已确定目标子动作具有依赖限制关系的待遮罩子动作。具体可包括以下三种情况:

一、存在一个已确定的目标子动作与待决策动作类型中待遮罩子动作存在依赖限制关系。

在一种可能的实施方式中, 在存在已确定的第 j 目标子动作, 且依赖限制情况指示第 j 目标子动作与第 i 动作类型中的至少一个第一子动作存依赖限制关系的情况下, 计算机设备对第 i 动作类型中的第一子动作执行动作遮罩, j 小于 i (即第 j 动作头串联在第 i 动作头之前)。

例如, 存在已确定的冲刺动作, 且依赖限制情况指示冲刺动作与下蹲动作存在依赖限制关系, 则在确定第一虚拟角色姿态这一动作类型的情况下, 计算机设备可以对姿态动作类型中的下蹲动作执行动作遮罩。

可选的, 计算机设备可以对前序动作输出头所确定的目标子动作执行独热处理, 随后根据用于表征后序动作类型中各个子动作与前序动作类型中的子动作的依赖限制关系的限制矩阵, 与前序目标子动作对应的独热矩阵, 通过矩阵乘法, 以获取后序动作类型对应的动作遮罩矩阵。

具体方式如下，计算机设备基于动作遮罩矩阵将前序动作输出头确定的动作传输至后序动作输出头。

示意性的，存在动作类型 a1 为发起进攻（包括进攻和不进攻）、动作类型 a2 为虚拟角色姿态（包括下蹲、跳跃、匍匐以及站立）以及动作类型 a7 为虚拟角色的移动状态（包括静步、快走、冲刺以及原地不动），动作类型 a1 对应动作输出头 H1、动作类型 a2 对应动作输出头 H2，动作类型 a7 对应动作输出头 H7。其中，进攻与冲刺存在依赖限制关系，且下蹲也与冲刺存在依赖限制关系，并且由于子动作的值在神经网络中以张量的形式传输，因此不能够直接采样目标子动作来避免梯度断开。

可选的，各个动作输出头对应的动作遮罩矩阵可以在动作决策模型的训练过程中确定。以动作输出头 H1、H2 以及 H7 为例，在存在三个输出的样本目标动作的情况下，计算机设备首先基于三个样本目标动作，构建独热（one-hot）矩阵 $\tau_{n \times m}$ ，其中， $n=3$ ，表示存在三个样本目标动作， $m=2$ ，表示 a1 的动作维度（即包含的子动作数量）。

可选的，在获取到的三个样本目标动作在动作输出头 H1 的输出样本目标子动作为{进攻，不进攻，进攻}的情况下，其对应的 $n \times m$ 的 one-hot 矩阵 $\tau_{n \times m}$ 可以表示为：

$$\tau_{n \times m} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}_{3 \times 2}$$

$\tau_{n \times m}$ 是样本目标动作中动作输出头 H1 输出的独热矩阵，行数为样本目标动作的个数 n ，列数代表动作类型 a2 中所包含的子动作数量 m 。

随后，计算机设备构造辅助映射矩阵 M ，即为构造能够表征动作类型 a1 中子动作与动作类型 a7 的子动作之间的依赖限制关系的映射矩阵。由于只是基于单个动作输出头 H1 输出的目标子动作采样从而构造与动作类型 a7 之间的映射关系，因此可以将动作输出头 H1 所有不同动作组合成的独热矩阵确定为一个 $m \times m$ 的单位矩阵，第一行映射目标子动作进攻，第二行映射目标子动作不进攻。

$$E = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}_{2 \times 2}$$

由于动作输出头 H7 对应维度 $P=4$ ，则为了体现出进攻的同时不冲刺，则可以得到辅助映射矩阵 $M_{m \times p}$ 为：

$$M = E \times \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 1 \end{bmatrix}_{2 \times 4}$$

$M_{m \times p}$ 矩阵的本质为由 H1 到 H7 的映射，行数代表前序动作头 H1 的维度 m ，列数代表后序被影响的动作头 H7 的维度 P ，每一行都与单位矩阵 $E_{m \times m}$ 中的每一行对应关联，每行的具体值是指前序动作输出头 H1 中每个子动作对后序输出动作头 H7 的影响。

最后，计算机设备再通过矩阵相乘，将 $\tau_{n \times m}$ 中发起进攻时不能冲刺的影响映射到动作输出头 H7，得到动作输出头 H7 对应的动作遮罩矩阵 X_7 为：

$$X_7 = \begin{bmatrix} 0 & 1 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}_{3 \times 2} \times \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 1 \end{bmatrix}_{2 \times 4} = \begin{bmatrix} 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 1 \end{bmatrix}_{3 \times 4}$$

动作遮罩矩阵 X_7 用于表征所有样本目标动作中前序动作输出头 H1 输出的样本目标子动作对后序动作输出头 H7 对应的动作类型中子动作的影响，可以看出， X_7 尺寸只和样本个数和动作输出头 H7 维度有关，行数代表采样了 3 个样本，列数代表动作类型 a7 中不同的子动作，上述过程已经将所有样本目标动作中确定的目标子动作为进攻时对应的冲刺子动作屏蔽掉，再对动作输出头 H7 确定的动作类型 a7 中子动作的概率分布执行概率分布转换，即可保证动作输出头 H7 不会采样冲刺作。

同样的，计算机设备也使用上述处理方式，得到动作类型 a2 中目标子动作对动作类型 a7 的动作遮罩矩阵 Y_7 ，由于动作类型 a1 和动作类型 a2 均独立影响动作类型 a7，则动作输出头 H7 叠加处理前序动作输出的目标子动作对动作类型 a7 中子动作的影响，从而得到动作输出

头 H7 对应的动作遮罩矩阵 L_7 为:

$$L_7 = X_7 \circ Y_7$$

L_7 矩阵是 X_7 和 Y_7 的哈达玛积, L_7 中所有值为 0 的元素所映射的动作类型 a7 中的子动作均会被遮罩。

基于上述方式, 通过将前序确定的目标子动作对于后序动作输出头对应的动作类型中子动作的影响映射到动作遮罩矩阵, 从而将该影响在 n 个动作输出头中依次向后传递, 能够在动作决策模型的训练过程中有效解决在高维度的动作空间下, 决策模型进行大量无效探索的问题。

二、存在至少两个已确定的目标子动作进行组合后, 与待决策动作类型中的待遮罩子动作存在依赖限制关系。

在另一种可能的实施方式中, 可能存在已确定的目标子动作, 这两个目标子动作对于后序动作类型中的子动作不产生独立影响, 在至少两个已确定的目标子动作进行结合后, 会与待决策的动作类型中的子动作存在依赖限制关系。

示意性的, 请参考图 9, 其示出了本申请一个示例性实施例提供的移动方向与朝向方向叠加影响移动状态的示意图。存在两个动作类型为移动方向以及水平转向角度, 这两个动作类型中子动作均不会单独对待确定的动作类型中的冲刺动作产生影响, 然而在确定出的目标子动作中水平转向角度与移动方向之间差距较大的情况下 (极端情况为 180°), 如图中, 出现第一虚拟角色注视方向与移动方向相差 180° , 由于在常规的设定中不支持第一虚拟角色倒退冲刺, 则此时第一虚拟角色不支持同时执行冲刺动作。

在这种情况下, 计算机设备可以将至少两个动作输出头输出的目标子动作结合, 从而确定至少两个目标子动作结合后对待决策的动作类型中子动作的影响, 并将该影响映射为待决策动作类型对应的动作遮罩矩阵。

具体的, 计算机设备可以先将至少两个动作输出头输出的目标子动作各自独热矩阵编码, 再将至少两个独热矩阵拼接, 形成合并独热矩阵, 再对合并独热矩阵经过逻辑转换, 使其映射到一个单位矩阵上。随后, 基于上述方式一中所示出的方式, 确定辅助映射矩阵, 并基于辅助映射矩阵, 确定至少两个目标子动作结合后对待决策动作类型的影响, 从而实现对待决策动作类型中子动作的动作遮罩。

示意性的, 存在动作类型 a3 为移动方向 (包括在 360° 内均匀分布的八种移动方向)、动作类型 a4 为水平朝向 (包括在 360° 内均匀分布的八种朝向角度) 以及动作类型 a7 为虚拟角色的移动状态 (包括静步、快走、冲刺以及原地不动), 动作类型 a3 对应动作输出头 H3、动作类型 a4 对应动作输出头 H4, 动作类型 a7 对应动作输出头 H7。

其中, 动作类型 a3 对应动作维度为 s, 动作类型 a4 对应动作维度为 h, 动作输出头 H3 所有不同采样子动作组合成独热矩阵为 $s \times s$ 的单位矩阵 $E_{s \times s}$, 动作输出头 H4 所有不同采样动作组合成的 one-hot 矩阵是一个 $h \times h$ 的单位矩阵 $E_{h \times h}$ 。计算机设备通过对单位矩阵 $E_{s \times s}$ 以及单位矩阵 $E_{h \times h}$ 特征拼接, 可以得到的合并独热矩阵 $T_{(s \times h) \times (s+h)}$ 并不是一个方阵, 行数为 $s \times h$ 代表所有动作组合情况, 列数是 $s+h$ 代表结合的维度, 即动作类型 a3 以及动作类型 a4 的子动作数量之和。随后, $T_{(s \times h) \times (s+h)}$ 将转换为一个维度是 $s \times h$ 的单位矩阵 $E_{(s \times h) \times (s \times h)}$, 则该单位矩阵的每一行代表一种动作结合的情况, 随后可以根据依赖限制关系确定映射辅助矩阵 M。

可选的, 假设所有样本目标动作中动作输出头 H3 和动作输出头 H4 输出的样本目标子动作对应的合并独热矩阵为 $\tau_{n \times (s+h)}$, 则通过以下计算可以确定动作输出头 H7 对应的动作遮罩矩阵:

首先计算样本目标动作中每行动作组合转化为近似独热矩阵的映射为:

$$U_{n \times (s \times h)} = \tau_{n \times (s+h)} \times X_{(s+h) \times (s \times h)}$$

其中, $X_{(s+h) \times (s \times h)}$ 为 $T_{(s \times h) \times (s+h)}$ 的广义逆矩阵, $U_{n \times (s \times h)}$ 为样本目标动作中每行动作组合转化为近似独热矩阵的映射, 由于 $X_{(s+h) \times (s \times h)}$ 不是 $T_{(s \times h) \times (s+h)}$ 精确的逆矩阵, 则需要 $U_{n \times (s \times h)}$ 转换为精确地独热矩阵:

$$Z_{n \times (s \times h)} = \text{one-hot}(\text{argmax}(U_{n \times (s \times h)}))$$

其中 $Z_{n \times (s \times h)}$ 的每一行代表一个样本的动作组合，每行存在 $s \times h$ 个矩阵值，只存在一个元素值为1，其余元素值均为0，元素值为1表征动作组合将会与映射辅助矩阵M对应行关联映射，从而影响动作输出头H7对应的动作遮罩矩阵对应行的矩阵值。 $F_{n \times p}$ 为动作输出头H7中的动作遮罩矩阵，每行表征动作类型a3以及动作类型a4会对动作类型a7中子动作产生的影响。

通过上述计算，能够在不同子动作之间的动作组合对待决策动作类型中子动作存在依赖限制关系的情况下，实现动作遮罩，从而解决强耦合高维度复杂动作空间下动作决策层探索效率低的问题。

三、角色状态信息以及环境感知信息指示当前虚拟环境的情况，与待决策动作类型中的待遮罩子动作存在依赖限制关系，即在当前虚拟环境下不适合执行待遮罩子动作。

在一种可能的实施方式中，第一虚拟角色可能具有攻击第二虚拟角色，或者躲避第二虚拟角色等任务目标，因此在动作决策时，计算机设备可能会基于第一虚拟角色对第二虚拟角色的可视情况对待决策动作类型中的子动作动作遮罩。

具体的，计算机设备以第一虚拟角色的眼部位置为起始点，向第一虚拟角色朝向方向上的第二虚拟角色的各个部位发射环境感知射线。根据环境感知射线的反射情况，确定虚拟环境中第二虚拟角色的各个部位的可视情况，可视情况用于表征第二虚拟角色各个部位是否被掩体遮挡。在环境感知射线被障碍物遮挡的情况下，则表示第二虚拟角色的该部位不可视；在环境感知射线未被障碍物遮挡，且碰撞到第二虚拟角色的情况下，则表示第二虚拟角色该部位可视。

请参考图10，其示出了本申请一个示例性实施例提供的可视情况感知的示意图。计算机设备从第一虚拟角色1001的眼部为起始点，向第二虚拟角色1002的各个部位发射环境感知射线，其中存在部分感知射线被障碍物1003遮挡，则这部分射线对应的第二虚拟角色1002的身体部位不可视，而存在感知射线1004未被障碍物遮挡，与第二虚拟角色1002的头部发生碰撞，则第二虚拟角色1002的头部可视。

在确定第二虚拟角色各个部位的可视情况后，计算机设备基于第二虚拟角色的各个部位的可视情况，对目标动作类型中的子动作动作遮罩。

其中，目标子动作与第二虚拟角色的各个部位的可视情况具有依赖限制关系。例如，目标动作类型为射击动作的情况下，若第二虚拟角色的各个部位的可视情况指示该第二虚拟角色不可视，则将目标动作中的“射击”子动作进行动作遮罩，能够避免第一虚拟角色在未瞄准敌人的情况下进行频繁射击。

在对目标动作类型中子动作进行遮罩后，计算机设备通过目标动作输出头，从目标动作类型中未被遮罩的子动作中确定子动作。

在另一种可能的实施方式中，在第一虚拟角色具有瞄准第二虚拟角色的需求的情况下，计算机设备可以先确定能够瞄准到第二虚拟角色的转向角度范围，再对转向动作类型中无法使第一虚拟角色瞄准第二虚拟角色的转向角度进行动作遮罩。

请参考图11，其示出了本申请一个示例性实施例提供的确定竖直方向有效瞄准范围的示意图。其中首先根据第二虚拟角色所处位置坐标以及第一虚拟角色所处位置坐标确定第一虚拟角色与第二虚拟角色之间的距离d，并且获取第二虚拟角色的高度h， $h = h_z - h_f$ 。则第一虚拟角色所持虚拟道具的准心在竖直方向上落在第一虚拟角色身高范围内的情况下，即可瞄准第二虚拟角色，则根据三角函数可以得到：

$$\begin{aligned}\theta_{\max} &= \arctan[(h_z - h_f)/d] \\ \theta_{\min} &= \arctan[(h_f - h_z)/d] \\ \tau_{pitch} &= \theta_{\max} - \theta_{\min}\end{aligned}$$

其中，以水平方向为当前朝向方向即为 0° ， $\theta_{\max}$ 为向上转动的有效瞄准范围， $\theta_{\min}$ 为向下转动的有效瞄准范围，则二者之差 τ_{pitch} 即为数值转向动作的有效瞄准范围。

对于水平转向角度，计算机设备可以根据第一虚拟角色与第二虚拟角色之间的距离 d ，以及虚拟角色的宽度确定有效瞄准范围，从而对水平转向动作类型中的子动作进行遮罩，使第一虚拟角色执行未被遮罩的子动作中的目标子动作后能够瞄准第二虚拟角色。

示意性的，请参考图 12，其示出了本申请一个示例性实施例提供的水平方向有效瞄准范围的示意图。第一虚拟角色当前朝向为 0° ，第一虚拟角色与第二虚拟角色之间的距离为 d ，第二虚拟角色的宽度为 W ，则确定有效瞄准范围为：

$$\theta = \arctan[(W/2)/d]$$

$$\tau_{pitch} = 2\theta$$

其中， θ 为第一虚拟角色向顺时针和逆时针旋转角度中一个有效旋转范围，将顺时针转动有效范围与逆时针转动有效范围进行结合，从而得到有效瞄准范围 τ_{pitch} 。

随后，计算机设备基于有效瞄准范围确定有效水平转向角度，且确定出的有效水平转向角度为第一虚拟角色在当前朝向下按照细粒度转向角度执行转向动作后的朝向在有效瞄准范围内。

进而，计算机设备再对水平转向动作中不属于有效水平转向角度的转向角度动作遮罩，并通过水平转向动作输出头，从水平转向动作中未被遮罩的有效水平转向角度中确定目标水平转向角度，最终使得第一虚拟角色执行目标水平转向动作从而瞄准第二虚拟角色。

在一种可能的实施方式中，为避免目标动作频繁做出某个动作，不符合人类行为逻辑，计算机设备可以为部分子动作设置冷却时间，在该子动作冷却未完成的情况下，对该子动作动作遮罩。从而避免出现第一虚拟角色反复执行同一动作造成抽搐现象。

本申请实施例中，基于三种不同的依赖限制情况以及已确定的目标子动作，对待决策的动作类型中的子动作动作遮罩，从而提升高维度的复杂动作空间下的动作决策效率在一定程度上节约计算资源。

请参考图 13，其示出了本申请一个示例性实施例提供的动作决策模型的示意图。其中包括动作遮罩层、信息处理层、特征提取层以及动作决策层。计算机设备将获取到的第一虚拟角色所处对局的状态信息（角色状态信息以及环境感知信息）通过标量编码器以及图像编码器进行编码，再经过全连接层以及嵌入层进行特征拼接后得到输入特征提取层（LSTM 网络）。通过 LSTM 网络对拼接后的状态特征进行特征提取得到融合特征。并且将融合特征输入动作决策层后，通过动作决策层确定目标动作，目标动作由多个目标子动作组成， n 个目标子动作分别通过 n 个动作输出头串行输出，并且在确定目标子动作的过程中通过动作遮罩层对存在依赖限制关系的子动作执行动作遮罩。

此外，在模型训练过程中动作决策模型中还包括价值网络，通过价值网络为确定的目标子动作进行价值评估，并根据评估得到的价值估计调整动作决策模型。此外，还可以将部分对局信息输入价值网络以使得价值网络能够根据当前对局情况判断样本目标子动作的价值。

在应用动作决策模型之前，应当在训练对局中训练该动作决策模型，下面将通过一个示例性实施例对动作决策模型的训练过程进行说明。

请参考图 14，其示出了本申请一个示例性实施例提供的动作决策模型训练过程的流程图，该过程包括以下步骤：

步骤 1401，获取样本状态信息，样本状态信息用于表征第一虚拟角色所处训练对局的对局状态。

本步骤可参照上述步骤 201 中获取状态信息的过程，在此不进行赘述。

步骤 1402，基于样本状态信息，通过强化学习的方式训练动作决策模型。

基于强化学习训练动作决策模型的过程包括如下内容。

步骤 1402a，将样本状态信息输入动作决策模型，得到动作决策模型输出的样本目标动作，样本目标动作由 n 个样本目标子动作构成。

实际上，对动作决策模型训练的过程即为对动作模型中动作决策层强化训练的过程。

此过程的具体实施方式可参照上述实施例确定目标动作的内容，本实施在此不做赘述。

步骤 1402b，控制第一虚拟角色执行样本目标动作，得到第一虚拟角色的样本动作执行结果。

该样本动作执行结果包括第一虚拟角色执行样本目标动作之后对所处对局的状态信息的变化。

步骤 1402c，基于第一虚拟角色的样本动作执行结果，确定动作执行奖励。

在训练的过程中，通常会为第一虚拟角色设定相应的动作执行任务，因此，为避免第一虚拟角色为完成目标任务而采取极致的操作，导致第一虚拟角色拟人化程度较低，在基于任务进度确定目标动作的执行奖励外，计算机设备还可以基于样本目标动作的拟人化程度判定样本目标动作的执行奖励。

首先，计算机设备确定执行各项奖惩项目的权重系数比例，其中，动作执行奖励可以分为主导奖励（即针对于目标任务进度的奖励）以及辅助拟人化奖励。在动作决策模型的训练初期，计算机设备将主导奖励的权重调高，辅助拟人化奖励的权重调低，以使得动作决策模型优先以完成任务为目标学习动作决策。在经过多轮训练后，动作决策模型的强度不断提升，从而计算机设备需要根据动作决策模型的表现和对应拟人化指标变化趋势动作调整辅助拟人化奖励的系数占比。其中，拟人化指标是指第一虚拟角色执行特定动作的概率与真实玩家执行特征动作的概率的匹配程度。例如，在第一虚拟角色在对局中无法合理使用掩体执行拉扯战斗的情况下，则可以认为主导奖励的权重过大，从而计算机设备需要降低主导奖励的权重，以保证动作决策模型在拥有一定强度的基础上能够确定出拟人化的目标动作。

可选的，信息处理层中还包括对局信息编码器，样本状态信息中还包括样本对局信息，样本对局信息用于表征第一虚拟角色所处的训练对局的实时情况。在一种可能的实施方式中，在将样本对局状态信息输入动作决策模型的情况下，计算机设备首先通过对局信息编码器对样本对局信息编码，得到对局状态编码，随后基于对局状态编码以及样本动作执行结果，确定动作执行奖励。通过这种方式，能够使动作决策模型根据全局的对局信息客观的判断当前目标动作带来的奖励，从而降低值估计的方差，从而提高训练效率。

本申请实施例中，动作执行奖励主要包括属性奖励、对局胜利奖励、对局失败奖励以及任务奖励中的至少一项。

属性奖励主要包括第一虚拟角色的生命值属性奖励，在第一虚拟角色的属性值降低的情况下，计算机设备可以将属性奖励确定为动作执行奖励。例如，在第一虚拟角色执行样本目标动作之后，第一虚拟角色的生命值属性降低，则确定该目标动作获得反向的属性奖励。或者，在第一虚拟角色的属性值提升的情况下，计算机设备可以将属性奖励确定为动作执行奖励。例如，在第一虚拟角色执行样本目标动作之后，第一虚拟角色的生命值属性提升，则确定该目标动作获得正向的属性奖励。

对局失败奖励是指在第一虚拟角色取得对局失败的情况下，会获取极大的反向奖励（即惩罚），则在第一虚拟角色取得对局失败的情况下，计算机设备可以将对局失败奖励确定为动作执行奖励。

对局成功奖励与对局失败奖励相反，在第一虚拟角色取得对局成功的情况下，会获得极大的正向奖励，则在第一虚拟角色的属性值降低的情况下，计算机设备可以将属性奖励确定为动作执行奖励。

在一些实施例中，第一虚拟角色在对局中存在目标任务，例如守护宝箱、躲避追击或者抵达指定地点等等，因此，在第一虚拟角色在训练对局中任务成功的情况下，计算机设备可以将任务奖励确定为动作执行奖励。

此外，为使动作决策模型能够实现拟人化的动作决策，动作执行奖励中还包括拟人化奖励。

可选的，计算机设备基于第一虚拟角色在训练对局中执行至少两次样本目标动作的样本

动作执行结果，确定至少两次样本目标动作的拟人化属性值。该拟人属性值可能为执行特定动作的次数、连续确定的至少两个样本目标动作的走位位置等等。在拟人化属性值低于拟人属性阈值的情况下，计算机设备可以将拟人化奖励确定为动作执行奖励。

示意性的，请参考表 3，其示出了本申请一个示例性实施例提供的动作执行奖励。

表 3

类别	子分类	奖惩设置
主导奖励	任务目标奖励	属性变化
		对局胜负
		任务是否成功
辅助拟人化奖励	姿态拟人化奖励	合理侧身
		合理蹲起
		合理静步
		大幅度转向惩罚
		走位分散
	战斗拟人化奖励	躲避三秒再进行进攻占比
		躲避 2.5 秒再进行进攻占比
		掩体利用
		危险区域停留
		射击后撤
		暗处伏击
		远程攻击
		对手靠近时进攻
		危险时刻探头
		攻击占比

在上表中，存在部分动作奖励需要根据连续多次样本目标动作确定，例如，合理侧身为拟人化奖励，真实玩家控制的虚拟角色通常不会在短时间内多次侧身，或者在周围没有掩体的情况下侧身，因此，在确定出的多次样本目标动作中，侧身动作的拟人化属性值（即侧身动作的次数）超过拟人属性阈值的情况下，计算机设备则将其确定为不合理侧身，基于动作决策模型反向奖励。

在表 3 中包括走位分散奖励、掩体利用奖励、暗处伏击奖励以及合理动作奖励（合理侧身、合理蹲起、合理静步等）。

一、走位分散奖励是指第一虚拟角色执行样本目标动作的情况下，连续多次走位的紧密程度降低的奖励。即第一虚拟角色不会连续多次在原地徘徊，在大规模地图中能够更加分散的走位，探索地图的不同位置。

可选的，样本动作执行结果包括第一虚拟角色执行前 $n+1$ 次动作后的位置点。计算机设备基于第一虚拟角色执行第 $n+1$ 次样本目标动作后的位置点，确定执行第 $n+1$ 次动作后的位置点与前 n 次执行动作后的位置点的第一紧密度，以及确定执行第 n 次动作后的位置点与前 $n-1$ 次执行动作后位置点的第二紧密度，在第一紧密度小于第二紧密度的情况下，计算机设备将走位分散奖励确定为动作执行奖励。

在对动作决策模型训练时，可能会导致第一虚拟角色陷入在局部区域内绕圈的现象，因此为了避免该现象，计算机设备可以基于紧密中心性算法，引导第一虚拟角色执行动作决策模型决策出的样本目标动作后，能够走位足够分散，走位路线尽可能遍及地图全局，以训练第一虚拟角色的走位能力。

具体的，在执行第 $n+1$ 次动作后的位置点与前 n 次执行动作后的位置点的紧密度小于执行第 n 次动作后的位置点与前 $n-1$ 次执行动作后位置点的紧密度的情况下，计算机设备确定

获取到正向走位分散奖励。或者，在执行第 $n+1$ 次动作后的位置点与前 n 次执行动作后的位置点的紧密度大于执行第 n 次动作后的位置点与前 $n-1$ 次执行动作后位置点的紧密度的情况下，计算机设备确定获取到反向的走位分散奖励。从而引导第一虚拟角色尽量远离最近去过的历史位置，避免出现虚拟角色原地徘徊的情况。

二、掩体利用奖励是指第一虚拟角色执行样本目标动作的情况下，所处位置与掩体相邻的奖励。在存在第二虚拟角色的情况下，期望第一虚拟角色能够尽量处于掩体旁边，以防止被第二虚拟角色攻击到。

可选的，样本动作执行结果包括第一虚拟角色执行样本目标动作后的样本状态信息，包括样本角色状态信息以及样本多环境感知信息，样本环境感知信息中包含二维深度图，二维深度图用于表征第一虚拟角色朝向方向上的障碍物对第一虚拟角色的遮掩情况。

计算机设备对二维深度图区域划分，得到至少两个深度区域，并确定第一深度区域的第一遮掩率，以及与第一深度区域相邻的至少两个第二深度区域的第二遮掩率，第一深度区域位于第一虚拟角色的视野中心。进而，计算机设备基于第一遮掩率与第二遮掩率的最小差值，确定掩体利用奖励，将掩体利用奖励确定为动作执行奖励，其中，掩体利用奖励的奖励程度与最小差值呈正相关关系。

本申请实施例中通过重塑改造深度图矩阵实现对掩体利用奖励的确定。在游戏场景中掩体是在第一虚拟角色战斗过程中为第一虚拟角色提供遮掩的区域，在虚拟环境中存在多个掩体，不同掩体之间的位置、朝向以及形状中至少有一种存在区别。而在二维深度图中，能够表征第一虚拟角色朝向方上的障碍物深度情况，因此，计算机设备可以通过二维深图的像素值分布情况确定掩体的像素分布区域。

可选的，计算机设备按照深度图数值方向进行划分，得到 1-19 共 19 个区域，并将深度图竖直划分成若干列，计算每一列的明亮率，明亮率越大表明前方越亮，存在障碍物遮挡的可能性越低，则遮掩率越低。在应用中，往往期望第一虚拟角色能够选择存在掩体又能够攻击到第二虚拟角色的区域，因此，需要使深度图中间列的明亮率最小，而中间列反应第一虚拟角色正前方的视野信息，明亮率越小则遮掩率越大，正前方为掩体区域的可能性越高。此外，期望第一虚拟角色能够处于掩体边缘，即能够通过侧身等动作对第二虚拟角色发起进攻，因此，要保证中间列的明亮率及其左右两列中较大明亮率的差值绝对值应尽可能大，即使得中间列的明亮率与其左右两列的明亮率（遮掩率）差值的最小值尽可能大。比如，期望第 10 列的遮掩率尽可能大，且第 9 列和第 11 列中存在一个区域的遮掩率尽可能小。

可选的，深度区域对应的明亮率 v_j 可以通过以下公式进行计算：

$$v_j = \frac{\sum_{i=1}^{i=n} x_{i,j}}{\sum_{i=1}^{i=n} 1}$$

其中 v_j 为每个深度区域中实际像素与每列理论最大像素和的比值，而深度区域的遮掩率为 $1-v_j$ 。则掩体利用奖励可以通过下式进行计算：

$$r = \alpha \cdot (1 - v_{10}) + \beta \cdot |1 - v_{10} - \min(1 - v_9, 1 - v_{11})|$$

其中， r 为掩体利用奖励，而 α 与 β 为奖励超参数。

需要说明的是，对二维深度图区域划分的方式还可以基于水平方向进行划分，或者，划分的区域数量可以基于具体的虚拟环境调整，在虚拟环境较为复杂的情况下，可以采用更多的深度区域，对二维深度图执行更精细的划分，从而获取更精细的掩体效果。在地图场景较大的情况下，控件结构较为简单，则可以粗粒度执行深度区域划分，本实施例对此不进行限定。

三、暗处伏击奖励即为可视性奖励，可视性奖励是指第一虚拟角色执行样本目标动作的情况下，所处位置为对第二虚拟角色的有效攻击位置的奖励。在存在第二虚拟角色的情况下，期望第二虚拟角色能够处于第一虚拟角色的可视位置，但第一虚拟角色不处于第二虚拟角色的可视位置，有利于发起进攻。

可选的，样本动作执行结果包括第一虚拟角色与第二虚拟角色之间的相互可视关系，相互可视关系用于表征第一虚拟角色以及第二虚拟角色相对于彼此的可视情况。

可选的，在相互可视关系指示第一虚拟角色不在第二虚拟角色的可视范围，且第二虚拟角色在第一虚拟角色的可视范围的情况下，计算机设备确定动作执行奖励为正向的可视性奖励。

在相互可视关系指示第一虚拟角色在第二虚拟角色的可视范围，且第二虚拟角色不在第一虚拟角色的可视范围的情况下，计算机设备确定动作执行奖励为反向的可视性奖励。

即第一虚拟角色相较于第二虚拟角色处于有利进攻位置的情况下确定获取正向的可视性奖励。反之，第一虚拟角色相较于第二虚拟角色处于不利的进攻位置的情况下，确定获取反向的可视性奖励。

可选的，相互可视关系可以通过射线探测的方式获取。计算机设备以第一虚拟角色眼部位置为起始点，向第一虚拟角色朝向方向上的第二虚拟角色的各个部位发射环境感知射线，并根据环境感知射线的反射情况，确定虚拟环境中第二虚拟角色的各个部位的可视情况，可视情况用于表征第二虚拟角色各个部位是否被掩体遮挡。在可视情况指示第二虚拟角色各个部位均被障碍物遮挡的情况下，确定第二虚拟角色不在第一虚拟角色的可视范围。同理，计算机设备以第二虚拟角色眼部位置为起始点，向第二虚拟角色朝向方向上的第一虚拟角色的各个部位发射环境感知射线，根据环境感知射线的反射情况，确定第一虚拟角色的各个部位的可视情况，在可视情况指示第一虚拟角色各个部位均被障碍物遮挡的情况下，确定第一虚拟角色不在第二虚拟角色的可视范围。

四、合理动作奖励是指第一虚拟角色执行样本目标动作的情况下，符合真实玩家操作逻辑的奖励。在模型训练的过程中，期望第一虚拟角色的行为能够尽可能的拟人化，执行更符合人类逻辑的行为。

在确定样本目标动作之前，计算机设备基于训练对局的对局信息（包括样本角色状态信息以及样本环境感知信息），确定第一虚拟角色的理想动作。例如，多环境感知信息表征第一虚拟角色前方存在较低的掩体，则可以为第一虚拟角色的理想动作可以为下蹲动作，再例如，环境感知信息表征第一虚拟角色右手边存在第二虚拟角色，则可以确定理想动作应当为向右转向。

随后，在动作决策模型对应的样本目标动作与理想动作一致的情况下，计算机设备则将合理动作奖励确定为动作执行奖励。

可选的，计算机设备以第一虚拟角色为起始点向四周发射环境感知射线，不同环境感知射线位于同一水平高度，并根据环境感知射线的反射情况，确定周围的障碍物分布。在存在一条最短的环境感知射线的长度小于距离阈值的情况下，则说明第一虚拟角色距离障碍物之间的距离较近，可以确定此时应当转向。其中，环境感知射线以最短射线作为基准向两侧扩散，计算机设备判定射线长度更早超过距离阈值的环境感知射线的射线方向为第一虚拟角色的理想转向方向。

而当动作决策模型确定出的转向方向与理想转向方向一致的情况下，计算机设备则将合理动作奖励确定为动作执行奖励；当动作决策模型确定出的转向方向与理想转向方向不一致的情况下，则第一虚拟角色可能出现卡墙的行为，计算机设备则给予一定的反向奖励。

步骤 1402d，基于动作执行奖励更新动作决策模型的模型参数。

基于动作执行奖励更新动作决策模型的过程即为，更新动作决策模型的模型参数以获取更大的动作执行奖励的过程，从而使得动作决策模型能够在执行目标任务的前提下尽可能的实现动作拟人化。

本申请实施例中，在对动作决策模型训练的过程中，通过增加辅助拟人化动作奖励，能够满足模型强度以及拟人性的双重需求，并且结合动作遮罩逻辑，使得虚拟角色的拟人性得到极大提升，更加符合真实玩家的行为逻辑。

本申请实施例中，第一虚拟角色所处的训练对局中存在至少一个第二虚拟角色，且第二虚拟角色执行第二动作决策模型输出的动作。在一些场景中，第二虚拟角色的目标任务与第一虚拟角色的目标任务相对应，例如，第一虚拟角色的目标任务为守护宝箱，而第二虚拟角色的目标任务可能为夺取宝箱。

第二虚拟角色作为对手虚拟角色，在对第一虚拟角色对应的第一动作决策模型训练后，同样需要对第二动作决策模型迭代训练，以获取不同强度的第二动作决策模型。

具体的，在第一虚拟角色对应的第一动作决策模型完成第 i 轮训练的情况下，基于训练后的第一动作决策模型对第二动作决策模型执行第 i 轮训练，并将完成第 i 轮训练的第二动作决策模型存储至评估模型池。

而对于第一虚拟角色对应的第一动作决策模型，在训练对局结束的情况下，计算机设备确定动作决策模型完成第 i 轮训练。随后，评估完成第 i 轮训练的动作决策模型，得到其对应的性能评估结果。

可选的，计算机设备从评估模型池中选择训练过的至少两个对照模型，再创建至少两个评估对局，评估对局中包括执行动作决策模型输出动作的第一虚拟角色，以及执行对照模型输出动作的一个第二虚拟角色，再根据至少两个评估对局的对局数据，确定性能评估结果。

其中，性能评估的指标可能为对局的胜负率，或者在对局中某个特定行为的比例等等，本实施例对此不做限定。

在性能评估结果指示动作决策模型未达到训练完成标准的情况下，计算机设备则需要控制动作决策模型进入第 $i+1$ 轮训练，即性能评估结果指示动作决策模型未达到训练完成标准的情况下，需要控制动作决策模型进入下一轮训练。

在性能评估结果指示动作决策模型达到训练完成标准的情况下，则确定动作决策模型完成训练。

示意性的，具体过程如下：

- (1) 基于客户端内置行为树控制对手虚拟角色，对第一动作决策模型执行第 0 轮训练；
- (2) 将第 0 轮训练的模型存入评估模型池并对其评估；
- (3) 根据完成第 0 轮训练的第一动作决策模型，对第二动作决策模型执行第 0 轮训练，并将其添加至对手模型池；
- (4) 从对手模型池中选择一个过往训练过的第二决策模型作为对手，对第一动作决策模型执行第 K 轮训练；
- (5) 将第 K 轮训练的第一动作决策模型存入评估模型池并对其评估；
- (6) 根据第 K 轮训练的第一动作决策模型对第二动作决策模型执行第 K 轮训练，并将其存入对手模型池；
- (7) 重复 (4) 至 (6) 中的步骤，直至第一动作决策模型的评估结果满足性能评估结果的情况下，确定第一动作决策模型完成训练。

本申请实施例中，通过对陪练的对手虚拟角色对应的第二动作决策模型以及第一虚拟角色对应的第一决策模型训练，从而提升第一决策模型的强度，使第一决策模型能够具有稳定的动作决策能力。

本申请实施例中，判断动作决策模型是否完成训练，需要结合对局指标以及拟人化指标共同判断。对局指标用于评判动作决策模型的强度，拟人化指标用于评判动作决策模型的拟人化程度。

请参考表 4，其示出了本申请一个示例性实施例提供的对局指标以及拟人化指标。

表 4

对局指标		拟人化指标（统计 500 局，计算平均每局的拟人性指标）	
平局率	双方平局方式结束对局	侧身率	合理侧身攻击第二虚拟角色
胜率	第一虚拟角色打败第二虚拟角色	蹲伏率	合理利用蹲起动作来隐藏自己不被对手发现的概率

败率	第一虚拟角色被第二虚拟角色打败	躲避率	平均一局游戏中不被发现的占比
遇敌进攻率	遇敌进攻帧数/遇敌总帧数	静步率	第一虚拟角色利用静步方式隐藏声音而不被发现的占比
证据开火率	遇敌打中敌人帧数/遇敌进攻总帧数	掩体利用率	第一虚拟角色利用掩体隐藏自身占比
存活时间	第一虚拟角色在一局游戏中存活的时长	走位分散率	是指第一虚拟角色走位灵活，不会长时间停留在同一个位置
卡墙次数	视野内 30° 内且与障碍物距离小于 0.5m 的帧数	危险区域停留率	第一虚拟角色在特定的战斗场景下处于没有掩体的危险空旷区域频率
		大幅转向率	第一虚拟角色的移动朝向不应当大幅度骤变
		进攻回退率	第一虚拟角色在射击第二虚拟角色时能主动一边射击一边远离第二虚拟角色
		暗中射击率	在第一虚拟角色在第二虚拟角色视野范围之外射击
		缠斗率	用于刻画利用掩体进行拉扯战斗的拟人行为

其中，在确定拟人性指标时，根据大量的对局数据确定平均的拟人性指标。

计算机设备在完成对动作决策模型的第 i 轮训练的情况下，获取第 i 轮训练对局的对局数据。基于至少两轮训练对局的对局数据，确定对局指标以及拟人化指标，拟人化指标是指特定动作在至少训练对局的对局数据中的实际执行占比。

在对局指标达到训练完成标准，且拟人化指标与真实玩家在对局中执行特定动作的目标执行占比相匹配的情况下，则确定动作决策模型完成训练。

本申请实施例中，基于对局指标以及拟人化指标共同作为动作决策模型的评估标准，从而使训练完成的动作决策模型能够兼顾强度以及动作模型的拟人化程度。

请参考图 15，其示出了本申请一个示例性实施例提供的训练过程中客户端与训练动作决策模型的服务器之间交互的示意图。其中，客户端 1501 为安装于计算机设备中用于提供虚拟环境的应用程序。服务器 1502 用于提供动作模型训练服务。在训练的过程中，客户端 1501 向服务器 1502 发送状态信息，服务器 1502 接收到状态信息之后，通过动作决策模型中的 Agent 组件对状态信息处理以及拼装成动作决策模型的输入特征，然后请求 Actor 组件获取当前策略下的样本目标动作。并将该样本目标动作映射生成发送给客户端 1501 的响应包。并且，将在此过程中产生的数据经过奖励计算以及样本处理，生成训练样本，批量发送至 Learner 组件进行策略参数优化。Learner 组件接收到的训练样本存放至本地的缓冲池中，每个训练步会依据策略从缓冲池中采用，当 Learner 组件训练若干步之后，将当前训练网络中的参数同步至 Actor 中的目标网络。可选的，本申请实施例中可以采用 PPO (Proximal Policy Optimization, 近端策略优化) 算法进行训练。

请参考图 16，其示出了本申请一个示例性实施例通过的训练过程中实现动作决策请求的决策模式示意图。

其中在训练过程中，客户端采用同步执决策模式，即将状态信息发送之后，等待服务器返回的动作包，基于动作包执行样本目标动作。在需要进行动作决策时，在一个游戏帧中，向第一服务器发送状态信息，并由第一服务器向提供动作决策服务的第二服务器中发送请求

数据。在动作决策后，第二服务器向第一服务器返回应答数据，并等待下一帧动作决策请求。由第一服务器获取样本目标动作，再返回动作指令至客户端控制目标虚拟角色执行样本目标动作。采用同步决策模式有利于保证模型训练效果。

请参考图 17，其示出了本申请一个示例性实施例通过的应用过程中实现动作决策请求的决策模式示意图。在应用阶段，采用异步决策模式，客户端向第一服务器发送状态信息后，不等待返回的动作包。在发送状态信息后周期检查是否接收到返回动作包的同时运行其他业务逻辑。在检查未接收到返回动作包的情况下，继续运行业务逻辑，在检查接收到返回动作包的情况下，控制目标虚拟角色执行目标动作后，等待下一次动作决策。在应用阶段，采用异步决策的模式，能够有效降低主线程的数据阻塞，保证动作决策模型推理耗时不影响其他业务逻辑，降低业务侧服务器（第一服务器）资源占用的情况，确保动作决策模型的安全性。

在一种可能的实施方式中，计算机设备可以采用监督学习与强化学习相结合的方式训练动作决策模型，在训练前期，先基于真实玩家的对局数据以监督学习的方式对动作决策模型训练，能够有效避免强化学习在动作决策模型训练前期无意义的探索，然后在后续的训练中采用强化学习的方式训练，能够有效提升动作决策模型的训练效率。

请参考图 18，其示出了本申请一个示例性实施例提供的虚拟角色的动作决策装置的结构示意图，该装置包括以下结构。

获取模块 1801，用于获取状态信息，所述状态信息用于表征第一虚拟角色所处对局的对局状态；

决策模块 1802，用于将所述状态信息输入动作决策模型，得到所述动作决策模型中 n 个动作输出头串行输出的 n 个目标子动作，其中，不同动作输出头对应不同动作类型，且 n 个所述动作输出头基于所述动作类型之间的依赖关系串行连接，所述依赖关系用于表征不同动作类型下子动作之间的依赖限制情况， n 为正整数；

控制模块 1803，用于控制所述第一虚拟角色执行由 n 个所述目标子动作构成的目标动作。

可选的，所述动作决策模型包括信息处理层、特征提取层以及动作决策层，所述动作决策层包括 n 个所述动作输出头；

所述决策模块 1802，用于：

将所述状态信息输入所述动作决策模型，通过所述信息处理层对所述状态信息编码，得到状态编码；

将所述状态编码输入所述特征提取层，通过所述特征提取层提取所述状态编码特征，得到融合特征；

基于所述融合特征，通过所述动作决策层中的 n 个所述动作输出头确定 n 个所述目标子动作。

可选的，所述决策模块 1802，用于：

将所述融合特征输入所述动作决策层中的第一动作输出头，通过所述第一动作输出头从第一动作类型中确定第一目标子动作；

将已确定的所述第一目标子动作至第 $i-1$ 目标子动作分别对应的嵌入编码向量，以及所述融合特征输入所述动作决策层，通过所述动作决策层中第 i 动作输出头从第 i 动作类型中确定第 i 目标子动作，其中，所述第一目标子动作至所述第 $i-1$ 目标子动作分别由所述第一动作输出头至第 $i-1$ 动作输出头确定， i 小于或等于 n ，且 i 大于 1。

可选的，所述动作决策模型还包括动作遮罩层，所述动作遮罩层与所述动作决策层中的 n 个所述动作输出头相连，所述动作遮罩层用于遮罩不同动作类型中的子动作；

所述装置还包括：

动作遮罩模块，用于基于已确定的所述第一目标子动作至所述第 $i-1$ 目标子动作，以及不同动作类型下子动作之间的所述依赖限制情况所指示的依赖限制关系，遮罩所述第 i 动作类

型中的子动作，其中，具有所述依赖限制关系的至少两个子动作不支持同时执行；

所述决策模块 1802，用于通过所述动作决策层中所述第 i 动作输出头，从所述第 i 动作类型中未被遮罩的子动作中确定所述第 i 目标子动作。

可选的，所述决策模块 1802，用于：

在存在已确定的第 j 目标子动作，且所述依赖限制情况指示所述第 j 目标子动作与所述第 i 动作类型中的至少一个第一子动作存在所述依赖限制关系的情况下，遮罩所述第 i 动作类型中的所述第一子动作， j 小于 i ；

或者，

在存在已确定的第 x 目标子动作以及第 y 目标子动作，且所述依赖限制情况指示所述第 x 目标子动作结合所述第 y 目标子动作后，与所述第 i 动作类型中的至少一个第二子动作存在所述依赖限制关系的情况下，遮罩所述第 i 动作类型中的所述第二子动作， x, y 小于 i 。

可选的，所述状态信息包括所述第一虚拟角色在所述对局中的角色状态信息，以及所述第一虚拟角色所处虚拟环境的环境感知信息，所述角色状态信息用于表征所述第一虚拟角色在所述对局中与第二虚拟角色之间的交互状态，以及所述第一虚拟角色在所述对局中与所述虚拟环境之间的交互状态，所述角色状态信息为一维向量，所述环境感知信息为二维图像，所述信息处理层中包括标量编码器以及图像编码器；

可选的，所述决策模块 1802，包括：

将所述状态信息输入所述动作决策模型，通过所述信息处理层中的所述标量编码器对所述角色状态信息编码，得到状态信息编码结果；

通过所述信息处理层中的所述图像编码器对所述环境感知信息编码，得到环境信息编码结果；

通过所述信息处理层拼接所述状态信息编码结果以及所述环境信息编码结果，得到所述状态编码。

可选的，所述动作决策模型还包括信息过滤层，所述信息过滤层与所述标量编码器的输出端相连，且所述信息过滤层与所述动作决策层的输入端相连；

所述装置还包括：

过滤模块，用于在获取到所述角色状态信息的情况下，通过所述信息过滤层过滤所述角色状态信息，得到过滤后的角色状态信息，所述过滤后的角色状态信息与第一动作输出头对应的第一动作类型具有相关关系；将过滤后的所述目标角色状态信息输入所述第一动作输出头。

可选的，所述装置还包括：

分解模块，用于正交分解所述第一虚拟角色的可执行动作，得到 n 个所述动作类型，不同动作类型中包括至少两个可执行子动作，不同动作类型下的子动作之间支持独立控制。

可选的，所述装置还包括：

训练模块，用于获取样本状态信息，所述样本状态信息用于表征所述第一虚拟角色所处训练对局的对局状态；

所述训练模块，还用于基于所述样本状态信息，通过强化学习的方式训练所述动作决策模型。

可选的，所述训练模块，用于：

将所述样本状态信息输入所述动作决策模型，得到所述动作决策模型输出的样本目标动作，所述样本目标动作由 n 个样本目标子动作构成；

控制所述第一虚拟角色执行所述样本目标动作，得到所述第一虚拟角色的样本动作执行结果；

基于所述第一虚拟角色的所述样本动作执行结果，确定动作执行奖励；

基于所述动作执行奖励更新所述动作决策模型的模型参数。

可选的，所述信息处理层中还包括对局信息编码器，所述样本状态信息中还包括样本对

局信息，所述样本对局信息用于表征所述第一虚拟角色所处的所述训练对局的实时情况；

所述训练模块，用于在将所述样本状态信息输入所述动作决策模型的情况下，通过所述对局信息编码器对所述样本对局信息编码，得到对局状态编码；

所述训练模块，还用于基于所述对局状态编码以及所述样本动作执行结果，确定所述动作执行奖励。

可选的，所述动作执行奖励包括属性奖励、对局胜利奖励、对局失败奖励以及任务奖励中的至少一项，所述对局胜利奖励为正向奖励，所述对局失败奖励为反向奖励；

所述训练模块，用于：

在所述第一虚拟角色取得对局胜利的情况下，将所述对局胜利奖励确定为所述动作执行奖励；

在所述第一虚拟角色取得对局失败的情况下，将所述对局失败奖励确定为所述动作执行奖励；

在所述第一虚拟角色的属性值降低的情况下，将所述属性奖励确定为所述动作执行奖励；

在所述第一虚拟角色的任务成功的情况下，将所述任务奖励确定为所述动作执行奖励。

可选的，所述动作执行奖励包括拟人化奖励；

所述训练模块，用于：

基于所述第一虚拟角色在所述训练对局中执行至少两次所述样本目标动作的样本动作执行结果，确定至少两次所述样本目标动作的拟人化属性值；

在所述拟人化属性值低于拟人属性阈值的情况下，将所述拟人化奖励确定为所述动作执行奖励。

可选的，所述训练模块，用于：

在所述训练对局结束的情况下，确定所述动作决策模型完成第*i*轮训练；

评估完成第*i*轮训练的所述动作决策模型，得到性能评估结果；

在所述性能评估结果指示所述动作决策模型未达到训练完成标准的情况下，控制所述动作决策模型进入第*i*+1轮训练；

在所述性能评估结果指示所述动作决策模型达到训练完成标准的情况下，确定所述动作决策模型完成训练。

可选的，所述第一虚拟角色所处的所述训练对局中存在至少一个第二虚拟角色，且第二虚拟角色执行第二动作决策模型输出的动作；

所述训练模块，用于在所述第一虚拟角色对应的第一动作决策模型完成所述第*i*轮训练的情况下，基于训练后的所述第一动作决策模型，控制所述第二动作决策模型进入第*i*轮训练，并将完成第*i*轮训练的所述第二动作决策模型存储至评估模型池；

所述训练模块，还用于：

从所述评估模型池中选择训练过的至少两个对照模型；

创建至少两个评估对局，所述评估对局中包括执行所述动作决策模型输出动作的所述第一虚拟角色，以及执行所述对照模型输出动作的一个第二虚拟角色；

根据所述至少两个所述评估对局的对局数据，确定所述性能评估结果。

可选的，所述训练模块，用于：

在完成对所述动作决策模型的第*i*轮训练的情况下，获取执行所述第*i*轮训练的所述训练对局的对局数据；

基于至少两轮训练的所述训练对局的所述对局数据，确定对局指标以及拟人化指标，所述拟人化指标是指特定动作在至少两轮所述训练对局中的实际执行占比；

在所述对局指标达到训练完成标准，且所述拟人化指标指示的实际执行占比与目标执行占比相匹配的情况下，确定所述动作决策模型完成训练，所述目标执行占比是真实玩家在真实对局中执行所述特征动作的占比。

综上所述，本申请实施例中，通过动作决策模型基于状态信息实现动作决策，从而能够

根据当前状态，确定出合理的目标动作。其中，动作决策模型中包含 n 个动作输出头串行连接， n 个动作输出头分别用于输出不同的目标子动作， n 个动作输出头基于动作类型之间的依赖关系串行连接，能够在动作决策时考虑到不同动作之间的依赖关系，并在串行输出 n 个目标子动作的过程中，使得 n 个目标子动作在时序上前后因果关联，从而提高目标动作的合理性，提升了虚拟角色的拟人性。

请参考图 19，其示出了本申请一个示例性实施例提供的计算机设备的结构示意图，计算机设备可实现为上述实施例中的终端或服务器。具体来讲：所述计算机设备 1900 包括中央处理单元（Central Processing Unit, CPU）1901、包括随机存取存储器 1902 和只读存储器 1903 的系统存储器 1904，以及连接系统存储器 1904 和中央处理单元 1901 的系统总线 1905。所述计算机设备 1900 还包括帮助计算机内的各个器件之间传输信息的基本输入/输出系统（Input/Output, I/O 系统）1906，和用于存储操作系统 1913、应用程序 1914 和其他程序模块 1915 的大容量存储设备 1907。

在一些实施例中，所述基本输入/输出系统 1906 包括有用于显示信息的显示器 1908 和用于用户输入信息的诸如鼠标、键盘之类的输入设备 1909。其中所述显示器 1908 和输入设备 1909 都通过连接到系统总线 1905 的输入输出控制器 1910 连接到中央处理单元 1901。所述基本输入/输出系统 1906 还可以包括输入输出控制器 1910 以用于接收和处理来自键盘、鼠标、或电子触控笔等多个其他设备的输入。类似地，输入输出控制器 1910 还提供输出到显示屏、打印机或其他类型的输出设备。

所述大容量存储设备 1907 通过连接到系统总线 1905 的大容量存储控制器（未示出）连接到中央处理单元 1901。所述大容量存储设备 1907 及其相关联的计算机可读介质为计算机设备 1900 提供非易失性存储。也就是说，所述大容量存储设备 1907 可以包括诸如硬盘或者驱动器之类的计算机可读介质（未示出）。

不失一般性，所述计算机可读介质可以包括计算机存储介质和通信介质。计算机存储介质包括以用于存储诸如计算机可读指令、数据结构、程序模块或其他数据等信息的任何方法或技术实现的易失性和非易失性、可移动和不可移动介质。计算机存储介质包括随机存取记忆体（Random Access Memory, RAM）、只读存储器（Read Only Memory, ROM）、闪存或其他固态存储其技术，只读光盘（Compact Disc Read-Only Memory, CD-ROM）、数字通用光盘（Digital Versatile Disc, DVD）或其他光学存储、磁带盒、磁带、磁盘存储或其他磁性存储设备。当然，本领域技术人员可知所述计算机存储介质不局限于上述几种。上述的系统存储器 1904 和大容量存储设备 1907 可以统称为存储器。

存储器存储有一个或多个程序，一个或多个程序被配置成由一个或多个中央处理单元 1901 执行，一个或多个程序包含用于实现上述方法的指令，中央处理单元 1901 执行该一个或多个程序实现上述各个方法实施例提供的方法。

根据本申请的各种实施例，所述计算机设备 1900 还可以通过诸如因特网等网络连接到网络上的远程计算机运行。也即计算机设备 1900 可以通过连接在所述系统总线 1905 上的网络接口单元 1911 接到网络 1912，或者说，也可以使用网络接口单元 1911 来连接到其他类型的网络或远程计算机系统（未示出）。

所述存储器还包括一个或者一个以上的程序，所述一个或者一个以上程序存储于存储器中，所述一个或者一个以上程序包含用于进行本申请实施例提供的方法中由计算机设备所执行的步骤。

本申请实施例还提供一种计算机可读存储介质，该可读存储介质中存储有至少一条计算机指令，至少一条计算机指令由处理器加载并执行以实现上述任一实施例所述的虚拟角色的动作决策方法。

本申请实施例提供了一种计算机程序产品，该计算机程序产品包括计算机指令，该计算

机指令存储在计算机可读存储介质中。计算机设备的处理器从计算机可读存储介质读取该计算机指令，处理器执行该计算机指令，使得该计算机设备执行上述方面提供的虚拟角色的动作决策方法。

本领域普通技术人员可以理解上述实施例的各种方法中的全部或部分步骤是可以通程序来指令相关的硬件来完成，该程序可以存储于一计算机可读存储介质中，该计算机可读存储介质可以是上述实施例中的存储器中所包含的计算机可读存储介质；也可以是单独存在，未装配入终端中的计算机可读存储介质。该计算机可读存储介质中存储有至少一条计算机指令，所述至少一条计算机指令由处理器加载并执行以实现上述任一方法实施例所述的虚拟角色的动作决策方法。

可选地，该计算机可读存储介质可以包括：ROM、RAM、固态硬盘（Solid State Drives, SSD）或光盘等。其中，RAM 可以包括电阻式随机存取记忆体（Resistance Random Access Memory, ReRAM）和动态随机存取存储器（Dynamic Random Access Memory, DRAM）。上述本申请实施例序号仅仅为了描述，不代表实施例的优劣。

本领域普通技术人员可以理解实现上述实施例的全部或部分步骤可以通过硬件来完成，也可以通过程序来指令相关的硬件完成，所述的程序可以存储于一种计算机可读存储介质中，上述提到的存储介质可以是只读存储器，磁盘或光盘等。

需要说明的是，本申请所涉及的信息（包括但不限于用户设备信息、用户个人信息等）、数据（包括但不限于用于分析的数据、存储的数据、展示的数据等）以及信号，均为经用户授权或者经过各方充分授权的，且相关数据的收集、使用和处理需要遵守相关国家和地区的相关法律法规和标准。

且本申请在收集用户的相关数据之前以及在收集用户的相关数据的过程中，都可以显示提示界面、弹窗或输出语音提示信息，该提示界面、弹窗或语音提示信息用于提示用户当前正在搜集其相关数据，使得本申请仅仅在获取到用户对该提示界面或者弹窗发出的确认操作后，才开始执行获取用户相关数据的相关步骤，否则（即未获取到用户对该提示界面或者弹窗发出的确认操作时），结束获取用户相关数据的相关步骤，即不获取用户的相关数据。

应当理解的是，在本文中提及的“多个”是指两个或两个以上。本文中提及的“第一”、“第二”等用于区别类似对象，而并不用于限定特定的顺序或先后次序。另外，本文中描述的步骤编号，仅示例性示出了步骤间的一种可能的执行先后顺序，在一些其它实施例中，上述步骤也可以不按照编号顺序来执行，如两个不同编号的步骤同时执行，或者两个不同编号的步骤按照与图示相反的顺序执行，本申请实施例对此不作限定。

以上所述仅为本申请的可选的实施例，并不用以限制本申请，凡在本申请的精神和原则之内，所作的任何修改、等同替换、改进等，均应包含在本申请的保护范围之内。

权利要求书

1. 一种虚拟角色的动作决策方法，所述方法由计算机设备执行，所述方法包括：

获取状态信息，所述状态信息用于表征第一虚拟角色所处对局的对局状态；

将所述状态信息输入动作决策模型，得到所述动作决策模型中 n 个动作输出头串行输出的 n 个目标子动作，其中，不同动作输出头对应不同动作类型，且 n 个所述动作输出头基于所述动作类型之间的依赖关系串行连接，所述依赖关系用于表征不同动作类型下子动作之间的依赖限制情况， n 为正整数；

控制所述第一虚拟角色执行由 n 个所述目标子动作构成的目标动作。

2. 根据权利要求 1 所述的方法，其中，所述动作决策模型包括信息处理层、特征提取层以及动作决策层，所述动作决策层包括 n 个所述动作输出头；

所述将所述状态信息输入动作决策模型，得到所述动作决策模型中 n 个动作输出头串行输出的 n 个目标子动作，包括：

将所述状态信息输入所述动作决策模型，通过所述信息处理层对所述状态信息编码，得到状态编码；

将所述状态编码输入所述特征提取层，通过所述特征提取层提取所述状态编码特征，得到融合特征；

基于所述融合特征，通过所述动作决策层中的 n 个所述动作输出头确定 n 个所述目标子动作。

3. 根据权利要求 2 所述的方法，其中，所述基于所述融合特征，通过所述动作决策层中的 n 个所述动作输出头确定 n 个所述目标子动作，包括：

将所述融合特征输入所述动作决策层中的第一动作输出头，通过所述第一动作输出头从第一动作类型中确定第一目标子动作；

将已确定的所述第一目标子动作至第 $i-1$ 目标子动作分别对应的嵌入编码向量，以及所述融合特征输入所述动作决策层，通过所述动作决策层中第 i 动作输出头从第 i 动作类型中确定第 i 目标子动作，其中，所述第一目标子动作至所述第 $i-1$ 目标子动作分别由所述第一动作输出头至第 $i-1$ 动作输出头确定， i 小于或等于 n ，且 i 大于 1。

4. 根据权利要求 3 所述的方法，其中，所述动作决策模型还包括动作遮罩层，所述动作遮罩层与所述动作决策层中的 n 个所述动作输出头相连，所述动作遮罩层用于遮罩不同动作类型中的子动作；

所述方法还包括：

基于已确定的所述第一目标子动作至所述第 $i-1$ 目标子动作，以及不同动作类型下子动作之间的所述依赖限制情况所指示的依赖限制关系，遮罩所述第 i 动作类型中的子动作，其中，具有所述依赖限制关系的至少两个子动作不支持同时执行；

所述通过所述动作决策层中第 i 动作输出头从第 i 动作类型中确定第 i 目标子动作，包括：

通过所述动作决策层中所述第 i 动作输出头，从所述第 i 动作类型中未被遮罩的子动作中确定所述第 i 目标子动作。

5. 根据权利要求 4 所述的方法，其中，所述基于已确定的所述第一目标子动作至所述第 $i-1$ 目标子动作，以及不同动作类型下子动作之间的所述依赖限制情况所指示的依赖限制关系，遮罩所述第 i 动作类型中的子动作，包括：

在存在已确定的第 j 目标子动作，且所述依赖限制情况指示所述第 j 目标子动作与所述

第 i 动作类型中的至少一个第一子动作存在所述依赖限制关系的情况下, 遮罩所述第 i 动作类型中的所述第一子动作, j 小于 i ;

或者,

在存在已确定的第 x 目标子动作以及第 y 目标子动作, 且所述依赖限制情况指示所述第 x 目标子动作结合所述第 y 目标子动作后, 与所述第 i 动作类型中的至少一个第二子动作存在所述依赖限制关系的情况下, 遮罩所述第 i 动作类型中的所述第二子动作, x, y 小于 i 。

6. 根据权利要求 2 至 5 任一所述的方法, 其中, 所述状态信息包括所述第一虚拟角色在所述对局中的角色状态信息, 以及所述第一虚拟角色所处虚拟环境的环境感知信息, 所述角色状态信息用于表征所述第一虚拟角色在所述对局中与第二虚拟角色之间的交互状态, 以及所述第一虚拟角色在所述对局中与所述虚拟环境之间的交互状态, 所述角色状态信息为一维向量, 所述环境感知信息为二维图像, 所述信息处理层中包括标量编码器以及图像编码器;

所述将所述状态信息输入所述动作决策模型, 通过所述信息处理层对所述状态信息编码, 得到状态编码, 包括:

将所述状态信息输入所述动作决策模型, 通过所述信息处理层中的所述标量编码器对所述角色状态信息编码, 得到状态信息编码结果;

通过所述信息处理层中的所述图像编码器对所述环境感知信息编码, 得到环境信息编码结果;

通过所述信息处理层拼接所述状态信息编码结果以及所述环境信息编码结果, 得到所述状态编码。

7. 根据权利要求 6 所述的方法, 其中, 所述动作决策模型还包括信息过滤层, 所述信息过滤层与所述标量编码器的输出端相连, 且所述信息过滤层与所述动作决策层的输入端相连;

所述方法还包括:

在获取到所述角色状态信息的情况下, 通过所述信息过滤层过滤所述角色状态信息, 得到过滤后的角色状态信息, 所述过滤后的角色状态信息与第一动作输出头对应的第一动作类型具有相关关系;

将过滤后的所述目标角色状态信息输入所述第一动作输出头。

8. 根据权利要求 1 至 7 任一所述的方法, 其中, 所述方法还包括:

正交分解所述第一虚拟角色的可执行动作, 得到 n 个所述动作类型, 不同动作类型中包括至少两个可执行子动作, 不同动作类型下的子动作之间支持独立控制。

9. 根据权利要求 1 至 8 任一所述的方法, 其中, 所述方法还包括:

获取样本状态信息, 所述样本状态信息用于表征所述第一虚拟角色所处训练对局的对局状态;

基于所述样本状态信息, 通过强化学习的方式训练所述动作决策模型。

10. 根据权利要求 9 所述的方法, 其中, 所述基于所述样本状态信息, 通过强化学习的方式训练所述动作决策模型, 包括:

将所述样本状态信息输入所述动作决策模型, 得到所述动作决策模型输出的样本目标动作, 所述样本目标动作由 n 个样本目标子动作构成;

控制所述第一虚拟角色执行所述样本目标动作, 得到所述第一虚拟角色的样本动作执行结果;

基于所述第一虚拟角色的所述样本动作执行结果, 确定动作执行奖励;

基于所述动作执行奖励更新所述动作决策模型的模型参数。

11. 根据权利要求 10 所述的方法, 其中, 所述信息处理层中还包括对局信息编码器, 所述样本状态信息中还包括样本对局信息, 所述样本对局信息用于表征所述第一虚拟角色所处的所述训练对局的实时情况;

所述方法还包括:

在将所述样本状态信息输入所述动作决策模型的情况下, 通过所述对局信息编码器对所述样本对局信息编码, 得到对局状态编码;

所述基于所述第一虚拟角色的所述样本动作执行结果, 确定动作执行奖励, 包括:

基于所述对局状态编码以及所述样本动作执行结果, 确定所述动作执行奖励。

12. 根据权利要求 10 所述的方法, 其中, 所述动作执行奖励包括属性奖励、对局胜利奖励、对局失败奖励以及任务奖励中的至少一项, 所述对局胜利奖励为正向奖励, 所述对局失败奖励为反向奖励;

所述基于所述第一虚拟角色的所述样本动作执行结果, 确定动作执行奖励, 包括:

在所述第一虚拟角色取得对局胜利的情况下, 将所述对局胜利奖励确定为所述动作执行奖励;

在所述第一虚拟角色取得对局失败的情况下, 将所述对局失败奖励确定为所述动作执行奖励;

在所述第一虚拟角色的属性值降低的情况下, 将所述属性奖励确定为所述动作执行奖励;

在所述第一虚拟角色的任务成功的情况下, 将所述任务奖励确定为所述动作执行奖励。

13. 根据权利要求 11 所述的方法, 其中, 所述动作执行奖励包括拟人化奖励;

所述基于所述第一虚拟角色的所述样本动作执行结果, 确定动作执行奖励, 包括:

基于所述第一虚拟角色在所述训练对局中执行至少两次所述样本目标动作的样本动作执行结果, 确定至少两次所述样本目标动作的拟人化属性值;

在所述拟人化属性值低于拟人属性阈值的情况下, 将所述拟人化奖励确定为所述动作执行奖励。

14. 根据权利要求 9 所述的方法, 其中, 所述方法还包括:

在所述训练对局结束的情况下, 确定所述动作决策模型完成第 i 轮训练;

评估完成第 i 轮训练的所述动作决策模型, 得到性能评估结果;

在所述性能评估结果指示所述动作决策模型未达到训练完成标准的情况下, 控制所述动作决策模型进入第 $i+1$ 轮训练;

在所述性能评估结果指示所述动作决策模型达到训练完成标准的情况下, 确定所述动作决策模型完成训练。

15. 根据权利要求 14 所述的方法, 其中, 所述第一虚拟角色所处的所述训练对局中存在至少一个第二虚拟角色, 且第二虚拟角色执行第二动作决策模型输出的动作;

所述方法还包括:

在所述第一虚拟角色对应的第一动作决策模型完成所述第 i 轮训练的情况下, 基于训练后的所述第一动作决策模型, 控制所述第二动作决策模型进入第 i 轮训练, 并将完成第 i 轮训练的所述第二动作决策模型存储至评估模型池;

所述评估完成第 i 轮训练的所述动作决策模型, 得到性能评估结果, 包括:

从所述评估模型池中选择训练过的至少两个对照模型;

创建至少两个评估对局, 所述评估对局中包括执行所述动作决策模型输出动作的所述第一虚拟角色, 以及执行所述对照模型输出动作的一个第二虚拟角色;

根据所述至少两个所述评估对局的对局数据，确定所述性能评估结果。

16. 根据权利要求9所述的方法，其中，所述方法还包括：

在完成对所述动作决策模型的第*i*轮训练的情况下，获取执行所述第*i*轮训练的所述训练对局的对局数据；

基于至少两轮训练的所述训练对局的所述对局数据，确定对局指标以及拟人化指标，所述拟人化指标是指特定动作在至少两轮所述训练对局中的实际执行占比；

在所述对局指标达到训练完成标准，且所述拟人化指标指示的实际执行占比与目标执行占比相匹配的情况下，确定所述动作决策模型完成训练，所述目标执行占比是真实玩家在真实对局中执行所述特征动作的占比。

17. 一种虚拟角色的动作决策装置，所述装置包括：

获取模块，用于获取状态信息，所述状态信息用于表征第一虚拟角色所处对局的对局状态；

决策模块，用于将所述状态信息输入动作决策模型，得到所述动作决策模型中*n*个动作输出头串行输出的*n*个目标子动作，其中，不同动作输出头对应不同动作类型，且*n*个所述动作输出头基于所述动作类型之间的依赖关系串行连接，所述依赖关系用于表征不同动作类型下子动作之间的依赖限制情况，*n*为正整数；

控制模块，用于控制所述第一虚拟角色执行由*n*个所述目标子动作构成的目标动作。

18. 一种计算机设备，所述计算机设备包括处理器和存储器，所述存储器中存储有至少一条计算机指令，所述至少一条计算机指令由所述处理器加载并执行以实现如权利要求1至16任一所述的虚拟角色的动作决策方法。

19. 一种计算机可读存储介质，所述可读存储介质中存储有至少一条计算机指令，所述至少一条计算机指令由处理器加载并执行以实现如权利要求1至16任一所述的虚拟角色的动作决策方法。

20. 一种计算机程序产品，所述计算机程序产品包括计算机指令，所述计算机指令存储在计算机可读存储介质中；计算机设备的处理器从所述计算机可读存储介质读取所述计算机指令，所述处理器执行所述计算机指令，使得所述计算机设备执行如权利要求1至16任一所述的虚拟角色的动作决策方法。