

语音回复方法、设备、系统、存储介质及产品

5 本公开要求于 2024 年 09 月 30 日提交中国专利局、申请号为 202411397888.7、申请名称为“语音回复方法、设备、系统、存储介质及产品”的中国专利申请的优先权，其全部内容通过引用结合在本公开中。

技术领域

10 本公开实施例涉及计算机技术领域，尤其涉及一种语音回复方法、设备、系统、存储介质及产品。

背景技术

15 智能对话是以一问一答的模式实现用户与智能设备之间的交互，其中，智能对话广泛应用于语音助手、智能客服等场景。

 目前，智能对话主要采用终端设备实现，例如，用户通过终端设备的界面输入语音信息或者文字信息，终端设备输出对应的回复信息，该方式存在对话效率低的问题。

20 发明内容

 本公开实施例提供一种语音回复方法、装置、设备、计算机可读存储介质及产品，用于解决现有技术采用终端设备实现智能对话存在效率低的技术问题。

 第一方面，本公开实施例提供一种语音回复方法，应用于携带摄像头的耳机设备，方法包括：

25 采集第一问询语音信息；
 获取摄像头采集的目标图像；
 得到第一问询语音信息的第一回复语音信息，其中，第一回复语音信息是与目标图像相关的信息，目标图像和第一问询语音信息被预先训练的对话模型进行处理以得到第一回复语音信息；
30 播放第一回复语音信息。

 第二方面，本申请实施例提供一种耳机设备，包括：耳机设备携带摄像头，耳机

设备包括：

采集单元，用于采集第一问询语音信息；

获取单元，用于获取摄像头采集的目标图像；

5 处理单元，用于得到第一问询语音信息的第一回复语音信息，其中，第一回复语音信息是与目标图像相关的信息，目标图像和第一问询语音信息被预先训练的对话模型进行处理以得到第一回复语音信息；

播放单元，用于播放第一回复语音信息。

第三方面，本申请实施例提供一种语音回复系统，包括第二方面的耳机设备和终端设备，其中：

10 耳机设备，用于采集第一问询语音信息；获取摄像头采集的目标图像，向终端设备发送目标图像和第一问询语音信息；

终端设备，用于通过预先训练的对话模型对目标图像和第一问询语音信息进行处理，得到第一问询语音信息的第一回复语音信息，向耳机设备发送第一回复语音信息；

15 耳机设备，用于接收终端设备发送的第一回复语音信息，并播放第一回复语音信息。

第四方面，本公开实施例提供一种电子设备，包括：处理器和存储器；

存储器存储计算机执行指令；

20 处理器执行存储器存储的计算机执行指令，使得至少一个处理器执行如上第一方面以及第一方面各种可能的语音回复方法。

第五方面，本公开实施例提供一种计算机可读存储介质，计算机可读存储介质中存储有计算机执行指令，当处理器执行计算机执行指令时，实现如上第一方面以及第一方面各种可能的语音回复方法。

25 第六方面，本公开实施例提供一种计算机程序产品，包括计算机程序，计算机程序被处理器执行时实现如上第一方面以及第一方面各种可能的语音回复方法。

30 本实施例提供的语音回复方法、设备、系统、存储介质及产品，包括：采集第一问询语音信息；获取摄像头采集的目标图像；得到第一问询语音信息的第一回复语音信息，其中，第一回复语音信息是与目标图像相关的信息，目标图像和第一问询语音信息被预先训练的对话模型进行处理以得到第一回复语音信息；播放第一回复语音信息。首先，本申请采用耳机实现智能对话，提高了智能对话的效率。此外，本申请采用对话模型能够快速确定第一回复语音信息。再者，通过获取目标图像以回复第一问询语音信息，可以实现对用户针对实时的环境提出的问题的回复。

附图说明

为了更清楚地说明本公开实施例或现有技术中的技术方案，下面将对实施例或现有技术描述中所需要使用的附图作一简单地介绍，显而易见地，下面描述中的附图是本公开的一些实施例，对于本领域普通技术人员来讲，在不付出创造性劳动性的前提下，还可以根据这些附图获得其他的附图。

- 图 1 为本申请实施例提供的一种语音回复方法的应用场景示例图；
- 图 2 为本申请实施例提供的一种语音回复方法的一个流程图；
- 图 3 为本申请实施例提供的另一种语音回复方法的一个流程图；
- 图 4 为本申请实施例提供的一种耳机设备的结构示意图；
- 图 5 为本公开实施例提供的电子设备的结构示意图。

具体实施方式

为使本公开实施例的目的、技术方案和优点更加清楚，下面将结合本公开实施例中的附图，对本公开实施例中的技术方案进行清楚、完整地描述，显然，所描述的实施例是本公开一部分实施例，而不是全部的实施例。基于本公开中的实施例，本领域普通技术人员在没有作出创造性劳动前提下所获得的所有其他实施例，都属于本公开保护的范围。

可以理解的是，在使用本公开各实施例公开的技术方案之前，均应当依据相关法律法规通过恰当的方式对本公开所涉及个人信息类型、使用范围、使用场景等告知目标用户并获得目标用户的授权。

可以理解的是，上述通知和获取目标用户授权过程仅是示意性的，不对本公开的实现方式构成限定，其它满足相关法律法规的方式也可应用于本公开的实现方式中。

在相关技术中，用户若需要问询周围环境中对象的相关信息，需要采用带有摄像头的移动终端拍摄该对象，得到图像，然后采用移动终端中的相关识别应用扫描该图像，在移动终端的界面上显示出该对象的相关信息，若用户需要了解该对象进一步的其他信息时，需要基于相关信息在应用中搜索得到，例如，用户处于当前环境中时，看到一束花，用户并不知道该花的名称，用户用移动终端拍摄该花，得到包含花的图像，然后用户采用移动终端上的识别应用扫描该图像，得到了该花的名称，然后用户需要知道该花的花期、种植地等地方，用户需要在应用中输入针对该花的问题，来得到相应的答案。可见，获取对象的相关信息的过程复杂，效率较低。

为了解决的技术问题，本公开提供了一种语音回复方法，通过在耳机设备上部署摄像头，控制摄像头实时获取目标图像，并且耳机设备采集第一问询语音信息，针对目标图像和第一问询语音信息得到第一回复语音信息，通过耳机设备播放该第一回复

语音信息，可以使用户快速简单的获取到想要的信息，提高对话效率。

需要说明的是，本公开提供的语音回复方法、设备、系统、存储介质及产品可应用在任意一种配置有摄像头的耳机设备中。

下面将结合附图详细介绍本申请实施例提供的语音回复方法。

5 图 1 为本申请实施例提供的一种语音回复方法的应用场景示例图。如图 1 所示，用户佩戴着耳机设备，该耳机设备包括：第一耳机和第二耳机，用户的一个耳朵佩戴着第一耳机，另一个耳朵佩戴着第二耳机。其中，耳机设备携带摄像头，在本申请实施例中，可以是第一耳机和/或第二耳机携带摄像头。此外，第一耳机和/或第二耳机携带麦克风，麦克风用于采集用户的问询语音信息。可以理解，本申请中耳机设备携带的摄像头采集目标图像，麦克风采集用户的问询语音信息，根据目标图像和用户的问询语音信息可以生成回复语音信息，该回复语音信息可以通过耳机设备播放，整个对话过程简单高效，提高了用户体验。

10 图 1 只是一种示例性的应用场景，可用于任意配置摄像头的耳机设备中，例如：入耳式耳机、半入耳式耳机、头戴式耳机、有线耳机、TWS（True Wireless Stereo，真无线耳机）等，本申请不限定耳机设备的类型。

图 2 为本申请实施例提供的一种语音回复方法的流程图，该语音回复方法可以包括以下几个步骤：

S201、采集第一问询语音信息。

20 其中，第一问询语音信息是耳机设备的麦克风采集的。可以理解，耳机设备的一个耳机（第一耳机或第二耳机）携带麦克风，该麦克风可以实时采集用户的第一问询语音信息。

可以理解，采集的第一问询语音信息是需要被预设应用采用对话模型处理的。一种实施例中，预设应用已经被唤醒，则第一问询语音信息为用户输入的任意的问询语音信息，例如“前面有什么？”，“这个花叫什么”，“这个花属于什么品种”或
25 “这个花有什么颜色”等等。

另一种实施例中，预设应用未被唤醒，第一问询语音信息包括：预设应用的预设唤醒词和问询语音信息，例如“XX，前面有什么”，“XX，XX，前面有什么”，“XX，看一下前面有什么”。

其中“XX”为预设应用的预设唤醒词。

30 S202、获取摄像头采集的目标图像。

其中，摄像头采集至少一张目标图像或者一系列连续的目标图像。

一种实施例中，获取摄像头采集的目标图像之前，还包括：响应对摄像头的唤醒操作，唤醒摄像头；获取摄像头采集的目标图像，包括：控制已唤醒的摄像头采集目标图像。

可以理解，在摄像头已经被唤醒的情况下，若采集到第一问询语音信息，则在采集到第一问询语音信息后，便控制摄像头采集目标图像。

其中，可以通过唤醒词或者触控耳机设备的方式唤醒摄像头。该摄像头的唤醒词可以和预设应用的唤醒词是相同的。例如，唤醒词如“XX”、或者“XX，看看”，
5 “XX，扫描”，“XX，再看一下”，“XX，再看一眼”，“XX，仔细看看”。

另一种实施例中，获取摄像头采集的目标图像，包括：通过识别模型确定第一问询语音信息的拍摄意图；根据拍摄意图控制摄像头采集目标图像。

可以理解，在采集到第一问询语音信息后，先识别第一问询语音信息中有摄像头的唤醒词，则唤醒摄像头，通过识别模型分析第一问询语音信息的拍摄意图，例如，
10 第一问询语音信息为“这个花的名称是什么”，则拍摄意图是拍摄花，然后可以控制摄像头对焦至花进行拍摄得到目标图像。其中，通过根据拍摄意图控制摄像头采集目标图像，可以得到准确的目标图像，进而可以用于后续准确的回复用户。

其中，该识别模型是部署在耳机设备上的网络模型。

在本申请实施例中，耳机设备包括：第一耳机和第二耳机，第一耳机携带第一摄像头，第二耳机携带第二摄像头，目标图像包括：第一摄像头采集的目标图像和第二摄像头采集的目标图像。
15

可以理解，一种实施例中，可以只在一个耳机设置至少一个摄像头，例如，只在第一耳机设置第一摄像头，则只需要第一摄像头采集目标图像。另一实施例中，可以在两个耳机均设置摄像头，则两个耳机的摄像头均采集目标图像。

其中，第一摄像头采集的目标图像和第二摄像头采集的目标图像是同一对象不同的角度，这样用于后续的第一回复语音信息的确定，可以提高第一回复语音信息的回复准确度。
20

S203、得到第一问询语音信息的第一回复语音信息。

其中，第一回复语音信息是与目标图像相关的信息，目标图像和第一问询语音信息被预先训练的对话模型进行处理以得到第一回复语音信息。
25

其中，对话模型为LLM（Large Language Model，大型语言模型），对话模型为预先训练好的多模态语言模型，可以对目标图像以及自然语言的第一问询语音信息处理，得到第一回复语音信息。其中，第一回复语音信息是与目标图像相关的信息。例如，若第一问询语音信息为“前面有什么？”，目标图像中包括的对应有“花”，则
30 第一回复语音信息可以是“前面有盆花”。若第一问询语音信息为“这个花是什么？”，目标图像包括“月季花”，则第一回复语音信息可以是“这是月季花”。

一种实施例中，耳机设备中部署对话模型，得到第一问询语音信息的第一回复语音信息包括：将目标图像和第一问询语音信息输入预先训练的对话模型进行处理，得到第一问询语音信息的第一回复语音信息。

可以理解，在耳机设备侧部署有对话模型，则耳机设备在采集到第一问询语音信息和目标图像后，耳机设备将第一问询语音信息和目标图像输入对话模型进行处理，得到第一问询语音信息的第一回复语音信息。

在本申请实施例中，对话模型输入的可以是语音格式的第一问询语音信息，也可以是5 将第一问询语音信息通过 STT（Speech To Text，语音到文本）技术转换成第一问询文本信息，将第一问询文本信息输入对话模型，对话模型还可以输出第一回复文本信息，耳机设备采用 TTS（Text To Speech，文本到语音）技术将第一回复文本信息从文本格式转换成语音格式，得到第一回复语音信息。

在本申请实施例中，还可以是在耳机设备侧安装有对话应用，在云端服务器中部署10 对话模型，耳机设备可以通过对话应用将目标图像和第一问询语音信息上传至云端服务的对话模型进行处理，得到第一回复语音信息。

另一种实施例中，得到第一问询语音信息的第一回复语音信息包括：向终端设备发送目标图像和第一问询语音信息，终端设备用于通过预先训练的对话模型对目标图像和第一问询语音信息进行处理，得到第一问询语音信息的第一回复语音信息；接收15 终端设备发送的第一回复语音信息。

可以理解，对话模型是部署在终端设备侧，则耳机设备将采集的目标图像和第一问询语音信息发送给终端设备，终端设备通过对话模型处理目标图像和第一问询语音信息，得到第一回复语音信息。

在本申请实施例中，还可以是在终端设备侧安装有对话应用，在云端服务器中部署20 对话模型，终端设备可以通过对话应用将目标图像和第一问询语音信息上传至云端服务的对话模型进行处理，得到第一回复语音信息，然后将第一回复语音信息发送至耳机设备进行播放。

综上，本申请可以采用多种方式，通过对话模型处理目标图像和第一问询语音信息得到第一回复语音信息。

25 S204、播放第一回复语音信息。

在本申请实施例中，可以在一个耳机（第一耳机或第二耳机）播放第一回复语音信息，也可以在第一耳机和第二耳机均播放第一回复语音信息。

综上，本申请通过在耳机设备上部署摄像头，控制摄像头实时获取目标图像，并且耳机设备采集第一问询语音信息，针对目标图像和第一问询语音信息得到第一回复30 语音信息，通过耳机设备播放该第一回复语音信息，可以使用户快速简单的获取到想要的信息，提高对话效率。

图 3 为本申请实施例提供的另一种语音回复方法的流程图，该语音回复方法可以包括以下几个步骤：

S301，采集第一问询语音信息。

该步骤的具体实现过程参照 S201，在此不再赘述。

S302，获取摄像头采集的目标图像。

该步骤的具体实现过程参照 S202，在此不再赘述。

S303，向终端设备发送目标图像和第一问询语音信息。

5 在本申请实施例中，耳机设备和终端设备通信连接，其中，耳机设备通过蓝牙或者 WIFI（移动热点）向终端设备发送目标图像和第一问询语音信息，在耳机设备通过 WIFI 向终端设备发送目标图像和第一问询语音信息时，耳机设备可以连接终端设备的热点。

10 其中，用户可以设置采用单侧耳机（第一耳机或第二耳机）的摄像头拍摄目标图像，用户也可以设置双侧耳机（第一耳机和第二耳机）的摄像头均拍摄目标图像。

一种实施例中，向终端设备发送目标图像，包括：采用第一耳机向终端设备发送第一摄像头采集的目标图像；采用第二耳机向终端设备发送第二摄像头采集的目标图像，在设置双侧耳机的摄像头均拍摄目标图像时，可以设置一个耳机（例如：第一耳机）为主耳机，另一个耳机（例如：第二耳机）为从耳机。

15 可以理解，第一耳机和终端设备通信连接，第二耳机和终端设备通信连接。则第一耳机和第二耳机可以单独向终端设备发送目标图像。

20 其中，若第一耳机侧携带麦克风，则第一耳机向终端设备发送第一摄像头采集的目标图像和第一耳机采集的第一问询语音信息，第二耳机向终端设备发送第二摄像头采集的目标图像。其中，若第二耳机侧携带麦克风，则第一耳机向终端设备发送第一摄像头采集的目标图像，第二耳机向终端设备发送第二摄像头采集的目标图像和第二耳机采集的第二问询语音信息。

在本申请实施例中，第一耳机和第二耳机单独和终端设备通信，不需要考虑耳机之间的连接，其中，携带麦克风的耳机为主耳机。

25 另一种实施例中，向终端设备发送目标图像，包括：第一耳机接收第二耳机发送的第二摄像头采集的目标图像；采用第一耳机向终端设备发送第一摄像头采集的目标图像和第二摄像头采集的目标图像。

30 可以理解，第一耳机和终端设备通信连接，第二耳机和终端设备无通信连接，第一耳机和第二耳机之间通信连接，该种实施例中，第一耳机为主耳机，第二耳机为从耳机。其中，第一耳机和第二耳机通过蓝牙通信连接。在该种情况下，第二耳机把第二摄像头采集的目标图像发送给第一耳机，第一耳机向终端设备发送第一摄像头采集的目标图像和第二摄像头采集的目标图像。

进一步地，在第一耳机和终端设备通信连接，第二耳机和终端设备无通信连接，第一耳机和第二耳机之间通信连接的情况下，若第一耳机携带麦克风，则第一耳机采集第一问询语音信息，第一耳机向终端设备发送第一问询语音信息。若第二耳机携带

麦克风，则第二耳机采集第一问询语音信息，第二耳机向第一耳机发送第一问询语音信息，第一耳机向终端设备发送第一问询语音信息。

在本申请实施例中，第一耳机接收到第二耳机发送的目标图像后，再向终端设备发送目标图像，只需要一个耳机和终端设备建立通信连接。一个耳机到终端设备传输数据的链路较短。

进一步地，本申请中，可以先向终端设备发送主耳机的摄像头拍摄的目标图像，再向终端设备发送从耳机的摄像头拍摄的目标图像，也可以同时发送，在此不加以限定。

此外，第一问询语音信息在编码后再发送至终端设备，例如，采用 OPUS 编码（一种声音编码方式）方式对第一问询语音信息编码。在终端设备接收到编码后的第一问询语音信息后，采用 OPUS 解码（一种声音解码方式）方式对第一问询语音信息解码后，输入问答模型进行处理。

S304，接收终端设备发送的第一回复语音信息。

一种实施例中，采用第一耳机接收终端设备发送的第一回复语音信息；第一耳机向第二耳机转发第一回复语音信息。

其中，在只有第一耳机和终端设备通信连接，第一耳机和第二耳机通信连接的情况下，终端设备将第一回复语音信息发送给第一耳机，第一耳机再将第一回复语音信息转发给第二耳机。

另一种实施例中，采用第一耳机接收终端设备发送的第一回复语音信息；采用第二耳机接收终端设备发送的第一回复语音信息。

其中，在第一耳机和第二耳机分别和终端设备建立了通信连接的情况下，终端设备将第一回复语音信息分别发送给第一耳机和第二耳机播放。

S305，播放第一回复语音信息。

在本申请实施例中，终端设备可以对第一回复语音信息编码后发送给耳机设备。在耳机设备接收到编码后的第一回复语音信息后，解码后播放第一回复语音信息。

其中，耳机设备通过喇叭播放第一回复语音信息。

S306，采集第二问询语音信息。

可以理解，用户可以针对耳机设备播放的第一回复语音信息继续询问，实现多轮对话。

S307，得到第二问询语音信息的第二回复语音信息。

其中，第二回复语音信息是与目标图像相关的信息，目标图像和第二问询语音信息被对话模型进行处理以得到第二回复语音信息。

在本申请实施例中，用户可以针对目标图像实现轮询对话，例如，若第一问询语音信息为“前面有什么？”，目标图像中包括的对应有“花”，则第一回复语音信息

可以是“前面有盆花”，第二问询语音信息为“这个花的名称”，然后将第二问询语音信息和目标图像输入对话模型进行处理，得到第二回复语音信息。

其中，第二回复语音信息的确定方式参照第一回复语音信息，在此不再赘述。

S308，播放第二回复语音信息。

- 5 一种实施例中，在针对目标图像的多轮对话结束后，耳机设备还可以采集其他问询语音信息和其他图像开启新一轮的对话。

在本申请实施例中，通过耳机设备与终端设备的通信连接，可以赋予耳机设备智能对话的功能，提高了实现智能对话的效率，提升用户体验。

- 10 图4为本申请实施例提供的一种耳机设备的结构示意图，耳机设备携带摄像头，该耳机设备40可以包括以下几个单元：

采集单元41，用于采集第一问询语音信息；

获取单元42，用于获取摄像头采集的目标图像；

- 15 处理单元43，用于得到第一问询语音信息的第一回复语音信息，其中，第一回复语音信息是与目标图像相关的信息，目标图像和第一问询语音信息被预先训练的对话模型进行处理以得到第一回复语音信息；

播放单元44，用于播放第一回复语音信息。

在一些实施例中，还包括：唤醒单元（未示出），用于在获取摄像头采集的目标图像之前，响应对摄像头的唤醒操作，唤醒摄像头；

获取单元42，具体用于控制已唤醒的摄像头采集目标图像。

- 20 在一些实施例中，获取单元42，具体用于通过识别模型确定第一问询语音信息的拍摄意图；根据拍摄意图控制摄像头采集目标图像。

在一些实施例中，耳机设备中部署对话模型，处理单元43，具体用于将目标图像和第一问询语音信息输入预先训练的对话模型进行处理，得到第一问询语音信息的第一回复语音信息。

- 25 在一些实施例中，处理单元43，具体用于向终端设备发送目标图像和第一问询语音信息，终端设备用于通过预先训练的对话模型对目标图像和第一问询语音信息进行处理，得到第一问询语音信息的第一回复语音信息；接收终端设备发送的第一回复语音信息。

- 30 在一些实施例中，耳机设备包括：第一耳机和第二耳机，第一耳机携带第一摄像头，第二耳机携带第二摄像头，目标图像包括：第一摄像头采集的目标图像和第二摄像头采集的目标图像。

在一些实施例中，处理单元43在向终端设备发送目标图像时，具体用于：采用第一耳机向终端设备发送第一摄像头采集的目标图像；采用第二耳机向终端设备发送第二摄像头采集的目标图像。

在一些实施例中，处理单元 43 在向终端设备发送目标图像时，具体用于：第一耳机接收第二耳机发送的第二摄像头采集的目标图像；采用第一耳机向终端设备发送第一摄像头采集的目标图像和第二摄像头采集的目标图像。

5 在一些实施例中，处理单元 43 在接收终端设备发送的第一回复语音信息时，具体用于：采用第一耳机接收终端设备发送的第一回复语音信息；第一耳机向第二耳机转发第一回复语音信息。

在一些实施例中，处理单元 43 在接收终端设备发送的第一回复语音信息时，具体用于：采用第一耳机接收终端设备发送的第一回复语音信息；采用第二耳机接收终端设备发送的第一回复语音信息。

10 在一些实施例中，播放第一回复语音信息之后，采集单元 41，还用于采集第二问询语音信息；

处理单元 43，还用于得到第二问询语音信息的第二回复语音信息，其中，第二回复语音信息是与目标图像相关的信息，目标图像和第二问询语音信息被对话模型进行处理以得到第二回复语音信息；

15 播放单元 44，还用于播放第二回复语音信息。

为了实现上述实施例，本公开实施例还提供了一种语音回复系统，包括上述的耳机设备和终端设备。

其中：

20 耳机设备，用于采集第一问询语音信息；获取摄像头采集的目标图像，向终端设备发送目标图像和第一问询语音信息；

终端设备，用于通过预先训练的对话模型对目标图像和第一问询语音信息进行处理，得到第一问询语音信息的第一回复语音信息，向耳机设备发送第一回复语音信息；

25 耳机设备，用于接收终端设备发送的第一回复语音信息，并播放第一回复语音信息。

为了实现上述实施例，本公开实施例还提供了一种计算机可读存储介质，计算机可读存储介质中存储有计算机执行指令，当处理器执行计算机执行指令时，实现如上述任一实施例的语音回复方法。

30 为了实现上述实施例，本公开实施例还提供了一种计算机程序产品，包括计算机程序，计算机程序被处理器执行时实现如上述任一实施例的语音回复方法。

为了实现上述实施例，本公开实施例还提供了一种电子设备，包括：处理器和存储器；

存储器存储计算机执行指令；

处理器执行存储器存储的计算机执行指令，使得处理器执行如上述任一实施例的

语音回复方法。

图 5 为本公开实施例提供的电子设备的结构示意图，该电子设备 50 可以为终端设备或服务器。其中，终端设备可以包括但不限于诸如移动电话、笔记本电脑、数字广播接收器、个人数字助理（Personal Digital Assistant，简称 PDA）、平板电脑

5 （Portable Android Device，简称 PAD）、便携式多媒体播放器（Portable Media Player，简称 PMP）、车载终端（例如车载导航终端）等等的移动终端以及诸如数字 TV、台式计算机等等的固定终端。图 5 出的电子设备仅仅是一个示例，不应对本公开实施例的功能和使用范围带来任何限制。

10 如图 5 所示，电子设备 50 可以包括处理装置（例如中央处理器、图形处理器等）51，其可以根据存储在只读存储器（Read Only Memory，简称 ROM）52 中的程序或者从存储装置 58 加载到随机访问存储器（Random Access Memory，简称 RAM）53 中的程序而执行各种适当的动作和处理。在 RAM 53 中，还存储有电子设备 50 操作所需的各种程序和数据。处理装置 51、ROM 52 以及 RAM 53 通过总线 54 彼此相连。输入/输出（I/O）接口 55 也连接至总线 54。

15 通常，以下装置可以连接至 I/O 接口 55：包括例如触摸屏、触摸板、键盘、鼠标、摄像头、麦克风、加速度计、陀螺仪等的输入装置 56；包括例如液晶显示器（Liquid Crystal Display，简称 LCD）、扬声器、振动器等输出装置 57；包括例如磁带、硬盘等的存储装置 58；以及通信装置 59。通信装置 59 可以允许电子设备 50 与其他设备进行无线或有线通信以交换数据。虽然图 5 示出了具有各种装置的电子设备 50，但是应理解的是，并不要求实施或具备所有示出的装置。可以替代地实施或具备更多或更少的装置。

特别地，根据本公开的实施例，上文参考流程图描述的过程可以被实现为计算机软件程序。例如，本公开的实施例包括一种计算机程序产品，其包括承载在计算机可读介质上的计算机程序，该计算机程序包含用于执行流程图所示的方法的程序代码。

25 在这样的实施例中，该计算机程序可以通过通信装置 59 从网络上被下载和安装，或者从存储装置 58 被安装，或者从 ROM 52 被安装。在该计算机程序被处理装置 51 执行时，执行本公开实施例的方法中限定的上述功能。

需要说明的是，本公开上述的计算机可读介质可以是计算机可读信号介质或者计算机可读存储介质或者是上述两者的任意组合。计算机可读存储介质例如可以是——

30 但不限于——电、磁、光、电磁、红外线、或半导体的系统、装置或器件，或者任意以上的组合。计算机可读存储介质的更具体的例子可以包括但不限于：具有一个或多个导线的电连接、便携式计算机磁盘、硬盘、随机访问存储器（RAM）、只读存储器（ROM）、可擦式可编程只读存储器（EPROM 或闪存）、光纤、便携式紧凑磁盘只读存储器（CD-ROM）、光存储器件、磁存储器件、或者上述的任意合适的组合。

在本公开中，计算机可读存储介质可以是任何包含或存储程序的有形介质，该程序可以被指令执行系统、装置或者器件使用或者与其结合使用。而在本公开中，计算机可读信号介质可以包括在基带中或者作为载波一部分传播的数据信号，其中承载了计算机可读的程序代码。这种传播的数据信号可以采用多种形式，包括但不限于电磁信

5 号、光信号或上述的任意合适的组合。计算机可读信号介质还可以是计算机可读存储介质以外的任何计算机可读介质，该计算机可读信号介质可以发送、传播或者传输用于由指令执行系统、装置或者器件使用或者与其结合使用的程序。计算机可读介质上包含的程序代码可以用任何适当的介质传输，包括但不限于：电线、光缆、RF（射频）等等，或者上述的任意合适的组合。

10 上述计算机可读介质可以是上述电子设备中所包含的；也可以是单独存在，而未装配入该电子设备中。

上述计算机可读介质承载有一个或者多个程序，当上述一个或者多个程序被该电子设备执行时，使得该电子设备执行上述实施例所示的方法。

可以以一种或多种程序设计语言或其组合来编写用于执行本公开的操作的计算机程序代码，上述程序设计语言包括面向对象的程序设计语言—诸如 Java、Smalltalk、15 C++，还包括常规的过程式程序设计语言—诸如“C”语言或类似的设计语言。程序代码可以完全地在目标用户计算机上执行、部分地在目标用户计算机上执行、作为一个独立的软件包执行、部分在目标用户计算机上部分在远程计算机上执行、或者完全在远程计算机或服务器上执行。在涉及远程计算机的情形中，远程计算机可以通过任意种类的网络——包括局域网（Local Area Network，简称 LAN）或广域网 20 （Wide Area Network，简称 WAN）—连接到目标用户计算机，或者，可以连接到外部计算机（例如利用因特网服务提供商来通过因特网连接）。

附图中的流程图和框图，图示了按照本公开各种实施例的系统、方法和计算机程序产品的可能实现的体系架构、功能和操作。在这点上，流程图或框图中的每个方框 25 可以代表一个模块、程序段、或代码的一部分，该模块、程序段、或代码的一部分包含一个或多个用于实现规定的逻辑功能的可执行指令。也应当注意，在有些作为替换的实现中，方框中所标注的功能也可以以不同于附图中所标注的顺序发生。例如，两个接连地表示的方框实际上可以基本并行地执行，它们有时也可以按相反的顺序执行，这依所涉及的功能而定。也要注意的，框图和/或流程图中的每个方框、以及 30 框图和/或流程图中的方框的组合，可以用执行规定的功能或操作的专用的基于硬件的系统来实现，或者可以用专用硬件与计算机指令的组合来实现。

描述于本公开实施例中所涉及到的单元可以通过软件的方式实现，也可以通过硬件的方式来实现。其中，单元的名称在某种情况下并不构成对该单元本身的限定，例如，第一获取单元还可以被描述为“获取至少两个网际协议地址的单元”。

本文中以上描述的功能可以至少部分地由一个或多个硬件逻辑部件来执行。例如，非限制性地，可以使用的示范类型的硬件逻辑部件包括：现场可编程门阵列（FPGA）、专用集成电路（ASIC）、专用标准产品（ASSP）、片上系统（SOC）、复杂可编程逻辑设备（CPLD）等等。

5 在本公开的上下文中，机器可读介质可以是有形的介质，其可以包含或存储以供指令执行系统、装置或设备使用或与指令执行系统、装置或设备结合地使用的程序。机器可读介质可以是机器可读信号介质或机器可读储存介质。机器可读介质可以包括但不限于电子的、磁性的、光学的、电磁的、红外的、或半导体系统、装置或设备，或者上述内容的任何合适组合。机器可读存储介质的更具体示例会包括基于一个或多个
10 个线的电气连接、便携式计算机盘、硬盘、随机存取存储器（RAM）、只读存储器（ROM）、可擦除可编程只读存储器（EPROM 或快闪存储器）、光纤、便捷式紧凑盘只读存储器（CD-ROM）、光学储存设备、磁储存设备、或上述内容的任何合适组合。

第一方面，根据本公开的一个或多个实施例，提供了一种语音回复方法，应用于
15 携带摄像头的耳机设备，方法包括：采集第一问询语音信息；获取摄像头采集的目标图像；得到第一问询语音信息的第一回复语音信息，其中，第一回复语音信息是与目标图像相关的信息，目标图像和第一问询语音信息被预先训练的对话模型进行处理以得到第一回复语音信息；播放第一回复语音信息。

根据本公开的一个或多个实施例，获取摄像头采集的目标图像之前，还包括：响
20 应对摄像头的唤醒操作，唤醒摄像头；获取摄像头采集的目标图像，包括：控制已唤醒的摄像头采集目标图像。

根据本公开的一个或多个实施例，获取摄像头采集的目标图像，包括：通过识别模型确定第一问询语音信息的拍摄意图；根据拍摄意图控制摄像头采集目标图像。

根据本公开的一个或多个实施例，耳机设备中部署对话模型，得到第一问询语音
25 信息的第一回复语音信息包括：将目标图像和第一问询语音信息输入预先训练的对话模型进行处理，得到第一问询语音信息的第一回复语音信息。

根据本公开的一个或多个实施例，得到第一问询语音信息的第一回复语音信息包
括：向终端设备发送目标图像和第一问询语音信息，终端设备用于通过预先训练的对话模型对目标图像和第一问询语音信息进行处理，得到第一问询语音信息的第一回复
30 语音信息；接收终端设备发送的第一回复语音信息。

根据本公开的一个或多个实施例，耳机设备包括：第一耳机和第二耳机，第一耳机携带第一摄像头，第二耳机携带第二摄像头，目标图像包括：第一摄像头采集的目标图像和第二摄像头采集的目标图像。

根据本公开的一个或多个实施例，向终端设备发送目标图像，包括：采用第一耳机向终端设备发送第一摄像头采集的目标图像；采用第二耳机向终端设备发送第二摄像头采集的目标图像。

5 根据本公开的一个或多个实施例，向终端设备发送目标图像，包括：第一耳机接收第二耳机发送的第二摄像头采集的目标图像；采用第一耳机向终端设备发送第一摄像头采集的目标图像和第二摄像头采集的目标图像。

根据本公开的一个或多个实施例，接收终端设备发送的第一回复语音信息，包括：采用第一耳机接收终端设备发送的第一回复语音信息；第一耳机向第二耳机转发第一回复语音信息。

10 根据本公开的一个或多个实施例，接收终端设备发送的第一回复语音信息，包括：采用第一耳机接收终端设备发送的第一回复语音信息；采用第二耳机接收终端设备发送的第一回复语音信息。

根据本公开的一个或多个实施例，播放第一回复语音信息之后，还包括：采集第二问询语音信息；得到第二问询语音信息的第二回复语音信息，其中，第二回复语音信息是与目标图像相关的信息，目标图像和第二问询语音信息被对话模型进行处理以得到第二回复语音信息；播放第二回复语音信息。

第二方面，根据本公开的一个或多个实施例，提供了一种耳机设备，耳机设备携带摄像头，耳机设备包括：

采集单元，用于采集第一问询语音信息；

20 获取单元，用于获取摄像头采集的目标图像；

处理单元，用于得到第一问询语音信息的第一回复语音信息，其中，第一回复语音信息是与目标图像相关的信息，目标图像和第一问询语音信息被预先训练的对话模型进行处理以得到第一回复语音信息；

播放单元，用于播放第一回复语音信息。

25 第三方面，根据本公开的一个或多个实施例，提供了一种语音回复系统，包括第二方面的耳机设备和终端设备，其中：

耳机设备，用于采集第一问询语音信息；获取摄像头采集的目标图像，向终端设备发送目标图像和第一问询语音信息；

30 终端设备，用于通过预先训练的对话模型对目标图像和第一问询语音信息进行处理，得到第一问询语音信息的第一回复语音信息，向耳机设备发送第一回复语音信息；

耳机设备，用于接收终端设备发送的第一回复语音信息，并播放第一回复语音信息。

第四方面，根据本公开的一个或多个实施例，提供了一种电子设备，包括：至少

一个处理器和存储器；

存储器存储计算机执行指令；

至少一个处理器执行存储器存储的计算机执行指令，使得至少一个处理器执行如上第一方面以及第一方面各种可能的设计的语音回复方法。

5 第五方面，根据本公开的一个或多个实施例，提供了一种计算机可读存储介质，计算机可读存储介质中存储有计算机执行指令，当处理器执行计算机执行指令时，实现如上第一方面以及第一方面各种可能的设计的语音回复方法。

第六方面，根据本公开的一个或多个实施例，提供了一种计算机程序产品，包括计算机程序，计算机程序被处理器执行时实现如上第一方面以及第一方面各种可能的
10 设计的语音回复方法。

以上描述仅为本公开的较佳实施例以及对所运用技术原理的说明。本领域技术人员应当理解，本公开中所涉及的公开范围，并不限于上述技术特征的特定组合而成的技术方案，同时也应涵盖在不脱离上述公开构思的情况下，由上述技术特征或其等同特征进行任意组合而形成的其它技术方案。例如上述特征与本公开中公开的（但不限
15 于）具有类似功能的技术特征进行互相替换而形成的技术方案。

此外，虽然采用特定次序描绘了各操作，但是这不应理解为要求这些操作以所示出的特定次序或以顺序次序执行来执行。在一定环境下，多任务和并行处理可能是有利的。同样地，虽然在上面论述中包含了若干具体实现细节，但是这些不应被解释为对本公开的范围的限制。在单独的实施例的上下文中描述的某些特征还可以组合
20 地实现在单个实施例中。相反地，在单个实施例的上下文中描述的各种特征也可以单独地或以任何合适的子组合的方式实现在多个实施例中。

尽管已经采用特定于结构特征和/或方法逻辑动作的语言描述了本主题，但是应当理解所附权利要求书中所限定的主题未必局限于上面描述的特定特征或动作。相反，上面所描述的特定特征和动作仅仅是实现权利要求书的示例形式。

权 利 要 求 书

- 1、一种耳机设备，其特征在于，所述耳机设备包括第一耳机和第二耳机，所述
第一耳机携带摄像头和/或所述第二耳机携带摄像头。
- 2、根据权利要求 1 所述的耳机设备，其特征在于，所述耳机设备包括入耳式耳
5 机、半入耳式耳机、头戴式耳机、有线耳机、真无线耳机中的一种。
- 3、根据权利要求 1 所述的耳机设备，其特征在于，所述耳机设备用于采集目标
图像，所述第一耳机携带第一摄像头，所述第二耳机携带第二摄像头，所述目标图像
包括：所述第一摄像头采集的目标图像和所述第二摄像头采集的目标图像。
- 4、根据权利要求 3 所述的耳机设备，其特征在于，所述第一耳机用于向终端设
10 备发送所述第一摄像头采集的目标图像，所述第二耳机用于向终端设备发送所述第二
摄像头采集的图像。
- 5、根据权利要求 3 所述的耳机设备，其特征在于，所述第一耳机用于接收所述
第二耳机发送的第二摄像头采集的目标图像，并向终端设备发送所述第一摄像头采集
的目标图像和所述第二摄像头采集的目标图像。
- 15 6、根据权利要求 1 所述的耳机设备，其特征在于，所述耳机设备用于采集目标
图像，所述第一耳机携带第一摄像头，所述目标图像包括所述第一摄像头采集的目标
图像；或，所述第二耳机携带第二摄像头，所述目标图像包括所述第二摄像头采集的
目标图像。
- 7、根据权利要求 3 或 6 所述的耳机设备，其特征在于，所述耳机设备还包括麦
20 克风，所述麦克风能够采集第一问询语音信息，所述目标图像和所述第一问询语音信
息被所述耳机输入至预训练的对话模型进行处理。
- 8、根据权利要求 7 所述的耳机设备，其特征在于，所述耳机设备部署有所述对
话模型；或，所述耳机设备安装有对话应用，云端服务器部署有所述对话模型，所述
耳机设备通过所述对话应用将所述目标图像和所述第一问询语音信息上传至所述云端
25 服务器的对话模型进行处理。
- 9、根据权利要求 7 所述的耳机设备，其特征在于，所述耳机设备与终端设备通
信连接，所述终端设备部署有所述对话模型，或，所述终端设备安装有对话应用，云
端服务器中部署有所述对话模型，所述终端设备将通过所述对话应用将所述目标图像
和所述第一问询语音信息上传至所述云端服务器的对话模型进行处理。
- 30 10、一种语音回复方法，其特征在于，应用于携带摄像头的耳机设备，所述方法
包括：
采集第一问询语音信息；
获取所述摄像头采集的目标图像；

得到所述第一问询语音信息的第一回复语音信息，其中，所述第一回复语音信息是与所述目标图像相关的信息，所述目标图像和所述第一问询语音信息被预先训练的对话模型进行处理以得到所述第一回复语音信息；

播放所述第一回复语音信息。

- 5 11、根据权利要求 10 所述的语音回复方法，其特征在于，所述获取所述摄像头采集的目标图像之前，还包括：

响应对所述摄像头的唤醒操作，唤醒所述摄像头；

所述获取所述摄像头采集的目标图像，包括：控制已唤醒的摄像头采集所述目标图像。

- 10 12、根据权利要求 10 所述的语音回复方法，其特征在于，所述获取所述摄像头采集的目标图像，包括：

通过识别模型确定所述第一问询语音信息的拍摄意图；

根据所述拍摄意图控制所述摄像头采集所述目标图像。

- 15 13、根据权利要求 10 所述的语音回复方法，其特征在于，所述耳机设备中部署所述对话模型，所述得到所述第一问询语音信息的第一回复语音信息包括：

将所述目标图像和所述第一问询语音信息输入预先训练的对话模型进行处理，得到所述第一问询语音信息的第一回复语音信息。

14、根据权利要求 10 所述的语音回复方法，其特征在于，所述得到所述第一问询语音信息的第一回复语音信息包括：

- 20 向终端设备发送所述目标图像和所述第一问询语音信息，所述终端设备用于通过预先训练的对话模型对所述目标图像和所述第一问询语音信息进行处理，得到所述第一问询语音信息的第一回复语音信息；

接收所述终端设备发送的第一回复语音信息。

- 25 15、根据权利要求 10 至 14 任一项所述的语音回复方法，其特征在于，所述耳机设备包括：第一耳机和第二耳机，所述第一耳机携带第一摄像头，所述第二耳机携带第二摄像头，所述目标图像包括：所述第一摄像头采集的目标图像和所述第二摄像头采集的目标图像。

16、根据权利要求 15 所述的语音回复方法，其特征在于，向终端设备发送所述目标图像，包括：

- 30 采用所述第一耳机向所述终端设备发送所述第一摄像头采集的目标图像；

采用所述第二耳机向所述终端设备发送第二摄像头采集的目标图像。

17、根据权利要求 15 所述的语音回复方法，其特征在于，向终端设备发送所述目标图像，包括：

所述第一耳机接收第二耳机发送的第二摄像头采集的目标图像；

采用所述第一耳机向所述终端设备发送所述第一摄像头采集的目标图像和所述第二摄像头采集的目标图像。

18、根据权利要求 15 所述的语音回复方法，其特征在于，所述接收所述终端设备发送的第一回复语音信息，包括：

- 5 采用所述第一耳机接收所述终端设备发送的所述第一回复语音信息；
 所述第一耳机向所述第二耳机转发所述第一回复语音信息。

19、根据权利要求 15 所述的语音回复方法，其特征在于，所述接收所述终端设备发送的第一回复语音信息，包括：

- 采用所述第一耳机接收所述终端设备发送的所述第一回复语音信息；
10 采用所述第二耳机接收所述终端设备发送的所述第一回复语音信息。

20、根据权利要求 10 至 14 任一项所述的语音回复方法，其特征在于，所述播放所述第一回复语音信息之后，还包括：

- 采集第二问询语音信息；
 得到所述第二问询语音信息的第二回复语音信息，其中，所述第二回复语音信息
15 是与所述目标图像相关的信息，所述目标图像和所述第二问询语音信息被所述对话模型进行处理以得到所述第二回复语音信息；
 播放所述第二回复语音信息。

21、一种耳机设备，其特征在于，所述耳机设备携带摄像头，所述耳机设备包括：

- 20 采集单元，用于采集第一问询语音信息；
 获取单元，用于获取所述摄像头采集的目标图像；
 处理单元，用于得到所述第一问询语音信息的第一回复语音信息，其中，所述第一回复语音信息是与所述目标图像相关的信息，所述目标图像和所述第一问询语音信息被预先训练的对话模型进行处理以得到所述第一回复语音信息；
25 播放单元，用于播放所述第一回复语音信息。

22、一种语音回复系统，其特征在于，包括权利要求 21 所述的耳机设备和终端设备，其中：

- 所述耳机设备，用于采集第一问询语音信息；获取所述摄像头采集的目标图像，向所述终端设备发送所述目标图像和所述第一问询语音信息；
30 所述终端设备，用于通过预先训练的对话模型对所述目标图像和所述第一问询语音信息进行处理，得到所述第一问询语音信息的第一回复语音信息，向所述耳机设备发送所述第一回复语音信息；

 所述耳机设备，用于接收所述终端设备发送的第一回复语音信息，并播放所述第一回复语音信息。

23、一种电子设备，其特征在于，包括：处理器和存储器；

所述存储器存储计算机执行指令；

所述处理器执行所述存储器存储的计算机执行指令，使得所述处理器执行如权利要求 10 至 20 任一项所述的语音回复方法。

- 5 24、一种计算机可读存储介质，其特征在于，所述计算机可读存储介质中存储有计算机执行指令，当处理器执行所述计算机执行指令时，实现如权利要求 10 至 20 任一项所述的语音回复方法。

25、一种计算机程序产品，包括计算机程序，其特征在于，所述计算机程序被处理器执行时实现如权利要求 10 至 20 任一项所述的语音回复方法。