

(12) 按照专利合作条约所公布的国际申请

(19) 世界知识产权组织
国际局

(43) 国际公布日
2024 年 12 月 5 日 (05.12.2024)



(10) 国际公布号
WO 2024/248736 A1

- (51) 国际专利分类号:
G06N 3/0455 (2023.01) *G06T 9/00* (2006.01)
G06N 3/08 (2023.01)
- (21) 国际申请号: PCT/SG2024/050361
- (22) 国际申请日: 2024 年 5 月 29 日 (29.05.2024)
- (25) 申请语言: 中文
- (26) 公布语言: 中文
- (30) 优先权:
202310618707.8 2023年5月29日 (29.05.2023) CN
- (71) 申请人: 阿里巴巴创新公司 (ALIBABA INNOVATION PRIVATE LIMITED) [SG/SG]; 新加坡新加坡勿拉士峇沙路 51 号来赞达一号大厦#03-06, Singapore 189554 (SG)。

- (72) 发明人: 陈大友 (CHEN, Dayou); 中国北京市朝阳区望京东园四区 6 号楼金辉大厦, Beijing 100102 (CN)。张士伟 (ZHANG, Shiwei); 中国浙江省杭州市西湖区三墩镇灯彩街 1008 号, Zhejiang 310030 (CN)。张迎亚 (ZHANG, Yingya); 中国北京市朝阳区望京东园四区 6 号楼金辉大厦, Beijing 100102 (CN)。赵德丽 (ZHAO, Deli); 中国浙江省杭州市西湖区三墩镇灯彩街 1008 号, Zhejiang 310030 (CN)。
- (81) 指定国(除另有指明, 要求每一种可提供的国家保护): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CV, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IQ, IR, IS, IT, JM, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, MG, MK, MN,

(54) Title: VIDEO GENERATION METHOD AND APPARATUS, AND METHOD AND APPARATUS FOR TRAINING VIDEO GENERATION MODEL

(54) 发明名称: 视频生成方法、训练视频生成模型的方法及装置

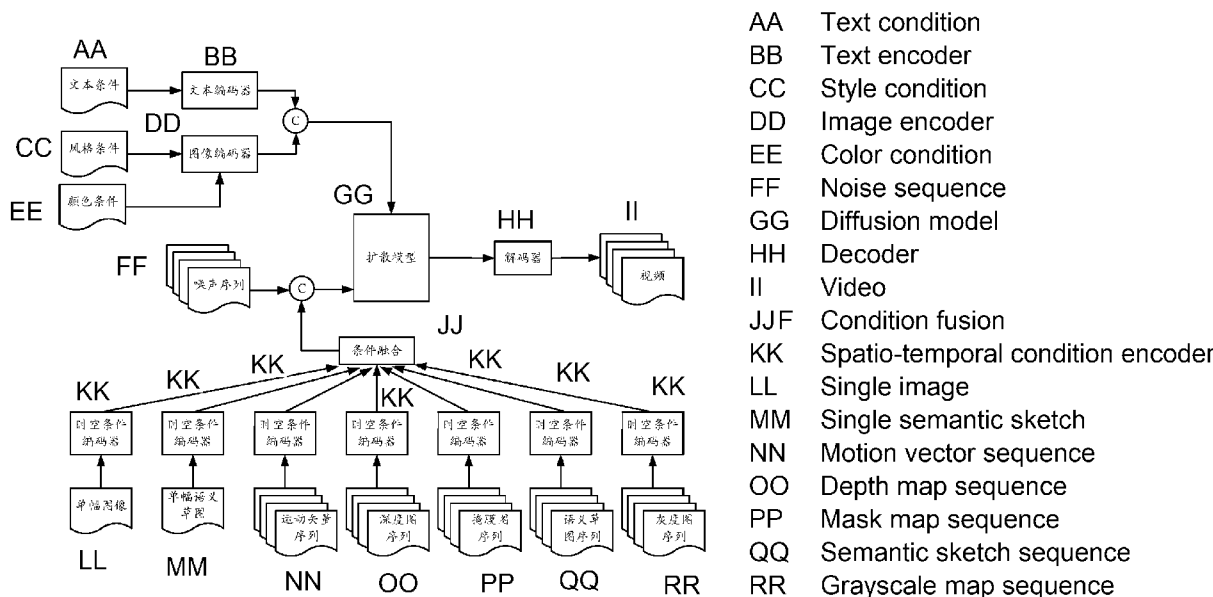


图 4

(57) Abstract: Disclosed in the embodiments of the present disclosure are a video generation method and apparatus, and a method and apparatus for training a video generation model. The main technical solution includes: acquiring a local condition, wherein the local condition comprises a space condition and/or a time condition; encoding the local condition, so as to obtain a local condition feature representation; integrating a noise sequence and the local condition feature representation, so as to obtain a noise hidden vector sequence; using a diffusion model to perform denoising processing on the noise hidden vector sequence, so as to obtain a denoised

MU, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, WS, ZA, ZM, ZW。

(84) 指定国(除另有指明，要求每一种可提供的地区保护): ARIPO (BW, CV, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SC, SD, SL, ST, SZ, TZ, UG, ZM, ZW), 欧亚 (AM, AZ, BY, KG, KZ, RU, TJ, TM), 欧洲 (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, ME, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG)。

本国际公布:
— 包括国际检索报告(条约第21条(3))。

hidden vector sequence; and using the denoised hidden vector sequence to perform decoding processing, so as to obtain a video.

(57) 摘要: 本公开实施例公开了一种视频生成方法、训练视频生成模型的方法及装置。其中主要技术方案包括: 获取局部条件, 局部条件包括空间条件和/或时间条件; 对局部条件进行编码处理得到局部条件特征表示; 将噪声序列与局部条件特征表示进行整合, 得到噪声隐向量序列; 利用扩散模型对噪声隐向量序列进行去噪声处理, 得到去噪后的隐向量序列; 利用去噪后的隐向量序列进行解码处理, 得到视频。

视频生成方法、训练视频生成模型的方法及装置

交叉援引

本公开要求于 2023 年 05 月 29 日提交中国专利局、优先权号为 202310618707.8、发明名称为“视频生成方法、训练视频生成模型的方法及装置”的中国专利公开的优先权，其全部内容通过引用结合在本公开中。

技术领域

本公开涉及计算机视觉技术领域，特别是涉及一种视频生成方法、训练视频生成模型的方法及装置。

背景技术

人工智能技术的不断发展给计算机视觉技术提供了新的契机，其中自动生成视频的能力越来越强大。基于文本的视频生成技术为很多内容创作者提供了全新的工具，让原本需要专业人员和昂贵设备制作的视频内容创作变得更加容易和低成本。然而，目前大多数技术以文本作为引导条件，生成的视频缺乏多样性和灵活性，导致视频的质量较低。

发明内容

有鉴于此，本公开提供了一种视频生成方法、训练视频生成模型的方法及装置，以便于提高生成的视频质量。

本公开提供了如下方案：

第一方面，提供了一种视频生成方法，所述方法包括：

获取局部条件，所述局部条件包括空间条件和/或时间条件；

对所述局部条件进行编码处理得到局部条件特征表示；

将噪声序列与所述局部条件特征表示进行整合，得到噪声隐向量序列；

利用扩散模型对所述噪声隐向量序列进行去噪声处理，得到去噪后的隐向量序列；

利用所述去噪后的隐向量序列进行解码处理，得到视频。

根据本公开实施例中一可实现的方式，所述方法还包括：获取全局条件，所述全局条件包括文本条件、风格条件和颜色条件中的至少一种；对所述全局条件进行编码处理得到全局条件特征表示；

所述利用扩散模型对所述噪声隐向量序列进行去噪声处理包括：所述扩散模型利用所述全局条件特征表示对所述噪声隐向量序列进行交叉注意力处理以预测噪声，并利用预测的噪声进行去噪声处理。

根据本公开实施例中一可实现的方式，所述空间条件包括单幅图像和单幅语义草

图中的至少一种；

所述时间条件包括运动矢量序列、深度图序列、掩膜图序列、语义草图序列和灰度图序列中的至少一种。

根据本公开实施例中一可实现的方式，对所述全局条件进行编码处理得到全局条件特征表示包括：对所述文本条件进行文本编码得到文本特征表示，以及对所述风格条件和/或颜色条件进行图像编码得到图像特征表示；对所述文本特征表示和所述图像特征表示进行整合，得到全局条件特征表示。

根据本公开实施例中一可实现的方式，对所述局部条件进行编码处理得到局部条件特征表示包括：

利用各时空条件编码器分别对各局部条件进行编码，得到各局部条件对应的特征张量，其中时空条件编码器与局部条件一一对应；

将各局部条件对应的特征张量进行融合处理，得到所述局部条件特征表示。

根据本公开实施例中一可实现的方式，所述利用时空条件编码器分别对各局部条件进行编码包括：

利用时空条件编码器对局部条件进行空间特征编码，得到该局部条件的空间特征表示；

若该局部条件为序列，则对该局部条件的空间特征表示进行时域自注意力处理，得到该局部条件对应的特征张量；否则，对该局部条件的空间特征表示在时序上复制产生时序上的空间特征表示，对所述时序上的空间特征表示进行时域自注意力处理，得到该局部条件对应的特征张量。

第二方面，提供了一种训练视频生成模型的方法，所述方法包括：

获取包括多个第一训练样本的第一训练数据，所述第一训练样本包括局部条件样本，所述局部条件样本包括空间条件样本和/或时间条件样本；

利用所述第一训练数据训练视频生成模型，所述视频生成模型包括：时空条件编码器、扩散模型和解码器；其中，时空条件编码器对所述局部条件样本进行编码处理得到局部条件特征表示；扩散模型对噪声隐向量序列进行去噪声处理，得到去噪后的隐向量序列，所述噪声隐向量序列是将噪声序列与所述局部条件特征表示进行整合后得到的；所述解码器利用所述去噪后的隐向量序列进行解码处理，得到视频；

所述训练的目标包括：最小化扩散模型进行去噪声处理时预测的噪声与高斯噪声之间的差异。

根据本公开实施例中一可实现的方式，所述第一训练样本还包括全局条件样本，所述全局条件样本包括文本条件样本、风格条件样本和颜色条件样本中的至少一种；

所述视频生成模型还包括全局编码器，所述全局编码器对所述全局条件样本进行编码处理得到全局条件特征表示；

所述扩散模型对噪声隐向量序列进行去噪声处理包括：所述扩散模型利用所述全局条件特征表示对所述噪声隐向量序列进行交叉注意力处理以预测噪声，并利用预测的噪声进行去噪声处理。

根据本公开实施例中一可实现的方式，所述全局编码器和所述解码器采用预训练得到的参数，在所述训练的每一轮迭代中，利用与所述训练的目标对应的损失函数更新所述时空条件编码器和扩散模型的参数。

根据本公开实施例中一可实现的方式，所述获取包括多个第一训练样本的第一训练数据包括：

获取视频样本；

获取所述视频样本的描述文本、风格图像和/或颜色直方图分别作为所述文本条件样本、风格条件样本和/或颜色条件样本；

提取单幅图像和单幅语义草图中的至少一种作为所述空间条件样本；

提取运动矢量序列、深度图序列、掩膜图序列、语义草图序列和/或灰度图序列中的至少一种作为所述时间条件样本。

根据本公开实施例中一可实现的方式，在利用所述第一训练数据训练视频生成模型之前，还包括预训练所述扩散模型；在预训练得到的扩散模型的参数基础上进一步利用所述第一训练数据训练视频生成模型；

其中，预训练所述扩散模型包括：

获取包括多个第二训练样本的第二训练数据，所述第二训练样本包括：从视频样本中提取描述文本作为文本样本；

利用全局编码器对所述文本样本进行编码处理得到文本特征表示；

将所述文本特征表示和噪声序列输入扩散模型以对所述扩散模型进行训练，所述扩散模型利用所述文本特征表示对所述噪声序列进行去噪声处理；所述训练的目标包括：最小化扩散模型在各时间步进行去噪声处理时预测的噪声与高斯噪声之间的差异。

根据本公开实施例中一可实现的方式，时空条件编码器对所述局部条件样本进行编码处理得到局部条件特征表示包括：

利用各时空条件编码器分别对各局部条件样本进行编码，得到各局部条件样本对应的特征张量，其中时空条件编码器与局部条件一一对应；

将各局部条件样本对应的特征张量进行融合处理，得到所述局部条件特征表示。

根据本公开实施例中一可实现的方式，所述利用各时空条件编码器分别对各局部

条件样本进行编码包括：

利用时空条件编码器对局部条件样本进行空间特征编码，得到该局部条件样本的空间特征表示；

若该局部条件样本为序列，则对该局部条件样本的空间特征表示进行时域自注意力处理，得到该局部条件样本对应的特征张量；否则，对该局部条件样本的空间特征表示在时序上复制产生时序上的空间特征表示，对所述时序上的空间特征表示进行时域自注意力处理，得到该局部条件样本对应的特征张量。

第三方面，提供了一种视频生成方法，由云端服务器执行，所述方法包括：

获取来自用户终端的局部条件，所述局部条件包括空间条件和/或时间条件；

对所述局部条件进行编码处理得到局部条件特征表示；

将噪声序列与所述局部条件特征表示进行整合，得到噪声隐向量序列；

利用扩散模型对所述噪声隐向量序列进行去噪声处理，得到去噪后的隐向量序列；

利用所述去噪后的隐向量序列进行解码处理，得到视频；

将所述视频发送给所述用户终端以进行展示。

第四方面，提供了一种视频生成装置，所述装置包括：

条件获取单元，被配置为获取局部条件，所述局部条件包括空间条件和/或时间条件；

视频生成单元，被配置为对所述局部条件进行编码处理得到局部条件特征表示；将噪声序列与所述局部条件特征表示进行整合，得到噪声隐向量序列；利用扩散模型对所述噪声隐向量序列进行去噪声处理，得到去噪后的隐向量序列；利用所述去噪后的隐向量序列进行解码处理，得到视频。

第五方面，提供了一种训练视频生成模型的装置，所述装置包括：

样本获取单元，被配置为获取包括多个第一训练样本的第一训练数据，所述第一训练样本包括局部条件样本，所述局部条件样本包括空间条件样本和/或时间条件样本；

模型训练单元，被配置为利用所述第一训练数据训练视频生成模型，所述视频生成模型包括：时空条件编码器、扩散模型和解码器；其中，时空条件编码器对所述局部条件样本进行编码处理得到局部条件特征表示；扩散模型对噪声隐向量序列进行去噪声处理，得到去噪后的隐向量序列，所述噪声隐向量序列是将噪声序列与所述局部条件特征表示进行整合后得到的；所述解码器利用所述去噪后的隐向量序列进行解码处理，得到视频；

所述训练的目标包括：最小化扩散模型进行去噪声处理时预测的噪声与高斯噪声之间的差异。

根据第六方面，提供了一种计算机可读存储介质，其上存储有计算机程序，该程序被处理器执行时实现上述第一方面至第三方面中任一项所述的方法的步骤。

根据第七方面，提供了一种电子设备，包括：

一个或多个处理器；以及

与所述一个或多个处理器关联的存储器，所述存储器被设置为存储程序指令，所述程序指令在被所述一个或多个处理器读取执行时，执行上述第一方面至第三方面中任一项所述的方法的步骤。

根据本公开提供的具体实施例，本公开公开了以下技术效果：

1) 本公开融入了空间条件和/或时间条件作为局部条件，利用局部条件特征表示与噪声序列进行整合得到噪声隐向量序列，并对噪声隐向量序列进行去噪声处理，得到去噪后的隐向量序列，进而解码得到视频。这种方式使得视频生成不再局限于在文本条件，而是引入空间条件和/或时间条件来引导视频生成，从而更加灵活、多样化地生成视频，提高视频的质量。

2) 本公开可以通过组合多种全局条件、多种局部条件中的空间条件和时间条件，来生成需要的视频；其中通过时空条件编码器实现多种局部条件的编码、对齐和融合，与噪声序列进行整合实现对视频的局部控制；通过全局编码器对全局条件进行编码和整合，在扩散模型中利用全局条件特征表示对噪声隐向量序列进行交叉注意力处理，实现对视频的全局控制。这种方式大大提高了视频生成的可控性，生成的视频更加灵活多样。

3) 本公开创新性地时间条件中引入运动矢量序列，通过融入运动矢量序列的控制信号，使得视频生成模型能够捕获帧间动态，从而实现对视频中内部运动的控制。

4) 本公开提供的时空条件编码器首先提取局部的空间信息，然后进行时序建模，促进了时域的显示嵌入。并且为不同的局部条件提供了统一的接口，增强了帧间的一致性。对于单幅图像、单幅语义草图等非时序型的空间条件，通过在时间维度进行复制以确保与时间条件的一致性，实现局部条件的融合。这种编码和融合的处理使得后续生成的视频在时间和空间感知方面都能够具有更好的可控性。

5) 本公开在训练视频生成装置之前，可以先利用文本生成视频的训练数据对扩散模型进行预训练，在预训练得到的扩散模型的参数基础上进一步训练视频生成模型，从而提高视频生成模型的效果和效率。

当然，实施本公开的任一产品并不一定需要同时达到以上所述的所有优点。

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，下面将对实施例中所需要使用的附图作简单地介绍，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的前提下，还可以根据这些附图获得其他的附图。

图 1 为是本公开实施例所适用的系统架构图；

图 2 为本公开实施例提供的视频生成方法流程图；

图 3 为本公开实施例提供的生成视频所需要条件的示意图；

图 4 为本公开实施例提供的视频生成模型的原理性示意图；

图 5 为本公开实施例提供的时空条件编码器的结构示意图；

图 6 为本公开实施例提供的扩散模型在各时间步的原理性示意图；

图 7a~图 7d 为本公开实施例提供的四个生成视频的实例图；

图 8 为本公开实施例提供的训练视频生成模型的方法流程图；

图 9 为本公开实施例提供的视频生成装置的示意性框图；

图 10 为本公开实施例提供的训练视频生成模型的装置的示意性框图；

图 11 为本公开实施例提供的电子设备的示意性框图。

具体实施方式

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所获得的所有其他实施例，都属于本公开保护的范围。

在本发明实施例中使用的术语是出于描述特定实施例的目的，而非旨在限制本发明。在本发明实施例和所附权利要求书中所使用的单数形式的“一种”、“所述”和“该”也旨在包括多数形式，除非上下文清楚地表示其他含义。

应当理解，本文中使用的术语“和/或”是一种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。另外，本文中字符“/”，一般表示前后关联对象是一种“或”的关系。

取决于语境，如在此所使用的词语“如果”可以被解释成为“在……时”或“当……时”或“响应于确定”或“响应于检测”。类似地，取决于语境，短语“如果确定”或“如果检测（陈述的条件或事件）”可以被解释成为“当确定时”或“响应于确定”或“当检测（陈述的条件或事件）时”或“响应于检测（陈述的条件或事件）”。

为了方便对本公开的理解，首先对本公开所适用的系统架构进行简单描述。图 1 示出了可以应用本公开实施例的示例性系统架构，如图 1 中所示，该系统架构包括用户终端，以及位于服务端的模型训练装置和视频生成装置。

其中，服务端的模型训练装置采用本公开实施例提供的方法，在离线阶段进行模型训练，得到视频生成模型。

用户终端可以通过网络与服务端的视频生成装置进行信息交互。例如，用户终端可以将用户输入的生成视频所使用的条件信息通过网络发送给视频生成装置，其中条件信息可以包括全局条件和局部条件，具体内容将在后续实施例中详述。

视频生成装置可以利用已经训练得到的视频生成模型，在全局条件和局部条件的限制下生成视频，并将生成的视频通过网络发送给用户终端，以供用户终端展示该视频。

上述用户终端可以包括但不限于诸如：智能移动终端、智能家居设备、可穿戴式设备、智能医疗设备、PC（Personal Computer，个人计算机）等。其中智能移动设备可以包括诸如手机、平板电脑、笔记本电脑、PDA（Personal Digital Assistant，个人数字助理）、互联网汽车等。智能家居设备可以包括智能家电设备，诸如智能电视、智能冰箱等等具备视频播放功能的家电设备。可穿戴式设备可以包括诸如智能手表、智能眼镜、智能手环、VR（Virtual Reality，虚拟现实）设备、AR（Augmented Reality，增强现实设备）、混合现实设备（即可以支持虚拟现实和增强现实的设备）等等。

模型训练装置和视频生成装置可以分别设置为独立的服务器，也可以设置于同一个服务器或服务器群组，还可以设置于独立的或者同一云服务器。云服务器又称为云计算服务器或云主机，是云计算服务体系中的一项主机产品，以解决传统物理主机与虚拟专用服务器（VPS，Virtual Private Server）服务中存在的管理难度大，服务扩展性弱的缺陷。模型训练装置和视频生成装置还可以设置于具有较强计算能力的计算机终端。

需要说明的是，上述视频生成装置除了在线上进行视频生成之外，也可以采用离线的方式进行视频生成。

应该理解，图1中的用户终端、模型训练装置、视频生成装置以及视频生成模型的数目是示意性的。根据实现需要，可以具有任意数目的用户终端、模型训练装置、视频生成装置以及视频生成模型。

图2为本公开实施例提供的视频生成方法流程图，该方法可以由图1所示系统架构中的视频生成装置执行。如图2中所示，该方法可以包括以下步骤：

步骤202：获取局部条件，局部条件包括空间条件和/或时间条件。

步骤204：对局部条件进行编码处理得到局部条件特征表示。

步骤206：将噪声序列与局部条件特征表示进行整合，得到噪声隐向量序列；利用扩散模型对噪声隐向量序列进行去噪声处理，得到去噪后的隐向量序列。

步骤208：利用去噪后的隐向量序列进行解码处理，得到视频。

由上述流程可以看出，本公开融入了空间条件和/或时间条件作为局部条件，利用局部条件特征表示与噪声序列进行整合得到噪声隐向量序列，并利用全局条件特征表示对噪声隐向量序列进行去噪声处理，得到去噪后的隐向量序列，进而解码得到视频。这种方式使得视频生成不再局限于在文本条件，而是引入空间条件和/或时间条件来引导视频生成，从而更加灵活、多样化地生成视频，提高视频的质量。

下面对上述流程中的各步骤分别进行详细描述。首先结合实施例对上述步骤 202 即“获取局部条件”进行详细描述。

现有的视频生成方法通常在文本条件的引导下，生成视频，灵活性和多样性差。本公开实施例中通过引入新的组合性条件来提高视频生成的可控性。从引入方式上来看，至少包括局部条件，还可以进一步包括全局条件。

其中，全局条件主要包括文本条件、风格条件和/或颜色条件。文本条件主要是对视频内容的描述，从全局上体现出要生成视频的内容。风格条件通常可以采用一幅风格图像，作为条件目的是生成具有与该风格图像所具有的风格特征一致的视频，即视频在整体上具有风格图像中所表达的风格。颜色条件可以采用颜色直方图等形式，作为条件的目标是为了对生成的视频的颜色分布进行指导。

局部条件主要包括空间条件和/或时间条件。其中，空间条件可以包括单幅图像、单幅语义草图等中的至少一种。其中单幅图像指的是单一图像，用以限制视频中图像帧的内容和结构。语义草图也称为手绘草图、素描图等等，是一种用来描述图像中对象基本语义的方式，是一种概括的表述手法，主要描绘出对象的边缘（或轮廓）特征，能够忽略对象的细节和冗余特征，保留主要信息。可以看出，空间条件主要是从空间特征上对视频内容进行指导。

时间条件包括运动矢量序列、深度图序列、掩膜图序列和语义草图序列中的至少一种。由于视频是图像序列，相应地，时间条件是沿着时间维度对视频进行的更精细的引导。

运动矢量序列包含各相邻帧之间的运动矢量，运动矢量指示各 Token（元素）在相邻帧之间的运动方向，包括水平方向和垂直方向，因此运动矢量可以表示为二维矢量。它显式地表达了相邻两个图像帧之间各 Token 的移动，用以使得视频生成模型生成的视频具有运动可控性。其中，图像的各 Token 指的是构成图像的元素。对于图像而言，将图像切分成不重叠的区块，则图像中的区块以及起始符等均为 Token。

深度图序列由各帧的深度图构成，深度图包含图像帧中各 Token 的深度信息，用以指导视频中各帧的深度信息。

掩膜图序列由各帧的掩膜图构成，掩膜图是将图像帧中部分区域的内容进行掩膜，从而使得模型具备预测掩膜部分内容的的能力。其中掩膜的区域可以是用户指定的，也

可以是随机确定的。

语义草图序列由各帧的语义草图构成，与空间条件中的单幅语义草图相比，草图序列能够提供更多的控制细节。

灰度图序列由各帧的灰度图构成，灰度图包含图像帧中各 Token 的灰度信息，用以指导视频中各帧的灰度信息。

上述时间条件中各序列的长度是一致的，都等于生成视频的长度，例如表示为 T1。

例如，文本条件为“一个站在圣诞树旁边的孩子”，风格条件为一幅风格图像，用以限制生成的视频的整体风格，即将风格图像的风格特征迁移至生成的视频，颜色条件为一幅颜色直方图，如图 3 中所示，用以限制所生成视频的颜色分布。空间条件中的单幅图像就可以是包含一个孩子和圣诞树的图像，用以限制生成的视频中孩子和圣诞树的样子等，如图 3 中所示。语义草图可以是描述孩子、圣诞树的基本线条的图，用以限制孩子和圣诞树的基本形状和位置等。时间条件中的运动矢量序列包含相邻帧之间的运动矢量，图中示意性的以小箭头表示，深度图序列包含各帧的深度图，掩膜图序列包含各帧的掩膜图，语义草图序列包含各帧的语义草图。各序列的长度均为 T1，T1 为要生成的视频的长度。

作为其中一种可实现的方式，可以向用户提供条件输入界面，用户可以在界面上选择输入上述全局条件和局部条件。例如，用户可以在文本输入框输入文本条件。再例如，用户可以通过界面上提供的图像绘制或编辑工具，输入单幅图像和/或单幅语义草图作为空间条件，输入运动矢量序列、深度图序列、掩膜图序列和语义图序列等中的至少一种作为时间条件，还可以通过上传自选图像或者从界面提供的图像库中选择图像的方式输入风格图像。其中在输入运动矢量序列时，用户可以通过框选图像中的特定区域，并通过特定手势、鼠标轨迹等方式产生移动轨迹来指示特定区域在视频中的运动路线，服务端依据该移动轨迹自动生成运动矢量序列。除了上述列举的条件输入方式之外，也可以采用其他条件输入方式，在此不做穷举。

下面结合实施例对上述步骤 204 即“对局部条件进行编码处理得到局部条件特征表示对全局条件进行编码处理得到全局条件特征表示，以及对局部条件进行编码处理得到局部条件特征表示”进行详细描述。

图 2 中涉及的步骤 204~步骤 208 可以通过预先训练得到的视频生成模型实现。如图 4 中所示，视频生成模型可以包括时空条件编码器、扩散模型和解码器，还可以进一步包括全局编码器。本步骤主要由时空条件编码器执行，如果使用了全局条件，则进一步全局编码器执行。

全局编码器用以对全局条件进行编码处理得到全局条件特征表示。

若全局条件中包含文本条件，则可以采用文本编码器对文本条件进行文本编码，

得到文本特征表示，将该文本特征表示作为全局条件特征表示。

若全局条件中包含文本条件和风格图像，则可以利用文本编码器对文本条件进行文本编码得到文本特征表示，以及利用图像编码器对风格条件和/或颜色条件进行图像编码得到图像特征表示；再对文本特征表示和图像特征表示进行整合，得到全局条件特征表示。

其中，文本编码器可以采用预训练语言模型实现，例如 BERT (Bidirectional Encoder Representation from Transformers, 基于转换的双向编码表示)、XLNet (一种通过排列语言模型实现双向上下文信息的自回归模型)、GPT (Generative Pre-Training, 生成式预训练) 模型、CLIP (是一种可以实现多模态编码的编码模型) 等。通过文本编码器实际上可以获得文本条件的语义嵌入即文本特征表示。

图像编码器可以使用诸如 VIT (Vision Transformer, 视觉转换器)、CLIP 等，通过图像编码器实际上可以获取风格图像的语义嵌入即图像特征表示。

对文本特征表示和图像特征表示进行的整合可以是将文本特征表示和图像特征表示在通道维度上进行拼接。图 4 中采用 “⊗” 表示整合处理。

时空条件编码器用以对局部条件进行编码处理得到局部条件特征表示。由于局部条件中包括空间条件和/时间条件。由于时间条件是序列型的条件，因此局部条件包含了丰富而复杂的时空关系，对视频的可控引导提出了挑战。有鉴于此，本公开实施例中针对各局部条件分别布设对应的时空条件编码器，作为其中一种可实现的方式，局部条件和时空条件编码器是一一对应的关系。

利用各时空条件编码器分别对各局部条件进行编码，得到各局部条件对应的特征张量，再将各局部条件对应的特征张量进行融合处理，得到局部条件特征表示。

其中，时空条件编码器的结构可以如图 5 中所示。时空条件编码器首先对局部条件进行空间特征编码，得到该局部条件的空间特征表示。如图 5 中所示，时空条件编码器可以由两个二维卷积 (Conv2D)、两个激活层 (例如可以采用 SiLU 激活函数) 和一个平均池化层组成。

若输入时空条件编码器的局部条件为序列，则对该局部条件 (即时序上) 的空间特征表示进行时域自注意力处理，得到该局部条件对应的特征张量，该部分由图 5 中所示的时域 Transformer (转换器) 执行。

若输入时空条件编码器的局部条件并非序列，例如是单幅图像、单幅语义草图等。则可以首先对该局部条件的空间特征表示在时序上复制产生时序上的空间特征表示，该部分在图 5 中由 “Ⓡ” 执行，使其与序列型的局部条件对应的空间特征表示在时序上对齐。然后通过时域 Transformer 在对时序上的空间特征表示进行时域自注意力处理，得到该局部条件对应的特征张量。

作为其中一种可实现的方式，将各局部条件对应的特征张量进行融合处理时，可以采用将各特征张量进行逐元素相加的方式。

可以看出，时空条件编码器实际上首先提取局部的空间信息，然后进行时序建模，促进了时域的显示嵌入。并且为不同的局部条件提供了统一的接口，增强了帧间的一致性。对于单幅图像、单幅语义草图等非时序型的空间条件，通过在时间维度进行复制以确保与时间条件的一致性，实现局部条件的融合。这种编码和融合的处理使得后续生成的视频在时间和空间感知方面都能够具有更好的可控性。尤其是在时间条件中引入的运动矢量序列，能够捕获帧间动态，从而实现对视频中内部运动的直接控制。

下面结合实施例对上述步骤 206 即“将噪声序列与局部条件特征表示进行整合，得到噪声隐向量序列；利用扩散模型对噪声隐向量序列进行去噪声处理，得到去噪后的隐向量序列”进行详细描述。

本公开实施例中可以随机产生噪声序列 ε ，随机产生的噪声序列符合正态分布，即高斯噪声，且 ε 的长度与要生成的视频长度一致，均为 $T1$ 。经过时空条件编码器对局部条件进行的编码处理以及整合后，局部条件特征表示与 ε 具有相同的空间形状。然后将噪声序列 ε 与局部条件特征表示进行整合，例如沿着通道维度将两者进行拼接，得到噪声隐向量序列 z_t ，该噪声隐向量序列作为视频生成的一路控制信号。

扩散模型因其具有更稳定的训练和生成灵活性，在图像生成领域已经得到了广泛的应用，但在视频生成领域尚未被很好地利用。为了有效地处理视频数据，本公开实施例中引入 IDM (Latent Diffusion Model, 隐在扩散模型) 来保持局部保真度，即将初始视频投影到隐在表示，然后再通过解码器将隐在表示映射回像素空间得到最终视频。其中，初始视频即噪声序列，最终视频即生成的视频。除了 IDM 之外，也可以采用其他类型的扩散模型。

扩散模型的处理可以理解预测正态分布的噪声，在每一时间步进行去噪声处理，旨在恢复真实的视频内容。这个过程模拟了长度为 T 的马尔科夫链的反向过程。其中 T 为扩散模型的总时间步， T 越长去噪效果越好，但对计算性能的影响越大，因此需要在两者之间进行权衡。可以取经验值或实验值，例如可以取 1000。

扩散模型可以在各时间步利用上一时间步得到的噪声隐向量序列（在第一个时间步则利用扩散模型被输入的噪声隐向量序列）预测当前时间步的噪声，利用预测的噪声对该噪声隐向量序列进行去噪声处理，得到当前时间步的噪声隐向量序列。

更进一步地，若输入的条件包括全局条件，则扩散模型利用全局条件特征表示对噪声隐向量序列进行交叉注意力处理以预测噪声，并利用预测的噪声进行去噪声处理。

具体地，如图 6 中所示，扩散模型在第一个时间步利用全局条件特征表示对噪声隐向量序列进行交叉注意力处理，以预测当前时间步的噪声，利用预测的噪声对噪声

隐向量序列进行去噪声处理，得到当前时间步的噪声隐向量序列。

在其他时间步利用全局条件特征表示对上一时间步得到的噪声隐向量序列进行交叉注意力处理，以预测当前时间步的噪声，利用预测的噪声对上一时间步得到的噪声隐向量序列进行去噪声处理，得到当前时间步的噪声隐向量序列。

其中，扩散模型可以将三维 UNet 作为骨干网络。UNet 网络是一种引入了跳跃连接的编码器-解码器网络，因其结构似字母 U 而得名，能够将深层高级特征与浅层低级特征相结合，鉴于其是一种已知网络，在此不做详述。

上述过程可以看做是，对噪声隐向量序列 z_t 和条件（例如包括全局条件和局部条件） c 应用去噪函数 $\varepsilon_\theta(\cdot, t)$ ，其中 $t \in \{1, \dots, T\}$ 。在去噪过程中，采用交叉注意力机制注入全局条件，使得全局条件在整体语义上对视频生成进行引导。

最终扩展模型最后一步得到的去噪后的隐向量序列在步骤 208 中输入解码器，由解码器利用去噪后的隐向量序列进行解码处理，得到视频。关于解码器的部分可以采用目前已有的视频生成模型的解码器，在此不做详述。

通过本公开实施例提供的上述方式，可以灵活组合各种条件来生成更多样化的视频，在此举几个实例说明本公开在视频生成上的可控性。

实例 1：可以组合全局条件中的文本条件和局部条件中的时间条件。

如图 7a 中所示，用户输入文本条件“一个站在森林中的长发女人的旋转视角”，以及语义草图序列，本实例中序列长度以 6 帧为例。最终生成视频如图 7a 中所示，图中示出的是视频中的各帧图像。

实例 2：可以组合局部条件中的空间条件和时间条件。

用户输入单幅图像和深度图序列，最终生成视频如图 7b 中所示。

实例 3：可以组合全局条件中的风格条件和局部条件中的深度图序列、语义草图序列。

用户输入风格图像，深度图序列和语义草图序列，最终生成视频如图 7c 中所示。

实例 4：可以组合全局条件中的文本条件、风格条件以及局部条件中的一幅语义草图、运动矢量序列。

用户输入文本条件“一轮移动的金色月亮”，一幅风格图像，一幅语义草图，并且用户可以在语义草图上手绘出月亮的运动方向，服务端的视频生成装置能够根据用户在语义草图上手绘的月亮的运动方向自动生成运动矢量序列，从而利用本公开实施例提供的方式生成视频。如图 7d 中所示，生成的视频中月亮沿用户手绘的运动方向发生移动。

下面结合实施例对上述视频生成模型的训练过程进行详细描述。图 8 为本公开实

施例提供的训练视频生成模型的方法流程图，该方法流程可以由图 1 所示系统架构中的模型训练装置执行。如图 8 中所示，该方法可以包括如下步骤：

步骤 802：获取包括多个第一训练样本的第一训练数据，第一训练样本包括局部条件样本，局部条件包括空间条件样本和/或时间条件样本。

更进一步地，第一训练数据样本还可以包括全局条件样本，全局条件样本可以包括文本条件样本、风格条件样本和颜色条件样本中的至少一种。

作为其中一种可实现的方式，可以首先获取视频样本，获取视频样本的描述文本作为文本条件样本。获取视频样本的风格图像作为风格条件样本，获取视频样本的颜色直方图作为颜色条件样本。例如，可以采用已有的视频描述文本生成模型来获取视频样本的描述文本，也可以从视频样本所来源网页上获取视频样本的描述文本，也可以人工添加视频样本的描述文本，等等。例如，可以将视频样本的第一帧图像作为风格图像。再例如，可以将视频样本的第一帧图像对应的颜色直方图作为颜色条件样本。

然后从视频样本中提取单幅图像和单幅语义草图中的至少一种作为空间条件样本。例如可以将视频样本的第一帧图像作为上述的单幅图像。再例如，可以对视频样本中的第一帧图像进行边缘提取，利用从图像中提取的边缘信息形成单幅语义草图。或者，可以利用手绘图生成工具对视频样本中的第一帧图像生成单幅语义草图。

再从视频样本中提取运动矢量序列、深度图序列、掩膜图序列、语义草图序列和灰度图序列中的至少一种作为时间条件样本。例如，可以从视频样本中提取各相邻帧之间的运动矢量构成运动矢量序列。从视频样本中提取各帧图像的深度图构成深度图序列。将视频样本中各帧进行掩膜处理，得到掩膜图序列。针对视频样本中的各帧图像生成语义草图，得到语义草图序列。针对视频样本中的各帧图像生成灰度图，得到灰度图序列。

除了上述方式之外，也可以采用其他方式来得到上述条件样本，在此不做一一列举。

需要说明的是，本公开实施例中涉及的“第一”、“第二”等限定并不具备大小、顺序和数量等方面的限制，用以在名称上加以区分，例如“第一训练样本”和“第二训练样本”用以在名称上区分两种训练样本。

步骤 804：利用第一训练数据训练视频生成模型，视频生成模型包括：时空条件编码器、扩散模型和解码器；其中，时空条件编码器对局部条件样本进行编码处理得到局部条件特征表示；扩散模型对噪声隐向量序列进行去噪声处理，得到去噪后的隐向量序列，噪声隐向量序列是将噪声序列与局部条件特征表示进行整合后得到的；解码器利用去噪后的隐向量序列进行解码处理，得到视频；训练的目标包括：最小化扩散模型进行去噪声处理时预测的噪声与高斯噪声之间的差异。

更进一步地，上述视频生成模型还可以包括全局编码器。全局编码器对全局条件样本进行编码处理得到全局条件特征表示。相应地，扩散模型可以利用全局条件特征表示对噪声隐向量序列进行交叉注意力处理以预测噪声，并利用预测的噪声进行去噪声处理。

作为其中一种可实现的方式，若全局条件样本中包含文本条件样本，则全局编码器可以采用文本编码器对文本条件样本进行文本编码，得到文本特征表示，将该文本特征表示作为全局条件特征表示。

作为另一种可实现的方式，若全局条件中包含文本条件样本之外，还包含风格条件样本和/或颜色条件样本，则全局编码器可以利用文本编码器对文本条件样本进行文本编码得到文本特征表示，以及利用图像编码器对风格条件样本和/或颜色条件样本进行图像编码得到图像特征表示；再对文本特征表示和图像特征表示进行整合，得到全局条件特征表示。

作为其中一种可实现的方式，可以利用各时空条件编码器分别对各局部条件样本进行编码，得到各局部条件样本对应的特征张量，其中时空条件编码器与局部条件一一对应；然后将各局部条件样本对应的特征张量进行融合处理，得到所述局部条件特征表示。

其中，可以首先利用时空条件编码器对局部条件样本进行空间特征编码，得到该局部条件样本的空间特征表示。若该局部条件样本为序列，则对该局部条件样本的空间特征表示进行时域自注意力处理，得到该局部条件样本对应的特征张量；否则，对该局部条件样本的空间特征表示在时序上复制产生时序上的空间特征表示，对时序上的空间特征表示进行时域自注意力处理，得到该局部条件样本对应的特征张量。

作为其中一种可实现的方式，扩散模型可以在第一个时间步利用全局条件特征表示对噪声隐向量序列进行交叉注意力处理，以预测当前时间步的噪声，利用预测的噪声对噪声隐向量序列进行去噪声处理，得到当前时间步的噪声隐向量序列。

扩散模型在其他时间步利用全局条件特征表示对上一时间步得到的噪声隐向量序列进行交叉注意力处理，以预测当前时间步的噪声，利用预测的噪声对上一时间步得到的噪声隐向量序列进行去噪声处理，得到当前时间步的噪声隐向量序列。

关于视频生成模型的具体结构和详细描述可以参见之前方法实施例中关于图 4、图 5 和图 6 的相关记载，在此不做赘述。

在本公开实施例中，可以依据上述训练目标构造损失函数，在训练视频生成模型的每一轮迭代中利用损失函数的取值，采用诸如梯度下降等方式更新模型参数，直至满足预设的训练结束条件。其中训练结束条件可以包括诸如损失函数的取值小于或等于预设的损失函数阈值，迭代次数达到预设的次数阈值等。

作为其中一种可实现的方式，构造的损失函数 L_{VLDM} 可以如下公式所示：

$$L_{VLDM} = E_{\varepsilon \in N(0,1), c, t} [\|\varepsilon - \varepsilon_\theta(z_t, c, t)\|_2^2] \quad (1)$$

其中， ε 表示高斯噪声，即满足正态分布的随机噪声。 $\varepsilon_\theta(z_t, c, t)$ 指的是在时间步 t 对噪声隐向量序列 z_t 和条件 c （包括上述实施例所涉及的全局条件和局部条件）预测得到的噪声，在每一轮迭代中可以随机选择一个时间步。 $\|\cdot\|_2^2$ 表示 2 范数平方的计算。 E 是取期望的处理， $E_{\varepsilon \in N(0,1), c, t}[\cdot]$ 指的是在 $\varepsilon \in N(0,1), c, t$ 约束下取 $[\cdot]$ 中内容的期望。 $N(0,1)$ 是正态分布。

另外，除了上述公式（1）中示出的损失函数之外，在此精神原则之内采用的其他损失函数同样在本公开的保护范围之内。例如在每一轮迭代中选择多个时间步预测得到的噪声与高斯噪声之间的差异得到损失函数。

作为其中一种可实现的方式，在视频生成模型的训练过程中，全局编码器和解码器可以采用预训练得到的参数，在训练的每一轮迭代中，利用与训练的目标对应的损失函数更新时空条件编码器和扩散模型的参数。

其中全局编码器的预训练可以采用其他任务，例如图像生成任务、文本分类任务等等。解码器的预训练也可以采用其他任务，例如图像生成任务、图像分类任务等等。这些是目前已有的训练任务，在此不做详述。

在视频生成模型的训练之前，可以使用图像生成模型中的扩散模型的参数进行初始化，这种方式在一定程度上缓解了训练难度，加快了训练速度，但仍在学习处理时序特征和多条件生成视频方面存在难度。有鉴于此，本公开提供了一种更为优选的实施方式，即采用两阶段的训练策略，在利用第一训练数据训练视频生成模型之前，首先对扩散模型进行预训练，该预训练是利用基于文本条件生成视频的过程来使得扩散模型进行学习，然后在预训练得到的扩散模型的参数基础上进一步利用第一训练数据（即包含全局条件和局部条件等多种条件）训练视频生成模型。

其中，预训练扩散模型的过程可以包括以下步骤：

首先获取包括多个第二训练样本的第二训练数据，第二训练样本包括：从视频样本中提取描述文本作为文本样本。其中可以采用已有的视频描述文本生成模型来获取视频样本的描述文本，也可以从视频样本所来源网页上获取视频样本的描述文本，也可以人工添加视频样本的描述文本，等等。

然后利用全局编码器对文本样本进行编码处理得到文本特征表示。再将得到的文本特征表示和噪声序列输入扩散模型以对扩散模型进行训练，其中扩散模型利用文本特征表示对噪声序列进行去噪声处理；训练的目标包括：最小化扩散模型在各时间步进行去噪声处理时预测的噪声与高斯噪声之间的差异。该部分预训练采用的损失函数与上述公式（1）相同，是包含的条件 c 不同，预训练过程中涉及的条件 c 包含文本条

件。

也就是说，输入条件有文本条件的情况下，利用生成视频的过程来训练扩散模型。然后将训练得到的扩散模型的参数作为步骤 804 中训练视频生成模型时扩散模型采用的初始化参数。

上述对本说明书特定实施例进行了描述。其它实施例在所附权利要求书的范围内。在一些情况下，在权利要求书中记载的动作或步骤可以按照不同于实施例中的顺序来执行并且仍然可以实现期望的结果。另外，在附图中描绘的过程不一定要求示出的特定顺序或者连续顺序才能实现期望的结果。在某些实施方式中，多任务处理和并行处理也是可以的或者可能是有利的。

根据另一方面的实施例，提供了一种视频生成装置。图 9 示出根据一个实施例的视频生成装置的示意性框图，该装置对应图 1 所示系统中的模型训练装置。如图 9 所示，该装置 900 包括：条件获取单元 901 和视频生成单元 902。其中各组成单元的主要功能如下：

条件获取单元 901，被配置为获取局部条件，局部条件包括空间条件和/或时间条件；

视频生成单元 902，被配置为对局部条件进行编码处理得到局部条件特征表示；将噪声序列与局部条件特征表示进行整合，得到噪声隐向量序列；利用扩散模型对噪声隐向量序列进行去噪声处理，得到去噪后的隐向量序列；利用去噪后的隐向量序列进行解码处理，得到视频。

更进一步地，条件获取单元 901，还可以被配置为获取全局条件，全局条件包括文本条件、风格条件和颜色条件中的至少一种。

相应地，视频生成单元 902，还可以别配置为对全局条件进行编码处理得到全局条件特征表示。扩散模型可以利用全局条件特征表示对噪声隐向量序列进行交叉注意力处理以预测噪声，并利用预测的噪声进行去噪声处理。

上述空间条件可以包括单幅图像和单幅语义草图中的至少一种。时间条件包括运动矢量序列、深度图序列、掩膜图序列、语义草图序列和灰度图序列中的至少一种。

其中，视频生成单元 902 可以通过图 3 中所示的视频生成模型实现。全局编码器用以对全局条件进行编码处理得到全局条件特征表示。时空条件编码器用以对局部条件进行编码处理得到局部条件特征表示。扩散模型用以对噪声隐向量序列进行去噪声处理，得到去噪后的隐向量序列。解码器利用去噪后的隐向量序列进行解码处理，得到视频。

作为其中一种可实现的方式，视频生成模型中的全局编码器对文本条件进行文本编码得到文本特征表示，以及对风格条件和/或颜色条件进行图像编码得到图像特征表

示；对文本特征表示和图像特征表示进行整合，得到全局条件特征表示。

作为其中一种可实现的方式，各时空条件编码器分别对各局部条件进行编码，得到各局部条件对应的特征张量，其中时空条件编码器与局部条件一一对应；将各局部条件对应的特征张量进行融合处理，得到局部条件特征表示。

具体地，时空条件编码器可以利用时空条件编码器对局部条件进行空间特征编码，得到该局部条件的空间特征表示；若该局部条件为序列，则对该局部条件的空间特征表示进行时域自注意力处理，得到该局部条件对应的特征张量；否则，对该局部条件的空间特征表示在时序上复制产生时序上的空间特征表示，对时序上的空间特征表示进行时域自注意力处理，得到该局部条件对应的特征张量。

根据另一方面的实施例，提供了一种训练视频生成模型的装置。图 10 示出根据一个实施例的训练视频生成模型的装置的示意性框图。如图 10 所示，该装置 1000 包括：样本获取单元 1001 和模型训练单元 1002，还可以进一步包括预训练单元 1003。其中各组成单元的主要功能如下：

样本获取单元 1001，被配置为获取包括多个第一训练样本的第一训练数据，第一训练样本包括局部条件样本，局部条件样本包括空间条件样本和/或时间条件样本。

模型训练单元 1002，被配置为利用第一训练数据训练视频生成模型，视频生成模型包括：时空条件编码器、扩散模型和解码器；其中，时空条件编码器对局部条件样本进行编码处理得到局部条件特征表示；扩散模型对噪声隐向量序列进行去噪声处理，得到去噪后的隐向量序列，噪声隐向量序列是将噪声序列与局部条件特征表示进行整合后得到的；解码器利用去噪后的隐向量序列进行解码处理，得到视频；

训练的目标包括：最小化扩散模型进行去噪声处理时预测的噪声与高斯噪声之间的差异。

更进一步地，上述第一训练样本还包括全局条件样本，全局条件样本包括文本条件样本、风格条件样本和颜色条件样本中的至少一种。

视频生成模型还包括全局编码器，全局编码器对全局条件样本进行编码处理得到全局条件特征表示。相应地，扩散模型利用全局条件特征表示对噪声隐向量序列进行交叉注意力处理以预测噪声，并利用预测的噪声进行去噪声处理。

作为其中一种可实现的方式，全局编码器和解码器采用预训练得到的参数，在训练的每一轮迭代中，利用与训练的目标对应的损失函数更新时空条件编码器和扩散模型的参数。

作为其中一种可实现的方式，样本获取单元 1001 可以获取视频样本；获取视频样本的描述文本、风格图像和/或颜色直方图分别作为文本条件样本、风格条件样本和/或颜色条件样本；提取单幅图像和单幅语义草图中的至少一种作为空间条件样本；提

取运动矢量序列、深度图序列、掩膜图序列、语义草图序列和/或灰度图序列中的至少一种作为时间条件样本。

作为其中一种可实现的方式，预训练单元 1003 可以被配置为采用如下方式预训练扩散模型：

获取包括多个第二训练样本的第二训练数据，第二训练样本包括：从视频样本中提取描述文本作为文本样本；

利用全局编码器对文本样本进行编码处理得到文本特征表示；

将文本特征表示和噪声序列输入扩散模型以对扩散模型进行训练，扩散模型利用文本特征表示对噪声序列进行去噪声处理；训练的目标包括：最小化扩散模型在各时间步进行去噪声处理时预测的噪声与高斯噪声之间的差异。

相应地，模型训练单元 1002 在预训练单元 1003 预训练得到的扩散模型的参数基础上进一步利用第一训练数据训练视频生成模型。

作为其中一种可实现的方式，时空条件编码器可以利用各时空条件编码器分别对各局部条件样本进行编码，得到各局部条件样本对应的特征张量，其中时空条件编码器与局部条件一一对应；将各局部条件样本对应的特征张量进行融合处理，得到局部条件特征表示。

作为其中一种可实现的方式，时空条件编码器可以对局部条件样本进行空间特征编码，得到该局部条件样本的空间特征表示；若该局部条件样本为序列，则对该局部条件样本的空间特征表示进行时域自注意力处理，得到该局部条件样本对应的特征张量；否则，对该局部条件样本的空间特征表示在时序上复制产生时序上的空间特征表示，对时序上的空间特征表示进行时域自注意力处理，得到该局部条件样本对应的特征张量。

本说明书中的各个实施例均采用递进的方式描述，各个实施例之间相同相似的部分互相参见即可，每个实施例重点说明的都是与其他实施例的不同之处。尤其，对于装置实施例而言，由于其基本相似于方法实施例，所以描述得比较简单，相关之处参见方法实施例的部分说明即可。以上所描述的装置实施例是示意性的，其中所述作为分离部件说明的单元可以是或者也可以不是物理上分开的，作为单元显示的部件可以是或者也可以不是物理单元，即可以位于一个地方，或者也可以分布到多个网络单元上。可以根据实际的需要选择其中的部分或者全部模块来实现本实施例方案的目的。本领域普通技术人员在不付出创造性劳动的情况下，即可以理解并实施。

需要说明的是，本公开所涉及的用户信息（包括但不限于用户设备信息、用户个人信息等）和数据（包括但不限于用于分析的数据、存储的数据、展示的数据等），均为经用户授权或者经过各方充分授权的信息和数据，并且相关数据的收集、使用

处理需要遵守相关国家和地区的相关法律法规和标准，并提供有相应的操作入口，供用户选择授权或者拒绝。

另外，本公开实施例还提供了一种计算机可读存储介质，其上存储有计算机程序，该程序被处理器执行时实现前述方法实施例中任一项所述的方法的步骤。

以及一种电子设备，包括：

一个或多个处理器；以及

与所述一个或多个处理器关联的存储器，所述存储器被设置为存储程序指令，所述程序指令在被所述一个或多个处理器读取执行时，执行前述方法实施例中任一项所述的方法的步骤。

本公开还提供了一种计算机程序产品，包括计算机程序，该计算机程序在被处理器执行时实现前述方法实施例中任一项所述的方法的步骤。

其中，图 11 示例性的展示出了电子设备的架构，具体可以包括处理器 1110，视频显示适配器 1111，磁盘驱动器 1112，输入/输出接口 1113，网络接口 1114，以及存储器 1120。上述处理器 1110、视频显示适配器 1111、磁盘驱动器 1112、输入/输出接口 1113、网络接口 1114，与存储器 1120 之间可以通过通信总线 1130 进行通信连接。

其中，处理器 1110 可以采用通用的 CPU、微处理器、应用专用集成电路（Application Specific Integrated Circuit，ASIC）、或者一个或多个集成电路等方式实现，被设置为执行相关程序，以实现本公开所提供的技术方案。

存储器 1120 可以采用 ROM（Read Only Memory，只读存储器）、RAM（Random Access Memory，随机存取存储器）、静态存储设备，动态存储设备等形式实现。存储器 1120 可以存储被设置为控制电子设备 1100 运行的操作系统 1121，被设置为控制电子设备 1100 的低级别操作的基本输入输出系统(BIOS) 1122。另外，还可以存储网页浏览器 1123，数据存储管理系统 1124，以及视频生成装置/模型训练装置 1125 等等。上述视频生成装置/模型训练装置 1125 就可以是本公开实施例中具体实现前述各步骤操作的应用程序。总之，在通过软件或者固件来实现本公开所提供的技术方案时，相关的程序代码保存在存储器 1120 中，并由处理器 1110 来调用执行。

输入/输出接口 1113 被设置为连接输入/输出模块，以实现信息输入及输出。输入输出/模块可以作为组件配置在设备中（图中未示出），也可以外接于设备以提供相应功能。其中输入设备可以包括键盘、鼠标、触摸屏、麦克风、各类传感器等，输出设备可以包括显示器、扬声器、振动器、指示灯等。

网络接口 1114 被设置为连接通信模块（图中未示出），以实现本设备与其他设备的通信交互。其中通信模块可以通过有线方式（例如 USB、网线等）实现通信，也可以通过无线方式（例如移动网络、WIFI、蓝牙等）实现通信。

总线 1130 包括一通路，在设备的各个组件（例如处理器 1110、视频显示适配器 1111、磁盘驱动器 1112、输入/输出接口 1113、网络接口 1114，与存储器 1120）之间传输信息。

需要说明的是，尽管上述设备示出了处理器 1110、视频显示适配器 1111、磁盘驱动器 1112、输入/输出接口 1113、网络接口 1114，存储器 1120，总线 1130 等，但是在具体实施过程中，该设备还可以包括实现正常运行所必需的其他组件。此外，本领域的技术人员可以理解的是，上述设备中也可以包含实现本公开方案所必需的组件，而不必包含图中所示的全部组件。

通过以上的实施方式的描述可知，本领域的技术人员可以清楚地了解到本公开可借助软件加必需的通用硬件平台的方式来实现。基于这样的理解，本公开的技术方案本质上或者说对现有技术做出贡献的部分可以以计算机程序产品的形式体现出来，该计算机程序产品可以存储在存储介质中，如 ROM/RAM、磁碟、光盘等，包括若干指令用以使得一台计算机设备（可以是个人计算机，服务器，或者网络设备等）执行本公开各个实施例或者实施例的某些部分所述的方法。

以上对本公开所提供的技术方案进行了详细介绍，本文中应用了具体个例对本公开的原理及实施方式进行了阐述，以上实施例的说明只是用于帮助理解本公开的方法及其核心思想；同时，对于本领域的一般技术人员，依据本公开的思想，在具体实施方式及应用范围上均会有改变之处。综上所述，本说明书内容不应理解为对本公开的限制。

工业实用性

本公开实施例提供了一种视频生成方法，包括：获取局部条件，局部条件包括空间条件和/或时间条件，对局部条件进行编码处理得到局部条件特征表示，将噪声序列与局部条件特征表示进行整合，得到噪声隐向量序列，利用扩散模型对噪声隐向量序列进行去噪声处理，得到去噪后的隐向量序列，利用所述去噪后的隐向量序列进行解码处理，得到视频。本公开融入了空间条件和/或时间条件作为局部条件，利用局部条件特征表示与噪声序列进行整合得到噪声隐向量序列，并对噪声隐向量序列进行去噪声处理，得到去噪后的隐向量序列，进而解码得到视频。这种方式使得视频生成不再局限于在文本条件，而是引入空间条件和/或时间条件来引导视频生成，从而更加灵活、多样化地生成视频，提高视频的质量。

权利要求书

1.一种视频生成方法，所述方法包括：

获取局部条件，所述局部条件包括空间条件和/或时间条件；

对所述局部条件进行编码处理得到局部条件特征表示；

将噪声序列与所述局部条件特征表示进行整合，得到噪声隐向量序列；

利用扩散模型对所述噪声隐向量序列进行去噪声处理，得到去噪后的隐向量序列；

利用所述去噪后的隐向量序列进行解码处理，得到视频。

2.根据权利要求1所述的方法，其中，所述方法还包括：获取全局条件，所述全局条件包括文本条件、风格条件和颜色条件中的至少一种；对所述全局条件进行编码处理得到全局条件特征表示；

所述利用扩散模型对所述噪声隐向量序列进行去噪声处理包括：所述扩散模型利用所述全局条件特征表示对所述噪声隐向量序列进行交叉注意力处理以预测噪声，并利用预测的噪声进行去噪声处理。

3.根据权利要求1所述的方法，其中，所述空间条件包括单幅图像和单幅语义草图中的至少一种；

所述时间条件包括运动矢量序列、深度图序列、掩膜图序列、语义草图序列和灰度图序列中的至少一种。

4.根据权利要求2所述的方法，其中，对所述全局条件进行编码处理得到全局条件特征表示包括：对所述文本条件进行文本编码得到文本特征表示，以及对所述风格条件和/或颜色条件进行图像编码得到图像特征表示；对所述文本特征表示和所述图像特征表示进行整合，得到全局条件特征表示。

5.根据权利要求1至4中任一项所述的方法，其中，对所述局部条件进行编码处理得到局部条件特征表示包括：

利用各时空条件编码器分别对各局部条件进行编码，得到各局部条件对应的特征张量，其中时空条件编码器与局部条件一一对应；

将各局部条件对应的特征张量进行融合处理，得到所述局部条件特征表示。

6.根据权利要求5所述的方法，其中，所述利用时空条件编码器分别对各局部条件进行编码包括：

利用时空条件编码器对局部条件进行空间特征编码，得到该局部条件的空间特征表示；

若该局部条件为序列，则对该局部条件的空间特征表示进行时域自注意力处理，得到该局部条件对应的特征张量；否则，对该局部条件的空间特征表示在时序上复制产生时序上的空间特征表示，对所述时序上的空间特征表示进行时域自注意力处理，得到该局部条件对应的特征张量。

7.一种训练视频生成模型的方法，所述方法包括：

获取包括多个第一训练样本的第一训练数据，所述第一训练样本包括局部条件样本，所述局部条件样本包括空间条件样本和/或时间条件样本；

利用所述第一训练数据训练视频生成模型，所述视频生成模型包括：时空条件编码器、扩散模型和解码器；其中，时空条件编码器对所述局部条件样本进行编码处理得到局部条件特征表示；扩散模型对噪声隐向量序列进行去噪声处理，得到去噪后的隐向量序列，所述噪声隐向量序列是将噪声序列与所述局部条件特征表示进行整合后得到的；所述解码器利用所述去噪后的隐向量序列进行解码处理，得到视频；

所述训练的目标包括：最小化扩散模型进行去噪声处理时预测的噪声与高斯噪声之间的差异。

8.根据权利要求7所述的方法，其中，所述第一训练样本还包括全局条件样本，所述全局条件样本包括文本条件样本、风格条件样本和颜色条件样本中的至少一种；

所述视频生成模型还包括全局编码器，所述全局编码器对所述全局条件样本进行编码处理得到全局条件特征表示；

所述扩散模型对噪声隐向量序列进行去噪声处理包括：所述扩散模型利用所述全局条件特征表示对所述噪声隐向量序列进行交叉注意力处理以预测噪声，并利用预测的噪声进行去噪声处理。

9.根据权利要求8所述的方法，其中，所述全局编码器和所述解码器采用预训练得到的参数，在所述训练的每一轮迭代中，利用与所述训练的目标对应的损失函数更新所述时空条件编码器和扩散模型的参数。

10.根据权利要求8所述的方法，其中，所述获取包括多个第一训练样本的第一训练数据包括：

获取视频样本；

获取所述视频样本的描述文本、风格图像和/或颜色直方图分别作为所述文本条件样本、风格条件样本和/或颜色条件样本；

提取单幅图像和单幅语义草图中的至少一种作为所述空间条件样本；

提取运动矢量序列、深度图序列、掩膜图序列、语义草图序列和/或灰度图序列中的至少一种作为所述时间条件样本。

11.根据权利要求7至10中任一项所述的方法，其中，在利用所述第一训练数据训练视频生成模型之前，还包括预训练所述扩散模型；在预训练得到的扩散模型的参数基础上进一步利用所述第一训练数据训练视频生成模型；

其中，预训练所述扩散模型包括：

获取包括多个第二训练样本的第二训练数据，所述第二训练样本包括：从视频样本中提取描述文本作为文本样本；

利用全局编码器对所述文本样本进行编码处理得到文本特征表示；

将所述文本特征表示和噪声序列输入扩散模型以对所述扩散模型进行训练，所述扩散模型利用所述文本特征表示对所述噪声序列进行去噪声处理；所述训练的目标包括：最小化扩散模型在各时间步进行去噪声处理时预测的噪声与高斯噪声之间的差异。

12.根据权利要求 7 至 10 中任一项所述的方法，其中，时空条件编码器对所述局部条件样本进行编码处理得到局部条件特征表示包括：

利用各时空条件编码器分别对各局部条件样本进行编码，得到各局部条件样本对应的特征张量，其中时空条件编码器与局部条件一一对应；

将各局部条件样本对应的特征张量进行融合处理，得到所述局部条件特征表示。

13.根据权利要求 12 所述的方法，其中，所述利用各时空条件编码器分别对各局部条件样本进行编码包括：

利用时空条件编码器对局部条件样本进行空间特征编码，得到该局部条件样本的空间特征表示；

若该局部条件样本为序列，则对该局部条件样本的空间特征表示进行时域自注意力处理，得到该局部条件样本对应的特征张量；否则，对该局部条件样本的空间特征表示在时序上复制产生时序上的空间特征表示，对所述时序上的空间特征表示进行时域自注意力处理，得到该局部条件样本对应的特征张量。

14.一种视频生成方法，由云端服务器执行，所述方法包括：

获取来自用户终端的局部条件，所述局部条件包括空间条件和/或时间条件；

对所述局部条件进行编码处理得到局部条件特征表示；

将噪声序列与所述局部条件特征表示进行整合，得到噪声隐向量序列；

利用扩散模型对所述噪声隐向量序列进行去噪声处理，得到去噪后的隐向量序列；

利用所述去噪后的隐向量序列进行解码处理，得到视频；

将所述视频发送给所述用户终端以进行展示。

15.一种视频生成装置，所述装置包括：

条件获取单元，被配置为获取局部条件，所述局部条件包括空间条件和/或时间条件；

视频生成单元，被配置为对所述局部条件进行编码处理得到局部条件特征表示；将噪声序列与所述局部条件特征表示进行整合，得到噪声隐向量序列；利用扩散模型对所述噪声隐向量序列进行去噪声处理，得到去噪后的隐向量序列；利用所述去噪后的隐向量序列进行解码处理，得到视频。

16.一种训练视频生成模型的装置，所述装置包括：

样本获取单元，被配置为获取包括多个第一训练样本的第一训练数据，所述第一训练样本包括局部条件样本，所述局部条件样本包括空间条件样本和/或时间条件样本；

模型训练单元，被配置为利用所述第一训练数据训练视频生成模型，所述视频生成模型包括：时空条件编码器、扩散模型和解码器；其中，时空条件编码器对所述局

部条件样本进行编码处理得到局部条件特征表示；扩散模型对噪声隐向量序列进行去噪声处理，得到去噪后的隐向量序列，所述噪声隐向量序列是将噪声序列与所述局部条件特征表示进行整合后得到的；所述解码器利用所述去噪后的隐向量序列进行解码处理，得到视频；

所述训练的目标包括：最小化扩散模型进行去噪声处理时预测的噪声与高斯噪声之间的差异。

17.一种计算机可读存储介质，其上存储有计算机程序，其中，该程序被处理器执行时实现权利要求 1 至 14 中任一项所述的方法的步骤。

18.一种电子设备，包括：

一个或多个处理器；以及

与所述一个或多个处理器关联的存储器，所述存储器被设置为存储程序指令，所述程序指令在被所述一个或多个处理器读取执行时，执行权利要求 1 至 14 中任一项所述的方法的步骤。

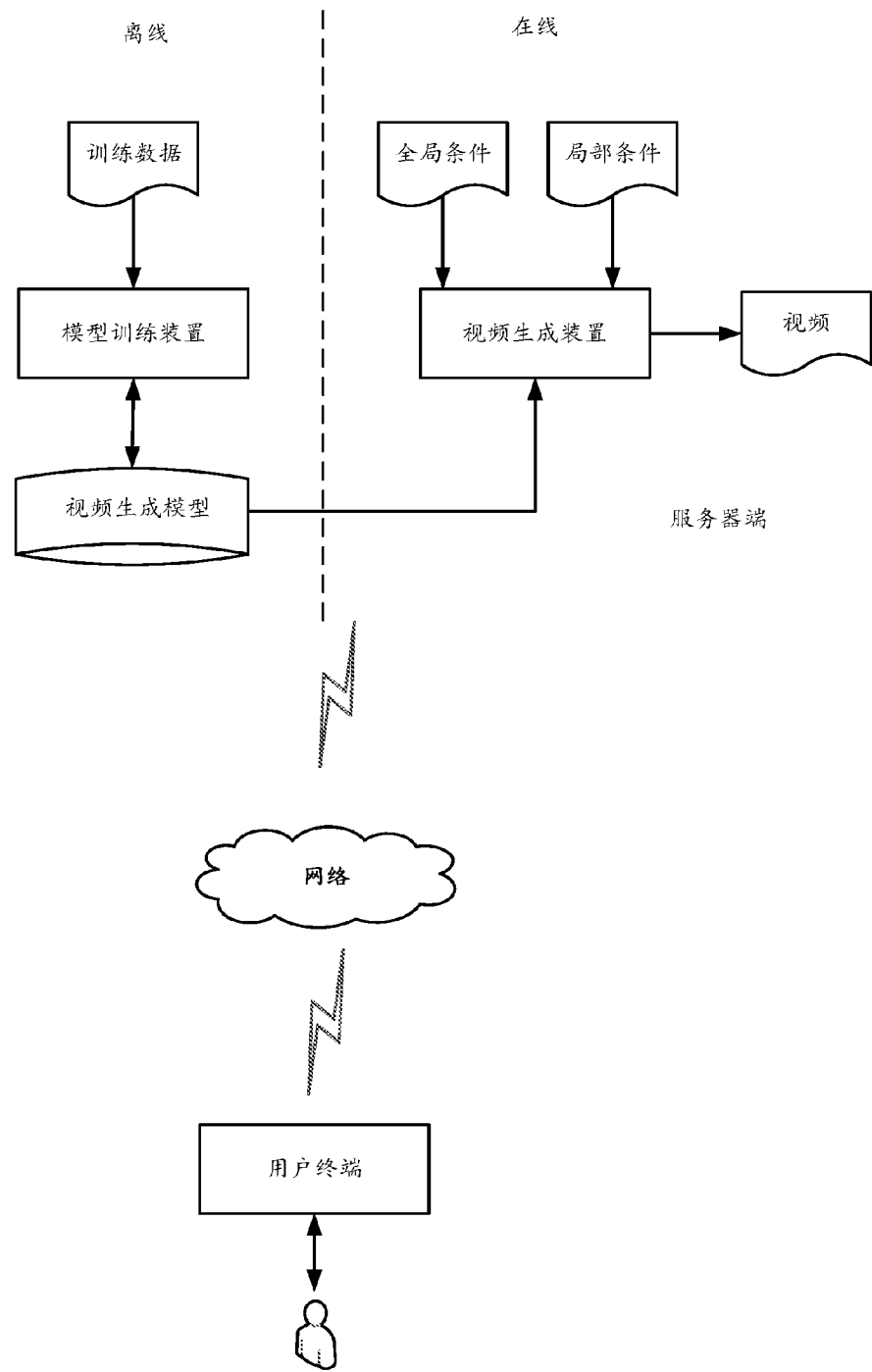


图 1

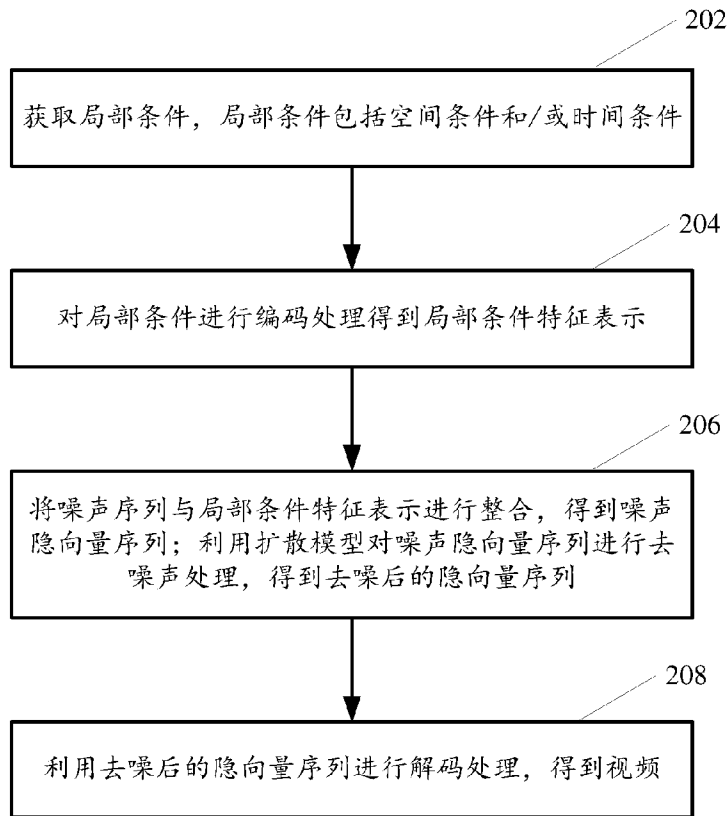


图 2

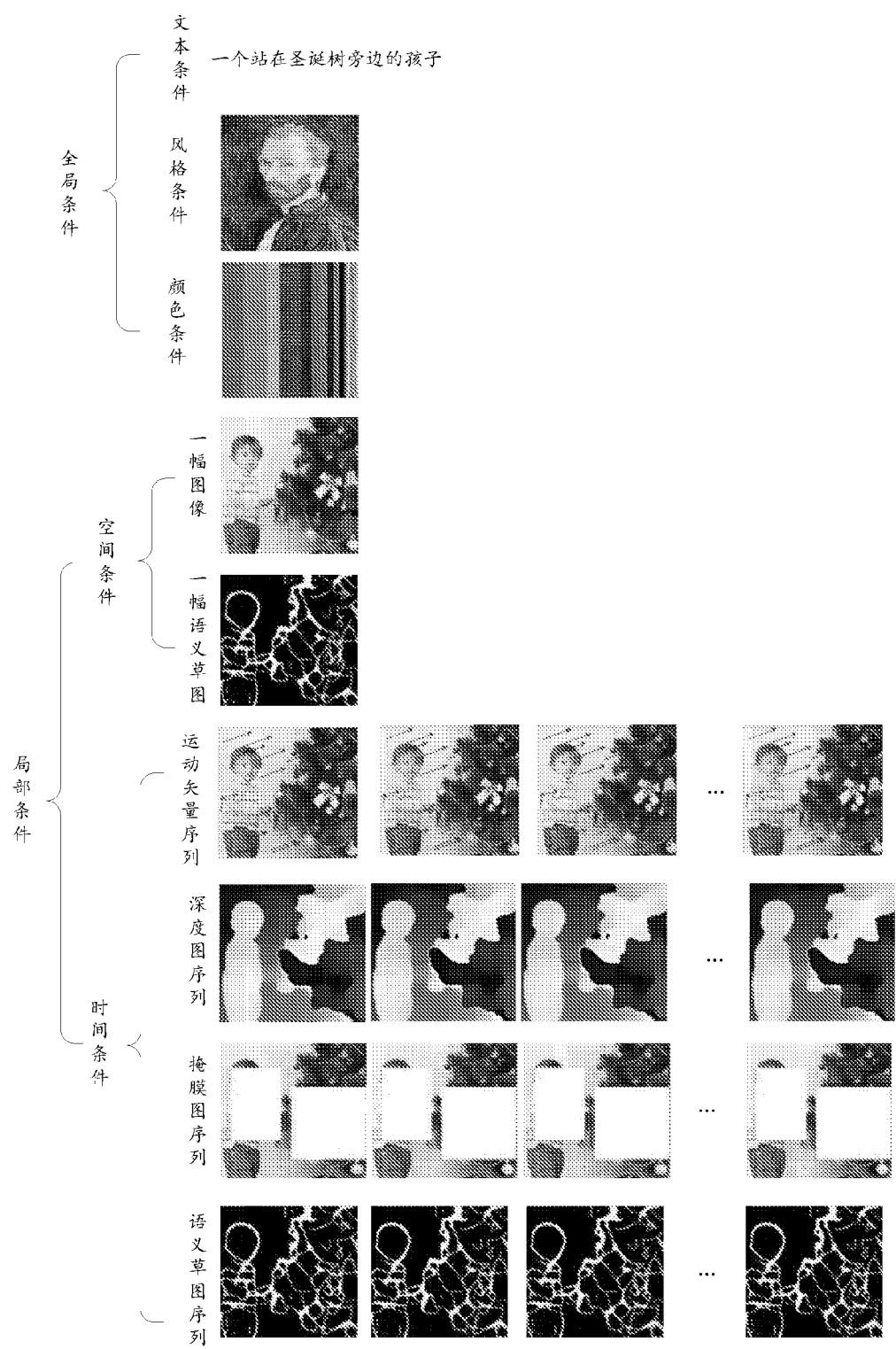


图 3

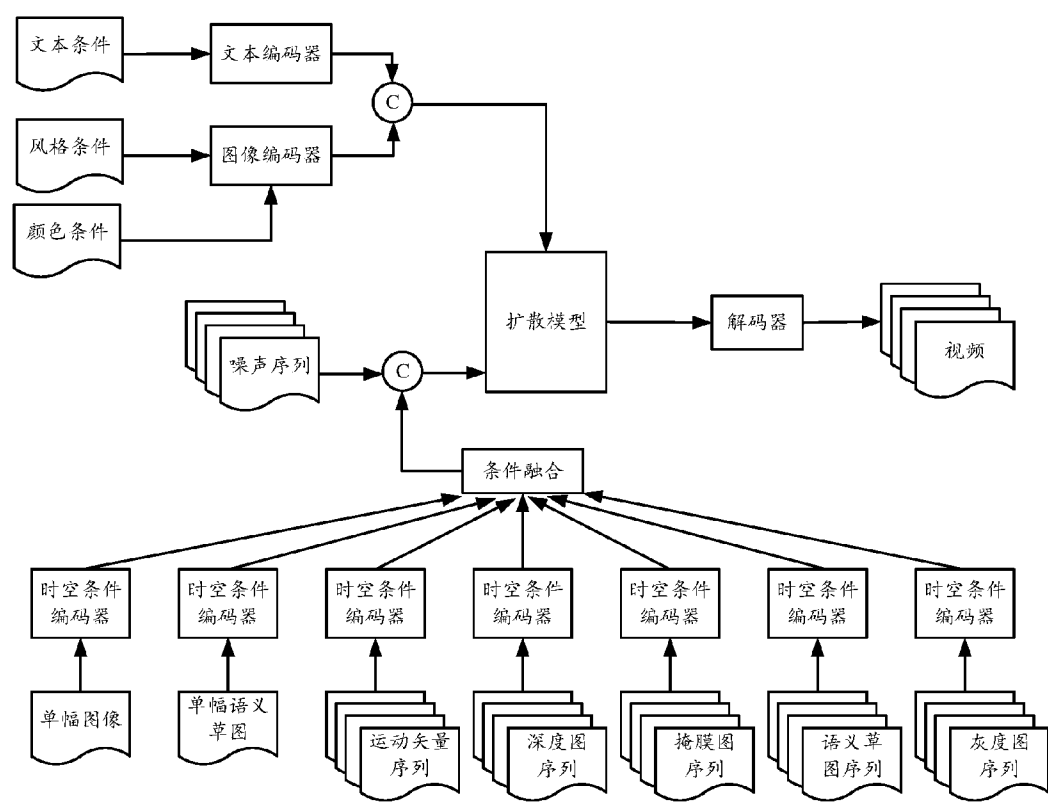


图 4

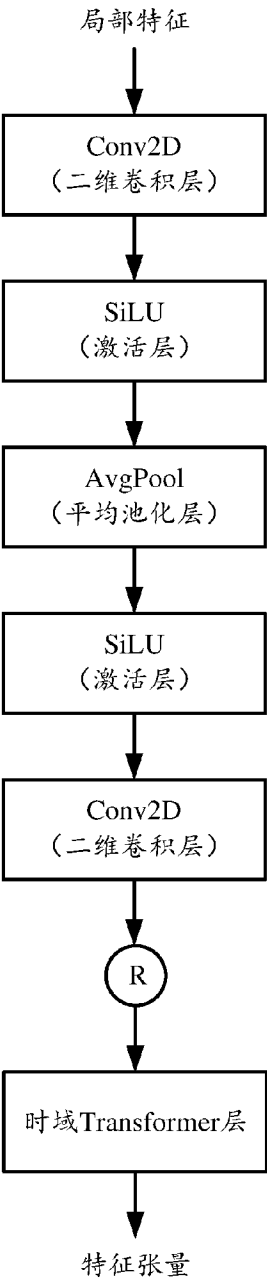


图 5

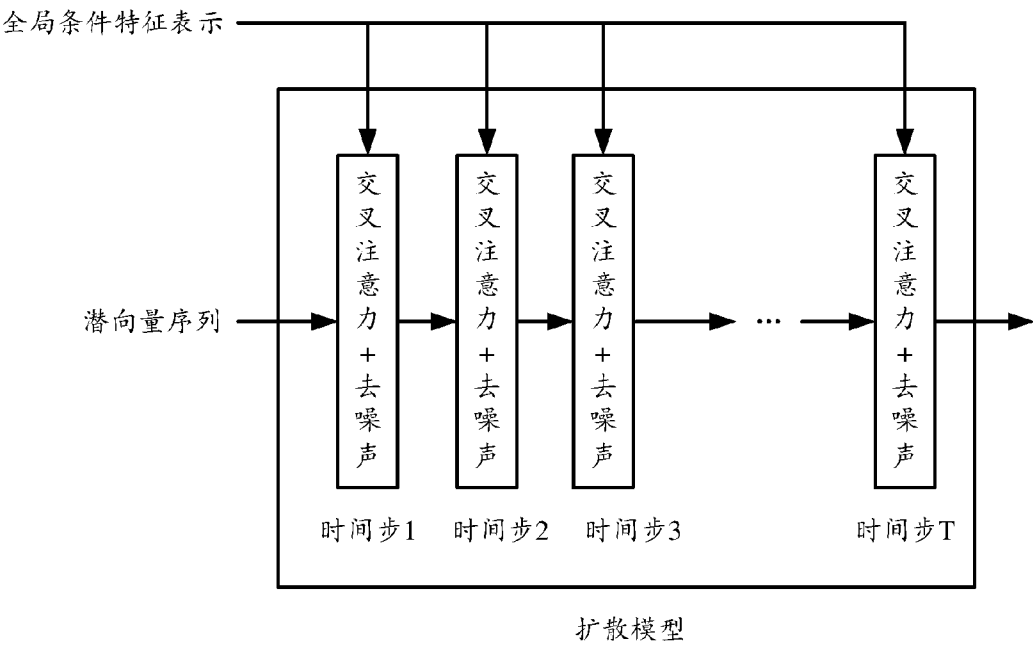


图 6

文本条件： 一个站在森林中的长发女人的旋转视角

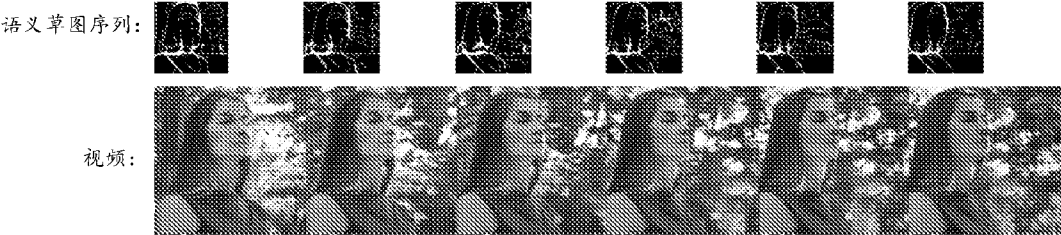


图 7a

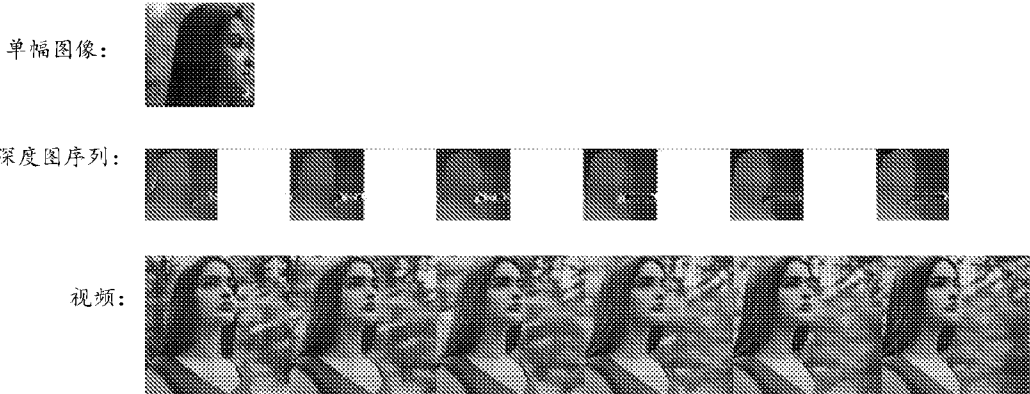


图 7b

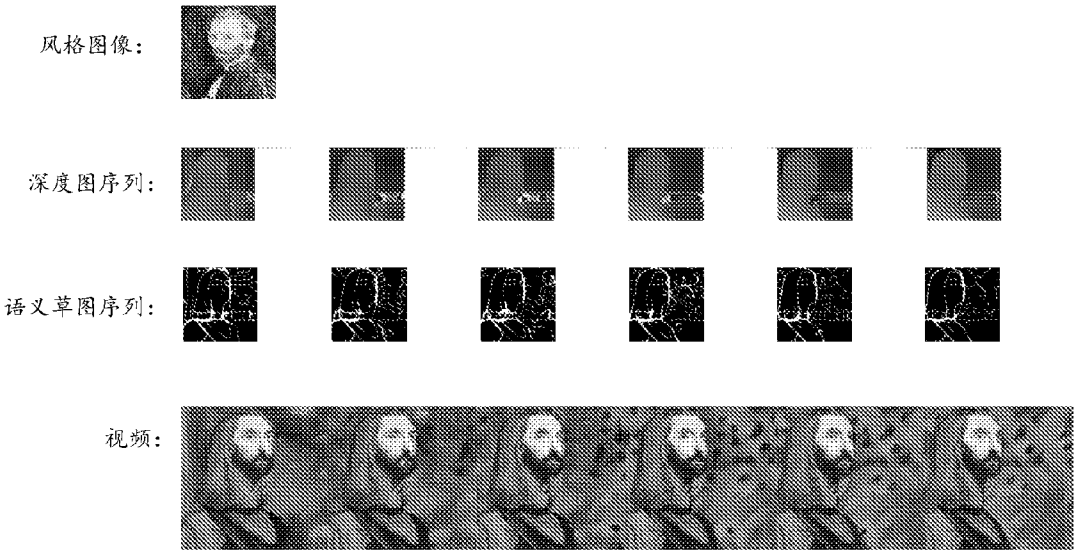


图 7c

文本条件: 一轮移动的金色月亮

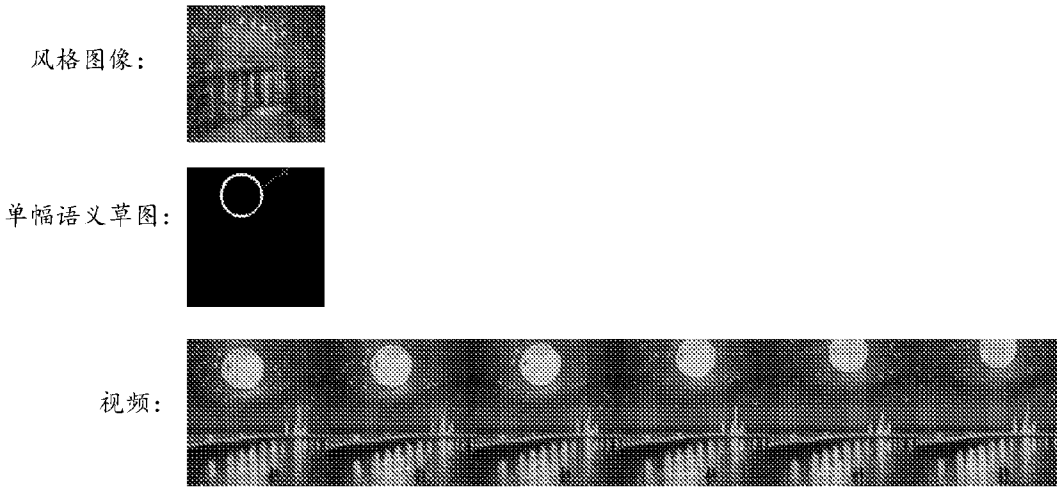


图 7d

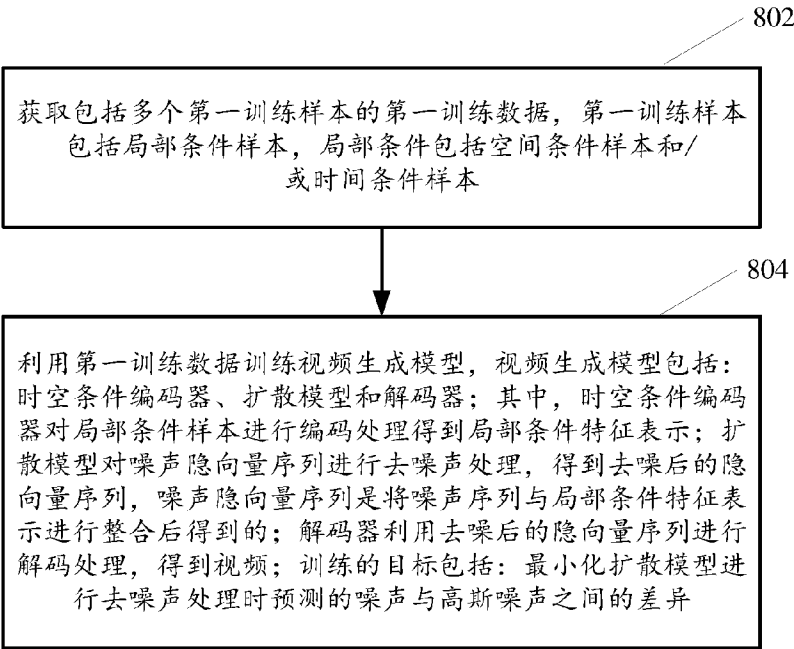


图 8

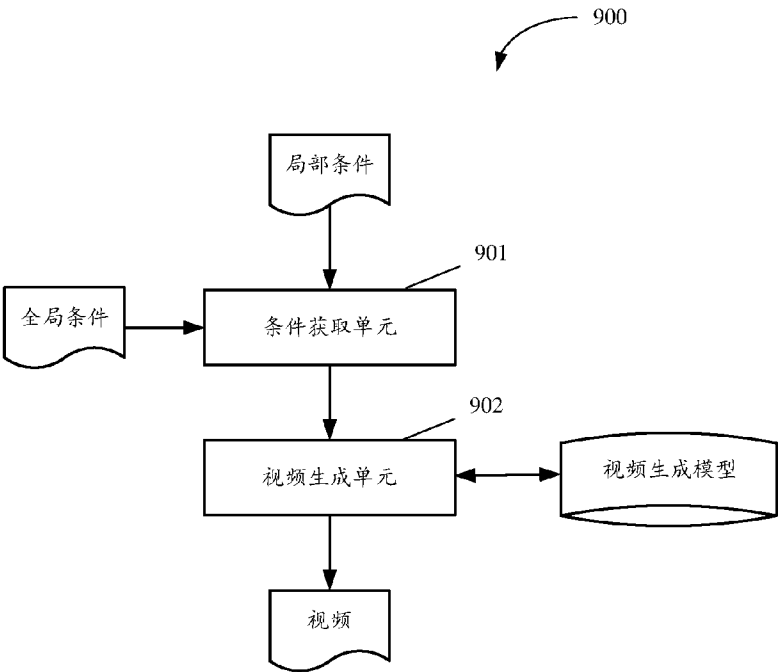


图 9

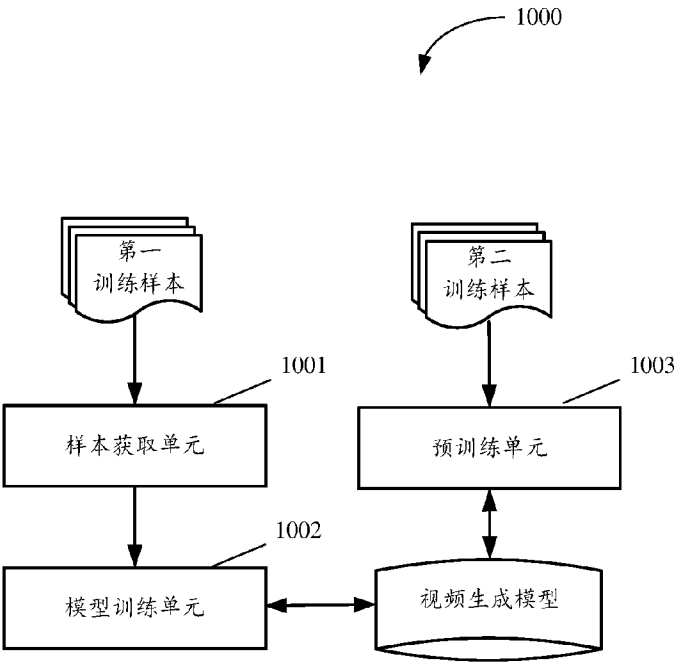


图 10

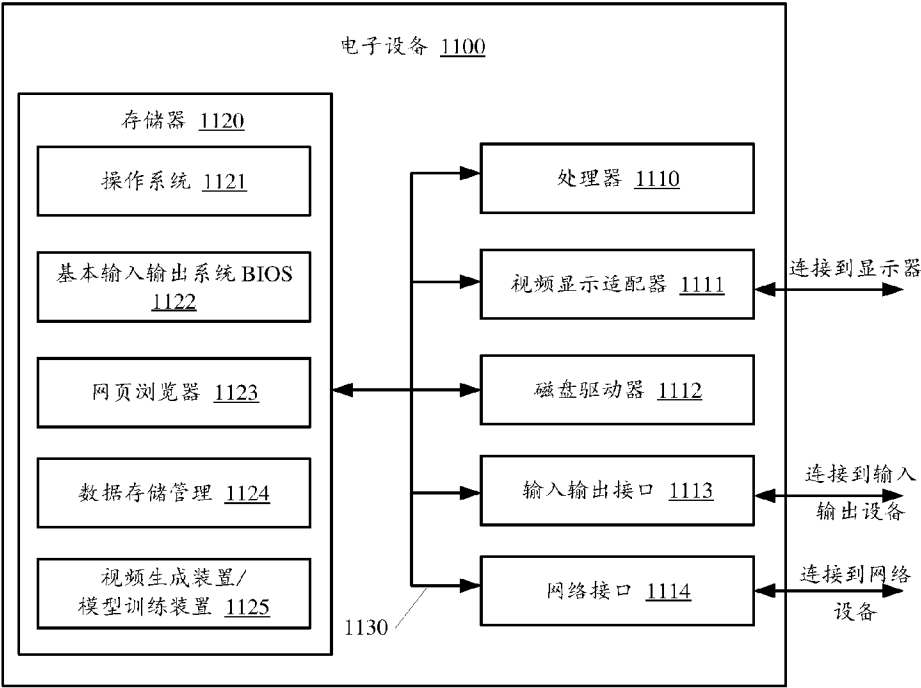


图 11

INTERNATIONAL SEARCH REPORT

International application No.
PCT/SG2024/050361

A. CLASSIFICATION OF SUBJECT MATTER				
G06N 3/0455 (2023.01) G06N 3/08 (2023.01) G06T 9/00 (2006.01) According to International Patent Classification (IPC)				
B. FIELDS SEARCHED				
Minimum documentation searched (classification system followed by classification symbols) G06N, G06T				
Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched				
Electronic data base consulted during the international search (name of data base and, where practicable, search terms used) FamPat, IEEE, Google Scholar: Video, 视频, generate, produce, 生成, diffusion, 扩散, model, 模型, noise, 噪声, condition, 条件, latent, 隐, implicit, sequence, 序列, vector, 矢量, variable, 变量, feature, attribute, 特征, 属性, denoise, 去噪, encode, 编码, decode,				
C. DOCUMENTS CONSIDERED TO BE RELEVANT				
Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.		
X	NI H. et al, Conditional Image-to-Video Generation with Latent Flow Diffusion Models. <i>Computer Science > Computer Vision and Pattern Recognition</i>, arXiv:2303.13744[cs.CV], 24 March 2023, pages 1-15 [searched on 29 August 2024] <DOI:HTTPS://DOI.ORG/10.48550/ARXIV.2303.13744>, figures 2-3, and sections 1-4	1-18		
☒ Further documents are listed in the continuation of Box C. ☒ See patent family annex.				
<table border="0"> <tr> <td style="vertical-align: top;"> * Special categories of cited documents: “A” document defining the general state of the art which is not considered to be of particular relevance “D” document cited by the applicant in the international application “E” earlier application or patent but published on or after the international filing date “L” document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified) “O” document referring to an oral disclosure, use, exhibition or other means “P” document published prior to the international filing date but later than the priority date claimed </td> <td style="vertical-align: top;"> “T” later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention “X” document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone “Y” document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art “&” document member of the same patent family </td> </tr> </table>			* Special categories of cited documents: “A” document defining the general state of the art which is not considered to be of particular relevance “D” document cited by the applicant in the international application “E” earlier application or patent but published on or after the international filing date “L” document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified) “O” document referring to an oral disclosure, use, exhibition or other means “P” document published prior to the international filing date but later than the priority date claimed	“T” later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention “X” document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone “Y” document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art “&” document member of the same patent family
* Special categories of cited documents: “A” document defining the general state of the art which is not considered to be of particular relevance “D” document cited by the applicant in the international application “E” earlier application or patent but published on or after the international filing date “L” document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified) “O” document referring to an oral disclosure, use, exhibition or other means “P” document published prior to the international filing date but later than the priority date claimed	“T” later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention “X” document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone “Y” document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art “&” document member of the same patent family			
Date of the actual completion of the international search 29 August 2024		Date of the issuance of the international search report 02 September 2024		
Name and mailing address of the ISA/SG  Intellectual Property Office of Singapore 1 Paya Lebar Link, #11-03 PLQ 1, Paya Lebar Quarter Singapore 408533 E-mail: pct@ipos.gov.sg		Authorized officer <u>Lu Liru</u> (Dr.) IPOS Telephone No. (+65) 6339 8616		

INTERNATIONAL SEARCH REPORT

International application No.

PCT/SG2024/050361

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	CN 115965791 A (BEIJING ZITIAO NETWORK TECHNOLOGY CO., LTD.), 14 April 2023, original document, paragraphs [0032]-[0072]	1, 14-15, 17-18
A	WO 2022/265992 A1 (GOOGLE LLC), 22 December 2022, entire document	-

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/SG2024/050361

Note: related known patent family members of the patent documents referred to by this international search report are listed in this annex. This authority only provides information on the related patent family members, and assumes no responsibility for the details of the patents.

Patent Documents referred in the Report	Publication Date/Announcement Date (day/month/year)	Patent Family	Publication Date/Announcement Date (day/month/year)
CN 115965791 A	14/04/2023	WO 2024/131597 A1	27/06/2024
WO 2022/265992 A1	22/12/2022	CN 117296061 A	26/12/2023
		EP 4295273 A1	27/12/2023
		US 2023/0267315 A1	24/08/2023
		JP 2024519657 A	21/05/2024

国际检索报告

国际申请号

PCT/SG2024/050361

A. 主题的分类

G06N 3/0455 (2023.01) G06N 3/08 (2023.01) G06T 9/00 (2006.01)

按照国际专利分类(IPC)

B. 检索领域

检索的最低限度文献(标明分类系统和分类号)

G06N, G06T

包含在检索领域中的除最低限度文献以外的检索文献

在国际检索时查阅的电子数据库(数据库的名称, 和使用的检索词(如使用))

FamPat, IEEE, Google Scholar: Video, 视频, generate, produce, 生成, diffusion, 扩散, model, 模型, noise, 噪声, condition, 条件, latent, 隐, implicit, sequence, 序列, vector, 矢量, variable, 变量, feature, attribute, 特征, 属性, denoise, 去噪, encode, 编码, decode, attention, 注意力, temporal, space 及类似关键词

C. 相关文件

类型*	引用文件, 必要时, 指明相关段落	相关的权利要求
X	NI H.等, Conditional Image-to-Video Generation with Latent Flow Diffusion Models. <i>Computer Science > Computer Vision and Pattern Recognition, arXiv:2303.13744 [cs.CV]</i> , 2023 年 03 月 24 日, 第 1-15 页 [于 2024 年 08 月 29 日 检索] <DOI: HTTPS://DOI.ORG/10.48550/ARXIV.2303.13744> 图 2-3, 第 1-4 节	1-18
☒ 其余文件在 C 栏的续页中列出。 ☒ 见同族专利附件。		

* 引用文件的具体类型:

"A" 认为不特别相关的表示了现有技术一般状态的文件	"T" 在申请日或优先权日之后公布, 与申请不相抵触, 但为了理解发明之理论或原理的在后文件
"D" 在此国际申请中引用的文件	"X" 特别相关的文件, 单独考虑该文件, 认定要求保护的发明不是新颖的或不具有创造性
"E" 在国际申请日的当天或之后公布的在先申请或专利	"Y" 特别相关的文件, 当该文件与另一篇或者多篇该类文件结合并且这种结合对于本领域技术人员为显而易见时, 要求保护的发明不具有创造性
"L" 可能对优先权要求构成怀疑的文件, 或为确定另一篇引用文件的公布日而引用的或者因其他特殊理由而引用的文件 (如具体说明的)	"&" 同族专利的文件
"O" 涉及口头公开、使用、展览或其他方式公开的文件	
"P" 公布日先于申请日但迟于所要求的优先权日的文件	

国际检索实际完成的日期

2024 年 08 月 29 日

国际检索报告发文日

2024 年 09 月 02 日

ISA/SG 的名称和邮寄地址

新加坡知识产权局
 巴耶利巴连路 1 号 #11-03
PLQ 1, 巴耶利峇中心
新加坡 408533
电子邮件账号: pct@ipos.gov.sg

受权官员

Lu Liru (Dr)

IPOS 客服电话号码: (+65) 6339 8616

国际检索报告

国际申请号

PCT/SG2024/050361

C (续). 相关文件		
类型*	引用文件, 必要时, 指明相关段落	相关的权利要求
X	CN 115965791 A (BEIJING ZITIAO NETWORK TECHNOLOGY) 2023 年 04 月 14 日 原文中的第[0032]-[0072]段	1, 14-15, 17-18
A	WO 2022/265992 A1 (GOOGLE LLC) 2022 年 12 月 22 日 全文	-

国际检索报告

关于同族专利的信息

国际申请号

PCT/SG2024/050361

备注: 此附件列出了本国际检索报告引用的专利文件的相关已知同族专利。本单位仅提供相关同族专利的信息, 并不对专利的详情承担任何责任。

检索报告中引用的专利文件	公布日或公告日	同族专利	公布日或公告日
CN 115965791 A	14/04/2023	WO 2024/131597 A1	27/06/2024
WO 2022/265992 A1	22/12/2022	CN 117296061 A	26/12/2023
		EP 4295273 A1	27/12/2023
		US 2023/0267315 A1	24/08/2023
		JP 2024519657 A	21/05/2024