(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
2025 年 6 月 5 日 (05.06.2025)

WIPO | PCT

(10) 国际公布号
WO 2025/112588 A1

(51) 国际专利分类号:
G06T 11/00 (2006.01)

(21) 国际申请号: PCT/CN2024/107883

(22) 国际申请日: 2024 年 7 月 26 日 (26.07.2024)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
202311613798.2 2023年11月28日 (28.11.2023) CN

(71) 申请人:阿里巴巴（中国）有限公司 **(ALIBABA (CHINA) CO., LTD)** [CN/CN]; 中国浙江省杭州市滨江区长河街道网商路699号4号楼5楼508室 310052 (CN)。

(72) 发明人:吴志凡**(WU, Zhifan)**; 中国北京市朝阳区广善路18号阿里巴巴北京朝阳科技园C栋 100102 (CN)。 黄梁华**(HUANG, Lianghua)**; 中国北京市朝阳区广善路18号阿里巴巴北京朝阳科技园C栋 100102 (CN)。 王威**(WANG, Wei)**; 中国北京市朝阳区广善路18号阿里巴巴北京朝阳科技园C栋 100102 (CN)。 魏延恒**(WEI, Yanheng)**; 中国北京市朝阳区广善路18号阿里巴巴北京朝阳科技园C栋 100102 (CN)。 刘宇**(LIU, Yu)**; 中国北京市朝阳区广善路18号阿里巴巴北京朝阳科技园C栋 100102 (CN)。

(74) 代理人:北京博浩百睿知识产权代理有限责任公司**(KSK IP AGENCY)**; 中国北京市海淀区知春路甲48号盈都大厦C座1单元10E 100098 (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, 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,

WO 2025/112588 A1

(54) **Title:** IMAGE GENERATION METHOD, ELECTRONIC DEVICE, AND COMPUTER READABLE STORAGE MEDIUM

(54) 发明名称: 图像生成方法、电子设备及计算机可读存储介质

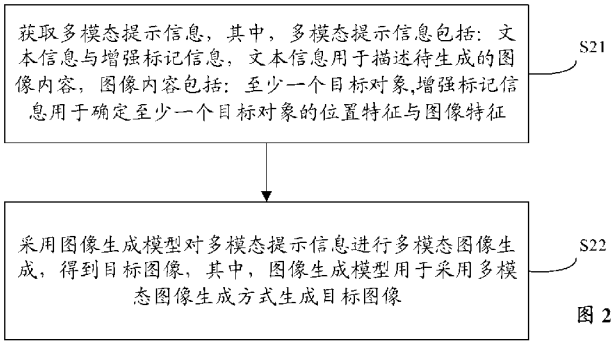

图 2

S21 Acquire multi-modal prompt information, wherein the multi-modal prompt information comprises: text information and augmented token information, the text information is used for describing image content to be generated, said image content comprises: at least one target object, and the augmented token information is used for determining position features and image features of the at least one target object

S22 Use an image generation model to perform multi-modal image generation on the multi-modal prompt information to obtain a target image, wherein the image generation model is used for using a multi-modal image generation mode to generate the target image

(57) **Abstract:** The present disclosure relates to the technical fields of large model technology and image processing. Disclosed are an image generation method, an electronic device, and a computer readable storage medium. The method comprises: acquiring multi-modal prompt information, wherein the multi-modal prompt information comprises: text information and augmented token information, the text information is used for describing image content to be generated, said image content comprises: at least one target object, and the augmented token information is used for determining position features and image features of the at least one target object; and using an image generation model to perform multi-modal image generation on the multi-modal prompt information to obtain a target image, wherein the image generation model is used for using a multi-modal image generation mode to generate the target image. The present disclosure solves the technical problems in the related art of low accuracy of generated images and limited image generation range caused by using only text prompt to generate the corresponding images.

[见续页]

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))。

(57) 摘要: 本公开公开了一种图像生成方法、电子设备及计算机可读存储介质，涉及大模型技术、图像处理技术领域。该方法包括：获取多模态提示信息，其中，多模态提示信息包括：文本信息与增强标记信息，文本信息用于描述待生成的图像内容，图像内容包括：至少一个目标对象，增强标记信息用于确定至少一个目标对象的位置特征与图像特征；采用图像生成模型对多模态提示信息进行多模态图像生成，得到目标图像，其中，图像生成模型用于采用多模态图像生成方式生成目标图像。本公开解决了相关技术中仅通过文本提示生成对应图像，导致生成的图像精确度较低，图像生成范围较局限的技术问题。

说 明 书

图像生成方法、电子设备及计算机可读存储介质

技术领域

本公开涉及大模型技术、图像处理技术领域，具体而言，涉及一种图像生成方法、电子设备及计算机可读存储介质。

背景技术

随着人工智能的发展，文本到图像生成领域取得了重大进展，能够基于文本提示生成对应的高质量图像，从而实现自动生成图像的功能。

目前，仅通过文本提示描述物体从而生成对应的图像，由于文本提示描述物理较为困难，因此生成的图像的精确度较低，并且该方法通常需要微调或仅支持使用单个物体作为约束条件，因此图像生成范围较为局限。

针对上述的问题，目前尚未提出有效的解决方案。

发明内容

本公开实施例提供了一种图像生成方法、电子设备及计算机可读存储介质，以至少解决相关技术中仅通过文本提示生成对应图像，导致生成的图像精确度较低，图像生成范围较局限的技术问题。

根据本公开实施例的一个方面，提供了一种图像生成方法，包括：获取多模态提示信息，其中，多模态提示信息包括：文本信息与增强标记信息，文本信息用于描述待生成的图像内容，图像内容包括：至少一个目标对象，增强标记信息用于确定至少一个目标对象的位置特征与图像特征；采用图像生成模型对多模态提示信息进行多模态图像生成，得到目标图像，其中，图像生成模型用于采用多模态图像生成方式生成目标图像。

根据本公开实施例的另一方面，还提供了另一种图像生成方法，通过终端设备提供一图形用户界面，图形用户界面所显示的内容至少部分地包含一图像生成场景，包括：响应对图形用户界面执行的第一控制操作，输入文本信息，其中，文本信息用于描述待生成的图像内容，图像内容包括：至少一个目标对象；响应对图形用户界面执行的第二控制操作，基于文本信息分别生成至少一个目标对象的位置特征与至少一个目标对象的图像特征以得到增强标记信息，以及采用图像生成模型对多模态提示信息进行多模态图像生成以得到目标图像，其中，增强标记信息用于确定至少一个目标对象的位置特征与图像特征，图像生成模型用于采用多模态图像生成方式生成目标图像，多模态提示信息包括：文本信息与增强标记信息；在图形用户界面内展示目标图像。

根据本公开实施例的另一方面，还提供了另一种图像生成方法，包括：接收当前

输入的多模态对话请求，其中，多模态对话请求中携带的信息包括：多模态对话文本信息与多模态对话增强标记信息，多模态对话文本信息用于描述待生成的多模态对话图像内容，多模态对话图像内容包括：至少一个对象，多模态对话增强标记信息用于确定至少一个对象的位置特征与图像特征；采用图像生成模型对多模态对话请求进行多模态图像生成，得到多模态对话图像，其中，图像生成模型用于采用多模态图像生成方式生成多模态对话图像；反馈多模态对话回复，其中，多模态对话回复中携带的信息包括：多模态对话图像。

根据本公开实施例的另一方面，还提供了一种电子设备，包括：存储器，存储有可执行程序；处理器，用于运行程序，其中，程序运行时执行任意一项上述的图像生成方法。

根据本公开实施例的另一方面，还提供了一种计算机可读存储介质，计算机可读存储介质包括存储的可执行程序，其中，在可执行程序运行时控制计算机可读存储介质所在设备执行任意一项上述的图像生成方法。

在本公开实施例中，通过获取包括文本信息与增强标记信息的多模态提示信息，能够根据文本信息确定待生成的图像内容、根据增强标记信息确定图像内容中至少一个目标对象的位置特征和图像特征，然后采用图像生成模型对多模态提示信息进行多模态图像生成，从而得到目标图像，由此达到了通过多模态提示信息准确生成包括多目标对象的高质量目标图像的目的。此外，能够对生成的图像实现更精准的控制，生成准确度更高的图像，且支持生成多目标对象的图像，使得图像的生成范围更广泛，从而实现了更精确、更多样化、支持多目标对象的高质量图像生成，以满足不同领域的图像生成需求的技术效果，进而解决了相关技术中仅通过文本提示生成对应图像，导致生成的图像精确度较低，图像生成范围较局限的技术问题。

容易注意到的是，上面的通用描述和后面的详细描述仅仅是为了对本公开进行举例和解释，并不构成对本公开的限定。

附图说明

此处所说明的附图用来提供对本公开的进一步理解，构成本公开的一部分，本公开的示意性实施例及其说明用于解释本公开，并不构成对本公开的不当限定。在附图中：

图 1 是根据本公开实施例 1 的一种图像生成方法的应用场景示意图；

图 2 是根据本公开实施例 1 的一种图像生成方法的流程图；

图 3 是根据本公开实施例 1 的另一种图像生成方法的流程图；

图 4 是根据本公开实施例 2 的一种图像生成方法的流程图；

图 5 是根据本公开实施例 3 的一种图像生成方法的流程图；

图 6 是根据本公开实施例 3 的一种人机对话场景示意图；

图 7 是根据本公开实施例 4 的一种图像生成装置的结构示意图;

图 8 是根据本公开实施例 4 的另一种图像生成装置的结构示意图;

图 9 是根据本公开实施例 4 的再一种图像生成装置的结构示意图;

图 10 是根据本公开实施例 5 的一种计算机终端的结构框图。

具体实施方式

为了使本技术领域的人员更好地理解本公开方案，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，所描述的实施例仅仅是本公开一部分的实施例，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做出创造性劳动前提下所获得的所有其他实施例，都应当属于本公开保护的范围。

需要说明的是，本公开的说明书和权利要求书及上述附图中的术语“第一”、“第二”等是用于区别类似的对象，而不必用于描述特定的顺序或先后次序。应该理解这样使用的数据在适当情况下可以互换，以便这里描述的本公开的实施例能够以除了在这里图示或描述的那些以外的顺序实施。此外，术语“包括”和“具有”以及他们的任何变形，意图在于覆盖不排他的包含，例如，包含了一系列步骤或单元的过程、方法、系统、产品或设备不必限于清楚地列出的那些步骤或单元，而是可包括没有清楚地列出的或对于这些过程、方法、产品或设备固有的其它步骤或单元。

本公开提供的技术方案主要采用大模型技术实现，此处的大模型是指具有大规模模型参数的深度学习模型，通常可以包含上亿、上百亿、上千亿、上万亿甚至十万亿以上的模型参数。大模型又可以称为基石模型/基础模型（Foundation Model），通过大规模无标注的语料进行大模型的预训练，产出亿级以上参数的预训练模型，这种模型能适应广泛的下游任务，模型具有较好的泛化能力，例如大规模语言模型（Large Language Model，简称 LLM）、多模态预训练模型（multi-modal pre-training model）等。

需要说明的是，大模型在实际应用时，可以通过少量样本对预训练模型进行微调，使得大模型可以应用于不同的任务中。例如，大模型可以广泛应用于自然语言处理（Natural Language Processing，简称 NLP）、计算机视觉、语音处理等领域，具体可以应用于如视觉问答（Visual Question Answering，简称 VQA）、图像描述（Image Caption，简称 IC）、图像生成等计算机视觉领域任务，也可以广泛应用于基于文本的情感分类、文本摘要生成、机器翻译等自然语言处理领域任务。因此，大模型主要的应用场景包括但不限于数字助理、智能机器人、搜索、在线教育、办公软件、电子商务、智能设计等。

首先，在对本公开实施例进行描述的过程中出现的部分名词或术语适用于如下解释:

扩散模型（Diffusion Model）：一种用于描述扩散过程的数学模型，是常用的生成

模型。本公开实施例中，扩散模型能够基于扩散过程将带噪声图片进行去噪，从而得到生成图片。

组成树（constituency tree）：一种用于表示自然语言句子结构的树状结构，由节点和边组成，其中节点代表句子中的词语或短语，边表示这些词语或短语之间的语法关系。在组成树中，句子被分解为不同的短语结构，如名词短语、动词短语、形容词短语等。这些短语结构可以进一步分解为更小的短语结构，直到最小的单词级别。组成树可以帮助理解句子的结构和语法关系，有助于进行句法分析和语义分析，还可以用于自然语言处理任务，如句法分析、机器翻译和信息抽取等。通过分析组成树，可以更好地理解句子的语法结构和语义含义。

自然语言处理（Natural Language Processing，NLP）解析器：一种用于自然语言处理的工具，可以将自然语言文本转换成结构化的数据，并进行语法分析、词性标注、命名实体识别等任务。NLP 解析器通常基于机器学习和深度学习技术，能够自动分析文本并提取其中的信息，可以帮助理解和处理大量的自然语言数据，例如语音识别、情感分析、文本分类等。

文本编码器（Text Encoder）：一种用于将文本数据转换为数值表示的模型或工具。在自然语言处理（NLP）领域中，文本编码器通常用于将文本数据转化为向量或矩阵形式，以便进行机器学习和深度学习任务，如文本分类、语义相似度计算、情感分析等。

微调（Fine-tune）：是指在使用预训练模型的基础上，通过在特定任务上进行进一步的训练和调整，以提高模型在该任务上的性能和适应能力。通常情况下，预训练模型是在大规模数据集上进行训练的，而 Fine-tune 则是在特定的小规模数据集上进行微调，以使模型更好地适应具体的任务需求。Fine-tune 通常用于自然语言处理、计算机视觉等领域的深度学习模型中。

在相关技术的文本到图像生成方法中，文本到图像模型通常使用从文本提示词中提取的词嵌入作为条件。然而，文本具有较高的抽象水平和有限的信息密度，使得准确描述物体变得困难，因此生成的图像的精确度较低，并且该方法通常需要微调或仅支持使用单个物体作为约束条件，因此图像生成范围较为局限。

目前，提出了 BLIP-Diffusion 模型，能够支持图文混合生成对应图像，但是仅支持单个物体的生成，且生成的图像的精确度较低。此外，还提出了 KOSMOS-G 模型，能够支持图文混合生成对应图像，但生成的图像的精确度仍然较低。

相关技术基于文本提示生成对应图像的方法存在如下缺陷。

缺陷 1：仅通过文本提示描述物体从而生成对应的图像，由于文本提示描述物理较为困难，因此生成的图像的精确度较低。

缺陷 2：通常需要微调或仅支持使用单个物体作为约束条件，因此模型训练成本

较高且图像生成范围较为局限;

缺陷 3: BLIP-Diffusion 模型和 KOSMOS-G 模型支持图文混合生成，但是生成图像的精确度较低，且 BLIP-Diffusion 模型仅支持单物体生成。

针对上述缺陷，在本公开之前尚未提出有效的解决方案。

实施例 1

根据本公开实施例，提供了一种图像生成方法，需要说明的是，在附图的流程图示出的步骤可以在诸如一组计算机可执行指令的计算机系统中执行，并且，虽然在流程图中示出了逻辑顺序，但是在某些情况下，可以以不同于此处的顺序执行所示出或描述的步骤。

考虑到大模型的模型参数量庞大，且移动终端的运算资源有限，本公开实施例提供的上述图像生成方法可以应用于如图 1 所示的应用场景，但不仅限于此。在如图 1 所示的应用场景中，大模型部署在服务器 10 中，服务器 10 可以通过局域网连接、广域网连接、因特网连接，或者其他类型的数据网络，连接一个或多个客户端设备 20，此处的客户端设备 20 可以包括但不限于: 智能手机、平板电脑、笔记本电脑、掌上电脑、个人计算机、智能家居设备、车载设备等。客户端设备 20 可以通过图形用户界面与用户进行交互，实现对大模型的调用，进而实现本公开实施例所提供的方法。

在本公开实施例中，客户端设备和服务器构成的系统可以执行如下步骤: 客户端设备执行获取用户在图形用户界面内输入的多模态提示信息，并将多模态提示信息发送给服务器等步骤，服务器执行采用图像生成模型对获取到的多模态提示信息进行多模态图像生成，从而得到目标图像，并将目标图像返回至客户端设备等步骤，需要说明的是，在客户端设备的运行资源能够满足大模型的部署和运行条件的情况下，本公开实施例可以在客户端设备中进行。

在上述运行环境下，本公开提供了如图 2 所示的图像生成方法。图 2 是根据本公开实施例 1 的一种图像生成方法的流程图。如图 2 所示，该方法可以包括如下步骤:

步骤 S21，获取多模态提示信息，其中，多模态提示信息包括: 文本信息与增强标记信息，文本信息用于描述待生成的图像内容，图像内容包括: 至少一个目标对象，增强标记信息用于确定至少一个目标对象的位置特征与图像特征;

步骤 S22，采用图像生成模型对多模态提示信息进行多模态图像生成，得到目标图像，其中，图像生成模型用于采用多模态图像生成方式生成目标图像。

文本信息可以理解为文本提示，即使用自然语言文字描述的图像内容，用于描述待生成的图像内容。示例性地，自然语言文字可以为中文、英文、日文等，此处不予限制。本公开实施例中，文本信息所描述的图像内容中可以包括至少一个目标对象，目标对象可以理解为待生成的图像内容中的人物、动物或物体等，即本公开能够支持多对象的图像生成。示例性地，目标对象可以包括男孩、女孩、医生、老师等真实人

物，可以包括游戏玩家、非玩家角色（Non-Player Character，NPC）等虚拟人物，可以包括猫、狗、孔雀、大象等动物，还可以包括桌子、汽车、草地、树木、火车站等物体，此处不予限制。

示例性地，文本信息可以为用户的图像生成需求对应的文本信息，若用户希望得到一只猫和一只狗在草地上的图像，则对应的文本信息可以为“一只猫和一只狗在草地上（即 a cat and a dog on the grass）”，该文本信息中即包括三个目标对象，分别为猫、狗和草地。可以理解的是，该文本信息还可以替换为“一只狗和一只猫在草地上”，或者“草地上有一只猫和一只狗”，本公开对文本信息的描述方式不予限制。

考虑到仅通过文本提示描述物体会导致图像生成不准确，因此本公开在提供文本信息的基础上，将增强标记信息作为附加信息，通过文本信息和增强标记信息共同生成目标图像，从而提高图像生成的准确率。

增强标记信息可以包括坐标信息和图像信息，其中，坐标信息可以为文本信息中包括的至少一个目标对象的坐标，用于确定至少一个目标对象的位置特征。示例性地，若用户希望得到一只猫和一只狗在草地上的图像，则坐标信息即包括猫的坐标（例如[坐标：12，15，100，200]）、狗的坐标（例如[坐标：100，150，220，240]）和草地的坐标（例如[坐标：500，500，500，500]）。

示例性地，至少一个目标对象的坐标可以是二维直角坐标系下的坐标，能够表示目标对象的坐标范围，也可以是二维坐标系或三维坐标系下的坐标，本公开对所采用的坐标系不予限制。

图像信息可以为文本信息中包括的至少一个目标对象的图像，用于确定文本信息中至少一个目标对象的图像特征。示例性地，若用户希望得到一只猫和一只狗在草地上的图像，则图像信息即包括用户所需要生成的一只猫的图像（例如[图像：一张用户上传的自己的猫的图像]）、一只狗的图像（例如[图像：一张用户上传的自己的狗的图像]）以及草地的图像（例如[图像：一张用户上传的草地的图像]）。

多模态提示信息可以理解为多种模态的提示信息，包括上述文本信息和增强标记信息。本公开实施例中，多模态提示信息包括文本模态、坐标模态以及图像模态的提示信息。

示例性地，可以将同一目标对象的多模态提示信息绑定在一起，以避免影响其他物体或全局条件，也即本公开中引入增强标记（augmented token），增强标记同时包含物体级的文本、坐标和图像信息，用于描述生成图像中的一个物体。

本公开实施例中，图像生成模型为适用于生成图像的模型，示例性地，图像生成模型可以为预训练的基于扩散过程的图像生成模型，即为扩散模型，图像生成模型还可以为自回归模型等，此处不予限制。图像生成模型能够基于多目标对象的多模态提示信息准确生成高质量目标图像，即能够基于多物体级别的多模态提示生成高质量图

像，从而实现对生成图像的精准控制。

示例性地，若将上述示例的多模态提示信息输入至图像生成模型，即输入文本信息：一只猫和一只狗在草地上、坐标信息：猫的坐标、狗的坐标和草地的坐标、图像信息：一只猫的图像、一只狗的图像以及草地的图像，则本公开的图像生成模型能够根据该多模态提示信息准确输出对应的目标图像，即一只猫和一只狗在草地上的图像。

可以理解的是，本公开实施例中的图像生成模型能够尽可能保持预训练的文本到图像模型的原始结构，以便集成对现有模型的扩展，并且本公开实施例中的图像生成模型仅改变了模型的输入，不需要改变模型的架构，因此能够保持在模型的基础上构建技术的可用性。

本公开实施例中，通过获取包括文本信息与增强标记信息的多模态提示信息，能够根据文本信息确定待生成的图像内容、根据增强标记信息确定图像内容中至少一个目标对象的位置特征和图像特征，然后采用图像生成模型对多模态提示信息进行多模态图像生成，从而得到目标图像，达到了通过多模态提示信息准确生成包括多目标对象的高质量目标图像的目的，且能够对生成的图像实现更精准的控制，生成准确度更高的图像，且支持生成多目标对象的图像，使得图像的生成范围更广泛。

本公开实施例提供的上述图像生成方法可以但不限于应用于剧本服务、设计服务、游戏服务、电商服务、教育服务、法律服务、医疗服务、会议服务、社交网络服务、金融产品服务、物流服务和导航服务等领域中涉及图像生成的应用场景中，例如：剧本服务中根据剧本或故事情节生成所需的图像素材、设计服务中根据用户的文字、图像、坐标描述生成设计图、游戏中根据角色形象的文字描述生成对应角色图、电商服务中根据商品的描述生成对应商品图等，此处不予限制。

采用本公开实施例，通过获取包括文本信息与增强标记信息的多模态提示信息，能够根据文本信息确定待生成的图像内容、根据增强标记信息确定图像内容中至少一个目标对象的位置特征和图像特征，然后采用图像生成模型对多模态提示信息进行多模态图像生成，从而得到目标图像，由此达到了通过多模态提示信息准确生成包括多目标对象的高质量目标图像的目的。此外，能够对生成的图像实现更精准的控制，生成准确度更高的图像，且支持生成多目标对象的图像，使得图像的生成范围更广泛，从而实现了更精确、更多样化、支持多目标对象的高质量图像生成，以满足不同领域的图像生成需求的技术效果，进而解决了相关技术中仅通过文本提示生成对应图像，导致生成的图像精确度较低，图像生成范围较局限的技术问题。

在一种可选的实施例中，在步骤 S21 中，获取多模态提示信息，包括如下方法步骤：

步骤 S211，获取文本信息；

步骤 S212，基于文本信息分别生成至少一个目标对象的位置特征与至少一个目标

对象的图像特征，得到增强标记信息；

步骤 S213，对文本信息与增强标记信息进行组合，得到多模态提示信息。

考虑到在推理过程中，获取所有目标对象的多模态提示信息可能较为困难，例如可能仅获取到了目标对象的文本信息，或者仅能够获取到部分目标对象的多模态提示信息（例如当用户希望在图像中合并真实物体和生成的物体的情况）。因此本公开实施例中，可以基于文本信息获取增强标记信息，即能够生成目标对象缺失的增强标记信息，从而实现多个模态的灵活组合。

本公开实施例中，在获取多模态提示信息时，可以先获取文本信息，然后基于文本信息分别生成至少一个目标对象的位置特征与至少一个目标对象的图像特征，即根据文本提示生成物体级别的坐标和图像，从而得到对应的增强标记信息，再将文本信息与增强标记信息进行组合，即可得到多模态提示信息。

在一种可选的实施例中，在步骤 S212 中，基于文本信息分别生成至少一个目标对象的位置特征与图像特征，得到增强标记信息，包括如下方法步骤：

步骤 S2121，对文本信息进行词性分析，选取至少一个目标分词，其中，至少一个目标分词满足预设词性要求，至少一个目标分词用于确定至少一个目标对象；

步骤 S2122，基于文本信息和至少一个目标分词生成至少一个目标对象的位置特征，以及基于文本信息和至少一个目标对象的位置特征生成至少一个目标对象的图像特征；

步骤 S2123，利用至少一个目标分词、至少一个目标对象的位置特征以及至少一个目标对象的图像特征确定增强标记信息。

可以理解的是，语言中的词性包括名词、动词、形容词、副词、代词、数词、量词、连词、介词、助词、叹词等，而通常名词用来表示人、事物、地点等，在生成图像时需要被生成出来，因此名词可以用来表示目标对象。

本公开实施例中，预设词性可以为名词，至少一个目标分词即为文本信息中的至少一个名词。通过对文本信息进行词性分析，选取文本信息中的为名词的至少一个目标分词，从而能够确定文本信息中所包含的至少一个目标对象，进而便于获取每个目标对象对应的多模态提示信息。

示例性地，可以使用组成树（constituency tree）来识别文本信息中的物体，即识别文本信息中的目标对象。例如文本信息为“一只猫和一只狗在草地上（即 a cat and a dog on the grass）”，通过组成树对文本信息进行词性分析，分析得到“一只（限定词）猫（名词）和（连词）一只（限定词）狗（名词）在（介词）草地（名词）上（即 a (determiner) cat (noun) and (conjunction) a (determiner) dog (noun) on (preposition) the (determiner) grass (noun)）”，识别出文本信息中的名词，即猫、狗和草地，从而将识别得到的名词猫、狗和草地作为选取出的多个目标分词。

此外，还可以使用条件随机场（Conditional Random Fields，CRF）、最大熵模型（Maximum Entropy Model）等，通过训练模型来对文本进行词性标注，或者使用循环神经网络（Recurrent Neural Networks，RNN）、长短时记忆网络（Long Short-Term Memory，LSTM）、注意力机制（Attention Mechanism）等进行词性标注和词性识别，此处不予限制。

本公开实施例中，在基于文本信息分别生成至少一个目标对象的位置特征与图像特征，得到增强标记信息时，可以对文本信息进行词性分析，从文本信息中选取出词性为名词的至少一个目标分词。然后基于文本信息和至少一个目标分词生成至少一个目标对象的位置特征以及图像特征。再根据至少一个目标分词、至少一个目标对象的位置特征以及至少一个目标对象的图像特征确定至少一个目标对象分别对应的增强标记信息。

在一种可选的实施例中，在步骤 S2122 中，基于文本信息和至少一个目标分词生成至少一个目标对象的位置特征，包括如下方法步骤：

步骤 S21221，采用位置特征生成模型对文本信息和至少一个目标分词进行位置特征生成，得到至少一个目标对象的位置特征。

本公开实施例中，位置特征生成模型为适用于生成位置特征的模型，例如：扩散模型、自回归模型等。位置特征生成模型可以为坐标模型，能够基于文本信息的文字内容和至少一个目标分词对应的物体，生成每个物体对应的位置坐标。

本公开实施例中，在基于文本信息和至少一个目标分词生成至少一个目标对象的位置特征时，可以将文本信息和至少一个目标分词输入至位置特征生成模型，采用位置特征生成模型对文本信息和至少一个目标分词进行位置特征生成，从而得到至少一个目标对象对应的位置特征，即得到至少一个目标对象对应的坐标信息。

示例性地，若文本信息为“一只猫和一只狗”，则多个目标分词为“猫”和“狗”，本公开通过将“一只猫和一只狗”、“猫”和“狗”这三个文本特征拼到一起，共同作为位置特征生成模型的输入，从而根据位置特征生成模型的输出得到猫的坐标和狗的坐标。

可以看出，本公开中当坐标模态信息缺失时，可以采用位置特征生成模型基于文本提示和多个物体自动生成多个物体对应的位置坐标，从而补全缺失的坐标模态信息。

在一种可选的实施例中，在步骤 S2122 中，基于文本信息和至少一个目标对象的位置特征生成至少一个目标对象的图像特征，包括如下方法步骤：

步骤 S21222，采用图像特征生成模型对文本信息和至少一个目标对象的位置特征进行图像特征生成，得到至少一个目标对象的图像特征。

本公开实施例中，图像特征生成模型为适用于生成图像特征的模型，例如：扩散模型、自回归模型等。图像特征生成模型可以为图像特征模型，能够基于文本信息的文字内容和至少一个目标对象的位置特征，生成至少一个目标对象对应的图像特征。

本公开实施例中，在基于文本信息和至少一个目标对象的位置特征生成至少一个目标对象的图像特征时，可以将文本信息和至少一个目标对象的位置特征输入至图像特征生成模型，采用图像特征生成模型对文本信息和至少一个目标对象的坐标进行图像特征生成，从而得到至少一个目标对象对应的图像特征，即得到至少一个目标对象对应的图像信息。

示例性地，若文本信息为“一只猫和一只狗”，且已获取到了猫的坐标和狗的坐标，本公开通过将“一只猫和一只狗（文本信息）”、“猫（文本）+猫坐标（位置特征）”、“狗（文本）+狗坐标（位置特征）”输入至图像特征生成模型，从而根据图像特征生成模型的输出得到猫的图像特征和狗的图像特征。

可以看出，本公开中当图像模态信息缺失时，可以采用图像特征生成模型基于文本提示和多个物体的位置坐标自动生成多个物体对应的图像信息，从而补全缺失的图像模态信息。

也即可以看出，本公开实施例中的图像生成模型不仅能够支持根据文字提示、坐标信息和图像信息生成目标图像，同时也支持仅根据文字提示生成目标图像。也即本公开可以基于多物体多模态提示生成图片，同时，当出现模态缺失问题时，也可以利用生成模型补全缺失的模态，因此应用范围更为广泛。

在一种可选的实施例中，该图像生成方法还包括如下方法步骤：

步骤 S2124，采用文本编码器对文本信息进行文本编码，得到全局文本特征，以及采用文本编码器对至少一个目标分词进行文本编码，得到对象文本特征。

考虑到文本编码器（Text Encoder）能够将文本数据转化为向量或矩阵形式，以便进行机器学习和深度学习任务，如词性标注等。因此本公开实施例中，在对文本信息进行处理时，可以采用文本编码器对文本信息进行文本编码处理，从而得到全局文本编码结果，即全局文本特征，同时也可以采用文本编码器对至少一个目标分词进行文本编码处理，从而得到名词部分编码结果，即对象文本特征。

可以理解的是，在对增强标记信息中的图像信息进行处理时，可以采用图像编码器（Image Encoder）对图像信息进行图像特征编码处理，从而得到图像特征，此处不过多赘述。

在一种可选的实施例中，在步骤 S2123 中，利用至少一个目标分词、至少一个目标对象的位置特征以及至少一个目标对象的图像特征确定增强标记信息，包括如下方法步骤：

步骤 S21231，对对象文本特征、至少一个目标对象的位置特征以及至少一个目标对象的图像特征进行组合，得到增强标记信息。

本公开实施例中，在利用至少一个目标分词、至少一个目标对象的位置特征以及至少一个目标对象的图像特征确定增强标记信息时，可以将对至少一个目标分词进行

文本编码处理得到的对象文本特征、至少一个目标对象的位置特征以及至少一个目标对象的图像特征进行组合，从而得到增强标记信息。示例性地，可以将对象文本特征、至少一个目标对象的位置特征以及至少一个目标对象的图像特征进行水平拼接，从而得到增强标记信息，此处不予限制。

示例性地，若文本信息为“一只猫和一只狗在草地上”，则可以对对象文本特征“猫”、“狗”和“草地”、至少一个目标对象的位置特征“猫的坐标”、“狗的坐标”和“草地的坐标”以及至少一个目标对象的图像特征“一只猫的图像”、“一只狗的图像”和“草地的图像”进行特征组合，从而得到增强标记信息。

在一种可选的实施例中，在步骤 S213 中，对文本信息与增强标记信息进行组合，得到多模态提示信息，包括如下方法步骤：

步骤 S2131，对全局文本特征、对象文本特征、至少一个目标对象的位置特征以及至少一个目标对象的图像特征进行组合，得到多模态提示信息。

本公开实施例中，在对文本信息与增强标记信息进行组合，得到多模态提示信息时，可以将全局文本特征、对象文本特征、至少一个目标对象的位置特征以及至少一个目标对象的图像特征进行组合，从而得到多模态提示信息。示例性地，可以将对象文本特征、至少一个目标对象的位置特征以及至少一个目标对象的图像特征进行水平拼接，再与全局文本特征进行垂直拼接，从而得到多模态提示信息，此处不予限制。

示例性地，若文本信息为“一只猫和一只狗在草地上”，则可以对全局文本特征“一只猫和一只狗在草地上”、对象文本特征“猫”、“狗”和“草地”、至少一个目标对象的位置特征“猫的坐标”、“狗的坐标”和“草地的坐标”以及至少一个目标对象的图像特征“一只猫的图像”、“一只狗的图像”和“草地的图像”进行特征组合，从而得到多模态提示信息。

在一种可选的实施例中，在步骤 S21 中，获取多模态提示信息，包括如下方法步骤：

步骤 S211，获取文本信息与附加信息，其中，附加信息包括以下至少之一：至少一个目标对象的位置信息、至少一个目标对象的图像信息；

步骤 S212，基于文本信息与附加信息确定增强标记信息；

步骤 S213，对文本信息与增强标记信息进行组合，得到多模态提示信息。

文本信息可以理解为用户输入的用于描述待生成的图像内容的文本提示。

附加信息可以理解为用户输入的坐标和/或图像，可以理解的是，附加信息包括用户输入的至少一个目标对象的坐标、用户输入的至少一个目标对象的图像中的至少之一。

本公开实施例中，通过获取用户输入的文本信息与附加信息，从而能够根据获取到的文本信息与附加信息能够确定出用于确定至少一个目标对象的位置特征与图像特

征的增强标记信息，进而能够通过对文本信息与增强标记信息进行组合，得到多模态提示信息。

示例性地，可以将文本信息与增强标记信息进行垂直拼接，从而得到多模态提示信息，此处不予限制。

图 3 是根据本公开实施例 1 的另一种图像生成方法的流程图，如图 3 所示，本公开的图像生成模型支持将文本信息、增强标记信息共同作为输入，从而生成目标图像，即支持将文本模态结合坐标模态和图像模态，基于多模态的提示信息生成目标图像。此外，本公开的图像生成模型还支持在模态缺失时，仅将文本信息作为输入，从而根据文本信息生成对应的增强标记信息，即根据文本模态生成对应的坐标模态和图像模态，进而根据文本模态和生成的坐标模态和图像模态生成目标图像。

进一步地，在基于多模态提示的图像生成过程中，可以根据输入的文本信息确定文本信息中的至少一个目标分词（即名词）、至少一个目标对象的坐标以及至少一个目标对象的图像，然后通过文本编码器确定的至少一个目标分词对应的对象文本特征，根据至少一个目标对象的坐标确定至少一个目标对象的位置特征，通过图像编码器确定至少一个目标对象的图像特征。再对对象文本特征、至少一个目标对象的位置特征以及至少一个目标对象的图像特征进行组合，得到增强标记信息，进而将文本信息嵌入和增强标记信息组合后输入至图像生成模型，最终得到目标图像。

示例性地，文本信息可以为“一只猫和一只狗在草地上”，因此能够确定多个目标分词包括“狗”、“猫”和“草地”，至少一个目标对象的坐标包括[狗的坐标]、[猫的坐标]以及[草地的坐标]，至少一个目标对象的图像包括狗的图像、猫的图像和草地的图像。然后通过文本编辑器确定多个目标分词对应的对象文本特征，根据至少一个目标对象的坐标确定至少一个目标对象的位置特征，通过图像编码器确定至少一个目标对象的图像特征。再对对象文本特征、至少一个目标对象的位置特征以及至少一个目标对象的图像特征进行组合，得到增强标记信息，进而将文本信息“一只猫和一只狗在草地上”和增强标记信息组合后输入至图像生成模型，最终得到目标图像。

在基于纯文本提示的图像生成过程中，为了应对模态缺失，可以仅输入文本信息，基于分词模型确定出文本信息中的至少一个目标分词，并根据文字编码器确定至少一个目标分词对应的对象文本特征。然后采用坐标模型对文本信息和至少一个目标分词进行位置特征生成，得到至少一个目标对象的位置特征，采用图像特征模型对文本信息和至少一个目标对象的位置特征进行图像特征生成，得到至少一个目标对象的图像特征。再对对象文本特征、生成的至少一个目标对象的位置特征以及至少一个目标对象的图像特征进行组合，得到增强标记信息，进而将文本信息嵌入和增强标记信息组合后输入至图像生成模型，最终得到目标图像。

示例性地，文本信息可以为“一只猫和一只狗在草地上”，因此能够根据分词模型

确定至少一个目标分词包括“狗”、“猫”和“草地”，并根据文字编码器确定至少一个目标分词对应的对象文本特征。然后采用坐标模型对文本信息和至少一个目标分词进行位置特征生成，得到至少一个目标对象的坐标（[狗的坐标]、[猫的坐标]以及[草地的坐标]）对应的位置特征，采用图像特征模型对文本信息和至少一个目标对象的位置特征进行图像特征生成，得到至少一个目标对象的图像（狗的图像、猫的图像和草地的图像）对应的图像特征。再对对象文本特征、生成的至少一个目标对象的位置特征以及至少一个目标对象的图像特征进行组合，得到增强标记信息，进而将文本信息嵌入和增强标记信息组合后输入至图像生成模型，最终得到目标图像。

可以看出，本公开给定了一个图像-文本对，通过获取物体级别（object-level）的文本、坐标和图像，并将这些信息整合到每个物体（object）的“增强标记（augmented token）”中。增强标记作为附加条件与文本提示一起在扩散模型中进行训练，使得本公开的图像生成模型能够处理多物体多模态的文本提示。

此外，为了在推理过程中处理模态缺失问题，即为了解决零样本图像生成的问题，本公开提出利用一个坐标模型和一个图像特征模型，根据文本提示生成物体级别的坐标和图像特征。因此，本公开可以仅通过文本提示或通过各种多模态提示的组合来生成目标图像，能够以灵活的方式结合各种模态生成目标图像。且通过大量的定性和定量实验证明，本公开不仅优于相关技术中的图像生成方法，还能够完成更广泛的图像生成任务。

容易理解的是，本公开提供的图像生成方法的有益效果包括以下几点。

有益效果（1），支持基于多物体级别的多模态提示生成高质量图像，能够对生成图像实现更精准的控制，应用范围更广；

有益效果（2），设计了坐标模型和图像特征模型，支持基于文本模态生成坐标和图像模态，从而能够克服可能出现的模态缺失问题；

有益效果（3），相关技术的图像生成模型大部分都是基于文本模态生成的，而本公开的图像生成模型可以同时基于文本、图像、坐标模态生成目标图像，生成的图像的精确度更高；

有益效果（4），相关技术的图像生成模型若想要生成一个给定的物体，例如狗，由于狗有很多品种，因此若想生成某种特定的狗，一般需要给出3~5张图片，对模型进行微调，模型才能学会，而本公开的图像生成模型不需要微调的过程，只需要在推断的时候给出一张狗的图片，就可以生成对应品种的狗，即能够实现零样本生成，在训练的时候不需要目标物体的参与，因此模型训练成本较低；

有益效果（5），本公开将增强标记添加到文本嵌入后，作为扩散模型生成图像的采样条件，从而使本公开不需要改变扩散模型的架构，只需改变扩散模型的输入，进而提高了在扩散模型的基础上构建技术的可用性。

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、使用和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，还需要说明的是，对于前述的各方法实施例，为了简单描述，故将其都表述为一系列的动作组合，但是本领域技术人员应该知悉，本公开并不受所描述的动作顺序的限制，因为依据本公开，某些步骤可以采用其他顺序或者同时进行。其次，本领域技术人员也应该知悉，说明书中所描述的实施例均属于优选实施例，所涉及的动作和模块并不一定是本公开所必须的。

通过以上的实施方式的描述，本领域的技术人员可以清楚地了解到根据上述实施例的方法可借助软件加必需的通用硬件平台的方式来实现，当然也可以通过硬件。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软件产品的形式体现出来，该计算机软件产品存储在一个存储介质（如 ROM/RAM、磁碟、光盘）中，包括若干指令用以使得一台终端设备（可以是手机，计算机，服务器，或者网络设备等）执行本公开各个实施例所述的方法。

实施例 2

在如实施例 1 中的运行环境下，本公开提供了如图 4 所示的一种图像生成方法，通过终端设备提供一图形用户界面，图形用户界面所显示的内容至少部分地包含一图像生成场景，图 4 是根据本公开实施例 2 的一种图像生成方法的流程图，如图 4 所示，该方法包括:

步骤 S41，响应对图形用户界面执行的第一控制操作，输入文本信息，其中，文本信息用于描述待生成的图像内容，图像内容包括：至少一个目标对象;

步骤 S42，响应对图形用户界面执行的第二控制操作，基于文本信息分别生成至少一个目标对象的位置特征与至少一个目标对象的图像特征以得到增强标记信息，以及采用图像生成模型对多模态提示信息进行多模态图像生成以得到目标图像，其中，增强标记信息用于确定至少一个目标对象的位置特征与图像特征，图像生成模型用于采用多模态图像生成方式生成目标图像，多模态提示信息包括：文本信息与增强标记信息;

步骤 S43，在图形用户界面内展示目标图像。

本公开实施例中的图形用户界面中至少显示有图像生成场景，用户能够通过执行控制操作在该图像生成场景中输入用于描述待生成的图像内容的文本信息，控制基于文本信息生成至少一个目标对象的位置特征与至少一个目标对象的图像特征以得到增强标记信息，控制采用图像生成模型对多模态提示信息进行多模态图像生成以得到目

标图像等步骤。可以理解的是，上述图像生成场景可以但不限于剧本、设计、游戏、电商、教育、医疗、会议、社交网络、金融产品、物流和导航等领域中涉及图像生成的应用场景。

上述图形用户界面还包括第一控件（或第一触控区域），当检测到作用于第一控件（或第一触控区域）的第一触控操作时，可以获取用户输入的文本信息。上述文本信息可以是用户通过第一触控操作从图形用户界面中的文本框进行输入的。上述第一触控操作可以是点选、框选、勾选、条件筛选等操作，此处不予限制。

文本信息可以理解为文本提示，即使用自然语言文字描述的图像内容，用于描述待生成的图像内容。示例性地，自然语言文字可以为中文、英文、日文等，此处不予限制。本公开实施例中，文本信息所描述的图像内容中可以包括至少一个目标对象，目标对象可以理解为待生成的图像内容中的人物、动物或物体等，即本公开能够支持多对象的图像生成。示例性地，目标对象可以包括男孩、女孩、医生、老师等真实人物，可以包括游戏玩家、非玩家角色（Non-Player Character，NPC）等虚拟人物，可以包括猫、狗、孔雀、大象等动物，还可以包括桌子、汽车、草地、树木、火车站等物体，此处不予限制。

示例性地，文本信息可以为用户的图像生成需求对应的文本信息，若用户希望得到一只猫和一只狗在草地上的图像，则对应的文本信息可以为“一只猫和一只狗在草地上（即 a cat and a dog on the grass）”，该文本信息中即包括三个目标对象，分别为猫、狗和草地。可以理解的是，该文本信息还可以替换为“一只狗和一只猫在草地上”，或者“草地上有一只猫和一只狗”，本公开对文本信息的描述方式不予限制。

上述图形用户界面还包括第二控件（或第二触控区域），当检测到作用于第二控件（或第二触控区域）的第二触控操作时，能够基于文本信息分别生成至少一个目标对象的位置特征与至少一个目标对象的图像特征以得到增强标记信息，以及采用图像生成模型对多模态提示信息进行多模态图像生成以得到目标图像。上述第二触控操作可以是点选、框选、勾选、条件筛选等操作，此处不予限制。

考虑到仅通过文本提示描述物体会导致图像生成不准确，因此本公开在提供文本信息的基础上，将增强标记信息作为附加信息，通过文本信息和增强标记信息共同生成目标图像，从而提高图像生成的准确率。

增强标记信息可以包括坐标信息和图像信息，其中，坐标信息可以为文本信息中包括的至少一个目标对象的坐标，用于确定至少一个目标对象的位置特征。示例性地，若用户希望得到一只猫和一只狗在草地上的图像，则坐标信息即包括猫的坐标（例如[坐标：12，15，100，200]）、狗的坐标（例如[坐标：100，150，220，240]）和草地的坐标（例如[坐标：500，500，500，500]）。

示例性地，至少一个目标对象的坐标可以是二维直角坐标系下的坐标，能够表示

目标对象的坐标范围，也可以是二维坐标系或三维坐标系下的坐标，本公开对所采用的坐标系不予限制。

图像信息可以为文本信息中包括的至少一个目标对象的图像，用于确定文本信息中至少一个目标对象的图像特征。示例性地，若用户希望得到一只猫和一只狗在草地上的图像，则图像信息即包括用户所需要生成的一只猫的图像（例如[图像：一张用户上传的自己的猫的图像]）、一只狗的图像（例如[图像：一张用户上传的自己的狗的图像]）以及草地的图像（例如[图像：一张用户上传的草地的图像]）。

多模态提示信息可以理解为多种模态的提示信息，包括上述文本信息和增强标记信息。本公开实施例中，多模态提示信息包括文本模态、坐标模态以及图像模态的提示信息。

示例性地，可以将同一目标对象的多模态提示信息绑定在一起，以避免影响其他物体或全局条件，也即本公开中引入增强标记（augmented token），增强标记同时包含物体级的文本、坐标和图像信息，用于描述生成图像中的一个物体。

本公开实施例中，图像生成模型为适用于生成图像的模型，示例性地，图像生成模型可以为预训练的基于扩散过程的图像生成模型，即为扩散模型，图像生成模型还可以为自回归模型等，此处不予限制。图像生成模型能够基于多目标对象的多模态提示信息准确生成高质量目标图像，即能够基于多物体级别的多模态提示生成高质量图像，从而实现对生成图像的精准控制。

示例性地，若将上述示例的多模态提示信息输入至图像生成模型，即输入文本信息：一只猫和一只狗在草地上、坐标信息：猫的坐标、狗的坐标和草地的坐标、图像信息：一只猫的图像、一只狗的图像以及草地的图像，则本公开的图像生成模型能够根据该多模态提示信息准确输出对应的目标图像，即一只猫和一只狗在草地上的图像。

可以理解的是，本公开实施例中的图像生成模型能够尽可能保持预训练的文本到图像模型的原始结构，以便集成对现有模型的扩展，并且本公开实施例中的图像生成模型仅改变了模型的输入，不需要改变模型的架构，因此能够保持在模型的基础上构建技术的可用性。

本公开实施例中，通过终端设备提供一图形用户界面，图形用户界面所显示的内容至少部分地包含一图像生成场景，若用户对图形用户界面执行了第一控制操作，则用户输入了用于描述待生成的至少一个目标对象的图像内容的文本信息，若用户对图形用户界面执行了第二控制操作，例如提交操作，则能够基于文本信息分别生成至少一个目标对象的位置特征与至少一个目标对象的图像特征，从而得到增强标记信息，同时能够采用图像生成模型对多模态提示信息进行多模态图像生成从而得到目标图像，进而将生成得到的目标图像在图形用户界面内展示，以反馈至用户。达到了通过多模态提示信息准确生成包括多目标对象的高质量目标图像的目的，且能够对生成的图像

实现更精准的控制，生成准确度更高的图像，且支持生成多目标对象的图像，使得图像的生成范围更广泛。

需要说明的是，上述第一触控操作和第二触控操作均可以是用户用手指接触上述终端设备的显示屏并触控该终端设备的操作。该触控操作可以包括单点触控、多点触控，其中，每个触控点的触控操作可以包括点击、长按、重按、划动等。上述第一触控操作和第二触控操作还可以是通过鼠标、键盘等输入设备实现的触控操作，此处不予限制。

本公开实施例提供的上述图像生成方法可以但不限于应用于剧本服务、设计服务、游戏服务、电商服务、教育服务、法律服务、医疗服务、会议服务、社交网络服务、金融产品服务、物流服务和导航服务等领域中涉及图像生成的应用场景中，例如：剧本服务中根据剧本或故事情节生成所需的图像素材、设计服务中根据用户的文字、图像、坐标描述生成设计图、游戏中根据角色形象的文字描述生成对应角色图、电商服务中根据商品的描述生成对应商品图等，此处不予限制。

采用本公开实施例，通过终端设备提供一图形用户界面，图形用户界面所显示的内容至少部分地包含一图像生成场景，若用户对图形用户界面执行了第一控制操作，则用户输入了用于描述待生成的至少一个目标对象的图像内容的文本信息，若用户对图形用户界面执行了第二控制操作，例如提交操作，则能够基于文本信息分别生成至少一个目标对象的位置特征与至少一个目标对象的图像特征，从而得到增强标记信息，同时能够采用图像生成模型对多模态提示信息进行多模态图像生成从而得到目标图像，进而将生成得到的目标图像在图形用户界面内展示，以反馈至用户，达到了通过多模态提示信息准确生成包括多目标对象的高质量目标图像的目的，且能够对生成的图像实现更精准的控制，生成准确度更高的图像，且支持生成多目标对象的图像，使得图像的生成范围更广泛，进而解决了相关技术中仅通过文本提示生成对应图像，导致生成的图像精确度较低，图像生成范围较局限的技术问题。

需要说明的是，本实施例的优选实施方式可以参见实施例 1 中的相关描述，此处不再赘述。

实施例 3

在如实施例 1 中的运行环境下，本公开提供了如图 5 所示的一种图像生成方法。图 5 是根据本公开实施例 3 的一种图像生成方法的流程图，如图 5 所示，该方法包括：

步骤 S51，接收当前输入的多模态对话请求，其中，多模态对话请求中携带的信息包括：多模态对话文本信息与多模态对话增强标记信息，多模态对话文本信息用于描述待生成的多模态对话图像内容，多模态对话图像内容包括：至少一个对象，多模态对话增强标记信息用于确定至少一个对象的位置特征与图像特征；

步骤 S52，采用图像生成模型对多模态对话请求进行多模态图像生成，得到多模

态对话图像，其中，图像生成模型用于采用多模态图像生成方式生成多模态对话图像；

步骤 S53，反馈多模态对话回复，其中，多模态对话回复中携带的信息包括：多模态对话图像。

多模态对话请求可以理解为用户向计算机或机器人发起的对话请求（request），该多模态对话请求中携带有多模态对话文本信息与多模态对话增强标记信息。其中，多模态对话文本信息可以理解为文本提示，即使用自然语言文字描述的待生成的多模态对话图像内容。示例性地，自然语言文字可以为中文、英文、日文等，此处不予限制。

本公开实施例中，多模态对话文本信息所描述的多模态对话图像内容中可以包括至少一个对象，对象可以理解为待生成的多模态对话图像内容中的人物、动物或物体等，即本公开能够支持多对象的图像生成。示例性地，对象可以包括男孩、女孩、医生、老师等真实人物，可以包括游戏玩家、非玩家角色（Non-Player Character，NPC）等虚拟人物，可以包括猫、狗、孔雀、大象等动物，还可以包括桌子、汽车、草地、树木、火车站等物体，此处不予限制。

示例性地，多模态对话文本信息可以为用户输入的多模态对话请求对应的文本信息，若用户希望得到一只猫和一只狗在草地上的图像，则对应的多模态对话文本信息可以为“一只猫和一只狗在草地上（即 a cat and a dog on the grass）”，该多模态对话文本信息中即包括三个对象，分别为猫、狗和草地。可以理解的是，该多模态对话文本信息还可以替换为“一只狗和一只猫在草地上”，或者“草地上有一只猫和一只狗”，本公开对多模态对话文本信息的描述方式不予限制。

考虑到仅通过多模态对话文本提示描述物体会导致多模态对话图像生成不准确，因此本公开在提供多模态对话文本信息的基础上，将多模态对话增强标记信息作为附加信息，通过多模态对话文本信息和多模态对话增强标记信息共同生成多模态对话图像，从而提高多模态对话图像生成的准确率。

多模态对话增强标记信息可以包括坐标信息和图像信息，其中，坐标信息可以为多模态对话文本信息中包括的至少一个对象的坐标，用于确定至少一个对象的位置特征。示例性地，若用户希望得到一只猫和一只狗在草地上的图像，则坐标信息即包括猫的坐标（例如[坐标：12，15，100，200]）、狗的坐标（例如[坐标：100，150，220，240]）和草地的坐标（例如[坐标：500，500，500，500]）。

示例性地，至少一个对象的坐标可以是二维直角坐标系下的坐标，能够表示目标对象的坐标范围，也可以是二维坐标系或三维坐标系下的坐标，本公开对所采用的坐标系不予限制。

图像信息可以为多模态对话文本信息中包括的至少一个对象的图像，用于确定多模态对话文本信息中至少一个对象的图像特征。示例性地，若用户希望得到一只猫和一只狗在草地上的图像，则图像信息即包括用户所需要生成的一只猫的图像（例如[图

像：一张用户上传的自己的猫的图像]）、一只狗的图像（例如[图像：一张用户上传的自己的狗的图像]）以及草地的图像（例如[图像：一张用户上传的草地的图像]）。

可以看出，本公开实施例中的多模态对话请求携带了多种模态的提示信息，包括上述多模态对话文本信息和多模态对话增强标记信息，也即多模态提示信息包括文本模态、坐标模态以及图像模态的提示信息。

示例性地，可以将同一对象的多种模态的提示信息绑定在一起，以避免影响其他物体或全局条件，也即本公开中引入增强标记（augmented token），增强标记同时包含物体级的文本、坐标和图像信息，用于描述生成图像中的一个物体。

本公开实施例中，图像生成模型为适用于生成图像的模型，示例性地，图像生成模型可以为预训练的基于扩散过程的图像生成模型，即为扩散模型，图像生成模型还可以为自回归模型等，此处不予限制。图像生成模型能够基于多模态对话请求准确生成高质量多模态对话图像，即能够基于多物体级别的多模态提示生成高质量图像，从而实现对生成多模态对话图像的精准控制。

示例性地，若采用图像生成模型对上述示例的多模态对话请求进行多模态图像生成，即向图像生成模型输入多模态对话文本信息：一只猫和一只狗在草地上、坐标信息：猫的坐标、狗的坐标和草地的坐标、图像信息：一只猫的图像、一只狗的图像以及草地的图像，则本公开的图像生成模型能够根据该多模态对话请求准确输出对应的多模态对话图像，即一只猫和一只狗在草地上的图像。

可以理解的是，本公开实施例中的图像生成模型能够尽可能保持预训练的文本到图像模型的原始结构，以便集成对现有模型的扩展，并且本公开实施例中的图像生成模型仅改变了模型的输入，不需要改变模型的架构，因此能够保持在模型的基础上构建技术的可用性。

多模态对话回复可以理解为计算机或机器人基于用户输入的多模态对话请求向用户反馈的答复内容（response），与用户输入的多模态对话请求相对应。多模态对话回复中携带了多模态对话图像，该多模态对话图像即为包括多模态对话请求中描述的待生成的多模态对话图像内容的图像。

上述步骤 S51 至步骤 S53 可以应用于人机对话场景，即用户与计算机或机器人之间进行对话的场景。如图 6 所示，图 6 是根据本公开实施例 3 的一种人机对话场景示意图，用户向计算机或机器人输入多模态对话请求，计算机或机器人能够向用户反馈该多模态对话请求对应的多模态对话回复。

可以理解的是，用户与计算机或机器人之间的对话可以是通过语音识别和自然语言处理技术实现的，也可以是通过文本交流的方式进行的，此处不予限制。

本公开实施例中，通过接收用户输入的包括多模态对话文本信息与多模态对话增强标记信息的多模态对话请求，能够根据多模态对话文本信息确定待生成的多模态对

话图像内容、根据多模态对话增强标记信息确定待生成的多模态对话图像内容中至少一个对象的位置特征和图像特征，然后采用图像生成模型对多模态对话请求进行多模态图像生成，从而得到多模态对话图像，进而向用户反馈将携带多模态对话图像的多模态对话回复。由此，达到了通过多模态对话请求准确生成包括多对象的高质量多模态对话图像的目的，且能够对生成的多模态对话图像实现更精准的控制，生成准确度更高的多模态对话图像，且支持生成多对象的多模态对话图像，使得多模态对话图像的生成范围更广泛。

本公开实施例提供的上述图像生成方法还可以但不限于应用于剧本服务、设计服务、游戏服务、电商服务、教育服务、法律服务、医疗服务、会议服务、社交网络服务、金融产品服务、物流服务和导航服务等领域中涉及图像生成的应用场景中，例如：设计服务中根据用户的文字、图像、坐标描述生成设计图、游戏中根据角色形象的文字描述生成对应角色图、电商服务中根据商品的描述生成对应商品图等，此处不予限制。

采用本公开实施例，通过接收用户输入的包括多模态对话文本信息与多模态对话增强标记信息的多模态对话请求，能够根据多模态对话文本信息确定待生成的多模态对话图像内容、根据多模态对话增强标记信息确定待生成的多模态对话图像内容中至少一个对象的位置特征和图像特征，然后采用图像生成模型对多模态对话请求进行多模态图像生成，从而得到多模态对话图像，进而向用户反馈将携带多模态对话图像的多模态对话回复。由此，达到了通过多模态对话请求准确生成包括多对象的高质量多模态对话图像的目的，且能够对生成的多模态对话图像实现更精准的控制，生成准确度更高的多模态对话图像，且支持生成多对象的多模态对话图像，使得多模态对话图像的生成范围更广泛，进而解决了相关技术中仅通过文本提示生成对应图像，导致生成的图像精确度较低，图像生成范围较局限的技术问题。

在一种可选的实施例中，在步骤 S51 中，接收当前输入的多模态对话请求，包括如下方法步骤：

步骤 S511，获取多模态对话文本信息；

步骤 S512，基于多模态对话文本信息分别生成至少一个对象的位置特征与至少一个对象的图像特征，得到多模态对话增强标记信息；

步骤 S513，对多模态对话文本信息与多模态对话增强标记信息进行组合，得到多模态对话请求。

本公开实施例中，在接收当前输入的多模态对话请求时，可以先获取多模态对话文本信息，然后基于多模态对话文本信息分别生成至少一个对象的位置特征与至少一个对象的图像特征，即根据多模态对话文本提示生成物体级别的坐标和图像，从而得到对应的多模态对话增强标记信息，再将多模态对话文本信息与多模态对话增强标记

信息进行组合，即可得到多模态对话请求。

需要说明的是，本实施例的优选实施方式可以参见实施例 1 中的相关描述，此处不再赘述。

实施例 4

根据本公开实施例，还提供了一种用于实施上述图像生成的装置实施例。图 7 是根据本公开实施例 4 的一种图像生成装置的结构示意图，如图 7 所示，该装置包括：

获取模块 701，被设置为获取多模态提示信息，其中，多模态提示信息包括：文本信息与增强标记信息，文本信息用于描述待生成的图像内容，图像内容包括：至少一个目标对象，增强标记信息用于确定至少一个目标对象的位置特征与图像特征；

图像生成模块 702，被设置为采用图像生成模型对多模态提示信息进行多模态图像生成，得到目标图像，其中，图像生成模型用于采用多模态图像生成方式生成目标图像。

可选地，上述获取模块 701 还被设置为：获取文本信息；基于文本信息分别生成至少一个目标对象的位置特征与至少一个目标对象的图像特征，得到增强标记信息；对文本信息与增强标记信息进行组合，得到多模态提示信息。

可选地，上述获取模块 701 还被设置为：对文本信息进行词性分析，选取至少一个目标分词，其中，至少一个目标分词满足预设词性要求，至少一个目标分词用于确定至少一个目标对象；基于文本信息和至少一个目标分词生成至少一个目标对象的位置特征，以及基于文本信息和至少一个目标对象的位置特征生成至少一个目标对象的图像特征；利用至少一个目标分词、至少一个目标对象的位置特征以及至少一个目标对象的图像特征确定增强标记信息。

可选地，上述获取模块 701 还被设置为：采用位置特征生成模型对文本信息和至少一个目标分词进行位置特征生成，得到至少一个目标对象的位置特征。

可选地，上述获取模块 701 还被设置为：采用图像特征生成模型对文本信息和至少一个目标对象的位置特征进行图像特征生成，得到至少一个目标对象的图像特征。

可选地，还包括：编码模块，被设置为：采用文本编码器对文本信息进行文本编码，得到全局文本特征，以及采用文本编码器对至少一个目标分词进行文本编码，得到对象文本特征。

可选地，上述获取模块 701 还被设置为：对对象文本特征、至少一个目标对象的位置特征以及至少一个目标对象的图像特征进行组合，得到增强标记信息。

可选地，上述获取模块 701 还被设置为：对全局文本特征、对象文本特征、至少一个目标对象的位置特征以及至少一个目标对象的图像特征进行组合，得到多模态提示信息。

可选地，上述获取模块 701 还被设置为：获取文本信息与附加信息，其中，附加

信息包括以下至少之一：至少一个目标对象的位置信息、至少一个目标对象的图像信息；基于文本信息与附加信息确定增强标记信息；对文本信息与增强标记信息进行组合，得到多模态提示信息。

采用本公开实施例，通过获取包括文本信息与增强标记信息的多模态提示信息，能够根据文本信息确定待生成的图像内容、根据增强标记信息确定图像内容中至少一个目标对象的位置特征和图像特征，然后采用图像生成模型对多模态提示信息进行多模态图像生成，从而得到目标图像，由此达到了通过多模态提示信息准确生成包括多目标对象的高质量目标图像的目的。此外，能够对生成的图像实现更精准的控制，生成准确度更高的图像，且支持生成多目标对象的图像，使得图像的生成范围更广泛，从而实现了更精确、更多样化、支持多目标对象的高质量图像生成，以满足不同领域的图像生成需求的技术效果，进而解决了相关技术中仅通过文本提示生成对应图像，导致生成的图像精确度较低，图像生成范围较局限的技术问题。

此处需要说明的是，上述获取模块 701 和图像生成模块 702 对应于实施例 1 中的步骤 S21 和步骤 S22，两个模块与对应的步骤所实现的实例和应用场景相同，但不限于上述实施例 1 所公开的内容。需要说明的是，上述模块或单元可以是存储在存储器中并由一个或多个处理器处理的硬件组件或软件组件，上述模块也可以运行在实施例 1 提供的服务器 10 中。

根据本公开实施例，还提供了另一种用于实施上述图像生成的装置实施例。图 8 是根据本公开实施例 4 的另一种图像生成装置的结构示意图，通过终端设备提供一图形用户界面，图形用户界面所显示的内容至少部分地包含一图像生成场景，如图 8 所示，该装置包括：

第一响应模块 801，被设置为响应对图形用户界面执行的第一控制操作，输入文本信息，其中，文本信息用于描述待生成的图像内容，图像内容包括：至少一个目标对象；

第二响应模块 802，被设置为响应对图形用户界面执行的第二控制操作，基于文本信息分别生成至少一个目标对象的位置特征与至少一个目标对象的图像特征以得到增强标记信息，以及采用图像生成模型对多模态提示信息进行多模态图像生成以得到目标图像，其中，增强标记信息用于确定至少一个目标对象的位置特征与图像特征，图像生成模型用于采用多模态图像生成方式生成目标图像，多模态提示信息包括：文本信息与增强标记信息；

展示模块 803，被设置为在图形用户界面内展示目标图像。

采用本公开实施例，通过终端设备提供一图形用户界面，图形用户界面所显示的内容至少部分地包含一图像生成场景，若用户对图形用户界面执行了第一控制操作，则用户输入了用于描述待生成的至少一个目标对象的图像内容的文本信息，若用户对

图形用户界面执行了第二控制操作，例如提交操作，则能够基于文本信息分别生成至少一个目标对象的位置特征与至少一个目标对象的图像特征，从而得到增强标记信息，同时能够采用图像生成模型对多模态提示信息进行多模态图像生成从而得到目标图像，进而将生成得到的目标图像在图形用户界面内展示，以反馈至用户，达到了通过多模态提示信息准确生成包括多目标对象的高质量目标图像的目的，且能够对生成的图像实现更精准的控制，生成准确度更高的图像，且支持生成多目标对象的图像，使得图像的生成范围更广泛，进而解决了相关技术中仅通过文本提示生成对应图像，导致生成的图像精确度较低，图像生成范围较局限的技术问题。

此处需要说明的是，上述第一响应模块 801、第二响应模块 802 和展示模块 803 对应于实施例 2 中的步骤 S41 至步骤 S43，三个模块与对应的步骤所实现的实例和应用场景相同，但不限于上述实施例 2 所公开的内容。需要说明的是，上述模块或单元可以是存储在存储器中并由一个或多个处理器处理的硬件组件或软件组件，上述模块也可以运行在实施例 1 提供的服务器 10 中。

根据本公开实施例，还提供了再一种用于实施上述图像生成的装置实施例。图 9 是根据本公开实施例 4 的再一种图像生成装置的结构示意图，如图 9 所示，该装置包括:

接收模块 901，被设置为接收当前输入的多模态对话请求，其中，多模态对话请求中携带的信息包括：多模态对话文本信息与多模态对话增强标记信息，多模态对话文本信息用于描述待生成的多模态对话图像内容，多模态对话图像内容包括：至少一个对象，多模态对话增强标记信息用于确定至少一个对象的位置特征与图像特征;

生成模块 902，被设置为采用图像生成模型对多模态对话请求进行多模态图像生成，得到多模态对话图像，其中，图像生成模型用于采用多模态图像生成方式生成多模态对话图像;

反馈模块 903，被设置为反馈多模态对话回复，其中，多模态对话回复中携带的信息包括：多模态对话图像。

可选地，上述接收模块 901 还被设置为：获取多模态对话文本信息；基于多模态对话文本信息分别生成至少一个对象的位置特征与至少一个对象的图像特征，得到多模态对话增强标记信息;对多模态对话文本信息与多模态对话增强标记信息进行组合，得到多模态对话请求。

采用本公开实施例，通过接收用户输入的包括多模态对话文本信息与多模态对话增强标记信息的多模态对话请求，能够根据多模态对话文本信息确定待生成的多模态对话图像内容、根据多模态对话增强标记信息确定待生成的多模态对话图像内容中至少一个对象的位置特征和图像特征，然后采用图像生成模型对多模态对话请求进行多模态图像生成，从而得到多模态对话图像，进而向用户反馈将携带多模态对话图像的

多模态对话回复。由此，达到了通过多模态对话请求准确生成包括多对象的高质量多模态对话图像的目的，且能够对生成的多模态对话图像实现更精准的控制，生成准确度更高的多模态对话图像，且支持生成多对象的多模态对话图像，使得多模态对话图像的生成范围更广泛，进而解决了相关技术中仅通过文本提示生成对应图像，导致生成的图像精确度较低，图像生成范围较局限的技术问题。

此处需要说明的是，上述接收模块 901、生成模块 902 和反馈模块 903 对应于实施例 3 中的步骤 S51 至步骤 S53，三个模块与对应的步骤所实现的实例和应用场景相同，但不限于上述实施例 3 所公开的内容。需要说明的是，上述模块或单元可以是存储在存储器中并由一个或多个处理器处理的硬件组件或软件组件，上述模块也可以运行在实施例 1 提供的服务器 10 中。

需要说明的是，本公开上述实施例中涉及到的优选实施方案与实施例 1 提供的方案以及应用场景、实施过程相同，但不仅限于实施例 1 所提供的方案。

实施例 5

本公开的实施例可以提供一种计算机终端，该计算机终端可以是计算机终端群中的任意一个计算机终端设备。可选地，在本实施例中，上述计算机终端也可以替换为移动终端等终端设备。

可选地，在本实施例中，上述计算机终端可以位于计算机网络的多个网络设备中的至少一个网络设备。

在本实施例中，上述计算机终端可以执行图像生成方法中以下步骤的程序代码：获取多模态提示信息，其中，多模态提示信息包括：文本信息与增强标记信息，文本信息用于描述待生成的图像内容，图像内容包括：至少一个目标对象，增强标记信息用于确定至少一个目标对象的位置特征与图像特征；采用图像生成模型对多模态提示信息进行多模态图像生成，得到目标图像，其中，图像生成模型用于采用多模态图像生成方式生成目标图像。

可选地，图 10 是根据本公开实施例 5 的一种计算机终端的结构框图。如图 10 所示，该计算机终端 A 可以包括：一个或多个（图中仅示出一个）处理器 1002、存储器 1004、存储控制器、以及外设接口，其中，外设接口与射频模块、音频模块和显示器连接。

其中，存储器可被设置为存储软件程序以及模块，如本公开实施例中的图像生成方法和装置对应的程序指令/模块，处理器通过运行存储在内的软件程序以及模块，从而执行各种功能应用以及数据处理，即实现上述的图像生成方法。存储器可包括高速随机存储器，还可以包括非易失性存储器，如一个或者多个磁性存储装置、闪存、或者其他非易失性固态存储器。在一些实例中，存储器可进一步包括相对于处理器远程设置的存储器，这些远程存储器可以通过网络连接至计算机终端 A。上述网络的实例

包括但不限于互联网、企业内部网、局域网、移动通信网及其组合。

处理器可以通过传输装置调用存储器存储的信息及应用程序，以执行下述步骤：获取多模态提示信息，其中，多模态提示信息包括：文本信息与增强标记信息，文本信息用于描述待生成的图像内容，图像内容包括：至少一个目标对象，增强标记信息用于确定至少一个目标对象的位置特征与图像特征；采用图像生成模型对多模态提示信息进行多模态图像生成，得到目标图像，其中，图像生成模型用于采用多模态图像生成方式生成目标图像。

可选地，上述处理器还可以执行如下步骤的程序代码：获取文本信息；基于文本信息分别生成至少一个目标对象的位置特征与至少一个目标对象的图像特征，得到增强标记信息；对文本信息与增强标记信息进行组合，得到多模态提示信息。

可选地，上述处理器还可以执行如下步骤的程序代码：对文本信息进行词性分析，选取至少一个目标分词，其中，至少一个目标分词满足预设词性要求，至少一个目标分词用于确定至少一个目标对象；基于文本信息和至少一个目标分词生成至少一个目标对象的位置特征，以及基于文本信息和至少一个目标对象的位置特征生成至少一个目标对象的图像特征；利用至少一个目标分词、至少一个目标对象的位置特征以及至少一个目标对象的图像特征确定增强标记信息。

可选地，上述处理器还可以执行如下步骤的程序代码：采用位置特征生成模型对文本信息和至少一个目标分词进行位置特征生成，得到至少一个目标对象的位置特征。

可选地，上述处理器还可以执行如下步骤的程序代码：采用图像特征生成模型对文本信息和至少一个目标对象的位置特征进行图像特征生成，得到至少一个目标对象的图像特征。

可选地，上述处理器还可以执行如下步骤的程序代码：采用文本编码器对文本信息进行文本编码，得到全局文本特征，以及采用文本编码器对至少一个目标分词进行文本编码，得到对象文本特征。

可选地，上述处理器还可以执行如下步骤的程序代码：对对象文本特征、至少一个目标对象的位置特征以及至少一个目标对象的图像特征进行组合，得到增强标记信息。

可选地，上述处理器还可以执行如下步骤的程序代码：对全局文本特征、对象文本特征、至少一个目标对象的位置特征以及至少一个目标对象的图像特征进行组合，得到多模态提示信息。

可选地，上述处理器还可以执行如下步骤的程序代码：获取文本信息与附加信息，其中，附加信息包括以下至少之一：至少一个目标对象的位置信息、至少一个目标对象的图像信息；基于文本信息与附加信息确定增强标记信息；对文本信息与增强标记信息进行组合，得到多模态提示信息。

采用本公开实施例，通过获取包括文本信息与增强标记信息的多模态提示信息，能够根据文本信息确定待生成的图像内容、根据增强标记信息确定图像内容中至少一个目标对象的位置特征和图像特征，然后采用图像生成模型对多模态提示信息进行多模态图像生成，从而得到目标图像，由此达到了通过多模态提示信息准确生成包括多目标对象的高质量目标图像的目的。此外，能够对生成的图像实现更精准的控制，生成准确度更高的图像，且支持生成多目标对象的图像，使得图像的生成范围更广泛，从而实现了更精确、更多样化、支持多目标对象的高质量图像生成，以满足不同领域的图像生成需求的技术效果，进而解决了相关技术中仅通过文本提示生成对应图像，导致生成的图像精确度较低，图像生成范围较局限的技术问题。

本领域普通技术人员可以理解，图 10 所示的结构仅为示意，计算机终端 A 也可以是智能手机（如 Android 手机、iOS 手机等）、平板电脑、掌上电脑以及移动互联网设备（MobileInternetDevices，MID）、PAD 等终端设备。图 10 其并不对上述电子装置的结构造成限定。例如，计算机终端 A 还可包括比图 10 中所示更多或者更少的组件（如网络接口、显示装置等），或者具有与图 10 所示不同的配置。

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过程序来指令终端设备相关的硬件来完成，该程序可以存储于一计算机可读存储介质中，存储介质可以包括：闪存盘、只读存储器（Read-Only Memory，ROM）、随机存取器（Random Access Memory，RAM）、磁盘或光盘等。

实施例 6

本公开的实施例还提供了一种计算机可读存储介质。可选地，在本实施例中，上述计算机可读存储介质可以用于保存上述实施例一所提供的图像生成方法所执行的程序代码。

可选地，在本实施例中，上述计算机可读存储介质可以位于计算机网络中计算机终端群中的任意一个计算机终端中，或者位于移动终端群中的任意一个移动终端中。

可选地，在本实施例中，计算机可读存储介质被设置为存储用于执行以下步骤的程序代码：获取多模态提示信息，其中，多模态提示信息包括：文本信息与增强标记信息，文本信息用于描述待生成的图像内容，图像内容包括：至少一个目标对象，增强标记信息用于确定至少一个目标对象的位置特征与图像特征；采用图像生成模型对多模态提示信息进行多模态图像生成，得到目标图像，其中，图像生成模型用于采用多模态图像生成方式生成目标图像。

可选地，在本实施例中，计算机可读存储介质被设置为存储用于执行以下步骤的程序代码：获取文本信息；基于文本信息分别生成至少一个目标对象的位置特征与至少一个目标对象的图像特征，得到增强标记信息；对文本信息与增强标记信息进行组合，得到多模态提示信息。

可选地，在本实施例中，计算机可读存储介质被设置为存储用于执行以下步骤的程序代码：对文本信息进行词性分析，选取至少一个目标分词，其中，至少一个目标分词满足预设词性要求，至少一个目标分词用于确定至少一个目标对象；基于文本信息和至少一个目标分词生成至少一个目标对象的位置特征，以及基于文本信息和至少一个目标对象的位置特征生成至少一个目标对象的图像特征；利用至少一个目标分词、至少一个目标对象的位置特征以及至少一个目标对象的图像特征确定增强标记信息。

可选地，在本实施例中，计算机可读存储介质被设置为存储用于执行以下步骤的程序代码：采用位置特征生成模型对文本信息和至少一个目标分词进行位置特征生成，得到至少一个目标对象的位置特征。

可选地，在本实施例中，计算机可读存储介质被设置为存储用于执行以下步骤的程序代码：采用图像特征生成模型对文本信息和至少一个目标对象的位置特征进行图像特征生成，得到至少一个目标对象的图像特征。

可选地，在本实施例中，计算机可读存储介质被设置为存储用于执行以下步骤的程序代码：采用文本编码器对文本信息进行文本编码，得到全局文本特征，以及采用文本编码器对至少一个目标分词进行文本编码，得到对象文本特征。

可选地，在本实施例中，计算机可读存储介质被设置为存储用于执行以下步骤的程序代码：对对象文本特征、至少一个目标对象的位置特征以及至少一个目标对象的图像特征进行组合，得到增强标记信息。

可选地，在本实施例中，计算机可读存储介质被设置为存储用于执行以下步骤的程序代码：对全局文本特征、对象文本特征、至少一个目标对象的位置特征以及至少一个目标对象的图像特征进行组合，得到多模态提示信息。

可选地，在本实施例中，计算机可读存储介质被设置为存储用于执行以下步骤的程序代码：获取文本信息与附加信息，其中，附加信息包括以下至少之一：至少一个目标对象的位置信息、至少一个目标对象的图像信息；基于文本信息与附加信息确定增强标记信息；对文本信息与增强标记信息进行组合，得到多模态提示信息。

上述本公开实施例序号仅仅为了描述，不代表实施例的优劣。

在本公开的上述实施例中，对各个实施例的描述都各有侧重，某个实施例中没有详述的部分，可以参见其他实施例的相关描述。

在本公开所提供的几个实施例中，应该理解到，所揭露的技术内容，可通过其它的方式实现。其中，以上所描述的装置实施例仅仅是示意性的，例如所述单元的划分，仅仅为一种逻辑功能划分，实际实现时可以有另外的划分方式，例如多个单元或组件可以结合或者可以集成到另一个系统，或一些特征可以忽略，或不执行。另一点，所显示或讨论的相互之间的耦合或直接耦合或通信连接可以是通过一些接口，单元或模块的间接耦合或通信连接，可以是电性或其它的形式。

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单元来实现本实施例方案的目的。

另外，在本公开各个实施例中的各功能单元可以集成在一个处理单元中，也可以是各个单元单独物理存在，也可以两个或两个以上单元集成在一个单元中。上述集成的单元既可以采用硬件的形式实现，也可以采用软件功能单元的形式实现。

所述集成的单元如果以软件功能单元的形式实现并作为独立的产品销售或使用时，可以存储在一个计算机可读取存储介质中。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或者该技术方案的全部或部分可以以软件产品的形式体现出来，该计算机软件产品存储在一个存储介质中，包括若干指令用以使得一台计算机设备（可为个人计算机、服务器或者网络设备等）执行本公开各个实施例所述方法的全部或部分步骤。而前述的存储介质包括:U盘、只读存储器(ROM,Read-Only Memory)、随机存取存储器（RAM，Random Access Memory)、移动硬盘、磁碟或者光盘等各种可以存储程序代码的介质。

以上所述仅是本公开的优选实施方式，应当指出，对于本技术领域的普通技术人员来说，在不脱离本公开原理的前提下，还可以做出若干改进和润饰，这些改进和润饰也应视为本公开的保护范围。

权 利 要 求 书

1. 一种图像生成方法，包括：

获取多模态提示信息，其中，所述多模态提示信息包括：文本信息与增强标记信息，所述文本信息用于描述待生成的图像内容，所述图像内容包括：至少一个目标对象，所述增强标记信息用于确定所述至少一个目标对象的位置特征与图像特征；

采用图像生成模型对所述多模态提示信息进行多模态图像生成，得到目标图像，其中，所述图像生成模型用于采用多模态图像生成方式生成所述目标图像。

2. 根据权利要求 1 所述的图像生成方法，其中，获取所述多模态提示信息包括：

获取所述文本信息；

基于所述文本信息分别生成所述至少一个目标对象的位置特征与所述至少一个目标对象的图像特征，得到所述增强标记信息；

对所述文本信息与所述增强标记信息进行组合，得到所述多模态提示信息。

3. 根据权利要求 2 所述的图像生成方法，其中，基于所述文本信息分别生成所述至少一个目标对象的位置特征与图像特征，得到所述增强标记信息包括：

对所述文本信息进行词性分析，选取至少一个目标分词，其中，所述至少一个目标分词满足预设词性要求，所述至少一个目标分词用于确定所述至少一个目标对象；

基于所述文本信息和所述至少一个目标分词生成所述至少一个目标对象的位置特征，以及基于所述文本信息和所述至少一个目标对象的位置特征生成所述至少一个目标对象的图像特征；

利用所述至少一个目标分词、所述至少一个目标对象的位置特征以及所述至少一个目标对象的图像特征确定所述增强标记信息。

4. 根据权利要求 3 所述的图像生成方法，其中，基于所述文本信息和所述至少一个目标分词生成所述至少一个目标对象的位置特征包括：

采用位置特征生成模型对所述文本信息和所述至少一个目标分词进行位置特征生成，得到所述至少一个目标对象的位置特征。

5. 根据权利要求 3 所述的图像生成方法，其中，基于所述文本信息和所述至少一个目标对象的位置特征生成所述至少一个目标对象的图像特征包括：

采用图像特征生成模型对所述文本信息和所述至少一个目标对象的位置特征进行图像特征生成，得到所述至少一个目标对象的图像特征。

6. 根据权利要求 3 所述的图像生成方法，其中，所述图像生成方法还包括：

采用文本编码器对所述文本信息进行文本编码，得到全局文本特征，以及采

用所述文本编码器对所述至少一个目标分词进行文本编码，得到对象文本特征。

7. 根据权利要求 6 所述的图像生成方法，其中，利用所述至少一个目标分词、所述至少一个目标对象的位置特征以及所述至少一个目标对象的图像特征确定所述增强标记信息包括：

对所述对象文本特征、所述至少一个目标对象的位置特征以及所述至少一个目标对象的图像特征进行组合，得到所述增强标记信息。

8. 根据权利要求 6 所述的图像生成方法，其中，对所述文本信息与所述增强标记信息进行组合，得到所述多模态提示信息包括：

对所述全局文本特征、所述对象文本特征、所述至少一个目标对象的位置特征以及所述至少一个目标对象的图像特征进行组合，得到所述多模态提示信息。

9. 根据权利要求 1 所述的图像生成方法，其中，获取所述多模态提示信息包括：

获取所述文本信息与附加信息，其中，所述附加信息包括以下至少之一：所述至少一个目标对象的位置信息、所述至少一个目标对象的图像信息；

基于所述文本信息与所述附加信息确定所述增强标记信息；

对所述文本信息与所述增强标记信息进行组合，得到所述多模态提示信息。

10. 一种图像生成方法，通过终端设备提供一图形用户界面，所述图形用户界面所显示的内容至少部分地包含一图像生成场景，所述图像生成方法包括：

响应对所述图形用户界面执行的第一控制操作，输入文本信息，其中，所述文本信息用于描述待生成的图像内容，所述图像内容包括：至少一个目标对象；

响应对所述图形用户界面执行的第二控制操作，基于所述文本信息分别生成所述至少一个目标对象的位置特征与所述至少一个目标对象的图像特征以得到增强标记信息，以及采用图像生成模型对多模态提示信息进行多模态图像生成以得到目标图像，其中，所述增强标记信息用于确定所述至少一个目标对象的位置特征与图像特征，所述图像生成模型用于采用多模态图像生成方式生成所述目标图像，所述多模态提示信息包括：所述文本信息与所述增强标记信息；

在所述图形用户界面内展示所述目标图像。

11. 一种图像生成方法，包括：

接收当前输入的多模态对话请求，其中，所述多模态对话请求中携带的信息包括：多模态对话文本信息与多模态对话增强标记信息，所述多模态对话文本信息用于描述待生成的多模态对话图像内容，所述多模态对话图像内容包括：至少一个对象，所述多模态对话增强标记信息用于确定所述至少一个对象的位置特征与图像特征；

采用图像生成模型对所述多模态对话请求进行多模态图像生成，得到多模态对话图像，其中，所述图像生成模型用于采用多模态图像生成方式生成所述多模

态对话图像;

反馈多模态对话回复，其中，所述多模态对话回复中携带的信息包括：所述多模态对话图像。

12. 根据权利要求 11 所述的图像生成方法，其中，接收当前输入的所述多模态对话请求包括:

获取所述多模态对话文本信息;

基于所述多模态对话文本信息分别生成所述至少一个对象的位置特征与所述至少一个对象的图像特征，得到所述多模态对话增强标记信息;

对所述多模态对话文本信息与所述多模态对话增强标记信息进行组合，得到所述多模态对话请求。

13. 一种电子设备，包括:

存储器，存储有可执行程序;

处理器，用于运行所述程序，其中，所述程序运行时执行权利要求 1 至 12 中任意一项所述的图像生成方法。

14. 一种计算机可读存储介质，所述计算机可读存储介质包括存储的可执行程序，其中，在所述可执行程序运行时控制所述计算机可读存储介质所在设备执行权利要求 1 至 12 中任意一项所述的图像生成方法。

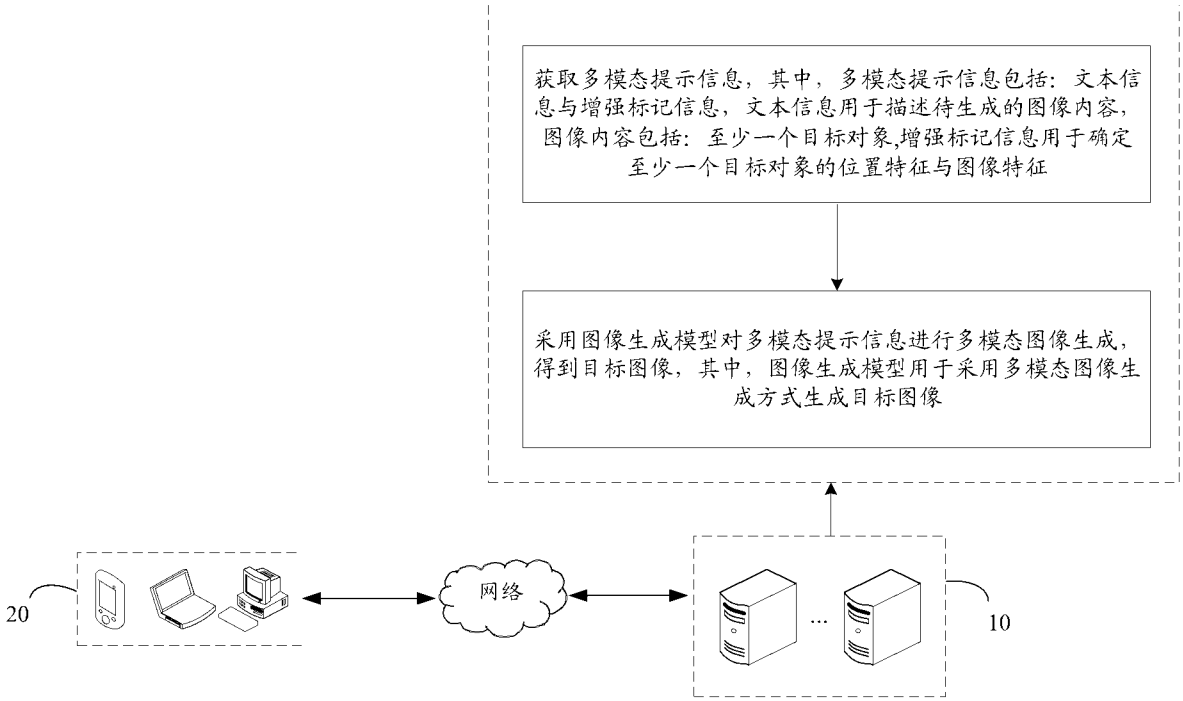
获取多模态提示信息，其中，多模态提示信息包括：文本信息与增强标记信息，文本信息用于描述待生成的图像内容，图像内容包括：至少一个目标对象,增强标记信息用于确定至少一个目标对象的位置特征与图像特征
采用图像生成模型对多模态提示信息进行多模态图像生成，得到目标图像，其中，图像生成模型用于采用多模态图像生成方式生成目标图像
20
网络
10

图 1

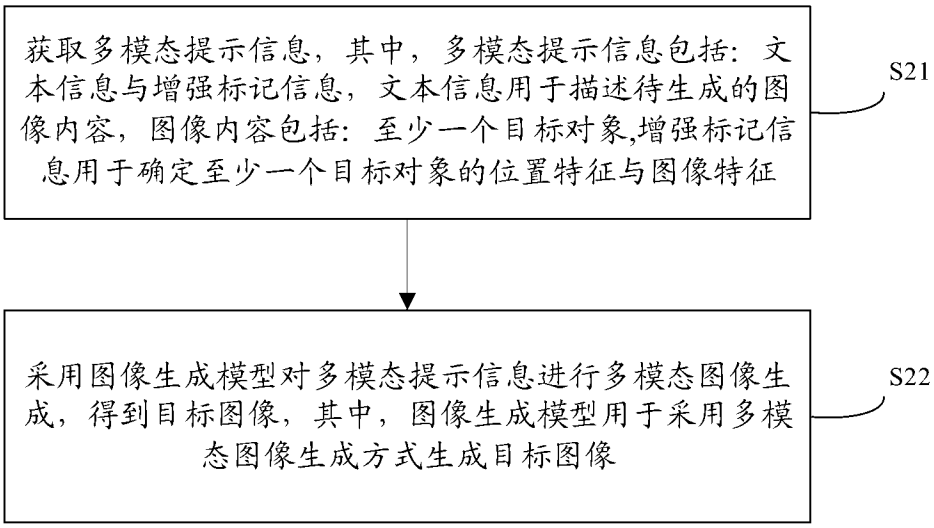
获取多模态提示信息，其中，多模态提示信息包括：文本信息与增强标记信息，文本信息用于描述待生成的图像内容，图像内容包括：至少一个目标对象,增强标记信息用于确定至少一个目标对象的位置特征与图像特征
S21
采用图像生成模型对多模态提示信息进行多模态图像生成，得到目标图像，其中，图像生成模型用于采用多模态图像生成方式生成目标图像
S22

图 2

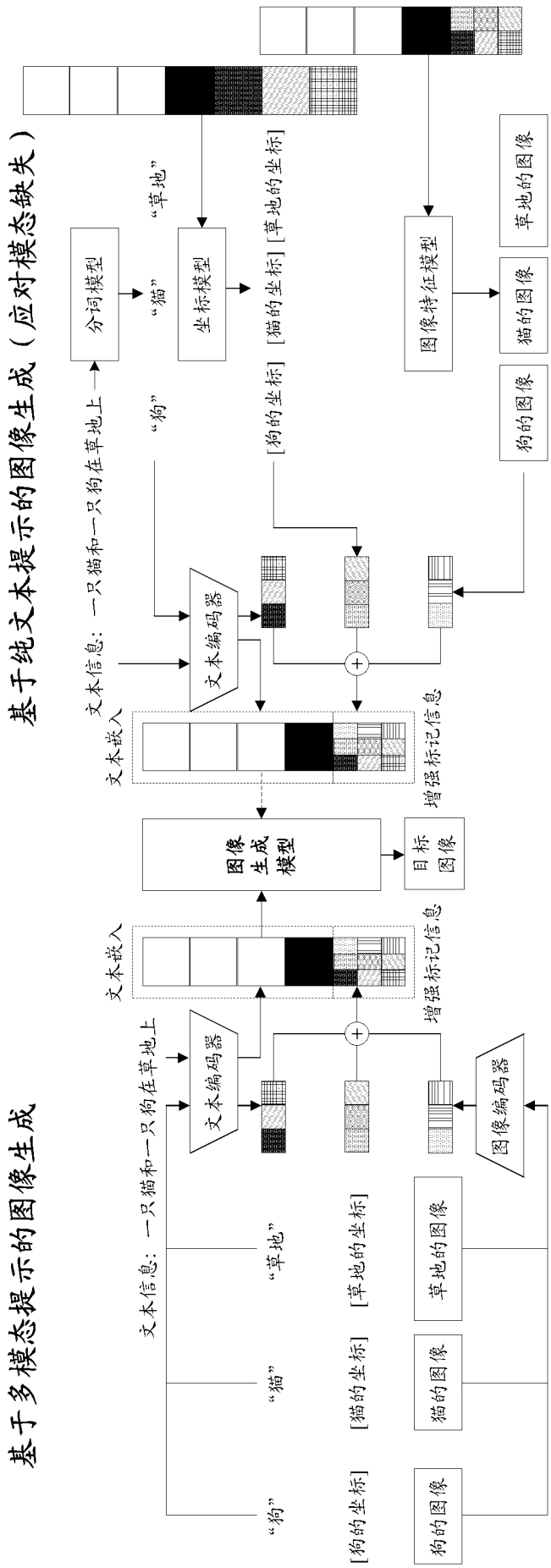
基于多模态提示的图像生成
文本信息：一只猫和一只狗在草地上
文本编码器
“狗”
“猫”
“草地”
[狗的坐标]
[猫的坐标]
[草地的坐标]
狗的图像
猫的图像
草地的图像
图像编码器
文本嵌入
增强标记信息
图像生成模型
目标图像
基于纯文本提示的图像生成（应对模态缺失）
文本信息：一只猫和一只狗在草地上
分词模型
坐标模型
图像特征模型
“狗”
“猫”
“草地”
[狗的坐标]
[猫的坐标]
[草地的坐标]
狗的图像
猫的图像
草地的图像
文本编码器
文本嵌入
增强标记信息

图3

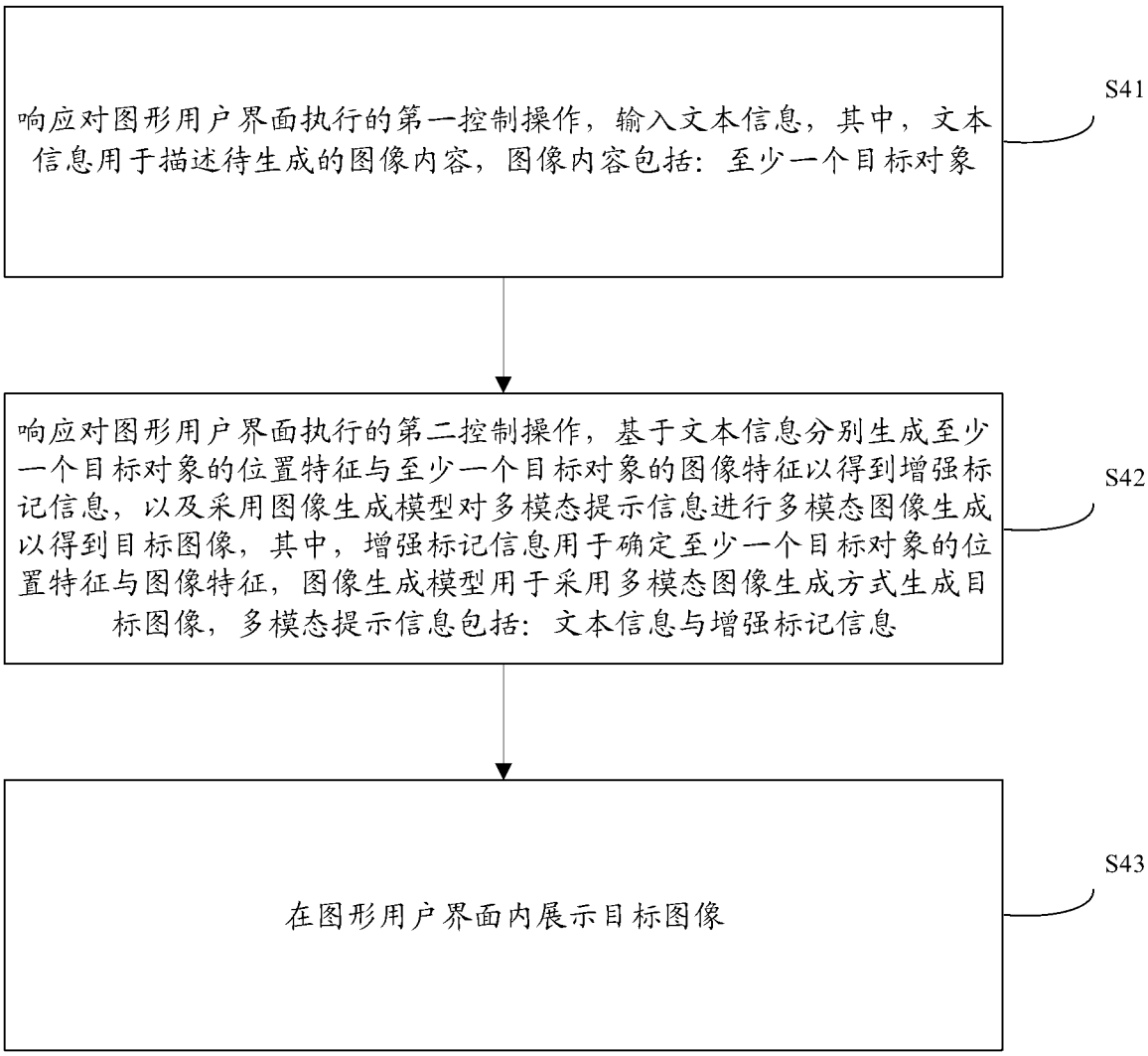
响应对图形用户界面执行的第一控制操作，输入文本信息，其中，文本信息用于描述待生成的图像内容，图像内容包括：至少一个目标对象
S41
响应对图形用户界面执行的第二控制操作，基于文本信息分别生成至少一个目标对象的位置特征与至少一个目标对象的图像特征以得到增强标记信息，以及采用图像生成模型对多模态提示信息进行多模态图像生成以得到目标图像，其中，增强标记信息用于确定至少一个目标对象的位置特征与图像特征，图像生成模型用于采用多模态图像生成方式生成目标图像，多模态提示信息包括：文本信息与增强标记信息
S42
在图形用户界面内展示目标图像
S43

图 4

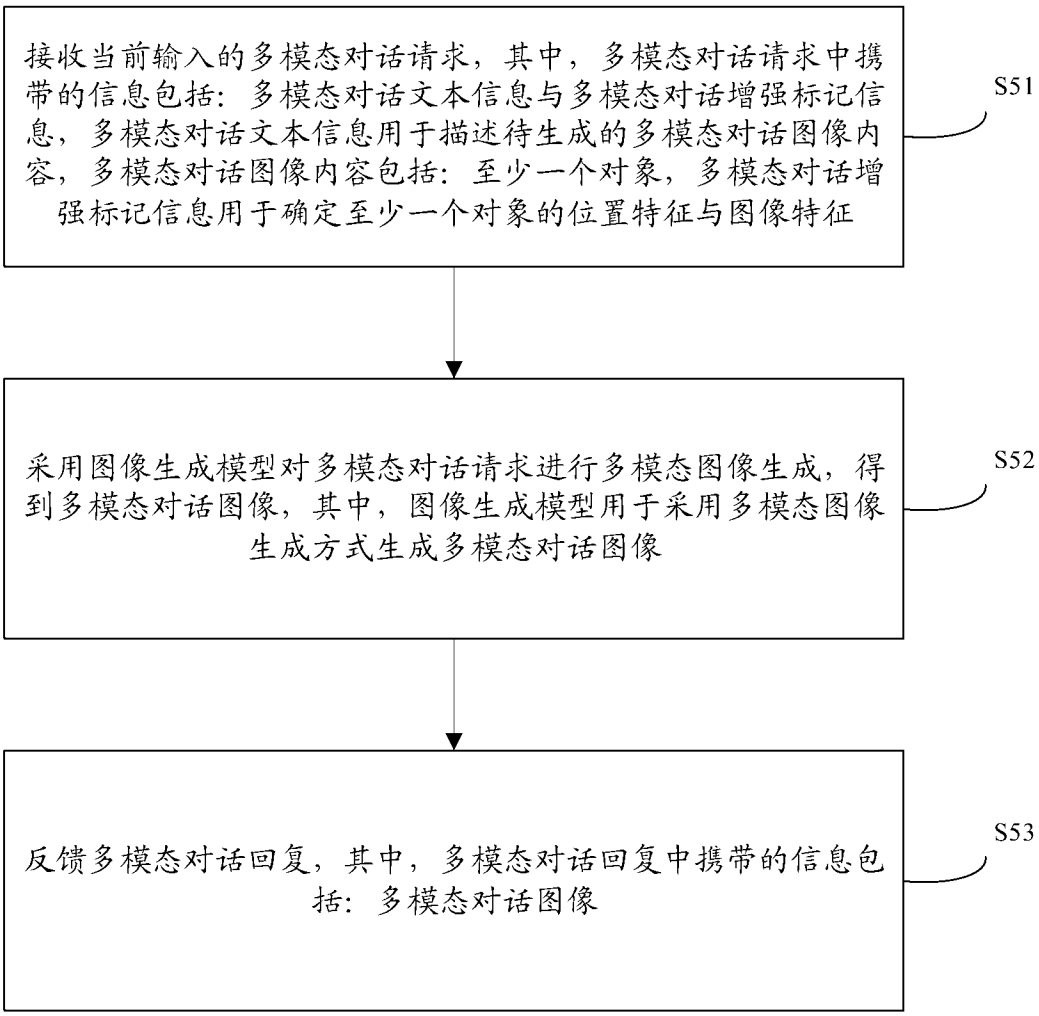
接收当前输入的多模态对话请求，其中，多模态对话请求中携带的信息包括：多模态对话文本信息与多模态对话增强标记信息，多模态对话文本信息用于描述待生成的多模态对话图像内容，多模态对话图像内容包括：至少一个对象，多模态对话增强标记信息用于确定至少一个对象的位置特征与图像特征
S51
采用图像生成模型对多模态对话请求进行多模态图像生成，得到多模态对话图像，其中，图像生成模型用于采用多模态图像生成方式生成多模态对话图像
S52
反馈多模态对话回复，其中，多模态对话回复中携带的信息包括：多模态对话图像
S53

图 5

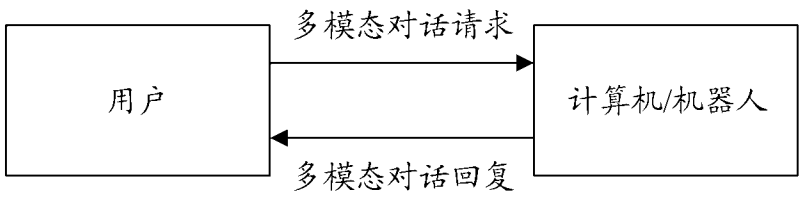
多模态对话请求
用户
计算机/机器人
多模态对话回复

图 6

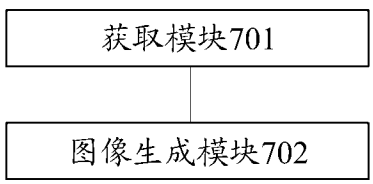
获取模块701
图像生成模块702

图 7


第一响应模块801
第一响应模块802
展示模块803

图 8


接收模块901
生成模块902
反馈模块903

图 9

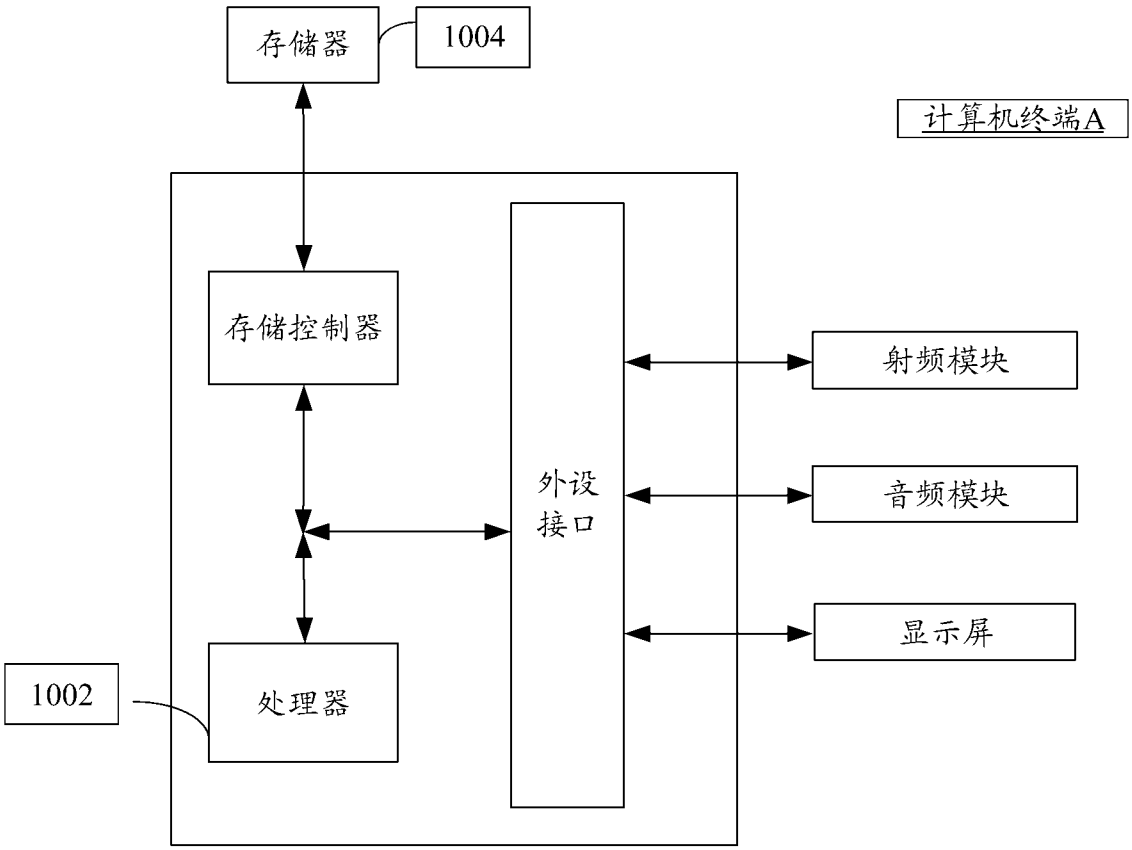
存储器
1004
计算机终端A
存储控制器
外设
接口
射频模块
音频模块
显示屏
1002
处理器

图 10

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2024/107883

A. CLASSIFICATION OF SUBJECT MATTER

G06T 11/00(2006.01)i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

IPC: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)

CNABS; CNTXT; VEN; USTXT; EPTXT; WOTXT; CNKI; IEEE; 必应, BING: 图像, 生成, 多模态, 模型, 文本, 位置, 增强标记, image, generate, multi-modal, model, text, position, augmented token

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
PX	CN 117689749 A (ZHEJIANG ALIBABA ROBOT CO., LTD.) 12 March 2024 (2024-03-12) claims 1-14	1-14
X	CN 114937181 A (STATE GRID INFORMATION AND TELECOMMUNICATION GROUP CO., LTD. et al.) 23 August 2022 (2022-08-23) description, paragraphs 3-46	1-14
X	CN 116704062 A (ALIPAY (HANGZHOU) INFORMATION TECHNOLOGY CO., LTD.) 05 September 2023 (2023-09-05) description, paragraphs 33-89	1-14
A	CN 116597039 A (ALIBABA (CHINA) CO., LTD.) 15 August 2023 (2023-08-15) entire document	1-14
A	US 2022156992 A1 (ADOBE INC.) 19 May 2022 (2022-05-19) entire document	1-14

☐ Further documents are listed in the continuation of Box C. ☑ See patent family annex.

* 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	Date of mailing of the international search report
25 October 2024	**01 November 2024**

Name and mailing address of the ISA/CN	Authorized officer
China National Intellectual Property Administration (ISA/CN) **China No. 6, Xitucheng Road, Jimenqiao, Haidian District, Beijing 100088**	Telephone No.

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2024/107883

Patent document cited in search report	Publication date (day/month/year)	Patent family member(s)	Publication date (day/month/year)
CN 117689749 A	12 March 2024	None	
CN 114937181 A	23 August 2022	None	
CN 116704062 A	05 September 2023	None	
CN 116597039 A	15 August 2023	CN 116597039 B	26 December 2023
US 2022156992 A1	19 May 2022	US 2023206525 A1	29 June 2023
		US 12008698 B2	11 June 2024
		US 11615567 B2	28 March 2023

国际检索报告

国际申请号
PCT/CN2024/107883

A. 主题的分类

G06T 11/00(2006.01)i

按照国际专利分类(IPC)或者同时按照国家分类和IPC两种分类

B. 检索领域

检索的最低限度文献(标明分类系统和分类号)

IPC: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））

CNABS;CNTXT;VEN;USTXT;EPTXT;WOTXT;CNKI;IEEE;必应:图像，生成，多模态，模型，文本，位置，增强标记，image，generate，multi-modal, model，text, position，augmented token

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
PX	CN 117689749 A (浙江阿里巴巴机器人有限公司) 2024年3月12日 (2024 - 03 - 12) 权利要求1-14	1-14
X	CN 114937181 A (国网信息通信产业集团有限公司 等) 2022年8月23日 (2022 - 08 - 23) 说明书第3-46段	1-14
X	CN 116704062 A (支付宝（杭州）信息技术有限公司) 2023年9月5日 (2023 - 09 - 05) 说明书第33-89段	1-14
A	CN 116597039 A (阿里巴巴（中国）有限公司) 2023年8月15日 (2023 - 08 - 15) 全文	1-14
A	US 2022156992 A1 (ADOBE INC.) 2022年5月19日 (2022 - 05 - 19) 全文	1-14

☐ 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

“D” 申请人在国际申请中引证的文件

“E” 在国际申请日的当天或之后公布的在先申请或专利

“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）

“O” 涉及口头公开、使用、展览或其他方式公开的文件

“P” 公布日先于国际申请日但迟于所要求的优先权日的文件

“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

“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

“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

“&” 同族专利的文件

国际检索实际完成的日期	国际检索报告邮寄日期
2024年10月25日	2024年11月1日
ISA/CN的名称和邮寄地址 中国国家知识产权局 中国北京市海淀区蓟门桥西土城路6号 100088	受权官员 翟紫伶 电话号码 (+86) 020-28958349

国际检索报告
关于同族专利的信息

国际申请号
PCT/CN2024/107883

检索报告引用的专利文件			公布日 (年/月/日)	同族专利			公布日 (年/月/日)
CN	117689749	A	2024年3月12日	无			
CN	114937181	A	2022年8月23日	无			
CN	116704062	A	2023年9月5日	无			
CN	116597039	A	2023年8月15日	CN	116597039	B	2023年12月26日
US	2022156992	A1	2022年5月19日	US	2023206525	A1	2023年6月29日
				US	12008698	B2	2024年6月11日
				US	11615567	B2	2023年3月28日