

案例分析：文化遗产图像自动描述和交互聊天 ——以广西高山村壁画为例

摘要

使用自然语言自动描述壁画并与壁画的局部内容进行交互对话是实现传统艺术数字化呈现和保护的重要方面。它可以帮助用户快速理解壁画内容，同时增加展览的趣味性。现有的研究工作缺乏对于中国壁画这一特殊的图像类型进行描述和交互，这可能由于缺乏相关领域的数据集，忽视传统艺术的数字化呈现等原因导致的。本文在前人研究基础上，结合中国传统文化知识，实现了对壁画的描述和交互。具体而言，我们利用在大规模图文上训练的模型对壁画进行整体和局部突出物体的表述，之后结合壁画上的题字信息检索出相似的古典诗词，最后通过设计提示使用大规模语言模型ChatGPT将这些信息融合到一起，生成人类可读的、融合了传统文化寓意和典故的文本描述。此外，本文也设计了人机交互的功能，即针对壁画的局部内容，与人工智能代理聊天，并输出特定类型的聊天内容，有助于增加大众对于图像细节的理解以及激发他们参观的兴趣。最后本研究分析了几类常见的描述错误，并总结了这一任务的挑战。用户可以在线对程序进行访问和体验¹。

关键词： 壁画描述；交互聊天；人工智能

Automatic Description and Interactive Chat for Cultural Heritage Images - A Case Study of Gao Shan Village Murals in Guangxi

Abstract

Automatically describing and engaging in interactive dialogue with murals using natural language is an important aspect of achieving digital presentation and preservation of traditional art. It can help users quickly understand the content of murals while increasing the enjoyment of exhibitions. Existing research lacks description and interaction for the specific image type of Chinese murals, which may be due to the absence of relevant datasets in this field and underestimation of digitizing traditional art. Building upon previous work, this paper focuses on providing descriptions and interactions of murals, combining them with Chinese traditional culture. Specifically, we utilize a model trained on large-scale image-text pairs to describe prominent objects in the entire mural as well as in specific areas. We retrieve similar classical poetry based on calligraphic inscriptions found in the murals. Furthermore, by designing prompts and

¹<https://modern-frigate-4f0.notion.site/AI-5c8f455ea4ad4801b02fd6b710276fc6>

incorporating a large-scale language model ChatGPT, we generate human-readable text that integrates all this information. Additionally, the paper presents a human-computer interaction feature, allowing users to chat with an AI agent specifically about the local content of the murals, thereby enhancing public understanding of image details and sparking interest. The showcased examples demonstrate the achievement of the primary goals of description and conversation. Finally, the study identifies and analyzes several common description errors and discusses the challenges associated with this task.

Keywords: mural description , interactive chatting , artificial intelligence

1 引言

在人工智能时代，数字化和智能化展示文化遗产艺术是文物保护和传承的重要手段，受到了很多研究机构和博物馆的高度重视，例如武汉大学数字人文研究中心²、北京大学数字人文研究中心³，上海博物馆⁴等。其中一个典型的案例是文化遗产图像交互式数字叙事，即从不同维度对图像中的概念和实体进行解构，通过语义增强的方法和数据叙事的思想，对图像进行深度解读。这样的数字化解读结果有利于大众对于图像内容进行深刻而全面的理解，同时交互式叙事则可以增加展览的趣味性，吸引更多的观众参与互动。(李本建, 谭阳, 2020)

壁画作为中国文化遗产图像的重要类型，兼具历史性和高雅性，对壁画的描述和解读有助于其传播和保护。为此，很多机构专门聘请壁画领域内的专家来解析壁画内容(霍嘉西, 李西林, 2020)。本次研究立足于如何让计算机自动地生成一段流畅、准确、全面的文字来描述壁画。一方面，对一副壁画的描述需要关照到图画中方方面面的信息，外加遣词造句、专家审核等。考虑到庞大的壁画数量，机器自动描述可以节省大量的人力物力资源。另一方面，借助于最近迅速发展的知识图谱和相关数据库资源建设⁵，机器可以利用更为全面的数据，拥有专家都可能忽略的知识。

在计算机和人工智能领域，与之相关的一个任务被称为自动图像描述(Karpathy and Fei-Fei, 2015)，它可以通过单句话进行描述、或者覆盖局部信息的密集描述，再或者对生成的描述进行控制(Chen et al., 2021; Vinyals et al., 2015)，例如文本长度、风格种类等等。然而，这些模型都是针对日常的图像，缺乏专门针对于传统文化浓厚的壁画描述。而壁画描述相比这些具有一下挑战：一是壁画中具有文字信息，而这些文字信息往往是理解一副壁画的关键线索；二是壁画中的事物往往和中国传统文化相关，具有大量的隐喻意义。这些都表明需要对针对一般图像的自动描述模型做出新的调整。

除了描述壁画内容外，交互式描述可以增强壁画传播和展示的趣味性。以往的针对遗产图像叙事的研究和系统都缺乏与用户的深度交互。这体现在一是缺少与机器持续多轮的对话。二是无法实现对生成内容的控制，例如从文本风格、长度等等。

本次研究试图解决上述问题，并设计出一个自动描述壁画的交互式程序。在生成描述方面，我们充分利用壁画上的题字信息，并检索了与其相关的古典诗词知识库，将它们作为壁画的传统文化知识信息，之后利用最新的生成式人工智能技术，将其融入到针对一般图像的预训练语言模型中，最终生成一段连贯流畅的段落。在交互式描述方面，我们一方面实现了对于生成文本的控制，另一方面，利用大语言模型技术，实现了和用户的多轮对话，有效提升了壁画描述的趣味性。

综上所述，本文的贡献如下：

- 一个与传统文化结合的壁画自动描述系统。它融入壁画的题字信息，并由此推断出壁画所传达的传统文化寓意，更有针对性对于壁画进行描述。
- 一个可交互的可控聊天系统。用户可以对壁画描述进行控制，并与机器背后的人工智能代理（即，ChatGPT）发生多轮对话。

²<http://dh.whu.edu.cn/dh/web/index.html>

³<https://pkudh.org/>

⁴<https://www.shanghaimuseum.net/mu/frontend/pg/index>

⁵<https://opendata.library.sh.cn/opendata/>

- 一个开源程序。将上述功能开发为程序，并将开源发布⁶。

2 相关工作

2.1 文化遗产交互式叙事

文化遗产交互式叙事是一种通过数字化技术，将文化遗产以交互式的方式呈现给公众，使观众能够更加直观、深入地了解 and 体验文化遗产的方式。这种方式不仅拓宽了文化遗产传播的途径，还增强了公众对文化遗产的参与感和认同感，以达到一种沉浸式的“心流体验”(徐迎庆, 图拉, 2023)。文化遗产交互式叙事主要包括以下几个方面：(1) 数字化技术：包括三维建模(Wachowiak and Karas, 2009)、虚拟现实(VR)(Guttentag, 2010)、增强现实(AR)(Haugstvedt and Krogstie, 2012)、人工智能(Pisoni et al., 2021)等技术，将实体文化遗产转化为数字化形式，以便进行传播和展示。(2) 交互设计：通过设计用户界面、互动方式(Maye et al., 2014)等，使观众能够自由地探索和体验文化遗产，提高观众的参与度和兴趣。(3) 叙事方式：将文化遗产背后的历史、故事、价值等以有趣、生动的方式呈现给观众，使观众对文化遗产有更深入的理解(徐迎庆, 图拉, 2023)。(4) 传播途径：通过互联网、移动设备、实体展示等多种途径，将文化遗产交互式叙事呈现给更广泛的观众(Pantano and Tavernise, 2009)。一个比较典型的案例可以参考由武汉大学数字人文研究中心开发的一款针对敦煌窟第61窟主室西壁的交互式叙事程序⁷。它通过可视化的方式将图中各类事物标识出来，并融入了多种叙事方式帮助观众理解壁画中的内容。但该程序仍然存在不足，例如，缺乏多轮的对观众对话的强交互形式，标注没有自动化，不能更迅速拓展到其他窟的壁画等等。

2.2 自动图像描述

自动图像描述(image captioning)通过分析图像以提取视觉信息，并将这些信息转换为自然语言文本以描述图像内容。主要技术包括图像特征提取、语言模型、评估等等。其中(1) 图像特征提取通过使用卷积神经网络(CNN)(Krizhevsky et al., 2012)等方法，从图像中提取出有助于描述图像内容的关键特征，如颜色、形状、纹理等。(2) 语言模型的目标是生成自然语言的描述文本，将提取的图像特征转换为文本表示。常用的语言模型包括循环神经网络(RNN)(Mikolov et al., 2010)、长短时记忆网络(LSTM)(Hochreiter and Schmidhuber, 1997)和Transformer(Vaswani et al., 2017)等。(3) 评估指标如BLEU分数(Papineni et al., 2002)、METEOR分数(Banerjee and Lavie, 2005)、CIDEr分数(Vedantam et al., 2015)等用于评价生成文本描述的性能。

根据图像描述的输出形式不同，这一任务分为：(1) 单句图像描述(Karpathy and Fei-Fei, 2015)：即通过单个语句描述目标图像，这时图像内容往往更为简单；(2) 多样图像描述(Wang et al., 2016)：生成多个候选的单个语句，作为可能备选句子集合，主要是考虑到图像内容的复杂性；(3) 可控图像描述(Chen et al., 2021)：根据可控信号，例如用户的鼠标轨迹、生成内容的长度或者语言的情感极性来生成针对性的语言；(4) 密集图像描述(Krishna et al., 2017)：从不同的局部生成对应描述，它们往往和位置信息关联，适合于内容复杂的图像。

本次壁画自动描述结合了单句以及密集图像描述的技术，用于产生一段解析壁画的段落；同时使用可控图像描述来进行交互。

2.3 生成式人工智能和大语言模型

生成式AI(生成式人工智能)是一种能够利用AI技术创作新内容的方法，涵盖了文本、图片、音乐、音频和视频等多种形式。最新的生成式AI技术基于大语言模型(Large Language Model)，这些模型能够执行多任务处理并完成开箱即用任务，如摘要、问答和分类等。相关的模型包括，ChatGPT⁸，GPT 4.0(OpenAI, 2023)，智源大模型(Zeng et al., 2022)。相比之前的预训练语言模型，大语言模型在预训练阶段需要的数据更多、模型的参数也成倍地增加，例如GPT-3(Brown et al., 2020)模型已经具有1750亿的参数。与拓展法则(scaling law)相对应的是模型表现出来一定的涌现能力(emergent abilities)(Wei et al., 2022)，例如GPT-3可以通过少量的样本进行上下文学习(in-context learning)就能达到不错的性能，这一过程不需要额

⁶<https://modern-frigate-4f0.notion.site/AI-5c8f455ea4ad4801b02fd6b710276fc6>

⁷<http://dh.whu.edu.cn/wutaihan/>

⁸<https://openai.com/blog/chatgpt/>

外的训练和更新参数，非常方便适配到其他下游任务。当前大语言模型技术已经产生较大影响，包括但不限于以下方面：改善聊天和搜索体验以提高客户互动度⁹；通过对话界面和汇总探索大量非结构化数据(Thoppilan et al., 2022)；协助处理重复性任务，如回复提案请求、将营销内容本地化为多种语言，以及检查客户合同是否合规等。本次研究所采用的大语言模型是ChatGPT，它被用于整合语句、作为知识库引入传统知识、个性化语言输出以及对话交互。

3 研究方法

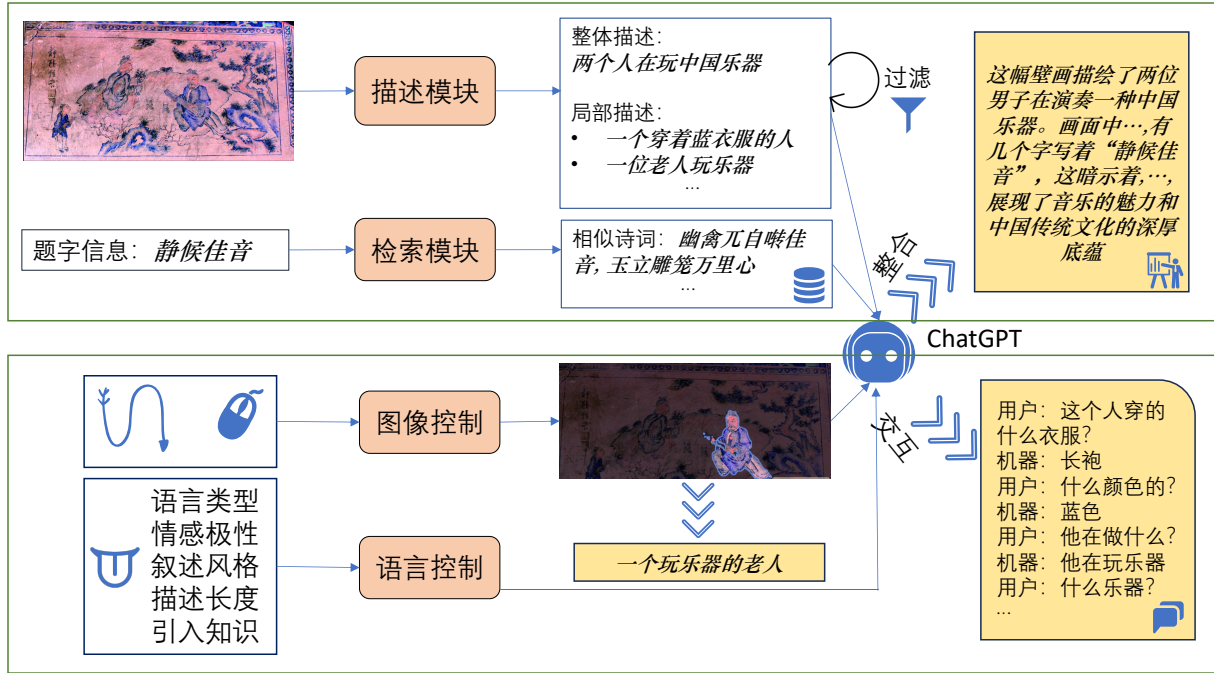


Figure 1: 本研究采用的模型框架和实现流程。其中上半部分展示了对于壁画的整体描述，下半部分展示了对于壁画的局部描述以及与之聊天交互。用于整合描述和交互的核心角色是大语言模型ChatGPT。

本研究采用的方法分为两大部分，一是实现对于整体壁画的描述，二是与对壁画的输出进行控制，同时实现聊天和交互。本次采用的模型参考开源项目CAT(Wang et al., 2023)，整体框架可以参考图 1。

3.1 融合背景知识的壁画整体描述

图 1上半部分展示了描述部分，其功能相当于“看图说话”，这一流程包括描述模块、检索模块和之后的整合模块。

描述模块 描述模块采用了在大量图文对预训练的图像描述模型(Li et al., 2023)，由于缺乏传统图像领域的图文数据库，没有在该模型上进行微调，直接“零样本”学习。首先，该模型对图像整体做出单句描述，这里记作 c_w ；同时它可以对局部突出物体进行分割，也生成相应的 i 句局部描述： $c_p = \{c_1, \dots, c_i\}$ ，它们分别对应图像区域 $r_p = \{r_1, \dots, r_i\}$ 。注意我们在该模块忽略了物体的位置信息。

考虑到局部描述往往含有较多噪声，即受限于描述的区域，产生一些不准确的描述，这里采用一些规则 d 对 c_p 进行过滤筛选：(1) 局部筛选。局部语句要与局部区域尽可能语义相似，这里采用Clip模型(Radford et al., 2021)计算二者的相似性，即 $Clip(c_j, r_j) \geq \alpha$ 。(2) 全局筛选。局部语句要与全局描述尽可能相似，这里将句子看作 n 元组的集合，并计算它们的Jaccard相似性，并选择前 k 个描述，即 $Topk(Jaccard(c_p, r_j))$ 。相关超参数在之后进行描述。之后剩下的符合要求的局部描述记做 c'_p 。

⁹<https://yiyuan.baidu.com/>

检索模块 考虑到壁画的文字信息对于理解整幅图像的重要性，我们将题字信息 c_h 手工录入到最终的系统中。文字信息往往包含中国文化典故或者场景，为了充分发掘这一信息，我们让文字信息从中国古典诗词库进行检索匹配¹⁰。其中，中国古典诗词库共收录了从先秦到当代共1,114,381首诗文，包罗万象，涵盖了各个类型、体裁、题材、风格的内容，是了解中国传统文化的一面重要窗口。为了尽可能多得发掘相关诗词，将题字信息是做 m 元组的集合，利用上述程序接口进行匹配。最终检索出的诗词语句记做 pm 。

整合模块 之后我们设计提示 \mathcal{P} ，利用大语言生成模型 \mathcal{G} 将上述信息融合在一起，得到最终的描述结果 c_f ：

$$c_f = \mathcal{G}(\mathcal{P}(c_w, c'_p, pm, c_h)). \quad (1)$$

其中提示 \mathcal{P} 如下：想象你是一位经验丰富的壁画描述专家，请整合下述信息 (c_w, c'_p, pm, c_h) ，并对壁画进行一段详细的、行文一致的描述，注意：（1）重点描绘画面中的壁画中突出的物体信息，例如相对位置、颜色、大小、形状、质地，而不是壁画所在的建筑信息；（2） c_h 和 pm 的信息对于壁画的描述很重要，它往往暗示了某个中国传统文化场景或者典故，请在最终描述中重点参考这一信息；（3）描述中不要出现数字，回答不要超过6个句子，不展示分辨率和绝对位置。最终结果请用中文显示。

3.2 可控壁画描述以及聊天交互

如图 1 下半部分所示，实现壁画的可控输出以及与之聊天交互是第二项功能。

可控描述 观众通过鼠标点击或者触碰轨迹来确定感兴趣的区域，之后采用一个现有的分割模型SAM (Kirillov et al., 2023)将这一区域分离出来。这就实现了图像的控制。同时用户也可以控制生成文本的特点，例如语言类型（哪一国家的语言）、情感极性（积极还是消极）、叙述风格（有想象力的还是事实描述的）、描述长度、是否引入外部知识等等，调整预期的语言风格。这部分语言控制通过ChatGPT来实现。例如，通过点击图中右半部分的人物，可以分割出他来，并且回应出“一个玩乐器的老人”这样的描述。

聊天交互 除了可控描述外，用户也可以针对指定的局部内容，与机器进行多轮的对话聊天。例如图 1 右下部分所显示的那样。这可以针对性地对图中的元素进行提问，提高用户的交互性并激发他们的兴趣。这一部分的功能是由ChatGPT来实现的。

4 实验和结果

4.1 壁画资源

本次研究所使用的壁画资源来自广西玉林的高山村。高山村是一个具有近六百年历史的岭南汉族移民村落，内有古宗祠12座，古民居60多幢500余间，是广西区保存完好、规模较大的明清古建筑群，被评为第二批“中国历史文化名村”、中国传统村落”、“全国文明村”。高山村古村落现存大小壁画有600幅，面积在500平方米左右，主要散布于承绪楼、致齐祠、聚星楼横屋、存理堂、绍德祠、思成祠、郎官祠、福斋祠附屋、叙祠、华彰祠、李垂宪祠、李拔谋故居、李氏新屋，其他民居也有部份。就单座数量而言，最多者可达100幅左右，少者也就大门口有几幅。

高山村壁画题材多样，内容丰富。既有传统题材，如福禄寿、科举、诗词名句、花鸟、山水、吉兽瑞禽，也有反映时代特征的轮船、都市码头、自鸣钟等，更有反映主人心境的诗词等。总体而言，高山村壁画可分为8部份内容。从人生追求、科举诉求、伦理教化、文学素养、生态家园、幸福愿景等多方面，尽情挥洒于墙端，教育当代，传之子孙。

目前，第一册《高山村壁画解析》已印刷完成，共收录70幅壁画。本次壁画描述也受到专家解析风格的影响，但值得注意的是，我们的描述更加侧重事实性的描写，还是以图中的事物为主体，专家对于部分壁画的描述有时更加注重壁画中所反映的传统文化。

¹⁰<https://open.cnkgraph.com/api/Writing/SimilarClauses/>

4.2 基础模型

本次研究建立在针对一般图像做描述和交互的模型CAT上。在图像描述部分，我们增加了对于密集描述的筛选（3.1）以及检索的匹配过程（3.1），以下是这两部分涉及到的超参数：Clip模型评分的阈值 α ：0.3；前 k 个与单一句子Jaccard相似的句子中 k ：3；检索时候题字的元组数量为4,5,6。

4.3 结果展示

表格 1展示了壁画描述的几个样例，可以看出，针对每一幅壁画，机器都可以自动生成较为正确、通顺的段落去描述该壁画。同时它也可以生成和传统文化相关的文字（用绿色表示）。例如在例子中的第一幅图中，可以识别出“鸟”，“树”等关键性物体。同时也阐发了图像想要传达的寓意，即“宁静”、“恬淡”的感觉。根据诗词语料库检索出来的诗词也可以作为专家描述的一个辅助。

值得注意的是，描述仍然会出现一些错误，例如一些事实性错误（红色标记），概念过度泛化（蓝色标记），一些不流畅的语言组织（紫色标记）。事实性错误大多出现在壁画本身不够清晰，或者拍摄角度造成的视野过小过偏所导致的；而对一些具有特定文化背景的事物，例如“鹤”等，模型容易将概念泛化为更大的类别，例如“鸟”。未来仍需要在模型中注入一些壁画相关的数据，同时建立与文化遗产相关的词库。生成流畅语言是大模型较为擅长的一部分，因而语言组织方面的错误最少。

图 2展示了我们的整个的功能界面。其中右下方展示了用户可以通过命令框针对图中的局部物体进行聊天交互。通过展示，可以看出我们的系统可以回答多轮对话，并且可以针对上一个问题中的回答进行提问（例如第三个问题针对“站”进行了进一步的追问），这得益于大模型可以生成流畅的对话。

5 挑战和结论

本次研究旨在利用最新的人工智能技术，对壁画进行自动描述并与之对话聊天。但仍存在如下的问题和挑战还未解决：第一，自动化水平还不够完善。作为描述关键的题字信息当前只能通过人为输入获取，而非自动文字识别。这主要考虑到壁画中存在字体各异、不够清晰等问题，普通的文字识别系统无法较为准确地识别。另一方面，拍摄过程中的视角也无法自动矫正，会导致一些壁画的视野较为歪曲或者狭窄的情况，这些也会影响识别的准确性。第二，传统文化的结合还不够。用于自动描述的模型尽管使用了网络上开源的大规模数据，仍缺乏针对遗产图像和传统文化知识的训练，导致模型不能像专家描述会挖掘这方面的信息。第三，缺乏一个全面、客观的量化标准。传统的针对自然语句的评分仅能刻画语句的流畅程度，同时，描述也具有相当的主观性，仅通过与单一专家的词汇匹配进行衡量具有一定的局限性。

然而，本次研究仍然给未来遗产图像的数字展示探索了一条可行的路径。以ChatGPT为代表的大语言模型技术极大地提升了自然语言的理解和生成的能力，也受到了各个领域的高度重视。这一技术对于改善文化遗产的交互性、趣味性、多样性都起到重要的作用，本次研究已经作为一个案例展示了这一点。

壁画	描述	检索诗词
	<p>这幅壁画描绘了一只栖息在树枝上的鸟。鸟的身体呈现出明亮的颜色，它的羽毛细腻而光滑，栩栩如生。在鸟的旁边，有一本书，书上写着东方文字。这本书和鸟形成了鲜明的对比，突出了东方文化的特色。壁画中还有一行文字：“池竹闲门教鹤守”，这句话暗示了中国传统文化中的典故或场景。根据类似的典故和诗词，我们可以推测这幅壁画可能描绘了一个池塘旁的竹林，其中有一只鹤守在闲门旁。整幅壁画给人一种宁静、恬淡的感觉，展现了中国传统文化的魅力。</p>	<p>题字：池竹闲门教鹤守</p> <ul style="list-style-type: none"> · 池竹闲门教鹤守， （《访含弘山僧不遇留题精舍》韦庄唐末至五代9764） · 却教闲鹤守， （《西湖寓次得云间张文敏墨梅册子欣然成诗十首其六》沈梧晚清784684）
	<p>这幅亚洲壁画描绘了两只大象和一棵树。画面中有两只大象在一片田野中，它们的相对位置是一只站立在前方，另一只则在后方。大象的颜色是灰色的，它们的大小和形状都非常逼真。树的树干上有一个巨大的洞，洞的颜色是黑色的，给人一种神秘的感觉。在壁画的一角，有一行文字写着“万象维新”。这句话暗示了壁画所描绘的场景与中国传统文化有关，可能是某个典故或者诗词的象征。根据相关典故和诗词的描述，这幅壁画可能代表着万象更新、繁荣昌盛的美好寓意。（《元旦试笔其一》弘历清804761）</p>	<p>题字：万象维新</p> <ul style="list-style-type: none"> · 万象维新嘉祉绵。（《元旦试笔其一》弘历清804761）
	<p>这幅壁画描绘了一棵树和一些动物。树的形状高大挺拔，树干粗壮，枝叶茂盛。动物们栩栩如生地栖息在树的周围，它们的颜色鲜艳，栩栩如生。在壁画的另一侧，有一张模糊的图像，似乎是一个沙滩上的冲浪板。在壁画的下方，有一块蓝色和黄色的标志，上面写着中文。标志上的文字是“三羊开泰”。这个典故暗示着中国传统文化中的吉祥和繁荣。</p>	<p>题字：三羊开泰</p>

Table 1: 壁画及其自动描述的示例，红色表示事实性错误，蓝色表示过度泛化，紫色表示生成的文字不通顺，绿色底纹表示文字与传统文化相关。

Caption-Anything

Gradio demo for Caption Anything, image to dense captioning generation with various language styles. To use it, simply upload your image, or click one of the examples to load them. Code: <https://github.com/ttengwang/Caption-Anything> Duplicate Space

The interface is divided into several functional areas:

- Image Area (Red border):** Displays the input image, a mural of two elephants. A bounding box highlights the elephants, and a caption reads "Two elephants are depicted in the image." The title "万象维新" is visible below the image.
- Control Panel (Yellow border):** Includes "Point Prompt" (Positive/Negative), "Clicking Mode" (Continuous/Single), "Clear Clicks", "Clear Image", "Language" (Chinese), "Sentiment" (Positive/Natural/Negative), "Factuality" (Factual/Imagination), "Generated Caption Length" (10), and "Enable Wiki" (Yes/No).
- Describe Everything (Green border):** Provides a detailed description of the image in Chinese and lists similar poems, such as "万象维新嘉祉绵。" by Hongliqing.
- Chatbox (Blue border):** Shows a sequence of prompts and generated captions: "caption: Two elephants are depicted in the image.", "它们是什么动物?", "The animals in the selected region are elephants.", "它们在干什么?", "The elephants in the selected region are standing.", "站在一起吗", "Yes, the elephants in the selected region of the image are standing together." It includes an input field and "Clear Text" and "Submit" buttons.

Figure 2: 壁画AI的整体界面和功能，其中左上角红色部分为壁画主体以及题字信息，右上角的绿色部分为整体描述和检索的相似诗词，左下角黄色部分为可控的功能选区，右下角蓝色部分为交互聊天的界面。

参考文献

- Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with high levels of correlation with human judgments. In *Proceedings of the ACL workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, volume 29, pages 65–72.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.
- Long Chen, Zhihong Jiang, Jun Xiao, and Wei Liu. 2021. Human-like controllable image captioning with verb-specific semantic roles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16846–16856.
- Daniel A Guttentag. 2010. Virtual reality: Applications and implications for tourism. *Tourism management*, 31(5):637–651.
- Anne-Cecilie Haugstvedt and John Krogstie. 2012. Mobile augmented reality for cultural heritage: A technology acceptance study. In *2012 IEEE international symposium on mixed and augmented reality (ISMAR)*, pages 247–255. IEEE.
- Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.
- Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137.
- Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. 2023. Segment anything. *arXiv preprint arXiv:2304.02643*.
- Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. 2017. Dense-captioning events in videos. In *Proceedings of the IEEE international conference on computer vision*, pages 706–715.
- Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, pages 1097–1105.
- Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*.
- Laura A Maye, Fiona E McDermott, Luigina Ciolfi, and Gabriela Avram. 2014. Interactive exhibitions design: What can we learn from cultural heritage professionals? In *Proceedings of the 8th Nordic Conference on Human-Computer Interaction: Fun, Fast, Foundational*, pages 598–607.
- Tomas Mikolov, Martin Karafiát, Lukáš Burget, Jan Cernocký, and Sanjeev Khudanpur. 2010. Recurrent neural network based language model. In *Eleventh Annual Conference of the International Speech Communication Association*.
- OpenAI. 2023. Gpt-4 technical report.
- Eleonora Pantano and Assunta Tavernise. 2009. Learning cultural heritage through information and communication technologies: a case study. *International Journal of Information Communication Technologies and Human Development (IJICTHD)*, 1(3):68–87.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318.
- Galena Pisoni, Natalia Díaz-Rodríguez, Hannie Gijlers, and Linda Tonolli. 2021. Human-centered artificial intelligence for designing accessible cultural heritage. *Applied Sciences*, 11(2):870.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

- Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, et al. 2022. Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.
- Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575.
- Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164.
- Melvin J Wachowiak and Basiliki Vicky Karas. 2009. 3d scanning and replication for museum and cultural heritage applications. *Journal of the American Institute for Conservation*, 48(2):141–158.
- Zhuhao Wang, Fei Wu, Weiming Lu, Jun Xiao, Xi Li, Zitong Zhang, and Yueting Zhuang. 2016. Diverse image captioning via grouptalk. In *IJCAI*, pages 2957–2964.
- Teng Wang, Jinrui Zhang, Junjie Fei, Yixiao Ge, Hao Zheng, Yunlong Tang, Zhe Li, Mingqi Gao, Shanshan Zhao, Ying Shan, et al. 2023. Caption anything: Interactive image description with diverse multimodal controls. *arXiv preprint arXiv:2305.02677*.
- Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. 2022. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*.
- Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, et al. 2022. Glm-130b: An open bilingual pre-trained model. *arXiv preprint arXiv:2210.02414*.
- 徐迎庆, 图拉. 2023. 沉浸式叙事视角下的中华文化数字化传承与传播. In *包装工程*, volume 44, pages 1–8+68.
- 李本建, 谭阳. 2020. 古村落数字化保护与传承之探讨——以广西高山村为例. *建筑与文化*, (07):59–61.
- 霍嘉西, 李西林. 2020. 韩休墓乐舞壁画解析. *交响 (西安音乐学院学报)*, 39(01):53–62.