文生图提示词中包含什么？Stable Diffusion 在视觉艺术教育中的潜力

Nassim Dehouche [1,*]、Kullathida Dehouche [2]

[1] 玛希隆大学国际学院工商管理系，泰国萨拉亚；nassim.deh@mahidol.edu[2] Rajamangala University of Technology Rattanakosin Poh-Chang艺术学院，泰国曼谷；kullathida.mee@rmutr.ac.th* 通讯作者：nassim.deh@mahidol.edu；玛希隆大学国际学院，泰国萨拉亚 Phutthamonthon 4 路 999 号，邮编 73170

摘要： 文生图人工智能（AI）近期迎来重大突破，DALL-E 及其开源对应程序 Stable Diffusion 相继发布。这些程序让任何人只需用自然语言（提示词）提供描述，就能创作原创视觉艺术作品。我们采用包含 72,980 条 Stable Diffusion 提示词的样本，对这种新型艺术创作媒介进行形式化界定，并评估其在艺术史、美学和技法教学中的潜力。研究结果表明，文生图人工智能有望彻底改变艺术教学方式，为实验和表达提供新的低成本可能性。然而，这也引发了有关艺术作品所有权的重要问题。随着越来越多的艺术作品通过这些程序创作，建立新的法律和经济模式以保护艺术家的权利将至关重要。

关键词： 人工智能；艺术；教育；计算创造力；知识产权。

> “首先，无论是人类艺术家还是神圣艺术家，都是‘凭借在智识中构想的言语’进行创作。”
> Ananda K. Coomaraswamy [1]

1. 引言

Coomaraswamy [1], 所倡导的传统艺术观认为，（人类）艺术是对（神圣创造）的模仿，而言语则是创作的起点。这一观点受到当代表现主义和形式主义观点 [2], 的显著挑战；随着人工智能（AI）近期取得进展，它也获得了新的技术表达。

事实上，人工智能在创造力领域取得了令人瞩目的进展，如今计算机能够根据简单的自然语言提示词生成相关且原创的文本 [3] 和图像 [4, 5], 。其中一些输出甚至与人类创作难以区分，因此获得了传统艺术竞赛的认可 [6]。

人工智能生成艺术仍是一个颇具争议的话题，尤其是关于它能否真正被视为艺术的辩论 [7], 。尽管学界对生成式人工智能模型的兴趣日益浓厚，但人们对其在视觉艺术教育中的潜在用途关注甚少。在我们看来，这些模型浓缩了数百年来人类艺术创作的成果，对艺术教育具有不容忽视的价值。因此，本文探讨将这些模型融入视觉艺术教育的可能性，尤其是用于艺术史、美学和技法教学。

在本引言部分之后，本文其余部分的结构如下。第2节将文生图领域的近期发展置于更广泛的

人工智能生成艺术史背景下。第3节专门聚焦于 Stable Diffusion——一种先进的开源文本生成图像系统，并介绍其基本功能。第4节阐述我们对 72,980 次 Stable Diffusion 交互进行分析所采用的方法与数据。基于这一分析，第5节提出关于 Stable Diffusion 提示词的形式化与程序化框架，可作为其在教育软件或课程中正式应用的基础；同时讨论其在艺术史、美学和技法等学科教学中的一些潜在用途，以及其对保护艺术家知识产权的影响。最后，第6节总结本文，并概述在我们看来，为促进文生图人工智能融入艺术教育，仍有待开展的工作。

2. 人工智能生成艺术简史

利用人工智能根据人类提示词创作连贯、原创内容的早期尝试可以追溯到20世纪50年代。当时，麻省理工学院人工智能实验室的研究人员开发了一个名为 ELIZA 的程序 [8]。ELIZA 能够运用模式匹配和自然语言处理技术，对文本输入生成简单的回应。虽然严格来说它并非艺术，但 ELIZA 是早期文本生成文本的一个例子：这种软件可以生成原创文本，供人类解读。最早的人工智能生成艺术实例之一，是艺术家 Harold Cohen 在20世纪70年代开发的名为 AARON 的程序 [9]。AARON 是一个能够生成复杂绘画作品的计算机程序。它运用一套规则与约束来创作艺术作品，并能从自身输出中学习，从而不断改进。

随着人工智能技术在20世纪80年代和20世纪90年代不断进步，更复杂、更精密的人工智能生成艺术开始涌现。例如，Karl Sims 运用进化算法生成了独特的三维图像和动画 [10]。近年来，深度学习的出现带来了更逼真的输出，因此人工智能生成艺术日益受到艺术界和公众的关注。2015年，谷歌的一个团队运用深度学习技术，在包含一万多幅绘画作品的数据集上训练神经网络，旨在根据输入图像生成原创艺术作品。由此产生的程序名为 DeepDream [11]，能够根据输入图像创作出超现实、视觉效果鲜明的图像（图像生成图像）。另一个值得一提的例子是巴黎艺术团体 Obvious 的作品：他们利用软件生成的一幅肖像画于2018年在佳士得拍卖会上以超过 432,000 美元的价格售出 [12]。

2020年，随着私营研究机构 OpenAI 发布第三代生成式预训练变换器（GPT-3），文生图能力实现了重大质的飞跃 [3]。GPT-3 在文本生成文本模型的通用性方面取得了重要进展，能够针对几乎任何自然语言提示词生成高度连贯的文本。这得益于模型庞大的规模，其包含1750亿个参数；比截至当时规模第二大的类似模型多出一个数量级。如此庞大的参数量使 GPT-3 能够理解它并未专门接受训练的语言任务，并开启了大语言模型时代。这些模型能够生成高质量、类似人类的文本，可用于多种应用，包括机器翻译、文本摘要和创意写作。GPT-3 的成功促成了 CLIP [13]，这一由 OpenAI 开发的另一突破性模型，该模型旨在将文本与图像关联起来。CLIP（对比式语言-图像预训练）是一种通用图文模型，使用来自互联网的4亿个文本-图像对进行训练，因此能够使用用户提供的任意标签执行图像分类。它还可以生成准确描述任意输入图像的文本（图像生成文本）。基于这些进展，OpenAI 发布了 DALL-E [4]，它能够生成逼真的

根据文本描述生成的图像（文生图）。尽管 DALL-E 仍是专有的闭源软件，但 CLIP 的代码已开源。这使人工智能公司 Stability AI 得以开发并训练 Stable Diffusion [5], 这一开源文生图模型，其性能可与 DALL-E 相媲美。Stable Diffusion 以宽松许可证发布，允许商业和非商业使用。

尽管 CLIP 及基于它的文生图系统代表了重要的技术突破，但也引发了重要的伦理与社会问题。由于 CLIP 是基于海量、未经筛选的互联网数据训练的，因此容易重现文化和社会中存在的偏见与不公平刻板印象 [14], ，而它可能对受保护作品的不公平使用也引起了法律专家的警觉 [15]。这些系统还可能被用于恶意目的，例如制造假新闻或传播虚假信息[16]。

3. Stable Diffusion

Stable Diffusion 是一种于 2022年 发布的文生图模型，它采用一种名为潜在扩散的深度学习技法 [5] ，根据文本描述生成图像。与一些早期的文生图模型不同，Stable Diffusion 的代码和模型权重均公开可用，并且可以在大多数消费级硬件上运行。

为了生成图像，Stable Diffusion 使用 CLIP [12] 将文本提示投射到共同的文本-图像嵌入空间中，并选取一张与输入提示在语义上接近、粗略且带噪声的图像。随后，系统会使用基于潜在扩散模型的去噪方法处理这张图像，以生成最终图像。除了文本提示外，Stable Diffusion 中的文生图生成脚本还允许用户输入各种参数，例如采样类型、输出图像尺寸和种子值。

这个潜在整数参数通常是随机设定的，但使用固定的种子值可以确保可复现性，并使生成图像的某些特征在不同提示词之间保持一致。例如，在图1(a)和图1(b)中，我们使用 Stable Diffusion 2.1版生成了两张图像，分别采用提示词“身穿皮夹克的年长女性/男性的细节照片，腰部以上镜头，森林背景，采用 Brandon Stanton、Humans of New York 的风格”，并将两张图像的种子值均设为 1034。我们可以看到，这个种子值使主体的某些面部特征在不同提示词和性别之间保持一致。此外，固定种子值也有助于在不同姿势和场景中保持主体外观一致。例如，图2展示了使用提示词“身穿皮夹克的年长女性/男性数字插画，维多利亚时代美学，腰部以上镜头，森林背景，采用 Magali Villeneuve 的风格”以及固定种子值 1242 生成的图像。需要注意的是，即使种子值固定，文本提示通常仍会产生一些随机瑕疵和不完美之处，例如图2(a)与图2(b)之间，以及图2(c)与图2(d)之间人物颈饰的差异。这些瑕疵可能需要通过额外的后期处理来修正生成的图像。



(**a**) (**b**)

(**c**) (**d**)

图1. 使用 Stable Diffusion 2.1 生成的图像，提示词为“身穿皮夹克的年长女性/男性的精细照片，腰部镜头，森林背景，Brandon Stanton、Humans of New York 的风格”。此外，还使用图像内补绘生成了图（c）和图（d）。

一些 Stable Diffusion 前端实现（如 DreamStudio[1]，）提供了额外的后处理任务功能，例如图像内补绘和图像外扩绘。图像内补绘是根据用户提供的提示词，在蒙版区域填入新内容，从而修改图像的特定部分。图像外扩绘则是根据用户提供的提示词，生成新内容，将图像扩展至原有尺寸之外。这两项功能都会使用 Stable Diffusion 模型生成新内容。例如，以图1(b)为起点，我们可以为主体添加配饰，如图1(c)所示；也可以更改图像的背景，如图1(d)所示，对应的图像内补绘提示词分别为“戴黄色帽子的男人”和“站在色彩缤纷的街角的男人”。

[1] hps://beta.dreamstudio.ai/



(**a**) (**b**)

(**c**) (**d**)

图2. 使用提示词“Magali Villeneuve 风格的维多利亚时代美学数字插画：一位年长女性/男性身穿皮夹克，腰部镜头，森林背景”在 Stable Diffusion 2.1 中生成的图像。

此外，图1和图2展示了 Stable Diffusion 复现当代在业艺术家风格的能力（分别为摄影师 Brandon Stanton 和插画师 Magali Villeneuve）。生成式人工智能这一备受争议的方面 [17] 将在第4.3节中得到更深入的分析。

4. 数据与方法

Stable Diffusion 生成的图像对文本提示的措辞高度敏感，因此我们着手考察这类输入的格式和语义内容。为此，我们从 Lexica[2], 这一搜索引擎收集了一个包含 72,980 条 Stable Diffusion 提示词的数据集。该搜索引擎收录了用户提交的精选 Stable Diffusion 生成结果及生成这些图像的提示词。我们的分析分为三个步骤：

1. 分词：每个提示词都会被拆分为“词元”；这些是基本的语言单位，可以是单词、短语、符号或提示词中的其他有意义元素。此步骤使用 BERT分词器 [18] 执行。2. 主题提取：此步骤旨在自动识别 72,980 条提示词中出现的主要主题或主题类别，并预先知道这些提示词是对图像的详细描述。此步骤使用 GPT-3 [3] API[3] 执行。

[2] hps://lexica.art/[3] h
ps://openai.com/api/

3. 分类：使用 GPT-3 API，将每条提示词中的词元归入步骤 1 中识别出的一种或多种语言主题。

此外，Stable Diffusion 准确复现特定艺术家风格的能力——这些艺术家的作品曾用于训练模型——一直以来都颇具争议。为了专门考察提示词中这一功能的使用情况，我们利用 BERT 的命名实体识别功能，识别出代表艺术家、品牌或团体名称的词元 [18] ，并统计了这些实体在所考察的 72,980 条提示词中各自出现的频率。

5. 结果与讨论

5.1. Stable Diffusion 提示词形式化

主题提取使我们能够识别表 1 所述的主要元素（即词元的语义类别）。这些是在所考察的 72,980 条提示词中出现频率最高的关键词类别。

表 1。72,980 条 Stable Diffusion 提示词中的主要元素

主题	描述
主体	图像中的角色和物体，例如“一个半机械人”，“两只狗”、“一辆汽车”、“一位巫师”等。
媒介	图像所属的视觉对象类型，例如“数字插画”、“照片”、“3D渲染图”、“概念艺术”，“海报”等。
技法	用于创作图像的工具和软件，例如“Blender”、“枕形镜头”、“Unreal Engine”、“Octane”等。
流派	描述图像艺术流派的美学特征，例如“动漫”、“超现实主义”、“巴洛克”、“照片级写实”、科幻，黑白、史诗奇幻、黑色电影等。
Mood	描述图像氛围和情绪的特征，例如“美丽”、“诡异”、“荒凉”等。
Tone	描述图像色彩构成的特征，例如“粉彩色”、“合成波色彩”、“空灵色彩”等。
光照	图像中光线和阴影的运用：“暗”、“电影式光照”、“写实阴影光照”、“影棚灯光”、光辉光线等。
分辨率	描述图像细节程度的特征，例如：“高度细致”、“照片级写实”、“100 mm”、“8K”、“16K”、“HQ”、“清晰对焦”等。
艺术参考	可用作灵感来源的艺术家或艺术作品，例如“Greg Rutkowski”、“吉卜力工作室”、“Artgerm”、“扎哈·哈迪德”等。
接受度/受欢迎程度	艺术类平台上的奖项、认可或热门状态，例如“在 ArtStation 上热门”、“杰作”、“获奖的”等。

出现频率较低的主题是对前述主要主题的延伸或补充细节，列于表2中。

表2. 72,980 条 Stable Diffusion 提示词中的次要元素

主题	示例
主体的物理特征	种族、年龄、服装、配饰、“可爱”、“迷人”，“胖乎乎的”等。
主体	
情感或心理主体的特征	“快乐”、“焦虑”、“得意”、“沉思”等。
环境/设置	时间、天气、“中世纪”、“后末日”等。
对称/重复	“对称”、“对称的”、“图案”、“母题”，“分形”，etc.
景深	“模糊的背景”、“深焦”、“光圈”、“F/4”，“F/2.8，“清晰对焦”、“散景”等。
角度	“超广角”、“天顶视角”、“电影视角”，“特写”等。
信息/意义	“宣传”、“宗教”、“广告”等。

5.2. 拟议的程序分类

识别出的提示词元素与传统摄影概念高度契合，并可按程序分类，如图3所示。

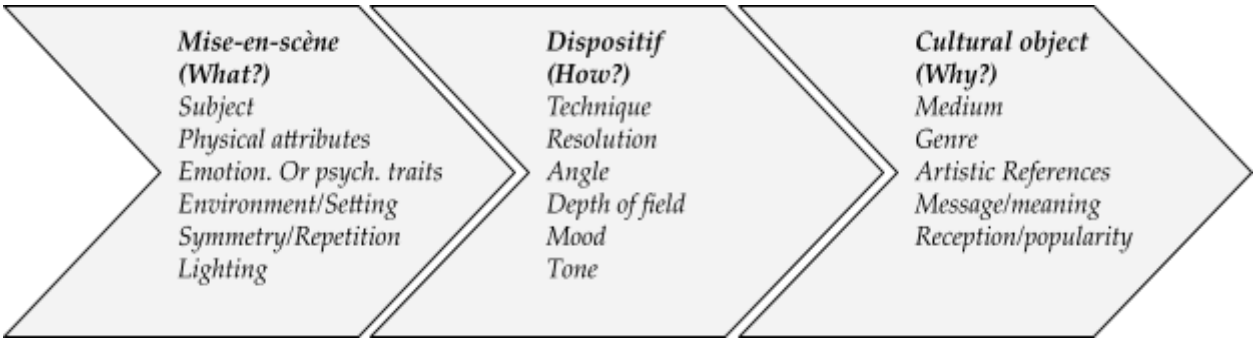


图3. 基于 72,980 条 Stable Diffusion 提示词中的语义元素提出的文生图提示词创作过程。

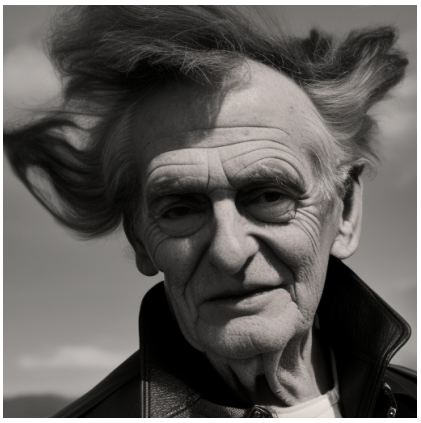
● **场面调度**： 场面调度是摄影、电影和戏剧研究中常用的术语，指镜头或场景中物体、布景和演员的安排 [20]。因此，这一类别包括画面中呈现的视觉和构图元素，用以创造预期的文化对象，例如“身穿皮夹克、神情桀骜不驯的年长女性/男性，置身于末日后城市，配以阴郁灯光”，如图4所示。




(**a**) (**b**)

图4. 使用 Stable Diffusion 2.1 生成的图像，提示词为“身穿皮夹克、神情坚定的年长女性/男性的数字绘画，背景为末日后城市，阴郁灯光，ArtStation 热门作品，Greg Rutkowski”

● 装置： 在摄影和电影中，装置这一概念涉及用于捕捉图像的物质技术配置 [19] 。在更广泛的分类中，这一类别还可能包括用于数字图像的软件工具和后期处理技术。如果场面调度指的是图像中呈现的内容，那么装置指的就是创作方式，例如“特写、黑白、大光圈、8K、锐利的边缘”，如图5所示。

(**a**) (**b**)

图5. 使用 Stable Diffusion 2.1 生成的图像，提示词为“快乐而若有所思的年长女性/男性肖像照片，身穿皮夹克，森林背景，特写，黑白，大光圈，8K，边缘锐利，RobertDoisneau”。

● 文化对象： 这些元素描述了艺术家创作的“对象”，兼具“人工制品”和“目的”两层含义；后一层包括对图像媒介和流派的描述，以及通过艺术参考（例如“安妮·莱博维茨拍摄的一张照片”或“米开朗基罗创作的一幅文艺复兴绘画”）说明其在艺术史中的定位；前一层则包括对信息/意义以及接受度/受欢迎程度的描述（例如“宗教主题、获奖作品”）。图6展示了这两种提示词组合。



(a)



(b)

图6. “一位穿着皮夹克的老妇人肖像，宗教主题，获奖的”，作为 (a) “安妮·莱博维茨拍摄的照片”和 (b)“米开朗基罗创作的文艺复兴绘画”。

需要注意的是，我们提出的程序分类中的各个元素并非彼此独立或互斥。例如，将某位艺术家的名字用作艺术参考，可能会影响生成图像的情绪和色调。出于创作目的，探索这些元素不同寻常或相互冲突的组合可能很有意思；但值得牢记的是，CLIP 根据文本关联出的初始图像只是一团充满噪点的像素汤，而提示词的作用是引导其去噪。因此，提示词越连贯，生成效果就越好。掌握文生图需要理解这些元素之间的相互作用，其中也包含一定程度的随机性，从而生成最连贯的艺术作品。

5.3. 视觉艺术需要新的经济模式

图7中的词云展示了在所研究的72,980条提示词中，作为艺术参考使用的命名实体的出现频率。我们发现，这些命名实体主要指代当代在职艺术家，他们经常在数字艺术平台ArtStation上发布自己的作品。例如，波兰画家格雷格·鲁特科夫斯基（包括其姓名的轻微拼写错误，以及与其他艺术家一同被提及的情况）出现在41.06%的提示词中，而作为接受度/受欢迎程度要素的ArtStation则出现在63.35%的提示词中。



图7. 在 72,980 条 Stable Diffusion 提示词中，用于获取灵感的艺术家、品牌或团体名称词云。

这些关键词之所以受欢迎，可以归因于诸如ArtStation这样的平台鼓励艺术家添加详细标签来描述其作品，以便残障人士更容易访问，这使得这些创作对于训练文生图人工智能模型特别有用。因此，ArtStation艺术家在某种程度上因其善举而受到惩罚。关于这种训练是否构成剽窃的法律问题仍悬而未决 [15] ，而且可能需要数年甚至数十年才能解决。除了可能不公平地使用这些艺术家的知识产权之外，Stable Diffusion的广泛使用还导致这些艺术家的原创作品在搜索引擎结果中被提示词中带有其姓名的人工智能生成作品所掩盖。

尽管这种方法并不完善，因为它没有考虑到创作图像时隐性使用的作品，但针对这些问题，一个简单的技术解决方案是：参照音乐流媒体经济模式，根据艺术家姓名作为风格参考出现在商业文生图应用中的频率，为艺术家设计补偿模式。

6. 结论

本文旨在提供一种结构化方法，帮助理解这种新的艺术创作媒介，并将其与既有的艺术教育概念联系起来，尽管 Stable Diffusion 提示词的输出在一定程度上具有随机性，且对措辞高度敏感。

Stable Diffusion 对艺术的“理解”无疑是肤浅的，本质上停留在噱头层面（正如 [21], 所指出的，噱头“仍是资本主义最成功的美学类别”）。然而，在教育者的恰当指导和策展下，它可以成为传授技术概念以及艺术流派、运动和美学等体验性概念的宝贵教学工具，这些概念构成了文化对象的特征。此外，在种子整数保持不变的情况下，对场面调度和装置元素进行调整，可以成为一种快速且低成本的实验与原型制作方法，从而避免占用昂贵的工作室时间。

尽管 Stable Diffusion 及类似软件潜力巨大，但要使其和谐融入艺术界，就必须明确相关的伦理与法律问题，尤其是这些软件引发的一个重要问题：如何公平补偿那些作品被用于训练这些模型的艺术家。

参考文献

1. Coomaraswamy, A.K.《基督教与东方艺术哲学》；Dover Publications：美国纽约，1956年；第154–196页。2. Braembussche, A.《思考艺术》；Springer：德国多德雷赫特，2009年。https://doi.org/10.1007/978-1-4020-5638-33. Brown, T. 等。《语言模型是少样本学习者》。《神经信息处理系统进展》。载于第33届神经信息处理系统会议（NeurIPS 2020）论文集。2020年。4. Ramesh, A.; Dhariwal, P.; Nichol, A.; Chu, C.; Chen, M.《基于 CLIP 潜变量的分层文本条件图像生成》，2021年，arXiv 预印本 arXiv:2204.06125。5. Rombach, R.; Blamann, A.; Lorenz, D.; Esser, P.; Ommer, B. 使用潜扩散模型进行高分辨率图像合成。载于 IEEE 计算机视觉与模式识别会议（CVPR）论文集。2022年。6. Roose, L. 一幅人工智能生成的图像赢得了艺术奖。艺术家们并不高兴。《纽约时报》。2022年。hps://www.nytimes.com/2022/09/02/technology/ai-artificial-intelligence-artists.html（访问于2022年12月9日）。7. Zylinska, J.《人工智能艺术：机器视觉与扭曲的梦境》；Open Humanities Press：英国伦敦，2020年。8. Weizenbaum, J.《计算机的力量与人类理性：从判断到计算》；W. H. Freeman and Company：美国纽约，1976年。ISBN 0-7167-0463-3。9. Cohen, P. Harold Cohen 与 AARON。《人工智能杂志》。2016年，37(4)，63-66。ISSN 0738-4602

10. Sims, K. 编排图像流。《可视化与计算机动画杂志》。1992年，3(1)，31-44。https://doi.org/10.1002/vis.434003010611. Mordvintsev, A.; Olah, C.; Tyka, M. DeepDream：用于可视化神经网络的代码示例。Google Research。2015年，2(5)。12. Cohn, G. 佳士得拍卖的人工智能艺术品以432,500美元成交。《纽约时报》。2018年，https://www.nytimes.com/2018/10/25/arts/design/ai-art-sold-christies.html （访问于2022年12月3日）。13. Radford, A. 等。通过自然语言监督学习可迁移的视觉模型，2021年，arXiv 预印本 arXiv:2103.00020。14. Dehouche, N. 预训练分类器中的隐性刻板印象。《IEEE Access》。2021年，9。https://doi.org/10.1109/ACCESS.2021.313689815. Franceschelli, G.; Musolesi, M. 生成式深度学习中的版权。《数据与政策》。2022年，4。https://doi.org/10.1017/dap.2022.1016. Fallis, D. 深度伪造的认知威胁。《哲学与技术》。2021年，34，623–643。https://doi.org/10.1007/s13347-020-00419-217. Zeilinger, M. 人工智能艺术、创作能动性与知识产权的局限；Meson Press：德国吕讷堡，2021年。https://doi.org/10.14619/183918. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT：用于语言理解的深度双向变换器预训练，2018年，arXiv 预印本 arXiv:1810.04805。19. Kessler, F.; Chateau, D.; Moure, J. 银幕与装置概念——一场对话，载于《银幕》；阿姆斯特丹大学出版社：荷兰阿姆斯特丹，2016年；第264-272页。20. Sikov, E. 场面调度：电影摄影，载于《电影研究》第二版，哥伦比亚大学出版社：美国纽约，2020年。https://doi.org/10.7312/siko19592-00521. Ngai, S.《噱头理论：审美判断与资本主义形式》；哈佛大学出版社，美国马萨诸塞州剑桥，2020年。