arXiv:2603.10578v1 [cs.CV] 11 Mar 2026

R4-CGQA：用于计算机图形图像质量评估的基于检索的视觉语言模型

Zhuangzi Li、Jian Jin、Shilv Cai、Weisi Lin，南洋理工大学计算与数据科学学院，新加坡

摘要

沉浸式计算机图形（*CG*）渲染已3.5K *CG*图像及其对应质量描述的数据集。在现代日常生活中无处不在。然而，全面评估计算机图形质量仍面临两项挑战：（*1*）现有计算机图形数据集缺乏对渲染质量的系统性描述；（*2*）现有计算机图形质量评估方法无法提供合理的基于文本的解释。为解决这些问题，我们首先从用户视角确定了计算机图形质量的六个关键感知维度，并构建了一个 *https://github.com/lizhuangzi/R4-CGQA*。

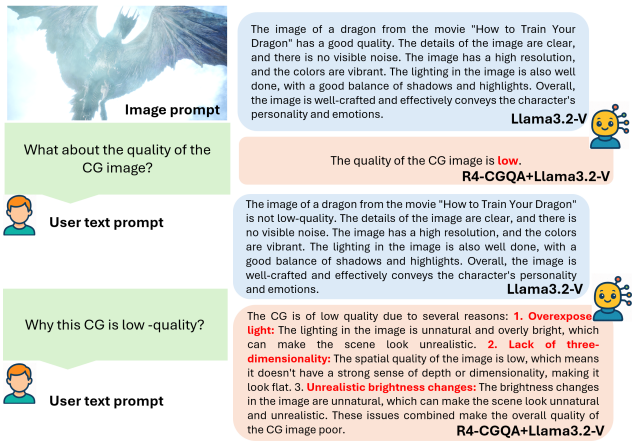

图 1. 展示所提出的 R4-CGQA 能力的示例。R4-CGQA 不仅使大型模型的回答更加简洁、准确，还能有效释放它们在 CGQA 任务上的潜力。

1. 引言

高质量计算机图形学（CG）渲染是许多工业应用中的关键技术，例如游戏设计 [7], 3D 动画 [10], 和电影特效 [29]。在这些场景中，高质量的视觉效果和沉浸式体验已成为满足用户期望的必要条件。因此，工业界需要智能质量评估算法来指导计算机图形内容的设计与渲染。尽管近期提出了许多图像质量评估方法和人工智能生成图像质量评估方法，但很少有研究专门考量 CG 图像的质量。直接将自然图像质量评估方法应用于 CG 图像并不恰当 [35],，因为 CG 图像完全由模拟构建而成，其中包括物体、纹理、光源和摄像机视角，因此它们在失真和感知特征方面都与自然图像存在显著差异。

为了解决这一问题，已有多项工作建立了计算机图形图像数据库，为CG质量评估（CGQA）提供判别标准。具体而言，Wang 等人 [35] 收集了1.2K张来自三维电影和三维游戏、带有人工评分的CG图像，并表明基于深度学习的方法比基于手工设计特征的方法更符合人类感知。Zhang 等人 [45] 构建了一个包含约6K张带评分图像的CG数据集，并提出了一种无参考的、基于深度学习的CGQA模型，该模型同时利用失真与美学质量表征。然而，这些数据库仅包含主观质量评分；换言之，其评估视角不可避免地受到限制。我们认为，一种好的CGQA方法应当揭示人类判断背后的原因，并进一步提

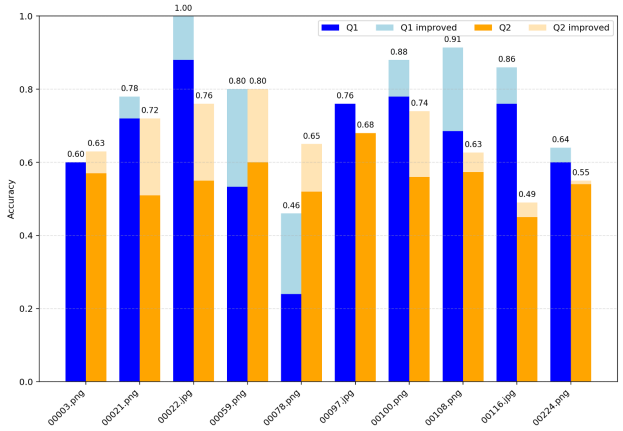

图 2. 我们准备了 10 张 CG图像，并为其配备了相关问题和答案（Q1：选择题，Q2：是非题），并在 LLaVA-13B 上评估答案准确率 [18]。深色柱表示由视觉语言模型直接提供评估时的准确率；浅色柱表示向视觉语言模型提供视觉上相似的 CG图像的参考描述时的准确率。

供预期CG渲染的清晰指导。因此，我们构建了一个新的数据集，其中包含3.5K张带有相应基于文本的质量描述的CG图像。我们请专业参与者从光照、材质、颜色、氛围、真实感和空间这六个维度中描述三个最显著的方面。该数据集使语言模型能够学习更准确、更全面的CGQA能力。

近年来，视觉语言模型（VLM）在质量描述和推理方面展现出令人瞩目的能力[43]。Chen 等人 [8]提出了一种结合视觉语言模型与 ChatGPT 的图像问答系统。Zhu 等人 [47] 提出了一种自适应图像问答方法，通过训练视觉语言模型进行比较判断。Q系列研究[20, 38, 39, 46]针对各种退化图像提出了基准测试、调优方法和评估方法。其他研究 [30, 34] 则专注于评估 AI 生成图像的质量。然而，由于在某些知识领域存在不确定性，视觉语言模型容易产生幻觉，尤其是在 CGQA 领域。微调视觉语言模型需要大量计算资源和训练数据，也不利于及时更新模型知识。

我们的初步研究表明，相似CG图像的质量描述能够为理解目标CG图像的质量提供有信息量的知识。具体而言，如图2所示，我们选取10张CG图像，并为每张图像向视觉语言模型提供视觉相似示例的描述，然后要求其回答多项选择题（Q1）和是非题（Q2）。图表清楚地表明，纳入来自那些CG图像的描述，这些图像在视觉上与目标CG相似，通常能够提高视觉语言模型回答相关问题的准确率。此外，我们注意到，纳入与目标图像无关的描述会损害视觉语言模型的理解。这些观察促使我们基于我们的CG数据库，利用基于检索的技术进行CGQA。然而，定义两张CG图像之间的相似度具有挑战性。在本文中，我们在检索过程中联合考虑内容相似度和质量相似度，以确保示例描述恰当。我们在多个视觉语言模型上开展实验，结果表明我们的方法带来了一致的提升。我们的方法具有通用性，可应用于许多视觉语言模型，例如Llama3.2-V，如图1所示。我们的贡献总结如下：

- 我们提出了一个全新的CG质量评估数据集，其中的描述涵盖CG质量的六个维度，为CGQA提供了合适的数据基础。据我们所知，这是首个专门设计用于系统阐释CG图像质量的数据集。

- 基于贝叶斯理论，我们提出了一种新颖且通用的CG质量评估框架，可用于现有视觉语言模型，并在检索阶段同时纳入内容相似度和质量相似度。

- 我们基于该数据集提供测试基准，并在多个具有代表性的视觉语言模型上开展实验，包括LLaVA、Llama 3.2-Vision和Qwen2.5-VL，实验结果证明了我们方法的有效性。

2. 相关工作

2.1. 传统图像质量评估

图像质量评估（IQA）方法通常根据是否有参考图像可用进行分类：全参考（FR） [26, 36], 简化参考（RR）[44], 和无参考（NR，也称为盲图像质量评估） [5, 21,28, 37]。到2010年代，IQA领域已确立全参考/简化参考/无参考分类体系，并形成了一套经过设计的算法（例如 SSIM、FSIM、VIF、GMSD、BLI-INDS、BRISQUE），与原始误差度量相比，这些算法显著提高了与人类判断的相关性 [42]。尽管这些手工设计技术对传统失真效果显著，但在复杂的真实场景图像上的性能往往会趋于饱和。基于深度学习的图像质量评估利用大规模标注数据训练网络，直接预测感知质量分数。Kim等人 [14]提出了 BIECON 模型，利用图像块级CNN特征预测局部质量分数，再将其汇总为全局主观分数。Su等人 [5] 为真实场景中的盲图像质量评估设计了一种自适应超网络架构，并将图像质量评估流程分解为三个阶段：内容理解、感知规则学习和质量预测。Varga [31] 采用 Inception模块来捕捉

多尺度纹理与结构，体现了同时处理不同频率和图像块尺寸下失真的必要性。

受 Transformer 在视觉任务中取得成功的启发，研究人员进一步将自注意力网络应用于图像质量评估。You 等人 [9, 41]提出了 TRIQ，这是一种基于 Transformer 的无参考图像质量评估模型，在卷积神经网络特征后接入一个浅层变换器编码器。TRIQ 的自适应位置编码能够处理任意分辨率的图像，其多头注意力机制比传统卷积神经网络更有效地捕获全局图像上下文。与此同时，Ke 等人 [13] 提出了 MUSIQ，这是一种多尺度图像质量 Transformer，通过融合多个尺度的图像块嵌入来处理全分辨率图像。Transformer 也被应用于全参考图像质量评估：Cheon 等人 [9]开发了图像质量变换器（IQT），它使用孪生编码器–解码器变换器，将参考图像与其失真版本进行比较。除了这些新颖的架构，研究人员还通过半监督学习和合成数据生成来应对数据受限问题。例如，Cao 等人 [6] 将正例–未标记学习与双分支网络相结合，利用未标记的退化图像进行全参考图像质量评估，并采用切片 Wasserstein 距离处理空间错位。总体而言，近期研究已确立基于深度学习的图像质量评估这一主导范式：基于卷积神经网络和 Transformer 的模型如今在许多基准测试中都能达到接近人类水平的相关性。引入注意力机制、多任务学习（例如联合进行失真识别）和大规模训练，显著提升了图像质量评估模型在不同失真类型和内容领域中的泛化能力。然而，这些方法通常无法为其预测分数提供具体、可供人类理解的原因，因而限制了它们指导后续图像修改的能力。

2.2. 使用大型视觉语言模型进行图像质量评估

最近，BLIP-2 [17], GPT-4 [23], 和 Qwen2.5-VL [4] 等大型视觉语言模型（VLM）相继涌现，它们通过大规模图文对学习丰富的视觉–语言相关性。随着这类 VLM 的兴起，图像质量评估研究也日益受到影响。研究人员不再从头训练专用的图像质量评估网络，而是利用在海量图像或图文数据集上训练的基础模型来提升感知质量评估。Wang 等人 [32] 探索了 CLIP 在零样本图像质量评估中的应用，并提出假设：CLIP 学到的视觉特征以及与人类描述性语言的对齐能力，可以作为人类质量判断的替代依据。他们证明，通过设计合适的文本提示（例如“高质量照片”与“低质量照片”），并在 CLIP 的嵌入空间中测量图像与这些描述的相似度，便

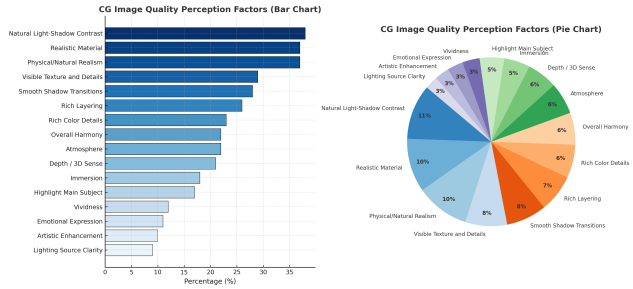

图3. 计算机图形图像质量感知因素。

可在无需针对特定任务进行训练的情况下获得质量预测。Chen 等人提出了 IQAGPT [8], , 这是一种创新的计算机断层扫描图像质量评估系统，它将图像质量描述视觉语言模型与 ChatGPT 相结合，以生成质量评分和文本报告。尽管这些 VLM 最初并非为传统图像质量评估而设计，但它们可以通过检查图像是否包含不真实物体或图文不匹配等方面，为质量评估提供补充 [33]。Zhu 等人 [47]将图像质量评估建模为适用于大型多模态模型（LMM）的比较推理任务，并训练视觉语言模型通过在训练期间生成指令提示（例如，给定两张输入图像后询问“哪张图像更好？”）来比较图像的成对质量。然而，这些方法对评分的可解释性仍有限，因此对于如何提升图像质量所提供的指导作用较弱。

Q-Bench [38] 提出了一个全面的基准测试，系统地评估多模态大语言模型（MLLMs）在三个方面的能力：低层次视觉感知、低层次视觉描述和整体视觉质量评估。在此基础上，Wu 等人提出了 Q-Instruct [39], , 利用图像质量描述和 ChatGPT 构建大规模问答对，用于微调视觉语言模型。Q-Adapt [20]旨在减少冲突，实现整体质量解释与属性级感知判断之间的协同。Zhang 等人 [46] 提出了 Q-Boost，以提升视觉语言模型在低层次图像质量评估任务中的性能。与此同时，近期关于人工智能生成图像的研究将“图像质量”的概念扩展为除失真保真度之外，还包括真实性和语义对齐 [3, 34]。相应地，研究者提出了多个面向视觉语言模型的人工智能生成图像质量评估数据集，例如 AIGIQA2K[16]和 AIGI-VC [30]。这些模型和数据集提供了强大的高层次表征和语义理解，帮助图像质量评估从低层次失真测量迈向整体感知评估。然而，在计算机图形学（CG）领域，仍然缺乏专门为基于视觉语言模型的质量评估而设计的 CG 数据集；一些现有数据集仅提供

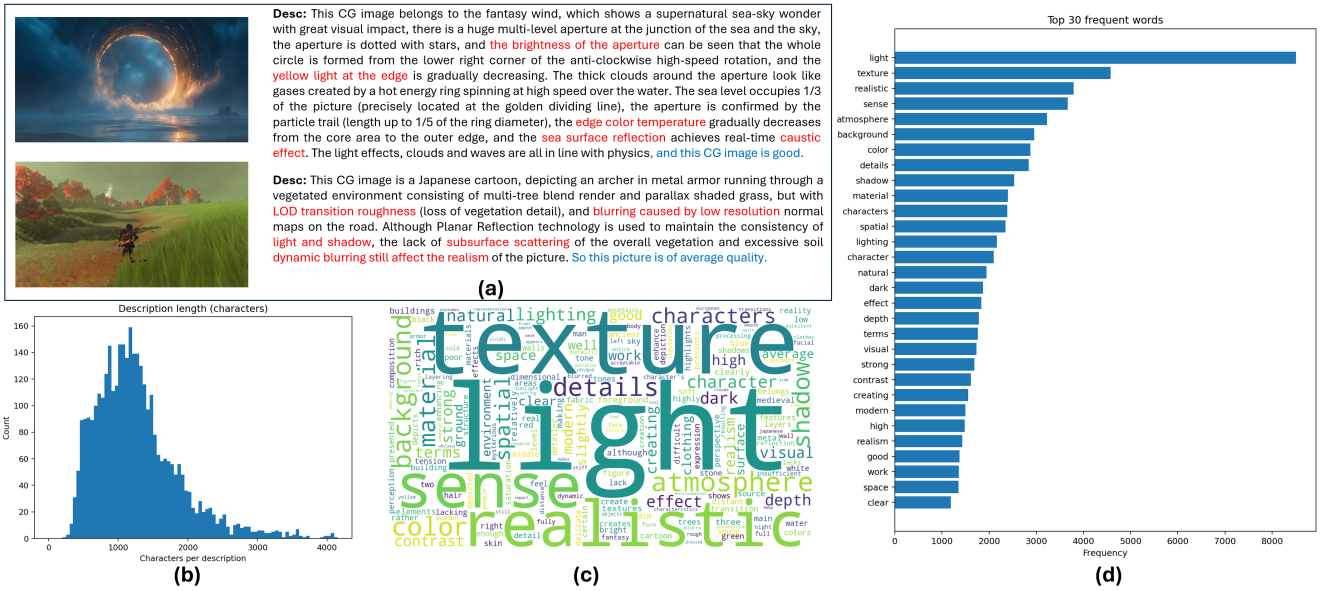

图4. 数据集特征概览。(a) 展示多样视觉内容的图像-描述对示例。(b) 数据集中描述长度的分布。(c) 高频描述性术语的词云可视化，突出显示纹理、光线和真实感等关键感知属性。(d) 频率最高的30个词及其对应频率。

标量质量分数 [35, 45]，并且缺乏对 CG 图像质量维度的详细分析。

3. 我们的 CGQA 数据集

近年来，计算机生成（CG）图像已广泛应用于游戏、虚拟现实、电影制作和数字内容创作等领域。随着这些行业的快速发展，人们越来越需要全面且结构良好的 CG图像数据集，以支持图像质量评估、真实感分析和视觉感知方面的研究。目前已有多个计算机图形学数据集，但其中许多质量有限，因为大多数图像的分辨率较低。此外，这些数据集通常只提供 CG图像的平均意见分（MOS）标签，却没有标注详细的感知维度，也没有描述计算机图形质量。因此，这类数据集并不适用于当今基于多模态大模型的智能评估系统。现有数据库 [45] 提供带有分数标签的 CG图像，但没有说明这些分数的依据或原因。我们认为，评估 CG图像的质量需要考虑光照、颜色和空间构图等多个方面。然而，给定一张 CG图像，要生成一致且合理的文本描述并非易事。

为解决这一问题，我们咨询了 CG 行业从业者，并总结了他们关注的关键术语，如图 3所示。基于这项分析，我们提炼出评估计算机图形质量的六个关键维度：光照质量、材质质量、色彩质量、氛围、真实感和空间。随后，我们招募了 15 名具有游戏经验或 CG 专业背景的参与者来构建数据集。评估前，每位参与者都接受了相关培训，以确保评分量表一致，并对描述形成统一的视角。标注过程中，我们要求参与者至少从三个显著维度描述计算机图形质量，并根据描述给出总体质量结论。本文讨论的 CG图像数据集包含使用多种渲染引擎和技术生成的高分辨率图像。该数据集共包含 3.5K 张风格多样的 CG图像，包括中世纪真实风格、现代真实风格、黑暗写实风格、奇幻风格和卡通等。这些 CG图像收集自 Wallpaper Engine、一些高清游戏 CG 截图（《战舰世界》、《僵尸世界大战》、《坦克世界》、《艾尔登法环》等）、CGIQA-6K [45]（MOS 评分大于 3.2 的图像，以及从评分低于 3.2 的图像中选出的一部分，作为低质量 CG 候选图像），以及若干从网上购买的计算机图形资源包。图像分辨率从 1080p 到 4K 不等。为确保标注质量，我们将图像分成若干子集并分发给标注者，使每位标注者都能看到内容多样的图像。

图4展示了我们数据集中的代表性样本及其统计特征。面板 (a) 展示了两张带标注的计算机生成图像示例及其详细的文本描述。每条描述中以红色突出显示的短语对应于关注不同感知维度的细粒度属性，例如光照效果、

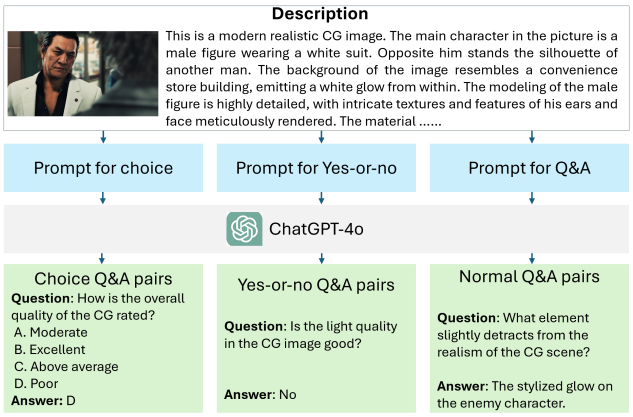

图5. GPT-4o用于为验证集和测试集生成三类问题。

色温、反射、纹理或材质逼真度；这些片段强调了所描述的具体方面。以蓝色突出显示的短语概括了整体质量评估，指出图像所感知到的逼真度或审美判断（例如“这张 CG图像质量不错”或“质量一般”）。面板 (b) 展示了描述长度的分布；大多数描述超过，000 个字符，表明多数条目包含丰富的细节。面板 (c) 通过词云展示出现频率最高的描述性术语，其中纹理、光线和逼真等高频词凸显了该数据集对视觉感知属性的高度关注。面板 (d) 列出了出现频率最高的 30 个词及其出现次数，进一步量化了语料中的主要词汇模式。

为便于使用我们的数据集，我们将其分为三部分：基础集、验证集和测试集。基础集包含3,190 张CG图像，可用于训练、微调和数据库构建。验证集和测试集分别包含90 和220 张图像。对于验证集和测试集中的每张图像，我们使用GPT-4o生成三类问题：选择题、是非题和常规问答，如图5所示，且每类至少包含五个问题。因此，验证集和测试集合计包含超过5K 个问答对。这些足以作为评估各种视觉语言模型的基准测试。

4. 我们的R4-CGQA

4.1. 基于贝叶斯理论的框架

我们考虑采用一种基于检索的方法来增强视觉语言模型的CGQA能力。与传统的微调相比，这种方法不仅性能开销更低，在处理未见过的新CG图像时也更加灵活。我们尝试挖掘大模型在回答与CG相关的问题方面的潜力：通过参考相似CG图像的信息来作答，这一过程可以用贝叶斯形式表示。令 x 表示一张查询CG图像， q 表示关于该图像的自然语言问题（例如：“这张CG图像的质量如何？”）， a表示视觉语言模型生成的答案。假设我们可以访问一个CG库 $\mathcal{D} = \{(x_i,\ t_i)\}_{i=1}^{N}$，其中 x_i 是一张库图像， t_i 是与之关联的人工撰写的描述（例如质量评论或场景说明）。在推理时，我们选择一个库索引 $I \in \{1, \ldots, N\}$，将 t_I 用作示例描述，并为视觉语言模型构建一个检索增强提示Prompt$(x,\ q,\ t_I)$。

给定参数为 ϕ的固定视觉语言模型，答案分布为

$$a \sim p_\phi\big(a \mid x, q, \mathrm{Prompt}(x, q, t_I)\big). \tag{1}$$

如果我们有一个任务特定效用函数 $U(a)$（例如，结合多项选择准确率和文本质量分数），理想的检索策略将最大化期望效用

$$I^\star = \arg\max_{i \in \{1,\ldots,N\}} \mathbb{E}_{a \sim p_\phi(\cdot \mid x, q, \mathrm{Prompt}(x,q,t_i))}\big[U(a)\big]. \tag{2}$$

然而，直接优化（2）是难以处理的。因此，我们采用贝叶斯检索视角，通过最大化代理后验来近似这一目标

$$P(I = i \mid x, q) \propto P(z(x) \mid I = i)\, \pi_i, \tag{3}$$

其中， $z(x)$ 表示从一个或多个嵌入空间中提取的查询潜在表示，而 π_i =$P(I = i)$ 是库索引上的先验。检索规则变为最大后验（MAP）估计量

$$I^\star = \arg\max_i \log P\big(z(x) \mid I = i\big) + \log \pi_i. \tag{4}$$

实际上，我们没有精确的生成模型 $P(z(x) \mid I = i)$。相反，我们假设 $z(x)$ 由 M 个不同的视图 $z^{(1)}(x) \ldots,$ $z^{(M)}(x)$ 组成（例如，内容和质量嵌入），并定义查询与库图像之间特定于视图的相似度 $s_m(x, x_i)$。一种简单且广泛使用的近似方法，是将每个视图的似然建模为相似度的指数族，

$$P\big(z^{(m)}(x) \mid I = i\big) \propto \exp\big(\beta_m s_m(x, x_i)\big), \tag{5}$$

其中权重 β_m 为非负数， $m = 1$、.....、 M。假设各视图之间具有条件独立性，则对数后验概率可表示为

$$\log P(I = i \mid x, q) = c + \sum_{m=1}^{M} \beta_m s_m(x, x_i) + \log \pi_i, \tag{6}$$

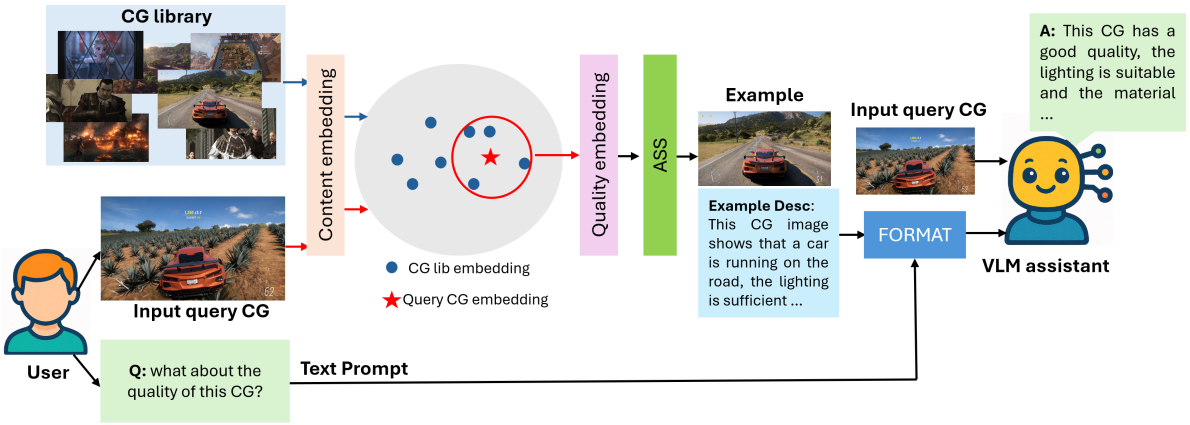

图 6. R4-CGQA 概览。查询CG图像和CG库被嵌入内容空间，并在查询周围形成一个局部邻域。在该邻域内计算并组合质量嵌入，以选出最相似的示例；该示例的文本描述将作为额外上下文，与输入CG一同用于查询VLM助手，从而生成最终的质量判断和解释。CG库由我们的**基础集**构建，并且具有可扩展性；ASS表示平均相似度搜索。

其中 c 表示常数，MAP规则（4）可简化为加权相似度融合：

$$I^\star = \arg\max_i \left[\sum_{m=1}^{M} \beta_m s_m(x, x_i) + \log \pi_i \right]. \quad (7)$$

我们接下来介绍的具体系统，是这一通用MAP检索规则的一种实例化，包含两个视图（$M = 2$）：图像内容和质量。我们采用均匀先验，因此 $\log \pi_i$ 成为常数，不会影响该方程。

4.2. 基于内容和质量的检索

我们先前的工作只考虑使用内容相似度检索，但内容相同的图像在质量上可能存在很大差异。具体而言，CLIP [24] 对图像退化并不敏感。如果“相似”图像之间的质量差异显著，将它们输入视觉语言模型可能会造成误导。因此，基于公式（7），我们为每幅图像结合基于CLIP的内容嵌入 [24] 和基于REIQA的质量嵌入 [25]，提出基于内容和质量的检索。如图6所示，对于任意图像 z，令 $f_c(z) \in \mathbb{R}^{d_c}$ 和 $f_q(z) \in \mathbb{R}^{d_q}$ 分别表示其内容嵌入和质量嵌入。我们使用 ℓ_2 归一化后的嵌入

$$\hat{f}_c(z) = \frac{f_c(z)}{\|f_c(z)\|_2}, \qquad \hat{f}_q(z) = \frac{f_q(z)}{\|f_q(z)\|_2}. \quad (8)$$

阶段1：基于 CLIP 的内容检索。对于查询图像 x 和每张库图像 x_i，我们在 CLIP空间中计算余弦相似度，

$$s_c(x, x_i) = \hat{f}_c(x)^\top \hat{f}_c(x_i). \quad (9)$$

为实现高效的全局搜索，所有 $f_c(x_i)$ 都由 FAISS [11] 建立索引。在推理时，我们会检索一小组候选最近邻 K

$$\mathcal{S}_K(x) = \underset{i \in \{1,\dots,N\}}{\text{TopK}} \; s_c(x, x_i), \quad (10)$$

对应于图中查询嵌入周围的红色圆圈6。

阶段2：质量嵌入与相似度融合。对于每个候选 $i \in \mathcal{S}_K(x)$，我们使用质量感知ResNet骨干网络计算质量嵌入 $f_q(x_i)$ 及相应的余弦相似度

$$s_q(x, x_i) = \hat{f}_q(x)^\top \hat{f}_q(x_i). \quad (11)$$

然后，我们通过简单平均融合内容相似度和质量相似度，

$$S(x, x_i) = \frac{1}{2} s_c(x, x_i) + \frac{1}{2} s_q(x, x_i), \qquad i \in \mathcal{S}_K(x). \quad (12)$$

选择融合得分最高的索引：

$$I^\star(x) = \arg\max_{i \in \mathcal{S}_K(x)} S(x, x_i). \quad (13)$$

实际上，我们还会应用相似度阈值 τ_{sim}：如果 $\max_{i \in \mathcal{S}_K(x)} S(x, x_i) < \tau_{\text{sim}}$，则不会为该查询使用示例描述，视觉语言模型仅使用查询图像和问题进行提示。图中展示了检索结果和答案的示例。7。

提示构建与VLM推理。选定示例索引 $I^\star$ 后，我们获取对应的描述 $t_{I^\star}$，并构建检索增强提示：

表 1. 在我们的测试集上评估最先进的视觉语言模型，“原始结果”部分表示直接使用视觉语言模型回答问题的结果，而 R4-CGQA 表示我们的方法所得的结果。分数中的红色表示提升。

VLMs	原始结果			R4-CGQA			↑平均值
	选择题	是非题	Q&A	选择题	是非题	Q&A	
LLava-1.6-7B	51。70%	49.73%	2.20	58.06% (+6.36%)	58.84% (+9.11%)	2.50 (+6.00%)	7.16%
LLava-1.6-13B	53。96%	51.34%	2.22	61.43% (+7.47%)	60.37% (+9.03%)	2.56 (+6.80%)	7.77%
Llama 3.2-V-11B	64。59%	56。87%	1.93	67.28% (+2.69%)	67.26% (+10.39%)	2.31 (+7.60%)	6.89%
Gemma3-4B	66.03%	53。55%	1.05	66.82% (+0.79%)	65。22% (+11.67%)	2.32 (+25.40%)	12.62%
MiniCPM-V-8B	60.05%	53.47%	1.92	67.63% (+7.58%)	61.98% (+8.51%)	2.34 (+8.40%)	8.16%
Bakllava-7B	43.72%	52.85%	1.67	55.97% (+12.25%)	61.17% (+8.32%)	1.96 (+5.80%)	8.79%
Pixtral-12B	70.11%	66.20%	2.78	73.12% (+3.01%)	69.10% (+2.90%)	2.89 (+2.20%)	2.70%
LlavaNext-8B	69。41%	61.92%	2.68	72.74% (+3.33%)	67。43% (+5.51%)	2.72 (+0.80%)	3.21%
LlavaNext-32B	75.00%	63.21%	2.83	75.84% (+0.84%)	68.72% (+5.51%)	2.90 (+1.40%)	2.58%
Qwen 2.5-VL-7B	66.73%	61.21%	1.30	67.76% (+1.03%)	63.87% (+2.66%)	2.46 (+23.20%)	8.96%
Qwen 2.5-VL-32B	77。71%	67.50%	2.79	79。21% (+1.50%)	70.24% (+2.74%)	2.87 (+1.60%)	1.95%

Prompt(x, q, t_{I^*}) = 格式模板(q, t_{I^*})，其中 FORMAT(·) 是一个固定模板，会先呈现示例描述，然后要求模型评价查询CG图像的质量。查询图像 x 和文本提示将输入 VLM助手，后者会输出标量质量判断（例如“非常好”）和自由形式的解释。

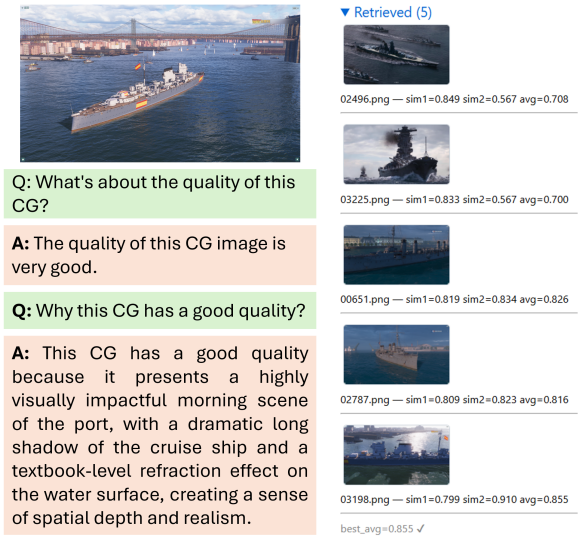

图7. R4-CGQA 系统的定性示例。左侧面板显示查询CG；右侧面板显示检索到的排名前5的图像及其 CLIP相似度（sim1）、基于REIQA的相似度（sim2）和平均分（avg）。底部文本展示了视觉语言模型对两个问题的回答，结合了标量质量评判和详细的自然语言解释。

5. 实验

5.1. 实验设置

我们在验证集上评估 R4-CGQA，以进行消融研究；并在测试集上评估当前最先进的视觉语言模型。基础集用作检索库。对于多项选择（选择题）和是非题，我们直接根据真实答案计算准确率。对于常规问答（问答），我们使用 GPT-4o-mini 在特定模板中将视觉语言模型的输出与真实答案进行比较并评分。一些视觉语言模型（LLaVA、Llama 3.2-V、Bakllava、Gemma3 和 MiniCPM-V）通过广受欢迎的 Ollama 工具进行访问 [22], ，而其他模型则使用 PyTorch 在 NVIDIA A800 80 GB GPU 上进行评估。我们用 10 表示公式（ K 中的最近邻数量，并用 T表示相似度阈值。更多结果和详细信息见补充材料。

5.2. 视觉语言模型评估

如表1所示，我们在视觉语言模型上评估了我们的方法，涉及 LLava1。6 [18], Llama3.2-Vision [2],Gemma3 [12], Mini-CPM [40], Bakllava [27], Pixtral [1],LlavaNext [15, 19], 以及 Qwen 2.5-VL[4], 在 R4-CGQA 的三类问题上：选择题、是非题和问答。R4-CGQA 在所有报告的指标上都能持续提升每个模型的表现。在选择题上，平均增益为 4.26 个百分点，其中 Bakllava 的得分从 43.72% 提升至 55.97%（+12.25 %），Mini-cpm 和 LLava-1.6-13B 分别提升了 +7.58% 和 +7.47%。对于是非题，平均提升幅度增至 6.94%（绝对值），Gemma3-4B 和 Llama 3.2-Vision 等模型分别提升了 11.67% 和 10.39%。在具有问答得分的九个模型中，R4-CGQA 在五分制下平均提升 0.32%（满分的 6.40%），其中 Gemma3-4B 从 1.05% 提升至 2.32%。结果表明，我们的方法更能发掘 CGQA 中相对较小模型的潜力。结果还表明，R4-CGQA 能为

表 2. 基于内容和质量的检索的消融研究。“w/o.”表示没有。

配置	Llava-1.6-7B 选择题	是非问答	选择题	Llama3.2-V-11B 是非问答		
Base	50.1%	48.8%	2.23	65.3%	55.8%	1.94
不使用质量	56.8%	59.0%	2.31	61.0%	68.3%	2.21
无内容	57.0%	**60.4%**	2.29	65.2%	68.9%	2.30
完整	**59.8%**	59.9%	**2.38**	**66.7%**	**69.0%**	**2.43**

表 3 12 $T=$ 。探索 Pixtral- B 上的多图像性能；以及。 0.9 $K=7$

问题	Base	多图像（仅）	多图像 + R4-CGQA	R4-CGQA
选择题	70。8%	68。5%	71。0%	73.5%
是非题	68.3%	71。6%	71。1%	71。7%

表 4. 探究不同 K 值对 Llava-1.6-7B 回答性能的影响； T 设为 0.8。

问题	$K=1$	$K=3$	$K=5$	$K=7$	$K=9$
选择题	57。1%	59.0%	59.8%	58.7%	57.2%
是非题	57.5%	58.0%	59.9%	56。5%	57.0%

较弱和较强的视觉语言模型都带来稳定且显著的提升。

5.3. 消融研究

探究内容与质量检索。 如表 2 所示，“不使用质量”表示在相似度搜索过程中移除质量嵌入模块；“不使用质量”表示使用 REIQA 而非 CLIP 进行内容嵌入，同时移除质量嵌入模块。对于 Llava-7B，所提出的完整流程进一步将选择题准确率提升至 59.8% ，同时保持较高的是非题准确率（59.9%），相较于基线模型分别绝对提升 +9.7% 和+11.1%。对于 Llama3.2-V，也呈现出相同趋势：仅质量检索在是非题上已表现出色，但完整框架仍取得最佳总体结果（选择题 / 是非题分别为 66.7% / 69.0% ），是非题准确率大幅提升 +13.2% 。此外，所有完整流程的 $Q\&A$ 分数均为最高。这些结果证实，基于内容和质量的检索比单独使用任一分支更加稳健。

探讨多图像输入。 一个自然的问题是，为什么我们不直接同时向视觉语言模型输入多张CG图像。总体而言，传统视觉语言模型无法对多张图像进行稳健的比较分析，因此输入多张图像往往会导致性能显著下降。尽管如此，我们也进行了

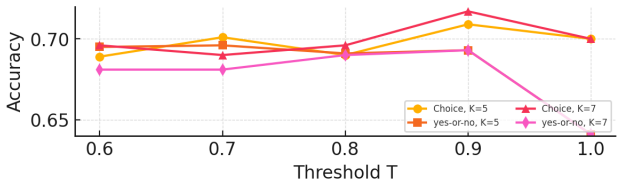

图 8. 在 LlavaNext-8B 上测试了阈值 T 和回答准确率的影响，并考察了 $K=5$ 和 $K=7$。

一项针对 Pixtral 的实验。Pixtral 能够处理多张图像，如表3所示。当我们将相似图像与当前图像一同输入时（仅多图像输入），选择题指标的表现下降了2.3%。即使将多张图像与我们的方法结合使用（多图像输入 + R4-CGQA），选择题表现仍比单独使用 R4-CGQA 2.5% 低。结果表明，直接向视觉语言模型输入多张图像并不是解决 CGQA 问题的合理方案。

考察 K。 表 4 展示了候选集大小 K 对 Llava-7B 的影响（相似度阈值固定） $T=0.8$。随着 K 从 1增至 5，准确率有所提升（最高达到59。8%/ 59。9%，分别对应选择题/是非题），但当 K 进一步增大时，准确率再次下降，这表明适量的邻居优于单个示例或过大且含噪声的候选集。

考察 T。 图 8 展示了阈值 T 对 LlavaNext-8B 的影响，以及 $K\in\{5,7\}$：当 T 介于 0。7 到0。9之间时，准确率相对稳定；配置 $K=5$ 始终比$K=7$更加稳定。当 $T=1.0$时，即不选择相似的图像描述作为示例，是非题上的性能会大幅下降。

6. 结论

本文针对当前计算机图形质量评估的局限性，提出了一个新的数据集，其中包含 3.5K 张计算机图形图像，并为其提供涵盖六个感知维度的详细基于文本的质量描述。在此资源基础上，我们进一步提出了一种检索增强的双流框架，通过利用内容相似度和质量相似度，在推理过程中向视觉语言模型提供相关的示例描述，该框架名为 R4-CGQA。在多个具有代表性的视觉语言模型上开展的大量实验表明，检索到的文本线索显著提升了它们生成准确且可解释的计算机图形质量判断的能力。我们的工作为在 CGQA 中增强视觉语言模型提供了一种可扩展且无需训练的解决方案，并将有力支持该方向未来的研究。

参考文献

[1] Pravesh Agrawal、Saurabh Garg 等。Pixtral 12b。arXiv 预印本 arXiv:2410.07073，2024 年。7[2] Meta AI。Llama 3.2 视觉：用于图像+文本理解的多模态基础模型，2024 年。7[3] Memoona Aziz、Umair Rehman、Muhammad Umair Danish 和 Katarina Grolinger。全局-局部图像感知分数（GLIPS）：评估 AI 生成图像的照片级真实感质量。IEEE 人机系统学报，55：233，2025 223 3：–。[4] Shuai Bai、Keqin Chen、Xuejing Liu 等。Qwen2.5-VL 技术报告。CoRR，abs/2502.13923，2025 年。3、7[5] Sebastian Bosse、Dominique Maniry、Klaus-Robert M¨uller、Thomas Wiegand 和 Wojciech Samek。用于无参考和全参考图像质量评估的深度神经网络。IEEE 图像处理学报，27：219，2018 1 206 2 ( )：–。[6] Yue Cao、Zhaolin Wan、Dongwei Ren、Zifei Yan 和 Wangmeng Zuo。结合半监督学习与正类-未标记学习以提升全参考图像质量评估。在 CVPR 上，页码 5841–5851。IEEE，2022 年。3[7] Wu CH、Chien YC、Chou MT 和 Huang YM。整合计算思维、游戏设计与设计思维：关于教育趋势、应用及影响的范围综述。《人文与社会科学》，12(1)：1–12，2025 年。1[8] Zhihao Chen、Bin Hu、Chuang Niu、Tao Chen、Yuxin Li、Hongming Shan 和 Ge Wang。IQAGPT：利用视觉语言模型和 ChatGPT 模型进行计算机断层扫描图像质量评估。《视觉计算、工业与生物医学艺术》，7(1)：20，2024 年。2、3[9] Manri Cheon、Sung-Jun Yoon、Byungyeon Kang 和 Jun-woo Lee。利用 Transformer 进行感知图像质量评估。在 CVPR 上，页码 433–442。Computer Vision Foundation / IEEE，2021 年。3[10] Y Han。基于计算机技术的 CG动画场景技术研究。《计算机科学进展》，(247)：1197–1206，2024 年。1[11] Johnson、Jeff、Douze、Matthijs 和 J´egou、Herv´e。Faiss：用于稠密向量高效相似度搜索和聚类的库。arXiv 预印本 arXiv:1702.08734，2017 年。6[12]Aishwarya Kamath、Slav Petrov、Noah Fiedel、Noam Shazeer、Oriol Vinyals、Jeff Dean、Demis Hassabis、Ko-ray Kavukcuoglu、Clement Farabet、Elena Buchatskaya、Jean-Baptiste Alayrac、Rohan Anil、Dmitry（“Dima”）Lepikhin、Sebastian Borgeaud、Olivier Bachem、Armand Joulin、Alek Andreev、Cassidy Hardin、Robert Dadashi、L´eonard Hussenot 等，以及 Gemma 团队。Gemma 3 技术报告。arXiv 预印本 arXiv:2503.19786，2025 年。7[13] Junjie Ke、Qifei Wang、Yilin Wang、Peyman Milanfar 和 Feng Yang。MUSIQ：多尺度图像质量 Transformer。在 ICCV 上，页码 5128–5137。IEEE，2021 年。3[14] Jongyoo Kim 和 Sanghoon Lee。全深度盲图像质量预测器。IEEE 精选信号处理主题学报，11(1)：206–220，2017 年。2[15] Bo Li、Yuanhan Zhang、Dong Guo、Renrui Zhang、Feng Li、Hao Zhang、Kaichen Zhang、Peiyuan Zhang、Yanwei Li、Ziwei Liu、Chunyuan Li。Llava-onevision：轻松迁移视觉任务。arXiv预印本 arXiv:2408.03326，2024年。7[16] Chunyi Li、Tengchuan Kou、Yixuan Gao、Yuqin Cao、Wei Sun、Zicheng Zhang、Yingjie Zhou、Zhichao Zhang、Weixia Zhang、Haoning Wu、Xiaohong Liu、Xiongkuo Min 和 Guangtao Zhai。Aigiqa-20k：用于 AI 生成图像质量评估的大型数据库。载于 CVPR Workshops，第 6327–6336 页。IEEE，2024年。3[17] Junnan Li、Dongxu Li、Silvio Savarese 和 Steven C. H. Hoi。BLIP-2：利用冻结的图像编码器和大型语言模型引导语言-图像预训练。载于 ICML，第 19730–19742 页。PMLR，2023年。3[18] Haotian Liu、Chunyuan Li、Qingyang Wu 和 Yong Jae Lee。视觉指令微调。载于神经信息处理系统大会，2023年。2，7[19] Haotian Liu、Chunyuan Li、Yuheng Li 和 Yong Jae Lee。Llava-next：改进推理、OCR 和世界知识。arXiv预印本或官方博客，2024年。7[20] Yiting Lu、Xin Li、Haoning Wu、Bingchen Li、Weisi Lin 和 Zhibo Chen。Q-adapt：通过渐进式指令微调调整 LMM，以进行视觉质量评估。arXiv预印本 arXiv:2504.01655，2025年。2，3[21] Kede Ma、Wentao Liu、Kai Zhang、Zhengfang Duanmu、Zhou Wang 和 Wangmeng Zuo。使用深度神经网络进行端到端盲图像质量评估。IEEE 图像处理汇刊，27(3):1202–1213，2018年。2[22] Ollama Inc. Ollama：用于运行大型语言模型的本地优先框架，2025年。7[23] OpenAI。GPT-4 技术报告。arXiv预印本 arXiv:2303.08774，2023年。3[24]Alec Radford、Jong Wook Kim、Chris Hallacy、Aditya Ramesh、Gabriel Goh、Sandhini Agarwal、Girish Sastry、Amanda Askell、Pamela Mishkin、Jack Clark、Gretchen Krueger 和 Ilya Sutskever。从自然语言监督中学习可迁移的视觉模型。载于 ICML，2021年。6[25] Avinab Saha、Sandeep Mishra 和 Alan C. Bovik。Re-iqa：在真实场景中进行图像质量评估的无监督学习。载于 IEEE/CVF 计算机视觉与模式识别会议（CVPR）论文集，第 5846–5855 页，2023年。6[26] Hamid R. Sheikh 和 Alan C. Bovik。图像信息与视觉质量。IEEE 图像处理汇刊，15(3):444–430，2006年。2[27] SkunkworksAI、LAION 和 Ontocord。Bakllava。https: // github.com/SkunkworksAI/BakLLaVA，2023年。开源模型发布。7[28] Shaolin Su、Qingsen Yan、Yu Zhu、Cheng Zhang、Xin Ge、Jinqiu Sun 和 Yanning Zhang。在自适应超网络的引导下评估真实场景中的图像质量。载于 CVPR，第 3664–3673 页，2020年。2[29] L Sun。三维动画特效在动画电影中的应用研究：以电影《阿凡达》为例。《科学编程》，2022(1):1928660，2022年。1[30] Yu Tian、Yixuan Li、Baoliang Chen、Hanwei Zhu、Shiqi Wang 和 Sam Kwong。视觉传播中的 AI 生成图像质量评估。载于 AAAI，第 7392–7400 页，2025年。2，3 – 。

[31] Domonkos Varga。用于无参考图像质量评估的多池化 Inception 特征。《应用科学》，10(2):2186，2020年。6[32] Jianyi Wang、Kelvin C. K. Chan 和 Chen Change Loy。探索 CLIP 在评估图像观感方面的应用。载于 AAAI，第 2555–2563 页。AAAI出版社，2023年。3[33] Jiarui Wang、Huiyu Duan、Guangtao Zhai 和 Xiongkuo Min。通过指令微调理解并评估人类对 AI 生成图像的偏好，2024年。3[34] Puyi Wang、Wei Sun、Zicheng Zhang、Jun Jia、Yanwei Jiang、Zhichao Zhang、Xiongkuo Min 和 Guangtao Zhai。大型多模态模型辅助的 AI 生成图像质量评估。载于 ACM MM，第 7803–7812 页。ACM，2024年。2，3[35] Tao Wang、Zicheng Zhang、Wei Sun、Xiongkuo Min、Wei Lu 和 Guangtao Zhai。计算机图形学生成图像的主观质量评估。载于第 24 届 IEEE 多媒体信号处理国际研讨会（MMSP 2022），第 1–5 页。IEEE，2022 年。1，4[36] Zhou Wang 和 Eero P. Simoncelli。基于小波域自然图像统计模型的简化参考图像质量评估。载于《人类视觉与电子成像 X》，美国加利福尼亚州圣何塞，2005年1月17日，159，2005年，第 149–2 页。[37] Zhou Wang、Alan C. Bovik、Hamid R. Sheikh 和 Eero P. Simoncelli。图像质量评估：从误差可见性到结构相似性。IEEE 图像处理汇刊，13(4):600–612，2004年。2[38] Haoning Wu、Zicheng Zhang、Erli Zhang、Chaofeng Chen、Liang Liao、Annan Wang、Chunyi Li、Wenxiu Sun、Qiong Yan、Guangtao Zhai 和 Weisi Lin。Q-bench：面向通用基础模型低层视觉能力的基准测试。载于 ICLR，2024年。2，3[39] Haoning Wu、Zicheng Zhang、Erli Zhang、Chaofeng Chen、Liang Liao、Annan Wang、Kaixin Xu、Chunyi Li、Jingwen Hou、Guangtao Zhai、Geng Xue、Wenxiu Sun、Qiong Yan 和 Weisi Lin。Q-instruct：提升多模态基础模型的低层视觉能力。载于 CVPR，第 25490–25500 页。IEEE，2024年。2，3[40] Yuan Yao、Tianyu Yu、Ao Zhang、Chongyi Wang、Junbo Cui、Hongji Zhu、Tianchi Cai、Haoyu Li、Weilin Zhao、Zhihui He、Qianyu Chen、Huarong Zhou、Zhensheng Zou、Haoye Zhang、Shengding Hu、Zhi Zheng、Jie Zhou、Jie Cai、Xu Han、Guoyang Zeng、Dahai Li、Zhiyuan Liu 和 Maosong Sun。MiniCPM-V：媲美 GPT-4V 的手机端多模态大语言模型。arXiv预印本 arXiv:2408.01800，2024年。7[41] Junyong You 和 Jari Korhonen。用于图像质量评估的 Transformer。载于 ICIP，第 1389–1393 页。IEEE，2020年。3[42] Junyong You 和 Jari Korhonen。用于图像质量评估的 Transformer。载于 ICIP，第 1389–1393 页。IEEE，2021年。2[43] Zhiyuan You、Jinjin Gu、Zheyuan Li、Xin Cai、Kaiwen Zhu、Tianfan Xue 和 Chao Dong。在真实场景中进行描述性图像质量评估。CoRR，abs/2405.18842，2024年。2[44] Richard Zhang、Phillip Isola、Alexei A. Efros、Eli Shechtman 和 Oliver Wang。……的惊人效力

将深度特征用作感知度量。在 *CVPR*，第 586–595 页。计算机视觉基金会 / IEEE计算机学会，2018年。2[45] Zicheng Zhang、Wei Sun、Yingjie Zhou、Jun Jia、Zhichao Zhang、Jing Liu、Xiongkuo Min 和 Guangtao Zhai。《野外计算机图形图像的主观与客观质量评估》。*ACM* 多媒体计算与通信应用汇刊，20(4):96:1–96:22，2024年。1，4[46] Zicheng Zhang、Haoning Wu、Zhongpeng Ji、Chunyi Li、Erli Zhang、Wei Sun、Xiaohong Liu、Xiongkuo Min、Fengyu Sun、Shangling Jui、Weisi Lin 和 Guangtao Zhai。《Q-boost：论低层多模态基础模型的视觉质量评估能力》。见 *IEEE*国际多媒体与博览会会议，*ICME 2024*——研讨会，尼亚加拉瀑布城，安大略省，加拿大，*2024*年*7*月*15*日至*19*日，第 1–6 页，2024年。2，3[47] Hanwei Zhu、Haoning Wu、Yixuan Li、Zicheng Zhang、Baoliang Chen、Lingyu Zhu、Yuming Fang、Guangtao Zhai、Weisi Lin 和 Shiqi Wang。《通过教大型多模态模型进行比较，实现自适应图像质量评估》。见 神经信息处理系统大会，第 32611–32629 页，2024年。2，3