视觉语言模型能否评估平面设计美学？基准、评估与数据集视角

Arctanx An[1,2*] **Shizhao Sun**[2†] **Danqing Huang**[3] **Mingxi Cheng**[3] **Yan Gao**[4] **Ji Li**[3] **Yu Qiao**[4] **Jiang Bian**[2]
[1]北京大学 [2]微软亚洲研究院 [3]微软 [4]中南大学

arXiv:2603.01083v1 [cs.CV] 1 Mar 2026

摘要

评估平面设计的审美质量是视觉传达的核心，但在视觉语言模型（VLMs）中仍未得到充分探索。我们研究视觉语言模型能否以与人类相当的方式评估设计美学。先前工作面临三个关键局限性：基准测试局限于狭窄原则和粗粒度评估协议、缺乏系统性的视觉语言模型比较，以及用于模型改进的训练数据有限。在本工作中，我们提出 AesEval-Bench，这是一个涵盖四个维度、十二项指标和三项完全可量化任务的综合基准：审美判断、区域选择和精确定位。随后，我们系统评估了专有、开源以及推理增强型视觉语言模型，揭示了它们与审美评估细微要求之间的明显性能差距。此外，我们构建了一个训练数据集，以针对该领域微调视觉语言模型，利用人工引导的视觉语言模型标注大规模生成任务标签，并借助基于指标的推理将抽象指标与具体设计区域联系起来。总体而言，我们的工作建立了平面设计审美质量评估的首个系统性框架。我们的代码和数据集将发布于：https://github.com/arctanxarc/AesEval-Bench。

1 引言

视觉语言模型（VLMs）（Yang 等，2025；Li 等，2024a；Hurst 等，2024）的快速发展，为理解和处理多模态信息开辟了新的机遇。尽管它们已在图像描述（Lin 等，2024b；Li 等，2025b）和视觉问答（An 等，2024；Zhang 等，2024；Lin 等，2025）等传统视觉任务中取得显著进展，但人们日益期待视觉语言模型能够为新的应用领域作出贡献。其中，平面设计——通过融合文本与视觉元素，在广告、品牌塑造和数字媒体中传达信息——因其广泛的社会和实践影响而成为一个充满希望的方向。至关重要的是，平面设计的成功在很大程度上取决于其审美质量，而审美质量受平衡、对比度和层级等原则影响。
在这项工作中，我们旨在探讨一个核心问题：视觉语言模型能否像人类一样理解并评估平面设计的审美质量？这一问题至少在三个方面具有重要意义。首先，它可以帮助人类设计师找出设计的不足之处并解释原因，从而更有效地加以改进。其次，对于生成式人工智能系统而言，这为自动反馈循环奠定了基础，使其能够在无需大量人工干预的情况下指导迭代优化。第三，这也意味着视觉语言模型有望从事实识别拓展至审美评估。

尽管这一方向十分重要，相关研究仍然有限（见表1）。首先，基准测试并不充分。为自然照片开发的基准（Zhou 等，2024；Huang 等，2024）往往忽略字体排印等设计特有因素，而早期针对平面设计的基准通常仅涵盖一小部分设计原则（Lin 等，2023；2024a；Jiang & Chen，2025）。此外，

*这项工作是在微软亚洲研究院实习期间完成的
†项目负责人

现有评估方案也存在局限。基于评分的方法无法指出美学问题出现在哪里（Haraguchi 等，2024），而基于描述的方法虽然能提供定性反馈，却难以量化（Jung 等，2025）。其次，缺少视觉语言模型之间的比较。在设计美学领域，尚未针对不同的视觉语言模型（无论是开源还是闭源）开展系统评估。第三，训练数据集匮乏，如何进一步提升视觉语言模型在这一领域的表现，仍未得到充分探讨。

为初步解决这些局限性，我们推出*AesEval-Bench*，这是一个用于评估平面设计审美质量的新基准（见图 1）。借鉴既有文献（Li 与 Chen，2009；Wangwiwattana 与 Meeyen，2024），我们确定了四个关键维度——字体排印、布局、颜色和图形，它们共同全面涵盖了设计美学。这些维度反映了人们评估视觉吸引力时一贯重视的主要因素。为实现更精细的分析，我们进一步定义了十二个指标，用于明确各维度中的具体方面，例如字体排印中的层级和易读性。针对每个指标，我们设计了三项具有挑战性的任务：1）审美判断，要求模型判断一项设计是否具有审美吸引力（是/否），以直观地衡量整体感知；2）区域选择，要求模型从候选区域中选出不悦目元素所在的位置，以检验模型能否超越整体判断、准确定位问题区域；3）精确定位，要求模型预测不悦目区域的精确边界框坐标，从而提供最详尽的诊断，并反映模型对美学更深入的理解。与以往的基准测试（）不同，AesEval-Bench 不仅通过其维度和指标涵盖了广泛的美学因素，还定义了可通过选择或边界框预测格式进行完全量化的评估任务，从而能够系统、可复现地评估设计美学。

借助 AesEval-Bench，我们系统评估了视觉语言模型评估平面设计审美质量的能力（见表 1）。对于每项设计，模型都要围绕十二项指标完成三项任务。我们以人工标注作为真实标注，并通过准确率（用于审美判断和区域选择）和边界框交并比（用于精确定位）衡量模型表现。总体而言，结果显示当前最先进的视觉语言模型与审美质量评估的细致要求之间仍存在明显差距。具体来说，专有视觉语言模型（如 GPT 系列（Achiam 等，2023））优于开源模型（如 Qwen-VL（Wang 等，2024）、Intern-VL（Zhu 等，2025）和 LLaVA（Liu 等，2023））。在开源模型中，较大的版本（32B、72B）通常比小型版本（7B）表现更好。令人意外的是，推理增强型视觉语言模型（如 GPT-o1（Jaech 等，2024）、GPT-o3 和 Gemini-2.5-Pro（Comanici 等，2025））并未明显优于不具备推理能力的对应模型。这些发现共同揭示了现有视觉语言模型的局限性，并凸显了针对审美质量评估开展领域专用训练的必要性。

基于这些发现，我们不再局限于评估，转而训练视觉语言模型以进行审美质量评估。为此，我们构建了一个由三个部分组成的训练数据集：任务 （模型被要求完成的内容）、任务标签 （期望的答案）以及推理路径（导向答案的解释）（见图2）。我们将推理路径视为至关重要，因为通用推理几乎没有带来收益。然而，构建此类数据面临两个关键挑战：大规模生成任务标签成本高昂，而生成真正能够提升性能的推理路径则需要新的方法。为应对这些挑战，我们提出两种解决方案。首先，人工引导的视觉语言模型标注，即以少量人工标注作为上下文示例，指导强大的视觉语言模型生成任务标签。该方法在与人类理解保持对齐的同时，降低了人工标注成本。其次， 基于指标的推理，即将抽象指标（例如层级或分层）明确关联到设计中的具体区域。每条推理路径均由与指标相关联的边界框坐标及其相关性的文本解释组成，从而提供细粒度且可解释的监督。我们以任务作为输入，并以推理路径和任务标签作为监督来微调视觉语言模型，随后在AesEval-Bench上进行评估。结果表明，所有任务和指标均取得了一致的性能提升（例如5.97%、2.70%和17.17%），证明人工引导的视觉语言模型标注能够产生可靠标签，而基于指标的推理则提供了有效的监督。

综上所述，我们的贡献如下：

- 我们提出 AesEval-Bench，这是一个用于评估平面设计审美质量的综合基准，涵盖四个维度、十二项指标和三项可量化任务。
- 我们系统评估了专有、开源及推理增强型视觉语言模型，揭示了它们在审美质量评估方面存在明显的性能差距。

表 1：AesEval-Bench 与现有图像美学和设计美学基准测试的比较。我们重点展示它们在规模、任务格式、源数据、所涵盖的设计维度以及是否包含推理路径方面的主要差异。

基准	#数据	Task 格式	来源	来源类型	维度				训练集	推理路径	开源
					Font	布局	图形	颜色			
图像美学基准											
AesBench（Huang 等，2024）	∼10k	自由形式	摄影图像	仅图像	×	×	✓	✓	×	×	✓
UNIAA-Bench（Zhou 等，2024）	∼6k	自由形式	摄影图像	仅图像	×	✓	×	✓	×	×	✓
FineArtBench (Jiang 和 Chen，2025)	-	选择自由形式	艺术作品 + 摄影图像	仅图像	×	×	×	✓	×	✓	×
设计美学基准											
DesignBench (Lin 等，2023年)	-	选择自由形式	平面设计	图像+JSON	✓	✓	×	✓	×	×	✓
DesignProbe (Lin 等，2024a)	∼1.6k	选择	平面设计	图像+Json	✓	✓	×	✓	×	×	×
GPT-Eval 基准 (Haraguchi 等，2024)	∼2k	评分	平面设计	图像+JSON	×	✓	×	×	×	×	×
UI-Bench (Jung 等，2025)	∼3k	选择	用户界面设计	仅图像	✓	✓	×	✓	×	×	×
UICrit（Jung 等，2025）	∼3k	描述自由形式边界框回归	用户界面设计	仅图像	✓	✓	×	✓	×	×	✓
AesEval-Bench（我们的方法）	∼4.5k	选择边界框回归	平面设计	图像+JSON	✓	✓	✓	✓	✓	✓	✓

- 我们构建了一个训练数据集，用于微调视觉语言模型，使其适用于这一领域。我们的方法引入了人工引导的视觉语言模型标注，以大规模生成任务标签，并采用基于指标的推理，将抽象指标与具体设计区域联系起来。在 AesEval-Bench 上进行的实验表明，该数据集能够持续提升模型在所有任务上的表现。

2 相关研究

审美质量评估。审美质量评估（Deng 等，2017）旨在自动评估视觉吸引力，作为人类判断的计算代理。在这一领域，已形成两条主要研究路线（见表 1）。图像美学评估 （Huang 等，2024；Zhou 等，2024；Jiang 和 Chen，2025）聚焦于摄影图像，其质量取决于色彩协调、光照和主体布局等因素。设计美学评估 （Lin 等，2023；2024a；Haraguchi 等，2024；Jung 等，2025）则针对海报、广告或用户界面等平面设计，这些设计的质量取决于字体排印、层级和对齐等设计相关因素。我们的工作属于设计美学评估。

尽管其重要性不言而喻，设计美学评估仍未得到充分探索。现有基准测试仅涵盖设计维度中的一小部分。例如，（Lin 等，2024a）忽略了与图形相关的因素，而（Haraguchi 等，2024）则同时忽略了字体和图形。此外，它们的任务表述缺乏严谨性。有些采用自由形式问答，难以量化（Lin 等，2023），而另一些仅提供整体评分，却不识别存在问题的区域，从而限制了可解释性与可操作性（Jung 等，2025）。我们的工作引入了一个基准，全面覆盖字体、布局、图形和颜色等方面与设计相关的美学因素，并以选择和边界框预测的形式定义结构良好且可量化的任务。

视觉语言模型。视觉语言模型（VLMs）（Wang 等，2024；Comanici 等，2025；Li 等，2024a；Zhang 等，2025c）在图像描述（Luo 等，2025a；Li 等，2024b；Zhang 等，2025d）和视觉问答（An 等，2024；Lin 等，2024b；Luo 等，2025b）等任务上取得了出色表现。然而，它们评估平面设计审美质量的能力在很大程度上仍未得到探索。先前工作通常仅评估一到两个视觉语言模型（例如，（Haraguchi 等，2024）研究了 GPT）。我们对包括专有、开源和推理增强模型在内的广泛视觉语言模型进行了系统比较。

近年来，人们越来越关注视觉语言模型的推理能力（Li 等人，2025a；Zhang 等人，2025b）。例如，Sarch 等人（2025）采用基于树的搜索来改进推理链，Li 等人（2025a）将推理轨迹可视化以提高透明度，而 Sun 等人（2024）、Shao 等人（2024）、Cao 等人（2025）和 Wu 等人（2025）则通过联合生成边界框和文本解释来探索基于视觉依据的推理。在我们的工作中，我们发现，当前视觉语言模型中的通用推理对于评估设计美学帮助有限。为此，我们构建了一个包含推理路径的训练数据集，明确将抽象设计指标与设计的具体区域

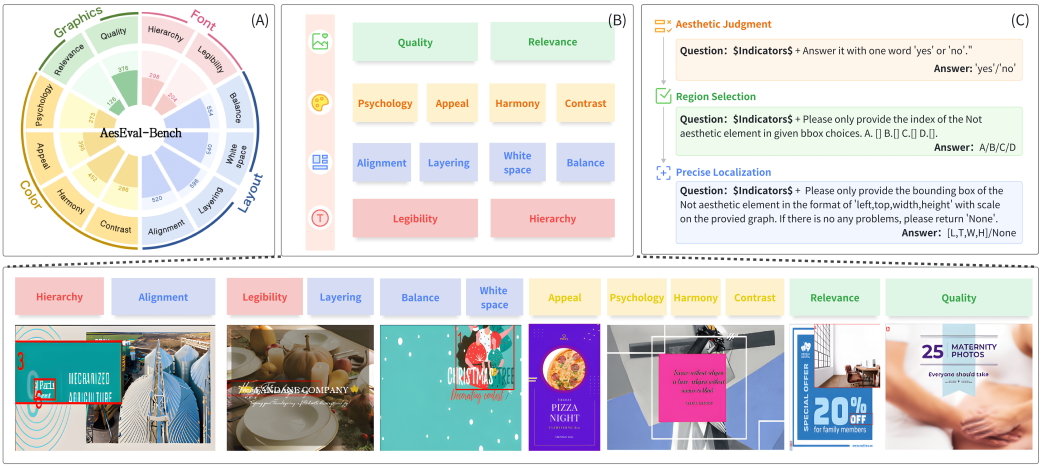


图 1：AesEval-Bench 概览。**(A)** 基准中考虑的四个维度和十二项指标。圆圈内的数字表示每个指标被标注为存在缺陷的设计数量。**(B)** 用于说明各项指标的示例设计，存在审美问题的区域以红框标出。所有指标的详细文本解释见附录。**(C)** 三项任务，以及示例问题和预期答案。

联系起来。与基于视觉依据的推理不同，后者会定位语义显著的实体（例如“狗”或“椅子”）；我们的区域以指标为中心，捕捉层级、对齐和间距等更高层次的概念，这些概念直接体现了设计原则。

3 基准构建

3.1 概述

AesEval-Bench 将设计美学评估表述为问答任务。输入包含任务 描述和设计图像，并可选择附带布局、字体或颜色信息等元数据（JSON 格式）。输出是与该任务对应的答案 。

为了涵盖设计美学评估的不同方面，我们引入了三种任务类型（图 1(C)）：1）审美判断 询问设计在美学上是否令人愉悦（是/否），从而衡量整体感知。2）区域选择 要求从出现审美问题的候选区域中进行选择，以检验在整体判断之外定位问题区域的能力。3）精确定位要求预测问题区域的精确边界框坐标，从而提供细粒度诊断。每项任务都附有一项指标的说明——即人类在评估视觉吸引力时始终强调的关键因素（例如层级、分层、对比度）。我们考虑十二项指标（图 1(B)），并将其归入四个维度（图 1(A)）。

对于设计图像，我们从 Crello 数据集（Yamaguchi，2021）的测试集划分中抽取 1200 个设计，该数据集包含专业平面设计及其设计图像和元数据。不同任务的预期答案各不相同。对于审美判断，答案为是或否。对于区域选择，答案是四个候选区域之一的索引。对于精确定位，答案是所识别区域的边界框坐标；如果设计不存在审美问题，则答案为 None。Ove总体而言，AesEval-Bench 包含 4500 个基础版问答对（在 1500 个设计上执行三项任务 - sign），并将每个任务进一步细分为十二项指标，以实现细粒度评估。

3.2 策展流程

建立维度与指标。 平面设计的审美质量本质上是多维的。为了定义严格的基准，我们首先全面回顾了经典与当代设计原则相关文献（McCormack & Lomas，2020年；Lou 等人，2022年；Lu 等人，2020年）。随后，我们咨询了专业设计师以完善这一分类体系，确保其

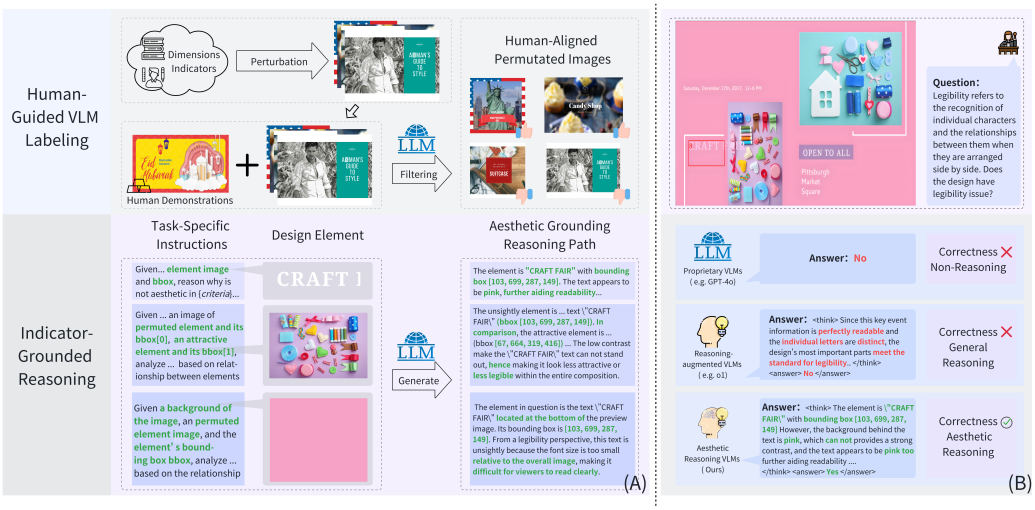


图2：**(A)** 训练数据构建中两个关键步骤的示意图。人工指导的视觉语言模型标注能够大规模判定设计是否存在审美问题。基于指标的推理会生成推理路径，明确将抽象指标与具体设计区域（以边界框坐标表示）关联起来。**(B)** 示例展示了非推理模型、通用推理模型与我们的基于指标的推理模型之间的差异。

兼顾理论基础与实践经验。经过这一过程，我们确定了四个核心维度——布局、字体、图形和颜色（图 1(A)）——每个维度又进一步细分为具体指标，这些指标在人类审美判断中始终受到重视。总之，我们归纳出十二项指标，共同涵盖了设计美学的关键因素（图 1(B)）。

构建潜在有缺陷的设计。 如第 3.1 节所述，AesEval-Bench 中的设计图像源自 Crello 数据集，该数据集包含专业品质的平面设计。为了有效评估设计美学，基准不仅必须包含设计精良的示例，也必须包含吸引力较弱的示例。因此，我们对 Crello 数据集进行改造，引入可控扰动，例如重新定位元素、更改字体选择或调整颜色。这些扰动可能降低视觉质量，也可能基本不影响视觉质量。例如，略微放大标题或许能保留层级，而将其向左移动则可能破坏平衡。每个基础版设计都会经历一至三次随机扰动，生成一系列变化，从美学上基本不变到明显有缺陷，同时仍显得真实。由于 Crello（Yamaguchi，2021）以 JSON 格式提供元素级元数据，并包含独立的设计图层，因此可以直接在 JSON 层面应用这些扰动，再将修改后的元数据与相应图层重新组合，渲染生成新的设计图像。

人在回路审美评审。 我们邀请人工标注员核验经过扰动的设计是否确实存在审美问题。标注开始前，所有标注员都会接受教程培训，其中包含设计精良和存在缺陷的案例示例，以及对其根本原因的详细说明。评审过程中，每位标注员都会看到一张设计图像及对焦点指标的描述，并需判断该设计是否存在相应缺陷（是或否）。对于每项设计，我们通过对多位标注员的结果进行多数投票，得出最终标签，以确保达成共识。

生成问答对。借助 JSON 格式的元数据、已应用扰动的记录和人工标注，我们可以系统地构建与各项任务对应的答案。对于审美判断，规则很简单：如果人工标注员将设计标记为良好，真实答案就是 no（即不存在审美问题）；否则就是 yes。对于区域选择，如果设计被标记为良好，四个答案选项由从元数据中随机抽取的三个边界框和一个 None 选项组成，真实答案为 None。如果设计被标记为有缺陷，四个选项则包括受扰动元素的边界框、从元数据中随机抽取的两个边界框，以及 None；真实答案设为受扰动元素的边界框。对于精确定位，如果设计被标记为良好，真实答案为 None；否则，真实答案就是受扰动元素的精确边界框。

3.3 评估方案

对于审美判断和区域选择，由于两者都被设定为选择问题，我们采用准确率作为指标，用于衡量模型预测与真实标注是否完全匹配。对于精确定位，任务包含两个部分：一个选择题（不存在审美问题时预测 None）和一个边界框回归问题（存在问题时预测精确边界框）。因此，当真实答案为 None 时，我们使用准确率；需要预测边界框时，则使用交并比（IoU）来量化预测边界框与真实标注边界框之间的重叠程度。

4 训练数据构建

对主流视觉语言模型的评估揭示了当前最先进视觉语言模型的能力与审美质量评估的细致要求之间存在明显差距。此外，增强推理能力的视觉语言模型并未表现出明确的性能提升（见第 5.1 节）。

为此，我们构建了一个名为 **AesEval-Train** 的训练数据集，用于微调视觉语言模型以适应这一领域。首先，我们采用与基准构建相同的流程来构建潜在有缺陷的设计（见第 3.2 节）。接下来，考虑到在大规模训练中，仅依靠人工标注来判定扰动后设计是否存在审美问题，既难以扩展，也不具成本效益，我们引入了人工引导的视觉语言模型标注。然后，我们按照基准构建流程生成问答对（见第 3.2 节）。最后，我们引入基于指标的推理，以生成特定领域的推理路径，从而提升任务表现。下文将详细介绍与基准构建不同的两个步骤。

人工引导的视觉语言模型标注。 我们将少量人工标注作为示例，并结合扰动区域的边界框坐标，输入强大的视觉语言模型。我们指示模型生成一个二元标签，以判断扰动后设计是否存在审美问题（见图2(A)）。通过纳入人工标注，我们在大幅降低人工标注成本的同时，确保与人类判断保持对齐。此外，将扰动区域作为先验信息提供（这一信息在真实场景中无法获取），能够简化标注流程并提高可靠性。借助这两种指导信息，尽管生成的标签可能并非完全准确，但其质量足以构成训练集，从而提升微调效果。

基于指标的推理。 如图2(B)所示，通用推理往往会解释或分析给定的指标和任务，却未将讨论与设计中的相关区域联系起来。为解决这一局限，我们提出将抽象指标与设计中的具体区域明确关联起来。具体而言，我们在推理路径中同时纳入相关区域的边界框（bbox）坐标，以及说明这些区域与该指标相关性的文本解释。
为了获得此类推理路径，我们通过向强大的视觉语言模型（例如我们实验中的 GPT）提供目标区域的边界框坐标以及相应的设计图层（图2(A)）来对其进行指导。模型需要输出所提供的坐标，并附上该区域与指标之间关系的解释，从而确保推理路径始终包含所需信息。我们进一步采用特定任务策略来确定感兴趣区域。对于审美判断，我们直接使用扰动区域的边界框。对于区域选择，我们同时纳入扰动区域和未扰动区域，以增强模型在候选区域之间进行区分的能力。对于精确定位，我们不仅突出扰动区域的边界框，还强调其与整体设计的关系，使模型能够在整体设计语境中更好地定位问题区域。

5 实验

5.1 在 AESEval-Bench 上对视觉语言模型进行基准测试

实验设置。我们对涵盖不同模型系列和参数规模的 10 个视觉语言模型进行了全面评估。对于非推理模型，我们选取了 LLaVA（Liu 等，2023）、Qwen2.5-VL（Bai 等，2025）和 Intern-VL3（Zhu 等，2025）等开源代表模型，以及闭源 GPT 模型（Jaech 等，2024）。对于推理增强模型，我们评估了

表 2：美学判断任务评估。总体准确率为所有指标的平均值。最佳和次佳结果分别以**粗体**和下划线标出。

模型	总体准确率	布局				颜色				Font		图形	
		平衡	分层	留白	对齐	协调	对比度	吸引力	心理学	易读性	层级	质量	相关性
非推理模型													
LLaVA-13B	0.5636	0.6506	0.5063	0.6975	0.6759	0.4411	0.2409	0.2615	0.2851	0.7660	0.8260	0.7386	0.6733
Qwen-VL-7B	0.6390	0.8272	0.4508	0.8076	0.8413	0.8223	0.9136	0.4009	0.1355	0.9430	0.3100	0.8183	0.3970
Qwen-VL-32B	0.6458	0.6762	0.5689	0.6862	0.6941	0.6209	0.5776	0.4948	0.1862	0.8023	0.7618	0.7708	0.9001
Qwen-VL-72B	0.6724	0.6752	0.6925	0.7131	0.6731	0.6921	0.7804	0.2620	0.2447	0.8739	0.7952	0.7729	0.8734
Intern-VL3-8B	0.6331	0.4617	0.6452	0.7577	0.7751	0.3487	0.3329	0.3082	0.5180	0.9486	0.6951	0.8571	0.9491
Intern-VL3-14B	0.6883	0.7406	0.3826	0.7563	0.7601	0.7706	0.8594	0.5276	0.1912	0.8309	0.8304	0.8373	0.7729
GPT-4o	0.7031	0.7588	0.6789	0.6597	0.4688	0.8129	0.8190	0.8237	0.3292	0.7506	0.7344	0.7857	0.8160
GPT-5	**0.7252**	0.8378	0.7832	0.7275	0.6510	0.6953	0.8375	0.4000	0.5237	0.7472	0.7419	0.9023	0.8551
推理增强模型													
GPT-o1	0.6705	0.7384	0.7531	0.5522	0.3398	0.7149	0.5439	0.7427	0.6750	0.7049	0.7266	0.7518	0.8030
GPT-o3	0.7105	0.7450	0.7597	0.6588	0.3964	0.7715	0.7005	0.7993	0.6316	0.7615	0.7332	0.8084	0.7596
Gemini-2.5-Pro	0.6368	0.7355	0.6924	0.5936	0.5089	0.7333	0.6495	0.6776	0.5604	0.5888	0.6997	0.5217	0.6803
图像美学评估专家模型													
AesExpert-7B	0.4056	0.5025	0.4142	0.3317	0.4636	0.3017	0.4147	0.3253	0.2318	0.2670	0.5073	0.6166	0.4904
UNIAA-LLaVA	0.2900	0.2393	0.2120	0.2471	0.2207	0.2733	0.3316	0.3041	0.3073	0.2893	0.5266	0.2410	0.2879

表 3：区域选择任务上的评估结果。总体准确率是所有指标的平均值。最佳和次佳结果分别以**粗体**和下划线标出。

模型	总体准确率	布局				颜色				Font		图形	
		平衡	分层	留白	对齐	协调	对比度	吸引力	心理学	易读性	层级	质量	相关性
非推理模型													
LLaVA-13B	0.6065	0.5823	0.5713	0.5612	0.5918	0.6171	0.5166	0.6319	0.6856	0.6130	0.7003	0.6329	0.5745
Qwen-VL-7B（基础版）	0.5795	0.5128	0.5370	0.5748	0.5443	0.5822	0.5433	0.5412	0.6384	0.5974	0.6379	0.6258	0.6190
Qwen-VL-32B	0.6311	0.5933	0.5397	0.5012	0.5252	0.5678	0.7065	0.6367	0.5735	0.7833	0.6278	0.7650	0.7533
Qwen-VL-72B	0.6626	0.5105	0.5839	0.4547	0.5348	0.5977	0.7225	0.6495	0.7940	0.7728	0.7934	0.7360	0.8015
Intern-VL3-8B	0.5799	0.5242	0.4948	0.5606	0.5363	0.5527	0.5568	0.5805	0.6342	0.6583	0.6157	0.6031	0.6415
Intern-VL3-14B	0.6378	0.5872	0.5204	0.5945	0.5745	0.6282	0.6997	0.6419	0.7034	0.7244	0.6870	0.6109	0.6818
GPT-4o	0.6745	0.4714	0.4894	0.5007	0.6011	0.7406	0.8591	0.7166	0.8135	0.6633	0.9080	0.6444	0.6865
GPT-5	**0.6989**	0.6484	0.5929	0.6630	0.6396	0.7229	0.7214	0.7510	0.6847	0.8038	0.7565	0.6953	0.7070
推理增强模型													
GPT-o1	0.6347	0.6319	0.5746	0.6397	0.5934	0.6178	0.6323	0.6880	0.7734	0.6320	0.7092	0.5936	0.5305
GPT-o3	0.6581	0.6483	0.6263	0.5325	0.3653	0.8272	0.5486	0.7113	0.7981	0.6586	0.7744	0.6601	0.7466
Gemini-2.5-Pro	0.6100	0.6810	0.6981	0.6096	0.2992	0.6050	0.6538	0.6678	0.6050	0.6164	0.6539	0.5696	0.6605
图像美学评估专家模型													
AesExpert-7b	0.2883	0.2954	0.2280	0.3174	0.2631	0.3426	0.2646	0.3176	0.3176	0.3176	0.2588	0.2765	0.2602
UNIAA-LLaVA	0.2418	0.1619	0.2915	0.1552	0.4075	0.1700	0.1516	0.1760	0.2479	0.2777	0.3796	0.2861	0.1968

GPT-o1、GPT-o3 和 Gemini-2.5-Pro（Comanici 等，2025）。此外，我们还纳入了专为图像美学评估设计的专家模型，即 AesExpert（Huang 等，2024）和 UNIAA-LLAVA（Zhou 等，2024）。所有模型均在相同的输入设置下进行评估，该设置包括一个问题（见图 1）、一张设计图像，以及 JSON 格式的元数据。

结果。 我们按照第 3.3 节介绍的方案评估视觉语言模型的表现。具体而言，除了报告各项指标的得分外，我们还提供总体评分，该评分为所有指标的平均值。

审美判断。 表 2 展示了结果。首先，在非推理模型中，GPT-5 表现最佳，总体准确率为 0.7252。这表明，即使是最强大的视觉语言模型，在设计美学评估方面仍面临困难。其次，推理增强模型的表现并未优于对应的非推理模型（例如，GPT-o1（Jaech 等，2024）和 GPT-o3 与 GPT-4o（Hurst 等，2024）及 GPT-5 相比），这说明通用推理对这一领域的帮助有限。第三，专为图像美学评估设计的专家模型整体表现更差，凸显了设计美学与图像美学之间的显著差距。最后，模型在不同指标上的表现各有差异。例如，与其他视觉语言模型相比，Qwen-VL 系列在易读性方面往往表现更好，但在心理学方面表现更差。

区域选择。 表3报告了结果。首先，视觉语言模型在此任务上的表现通常不如在审美判断任务上的表现，这可能是因为该任务不仅要求评估设计是否美观，还要求识别瑕疵出现的位置。其次，与审美判断任务的结果一致，GPT-5表现最佳，而推理增强模型并未显示出明显优势。最后，在不同模型系列中，较大的模型（如72B）通常优于较小的模型（如7B）。我们发现一些模型存在过拟合现象，因此在计算最终分数时采用加权和。

表 4：选择组件的精确定位任务评估，模型需要在不存在审美问题时预测 None。总体评分为所有指标准确率的平均值。最佳和次佳结果分别以**粗体**和下划线标出。

模型	总体评分	布局				颜色				Font		图形	
		平衡	分层	留白	对齐	协调	对比度	吸引力	心理学	易读性	层级	质量	相关性
非推理模型													
LLaVA-13B	0.4455	0.4523	0.6130	0.5699	0.4714	0.3356	0.3723	0.2898	0.2474	0.7301	0.2453	0.5611	0.4573
Qwen-VL-7B	0.5192	0.5104	0.5839	0.4546	0.5347	0.4376	0.5625	0.5495	0.5339	0.4528	0.5334	0.5360	0.5415
GPT-4o	0.5680	0.6417	0.2063	0.5954	0.1679	0.8626	0.6372	0.7200	0.6713	0.4594	0.5164	0.7762	0.5618
GPT-5	**0.6090**	0.6306	0.6142	0.6910	0.6247	0.6057	0.6170	0.5643	0.5989	0.5464	0.6134	0.6217	0.5804
推理增强模型													
GPT-o1	0.5295	0.4628	0.4870	0.5248	0.4943	0.5322	0.4933	0.4912	0.5884	0.5474	0.5879	0.6258	0.5190
GPT-o3	0.5800	0.5922	0.3868	0.5369	0.6570	0.7554	0.6595	0.6011	0.6985	0.7396	0.3952	0.5311	0.4063
Gemini-2.5-Pro	<u>0.6047</u>	0.6319	0.5746	0.5397	0.5934	0.6178	0.6323	0.6880	0.7734	0.5320	0.7092	0.4336	0.5305
图像美学评估专家模型													
AesExpert-7b	0.3377	0.3229	0.3146	0.3756	0.3172	0.4276	0.3582	0.3146	0.3267	0.3314	0.3803	0.3025	0.2804

表5：边界框预测组件在精确定位任务上的评估，模型应输出审美问题的坐标。总体评分是所有指标的平均交并比。最佳和次佳结果分别以**粗体**和下划线标出。

模型	总体评分	布局				颜色				Font		图形	
		平衡	分层	留白	对齐	协调	对比度	吸引力	心理学	易读性	层级	质量	相关性
非推理模型													
LLaVA-13B	0.0559	0.0653	0.0080	0.0302	0.0172	0.1024	0.0792	0.0427	0.0303	0.0188	0.0102	0.1510	0.1160
Qwen-VL-7B（基础版）	0.0514	0.0067	0.1669	0.0101	0.0259	0.0109	0.0036	0.0039	0.2306	0.0012	0.0452	0.0994	0.0063
GPT-4o	<u>0.1712</u>	0.1822	0.2974	0.1399	0.2552	0.0883	0.1353	0.0664	0.2186	0.2144	0.0586	0.2707	0.1270
GPT-5	**0.1993**	0.1866	0.1546	0.2077	0.1613	0.1829	0.1525	0.1529	0.3348	0.1745	0.1835	0.2667	0.2338
推理增强模型													
O1	0.1286	0.0907	0.0767	0.1412	0.1226	0.1719	0.0204	0.1546	0.1236	0.1440	0.1441	0.1912	0.1617
O3	0.1418	0.0619	0.1915	0.0552	0.3075	0.0700	0.0516	0.0760	0.1479	0.1777	0.2796	0.1861	0.0968
Gemini-2.5-Pro	0.0977	0.0518	0.1620	0.1052	0.0710	0.0760	0.0487	0.0490	0.2257	0.0963	0.0903	0.0945	0.1014
图像美学评估专家模型													
AesExpert-7b	0.0327	0.0440	0.0203	0.0093	0.1084	0.0861	0.0063	0.0049	0.0118	0.0001	0.0348	0.0601	0.0067

<u>精确定位。</u> 如第3.3节所述，该任务由两个部分组成，每个部分分别评估：选择题，即当不存在审美问题时模型预测为 None（表4）；以及边界框预测问题，即模型输出审美问题的确切边界框（表5）。我们排除了部分视觉语言模型（例如 Intern-VL 系列和小型 Qwen-VL 模型），因为它们未能产生有意义的边界框预测。对于选择题，视觉语言模型取得了合理的性能，最佳模型的总体评分达到 0.6090。对于边界框预测问题，即使表现最好的模型 GPT-5 得分也低于 0.20，这凸显了精确定位审美问题的巨大难度。

<u>关于输入组件的讨论。</u> 在对视觉语言模型进行基准测试时，输入由三个组件组成：（1）问题，其中包含对目标指标的详细解释；（2）设计图像；（3）JSON 格式的元数据，包含布局、颜色和字体信息。我们以 GPT-4o 为代表示例，分析各组件对模型性能的贡献。图3展示了结果，其中*完整模型* 表示使用全部三个组件的设置；*无图像*移除设计图像；*无解释* 省略详细的指标描述；*无元信息* 排除元数据。我们的发现揭示了三个关键见解。首先，在所有任务中，设计图像都不可或缺——移除它会导致最大的性能下降。其次，指标解释对更直观的指标（例如平衡）影响有限，但对主观指标（例如相关性或心理学）起着关键作用，这些指标需要更清晰的定义。最后，元数据对精确定位的影响最大，与审美判断或区域选择相比，缺少元数据会导致更大的性能下降。我们推测，这是因为元数据提供了明确的布局信息，有助于定位任务中的边界框预测。

5.2 使用 AesEval-Train 微调视觉语言模型

设置。 我们按照第4节所述的流程构建训练集，得到3万组问答对。在实验中，我们使用 Qwen2.5-VL-7B-Instruct（Bai 等，2025）作为代表性模型，并在构建的数据集上进行全参数微调。学习率设为 1e-6，使用余弦调度器，预热比例为 3%。为提高计算效率，

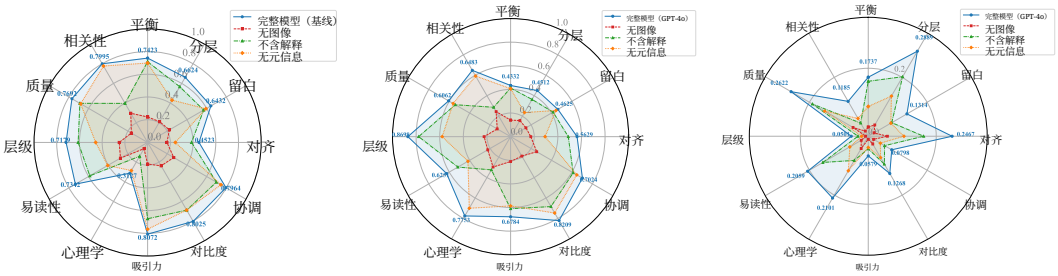


图3：使用不同输入组件的模型变体的结果。

表6：使用我们构建的训练集对视觉语言模型进行微调的结果及消融研究。

模型设置	总体评分	总体增益	总体排名	布局				颜色				Font		图形	
				平衡	分层	留白	对齐	协调	对比度	吸引力	心理学	易读性	层级	质量	相关性
审美判断（准确性）															
Qwen-VL-7B（基础版）	0.6390	-	9	0.8272	0.4508	0.8076	0.8413	0.8223	0.9136	0.4009	0.1355	0.9430	0.3100	0.8183	0.3970
+ AesEval-Train	0.6987	+ 5.97%	4	0.7123	0.6789	0.7215	0.6868	0.7031	0.6654	0.7329	0.6577	0.7436	0.6482	0.7096	0.7244
- 推理路径	0.6576	-	-	0.6511	0.6589	0.6413	0.6687	0.6309	0.6791	0.6207	0.6893	0.6115	0.6985	0.6508	0.6904
- 正样本	0.2072	-	-	0.2101	0.1999	0.2202	0.1893	0.2058	0.2147	0.1955	0.2246	0.1855	0.2296	0.2004	0.2108
区域选择（准确率）															
Qwen-VL-7B（基础版）	0.5795	-	10	0.5128	0.5370	0.5748	0.5443	0.5822	0.5433	0.5412	0.6384	0.5974	0.6379	0.6258	0.6190
+ AesEval-Train	0.6065	+ 2.70%	8	0.5827	0.6108	0.5963	0.6289	0.5714	0.6236	0.6541	0.6412	0.6389	0.6487	0.5899	0.5915
- 推理路径	0.5795	-	-	0.5732	0.5279	0.5697	0.5322	0.5741	0.5322	0.5341	0.6283	0.5953	0.6318	0.6667	0.5885
- 正样本	0.5327	-	-	0.5089	0.5411	0.5257	0.5543	0.5013	0.5587	0.5291	0.5709	0.5212	0.5788	0.5146	0.4878
精确定位（交并比）															
Qwen-VL-7B（基础版）	0.0514	-	8	0.0067	0.1669	0.0101	0.0259	0.0109	0.0036	0.0039	0.2306	0.0012	0.0452	0.0994	0.0063
+ AesEval-Train	0.2231	+ 17.17%	1	0.2518	0.1982	0.3103	0.0857	0.2204	0.2846	0.1552	0.3901	0.0607	0.2313	0.1152	0.2745
- 推理路径	0.0782	-	-	0.1523	0.0211	0.1987	0.0095	0.0750	0.1204	0.0348	0.0812	0.0159	0.0555	0.0601	0.1139
- 正样本	0.0641	-	-	0.1866	0.1546	0.2077	0.1613	0.1829	0.1525	0.1529	0.3348	0.1745	0.1835	0.2667	0.2338

训练采用 bfloat16 混合精度和 FlashAttention-2（Dao等人，2022年）。视觉编码器保持冻结，而语言模型参数则进行微调。

主要结果。 表6展示了结果，其中 *Qwen-VL-7B*（基础版）表示未经微调的基础模型，而 + *AesEval-Train*指在我们构建的训练集上微调后的模型。首先，在全部三项任务中，使用 AesEval-Train 进行微调都带来了显著的性能提升。此外，在审美判断方面，微调后的模型甚至超过了参数量最大的 Qwen-VL 变体（72B参数）；在精确定位方面，尽管 GPT-5 的参数量远大于该模型，该模型仍优于 GPT-5。这些结果表明，我们提出的流程能够有效构建训练数据，从而显著提升模型性能。

消融研究。 我们考察了不同数据配方对模型性能的影响。表6报告了结果，其中 -推理路径 表示仅使用普通问答对进行训练，不采用所提出的基于指标的推理；-正样本表示仅使用有缺陷的设计进行训练。我们发现，-推理路径 仍能提升全部三项任务的性能，这表明融入设计美学领域专属知识是有益的。然而，其表现仍明显逊于采用推理路径的完整变体（+ *AesEval-Train*），凸显了基于指标的推理的有效性。此外，- 正样本 的表现逊于 + *AesEval-Train* 和-推理路径，这凸显了维持训练集标签平衡的重要性。

6 结论

在本研究中，我们提出了用于设计美学评估的 AesEval-Bench，涵盖四个维度、十二项指标和三项可量化任务。基于该基准，我们系统评估了专有、开源及推理增强型视觉语言模型，发现它们在设计美学评估方面存在明显差距。此外，我们构建了一个训练数据集，用于对该领域的视觉语言模型进行微调。实验表明，该数据集显著提升了模型在所有任务上的表现。

局限性。 首先，由于 Crello 是源数据集，该基准未涵盖所有类型的平面设计，例如信息图表或移动端用户界面。其次，目前尚无针对指标的完全解耦分类体系。未来若能提出更严格的解耦分类体系，采用该体系并相应更新我们的指标将很有价值。第三，高度

设计中主观性较强的方面（如创造力）并未纳入其中。最后，利用强化学习进一步增强推理能力，则留待未来探索。

参考文献

Josh Achiam、Steven Adler、Sandhini Agarwal、Lama Ahmad、Ilge Akkaya、Florencia Leoni Aleman、DiogoAlmeida、JankoAltenschmidt、Sam Altman、Shyamal Anadkat 等。GPT-4 技术报告。*arXiv* 预印本 *arXiv:2303.08774*，2023年。

Ruichuan An、Sihan Yang、Ming Lu、Renrui Zhang、Kai Zeng、Yulin Luo、Jiajun Cao、Hao Liang、Ying Chen、Qi She 等。Mc-LLaVA：多概念个性化视觉语言模型。*arXiv* 预印本 *arXiv:2411.11706*，2024。

Ruichuan An、Sihan Yang、Ziyu Guo、Wei Dai、Zijun Shen、Haodong Li、Renrui Zhang、Xinyu Wei、Guopeng Li、Wenshan Wu 等。Genius：生成式流体智力评测套件。*arXiv* 预印本 *arXiv:2602.11144*，2026。

Shuai Bai、Keqin Chen、Xuejing Liu、Jialin Wang、Wenbin Ge、Sibo Song、Kai Dang、Peng Wang、Shijie Wang、Jun Tang 等。Qwen2. 5-vl 技术报告。*arXiv* 预印本 *arXiv:2502.13923*，2025。

Meng Cao、Haoze Zhao、Can Zhang、Xiaojun Chang、Ian Reid 和 Xiaodan Liang。Ground-r1：通过强化学习激励基于视觉依据的推理。*arXiv* 预印本 *arXiv:2505.20272*，2025。

Gheorghe Comanici、Eric Bieber、Mike Schaekermann、Ice Pasupat、Noveen Sachdeva、Inderjit Dhillon、Marcel Blistein、Ori Ram、Dan Zhang、Evan Rosen 等。Gemini 2.5：凭借先进的推理、多模态、长上下文和新一代智能体能力，拓展前沿。*arXiv* 预印本 *arXiv:2507.06261*，2025。

TriDao、Dan Fu、Stefano Ermon、Atri Rudra 和 Christopher R´e。Flashattention：借助 IO 感知实现快速、节省内存的精确注意力。神经信息处理系统进展，35：16344–16359，2022年。

Yubin Deng、Chen Change Loy 和 Xiaoou Tang。图像美学评估：一项实验调查。*IEEE* 信号处理杂志，34(4)：80–106，2017年。

Ziyu Guo、Xinyan Chen、RenruiZhang、Ruichuan An、Yu Qi、Dongzhi Jiang、Xiangtai Li、Manyuan Zhang、Hongsheng Li 和 Pheng-Ann Heng。视频模型是否已具备零样本推理能力？一项基于 MME-COF 基准的实证研究。*arXiv* 预印本 *arXiv:2510.26802*，2025。

Daichi Haraguchi、Naoto Inoue、Wataru Shimoda、Hayato Mitani、Seiichi Uchida 和 Kota Yam- aguchi。GPT 能否依据设计原则评估平面设计？载于*SIGGRAPH Asia 2024* 技术交流，第 1–4 页，2024。

Yipo Huang、Quan Yuan、Xiangfei Sheng、Zhichao Yang、Haoning Wu、Pengfei Chen、Yuzhe Yang、Leida Li 和 Weisi Lin。Aesbench：用于评测多模态大语言模型图像美学感知能力的专家基准。*arXiv* 预印本 *arXiv:2401.08276*，2024。

Aaron Hurst、Adam Lerer、Adam P Goucher、Adam Perelman、Aditya Ramesh、AidanClark、AJ Ostrow、Akila Welihinda、Alan Hayes、Alec Radford 等。GPT-4o 系统卡。*arXiv* 预印本 *arXiv:2410.21276*，2024。

Aaron Jaech、Adam Kalai、Adam Lerer、Adam Richardson、AhmedEl-Kishky、AidenLow、Alec Helyar、Aleksander Madry、Alex Beutel、Alex Carney 等。OpenAI o1 系统卡。*arXiv* 预印本 *arXiv:2412.16720*，2024。

Ruixiang Jiang与Changwen Chen。多模态大语言模型能够以零样本方式对美学进行推理。*arXiv* 预印本 *arXiv:2501.09012*，2025。

Sam Jung、Agustin Garcinuno 和 Spencer Mateega。UI-Bench：用于评估 AI 文本转应用工具设计能力的基准。*arXiv* 预印本 *arXiv:2508.20410*，2025。

Hyolim Kang、Yunsu Park、Youngbeom Yoo、Yeeun Choi 和 Seon Joo Kim。基于视觉语言模型的开放式分层流式视频理解。载于*IEEE/CVF*国际计算机视觉会议论文集，第 20715–20725 页，2025。

Bo Li、Yuanhan Zhang、Dong Guo、Renrui Zhang、Feng Li、Hao Zhang、Kaichen Zhang、Peiyuan Zhang、Yanwei Li、Ziwei Liu 等。Llava-onevision：轻松实现视觉任务迁移。*arXiv* 预印本 *arXiv:2408.03326*，2024a。

Chengzu Li、Wenshan Wu、Huanyu Zhang、Yan Xia、Shaoguang Mao、Li Dong、Ivan Vuli´c 和 Furu Wei。在空间中推理时进行想象：多模态思维可视化。*arXiv* 预印本 *arXiv:2501.07542*，2025a。

Congcong Li 和 Tsuhan Chen。绘画作品的审美视觉质量评估。*IEEE*信号处理选题期刊，3(2):236–252，2009。

Jian Li、Weiheng Lu、Hao Fei、Meng Luo、Ming Dai、Min Xia、Yizhang Jin、Zhenye Gan、Ding Qi、Chaoyou Fu 等。多模态大语言模型基准测试综述。*arXiv* 预印本 *arXiv:2408.08632*，2024b。

Junxian Li, Di Zhang, Xunzhi Wang, Zeying Hao, Jingdi Lei, Qian Tan, Cai Zhou, Wei Liu, Yaotian Yang, Xinrui Xiong 等。Chemvlm：探索多模态大语言模型在化学领域的能力。载于*AAAI*人工智能会议论文集，第39卷，第415–423页，2025b。

Jieru Lin, Danqing Huang, Tiejun Zhao, Dechen Zhan 和 Chin-Yew Lin。Designprobe：面向多模态大语言模型的平面设计基准测试。*arXiv* 预印本 *arXiv:2404.14801*，2024a。

Kevin Lin, Zhengyuan Yang, Linjie Li, Jianfeng Wang 和 Lijuan Wang。Designbench：探索并评测 DALL·E 3 在视觉设计构想方面的能力。*arXiv* 预印本 *arXiv:2310.15144*，2023年。

Weifeng Lin、Xinyu Wei、Ruichuan An、Peng Gao、Bocheng Zou、Yulin Luo、Siyuan Huang、Shang- hang Zhang 和 Hongsheng Li。Draw-and-understand：借助视觉提示，让多模态大语言模型理解你的意图。*arXiv* 预印本 *arXiv:2403.20271*，2024b。

Weifeng Lin、Xinyu Wei、Ruichuan An、Tianhe Ren、Tingwei Chen、Renrui Zhang、Ziyu Guo、Wen- tao Zhang、Lei Zhang 和 Hongsheng Li。感知一切：识别、解释、描述并分割图像和视频中的一切。*arXiv* 预印本 *arXiv:2506.05302*，2025年。

Haotian Liu、Chunyuan Li、Qingyang Wu 和 Yong Jae Lee。视觉指令调优。神经信息处理系统进展，36：34892–34916，2023年。

Yun Lou、Weiyue Gao、Pei Chen、Xuanhui Liu、Changyuan Yang和Lingyun Sun。Designeva：一种具有多方面感知评估的设计支持工具。见国际人机交互会议，第508–519页。Springer，2022年。

Yue Lu、Chao Guo、Yilun Lin、Fan Zhuo和FY Wang。美术绘画的计算美学：研究现状与展望。自动化学报，46(11)：2239–2259，2020年。

Meng Luo、Hao Fei、Bobo Li、Shengqiong Wu、Qian Liu、Soujanya Poria、Erik Cambria、Mong-Li Lee和Wynne Hsu。Panosent：面向多模态对话式基于方面的情感分析的全景六元组抽取基准。见第*32*届*ACM*国际多媒体会议论文集，第7667–7676页，2024。

Meng Luo、Shengqiong Wu、Liqiang Jing、Tianjie Ju、Li Zheng、Jinxiang Lai、Tianlong Wu、Xinya Du、Jian Li、Siyuan Yan、Jiebo Luo、William Yang Wang、Hao Fei、Mong-Li Lee 和 Wynne Hsu。Dr.v：一种分层的感知—时间—认知框架，通过细粒度时空定位诊断视频幻觉，2025a。URL：https://arxiv.org/abs/2509. 11866。

Meng Luo、Shengqiong Wu、Liqiang Jing、Tianjie Ju、Li Zheng、Jinxiang Lai、Tianlong Wu、Xinya Du、Jian Li、Siyuan Yan 等。《Dr. v：通过细粒度时空定位诊断视频幻觉的分层感知—时间—认知框架》。*arXiv* 预印本 *arXiv:2509.11866*，2025b。

Jon McCormack 和 Andy Lomas。《运用深度学习理解审美评估》。载于音乐、声音、艺术与设计计算智能国际会议（*EvoStar* 的组成部分），第 118–133 页。Springer，2020年。

Joonhyung Park、Peng Tang、Sagnik Das、Srikar Appalaraju、Kunwar Yashraj Singh、R Manmatha 和 Shabnam Ghadar。《R-vlm：面向精确图形用户界面定位的区域感知视觉语言模型》。载于计算语言学协会年会论文集：*ACL 2025*，第 9669–9685 页，2025。

Gabriel Sarch、SnigdhaSaha、Naitik Khandelwal、Ayush Jain、Michael J Tarr、Aviral Kumar 和 Katerina Fragkiadaki。《用于视觉推理的基于定位的强化学习》。*arXiv* 预印本 *arXiv:2505.23678*，2025。

Hao Shao、ShengjuQian、Han Xiao、Guanglu Song、Zhuofan Zong、Letian Wang、Yu Liu 和 Hong- sheng Li。《视觉思维链：通过全面的数据集和基准推进多模态语言模型的思维链推理》。神经信息处理系统进展，37：8612–8642，2024。

KunalSingh、Shreyas Singh 和 MukundKhanna。Trishul：面向大型视觉语言模型 GUI 智能体的区域识别与屏幕层级理解。载于计算机视觉与模式识别会议论文集，第 170–179 页，2025 年。

Guangyan Sun、Mingyu Jin、ZhentingWang、Cheng-Long Wang、Siqi Ma、Qifan Wang、Tong Geng、Ying Nian Wu、Yongfeng Zhang 和 Dongfang Liu。视觉智能体作为快思考与慢思考者。*arXiv* 预印本 *arXiv:2408.08862*，2024 年。

Peng Wang、Shuai Bai、Sinan Tan、Shijie Wang、Zhihao Fan、Jinze Bai、Keqin Chen、Xuejing Liu、Jialin Wang、Wenbin Ge 等。Qwen2-vl：提升视觉语言模型在任意分辨率下对世界的感知能力。*arXiv* 预印本 *arXiv:2409.12191*，2024 年。

Chatchai Wangwiwattana 和 Areerat Meeyen。《人工智能视角下的设计：自动化美学评估以学习平面设计概念》。载于人工智能多学科趋势国际会议，第 336–347 页。Springer，2024。

QiongWu、XiangcongYang、YiyiZhou、Chenxin Fang、Baiyang Song、Xiaoshuai Sun 和 Ron-grong Ji。《多模态大语言模型的基于定位的思维链》。《style id='1'>arXiv 预印本 arXiv:2503.12799，2025。

Kota Yamaguchi。《Canvasvae：学习生成矢量图形文档》。载于论文集 *其 IE EE/CVF* 国际计算机视觉会议，第 5481–5489 页，2021。

An Yang、Anfeng Li、Baosong Yang、Beichen Zhang、Binyuan Hui、Bo Zheng、Bowen Yu、ChangGao、ChengenHuang、Chenxu Lv 等。《Qwen3 技术报告》。《style id='1'>arXiv 预印本 arXiv:2505.09388，2025。

Di Zhang、Jingdi Lei、Junxian Li、Xunzhi Wang、Yujie Liu、Zonglin Yang、Jiatong Li、Weida Wang、Suorong Yang、Jianbo Wu 等。Critic-v：VLM 评论者助力发现多模态推理中的 VLM 错误。发表于计算机视觉与模式识别会议论文集，第 9050–9061 页，2025a。

Huanyu Zhang、Chengzu Li、Wenshan Wu、Shaoguang Mao、Yifan Zhang、Haochen Tian、Ivan Vuli´c、Zhang Zhang、Liang Wang、Tieniu Tan 等。扩展及更进一步：推进多模态大语言模型中的空间推理需要新的方案。*arXiv* 预印本 *arXiv:2504.15037*，2025b。

Huanyu Zhang、Wenshan Wu、Chengzu Li、Ning Shang、Yan Xia、Yangyu Huang、Yifan Zhang、Li Dong、Zhang Zhang、Liang Wang 等。潜在草图板：通过绘制视觉思维草图，激发多模态大语言模型的多模态推理。*arXiv* 预印本 *arXiv:2510.24514*，2025c。

Qintong Zhang、Bin Wang、Victor Shea-Jay Huang、Junyuan Zhang、Zhengren Wang、Hao Liang、Conghui He 和 Wentao Zhang。《文档解析揭秘：结构化信息抽取的技术、挑战与展望》。*arXiv* 预印本 *arXiv:2410.21169*，2024。

YiFan Zhang、Huanyu Zhang、Haochen Tian、Chaoyou Fu、Shuangqing Zhang、Junfei Wu、Feng Li、Kun Wang、Qingsong Wen、Zhang Zhang 等。Mme-realworld：你的多模态大语言模型能否挑战人类难以应对的高分辨率真实场景？载于第十三届国际学习表征会议，2025d。

Ming Zhong、Yuanlei Wang、Liuzhou Zhang、Arctanx An、Renrui Zhang、Hao Liang、Ming Lu、Ying Shen 和 Wentao Zhang。Vcu-bridge：通过语义桥接实现层次化视觉内涵理解。*arXiv* 预印本 *arXiv:2511.18121*，2025。

Zhaokun Zhou、Qiulin Wang、Bin Lin、Yiwei Su、Rui Chen、Xin Tao、Amin Zheng、Li Yuan、Pengfei Wan 和 Di Zhang。Uniaa：统一的多模态图像美学评估基线与基准。*arXiv* 预印本 *arXiv:2404.09619*，2024。

Jinguo Zhu、Weiyun Wang、Zhe Chen、Zhaoyang Liu、Shenglong Ye、Lixin Gu、Hao Tian、Yuchen Duan、Weijie Su、Jie Shao 等。Internvl3：探索开源多模态模型的先进训练与测试时方案。*arXiv* 预印本 *arXiv:2504.10479*，2025。

维度指标提示：

图形质量： 分辨率是衡量图像细节的指标。合适的分辨率能确保最佳图像质量和可读性。该元素是否存在质量问题？**图形相关性：** 相关性指图形元素与其所传达含义之间的直接联系。该元素是否存在相关性问题？

色彩和谐： 色彩和谐是指图像中存在两种或两种以上颜色时，整体色彩所呈现的协调、愉悦与美感。
色彩对比： 色彩对比是指各种颜色之间存在的对比、对立和差异。

色彩吸引力： 色彩吸引力是指色彩的选择与搭配能够吸引观众的注意力。

色彩心理学： 色彩心理学是指颜色能够引发主观心理体验，并影响情绪、感受和行为。

布局-平衡： 平衡是指设计中视觉重量的分布。它可以是对称的（两侧重量相等），也可以是非对称的（两侧重量不等，但仍能实现视觉平衡）。

布局-层叠： 运用尺寸、颜色、对比度及其他视觉线索，建立设计元素的层级结构，引导观众的视线。 **布局-留白：** 留白是指设计元素周围的空白区域。有效利用留白可以营造平衡、视觉层级和清晰感。**布局对齐：** 对齐是指设计元素相对于彼此或特定轴线或网格的排列方式。恰当的对齐能够营造秩序感和条理性，使设计更易于理解和浏览。

字体层级： 字体的呈现具有层级结构，使用户能够浏览文本并获取关键信息。

字体易读性： 易读性是指识别单个字符以及它们并排排列时彼此之间关系的难易程度。

表7：各指标描述展示。

任务推理指令：

审美判断：给定预览图像、元素图像以及元素边界框 bbox，请推理该元素在标准方面为何不具美感。你的理由应仅包含关于标准的分析。请思考并推理。一旦提及该元素，你应在元素之后添加其边界框（仅边界框）。

区域选择：给定一张预览图像、一张不美观元素的图像及其边界框 bbox[0], 以及一张美观元素的图像及其边界框 bbox[1],请基于元素之间的关系，分析第一个元素在标准方面为何不美观。你的推理应仅包含对标准的分析。请思考并推理。一旦提及某个元素，你应在该元素之后添加其边界框（仅边界框）。

精确定位：给定一张预览图像、一张元素图像以及该元素的边界框 bbox，请根据它们之间的关系，从标准的角度分析该元素为何显得不美观。你的推理应仅包含对标准的分析。请思考并推理。提及该元素后，应紧接着给出其边界框（仅边界框）。

表8：任务推理指令展示。

A LLM 使用声明

在实施和调试过程中，我们借助大语言模型（LLMs）获取技术指导；在协作完成稿件初稿后，我们进一步使用大语言模型润色文稿并改善整体表达。

B 提示与指令

首先，维度-指标提示建立了一套明确的评估标准。这些标准分为四个核心维度：图形、颜色、布局和字体，每个维度都包含特定指标，例如字体易读性。如表7所示，每项指标都有明确的定义，并配有一个引导性问题，以规范分析过程。

其次，任务推理指令（表8）为创建推理路径提供了操作指南。指令要求分析聚焦于某个元素本身的缺陷、它与其他元素的关系或其直接上下文，同时明确要求纳入该元素的边界框（bbox），以便基于精确的空间证据进行推理。

C 补充相关工作

视觉语言模型的评估。近年来，旨在评估视觉语言模型的基准测试（Li 等，2024b）显著增多，涵盖了对感知（Luo et al., 2024）和推理（Guo et al., 2025; Zhang et al., 2025a）的通用评估，包括区域级评估（Lin 等，2024b; 2025; Park et al., 2025）和层次化理解评估（Kang et al., 2025; Singh et al., 2025; Zhong et al., 2025），以及化学（Li 等，2025b）或图像生成（An et al., 2026）等领域的专项评估。现有基准测试虽能有效评估通用能力，却往往忽视视觉理解中细腻且主观的维度。为此，我们提出了一项专门聚焦审美维度的基准，旨在评估视觉语言模型如何理解艺术质量和视觉吸引力。

D 数据来源展示

为回应关于 Crello 数据集多样性的疑虑，我们在图 4 中展示了其典型数据的视觉样本。如图所示，该数据集涵盖了丰富多样的设计类别与美学风格，包括文字手写风格、摄影主导风格以及各种艺术风格的设计。这种多样性确保我们的基准能够在符合现实情况的平面设计任务分布上评估模型，从而避免偏向任何单一视觉风格。

E 缺陷注入流程

F 现实世界设计数据展示

为验证我们的模型在 Crello 合成分布之外的泛化能力，我们编制了一个独立的分布外（OOD）测试集，其中包含现实世界中的有缺陷设计。如图 5 所示，这些样本直接取自中间设计草稿，并由专业设计师进行标注，体现了真实设计细腻且往往复杂的特性。

将现实缺陷映射为合成扰动。 为了进一步验证数据构建流水线的设计，我们分析了收集到的现实世界中的有缺陷设计（图5），并将其中的缺陷映射到算法1中的操作。如“创业招聘会”海报所示，文字与深色背景融为一体，难以辨认。我们通过

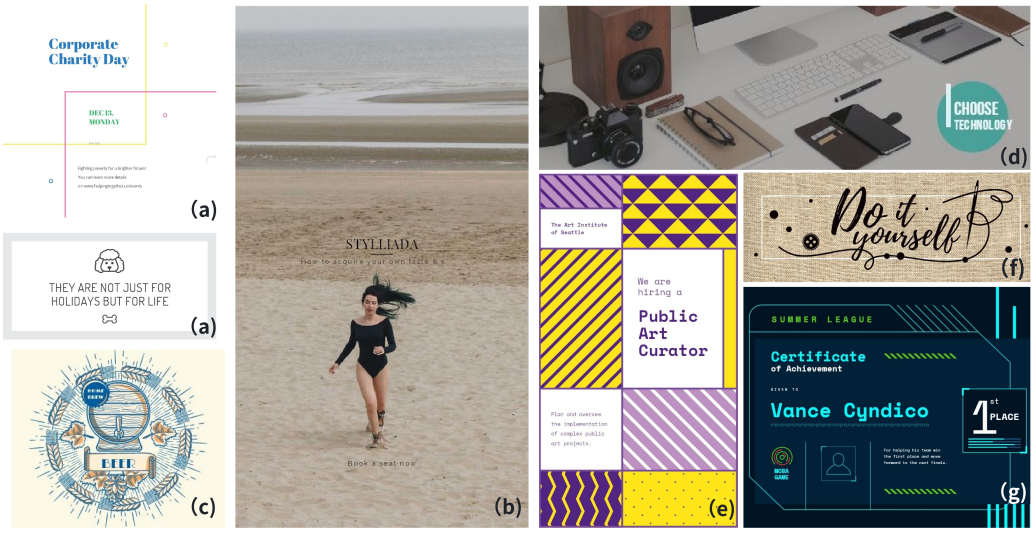


图 4：**来自 Crello 数据集的多样化设计样本**。该集合展现了丰富多样的视觉风格和结构布局，包括 (a) 极简主义以排版为中心的设计（例如“Corporate Charity”）、(b) 以摄影为主、呈现真实人物的时尚专题、(c) 复古插画、(d) 照片级真实感科技样机、(e) 几何抽象艺术、(f) 具有纹理效果的艺术字体排印，以及 (g) 赛博朋克主题证书。这些视觉证据驳斥了风格同质性的担忧，证实该数据集对各设计领域具有全面且扎实的覆盖。

流水线中的 **颜色扰动** 模块模拟了这一特定错误。同样，“End Violence”海报存在严重的图层问题，图形元素遮挡了关键信息。我们的**布局扰动** 模块通过对元素应用随机坐标偏移（位置偏移），有效地再现了这类缺陷，从而造成意外重叠和布局冲突。这种结构上的对应关系证实，我们的合成流水线能够生成与现实世界中的有缺陷设计相吻合的有意义负例。

关于数据集多样性和基准展示的说明。需要指出的是，在基准展示（第 I 节）中，我们有意选择了存在单一且明显缺陷的示例。这种选择策略纯粹出于教学目的——清晰、明确地展示每一种特定美学指标（例如，单独呈现“平衡”失调的具体表现）。读者应注意，完整的 AesEval-Bench 和 AesEval-Train 数据集要丰富得多，也更具挑战性。它们涵盖了广泛的难度层次，从展示中呈现的明显单一缺陷示例，到复杂的多缺陷设计（类似图 5 中的真实案例）；在这些设计中，多个指标（例如对齐、易读性和色彩协调）可能同时存在问题。

G 任务专用提示示例。

为确保对推理增强模型（如 GPT-o1、GPT-o3）进行严格评估，我们还制定了一套专门的优化提示，以激发其思维链能力。如表 9 所述，这些优化提示不同于用于标准视觉语言模型的直接提问（“原始提示”），而是明确指示模型进行逐步分析。然而，即使是针对特定任务设计的提示，也无法提升模型的审美理解能力，这揭示了通用推理在此任务中的局限性。

H 统计检验

我们进行了严格的假设检验，以确保表6中报告结果的可靠性。

算法1 通过合成扰动进行数据构建流水线

要求： 原始设计 D、元数据 M （JSON 格式）、扰动库 $\mathcal{P}$、
受扰动元素数量 n。
确保： 扰动后设计 D'、更新后的元数据 M'。
初始化： $M' \leftarrow M$
元素选择： 从 M'中随机选择一部分元素 $E = \{e_1, e_2, \ldots, e_n\}$。
对所有 $e_i \in E$ **执行**
  扰动选择： 随机抽取一个 e_i类型适用的操作 $op \in \mathcal{P}$。
  参数采样： 采样扰动强度 δ 或目标属性。
  如果 op是 布局扰动**则**
    e_i.位置 $\leftarrow e_i$.位置 $+\mathcal{U}(-\delta_{pos},\ \delta_{pos})$ {移动位置}
  否则，如果 op 是 字体扰动**则**
    如果 op是尺寸变化**则**
      e_i.尺寸 $\leftarrow e_i$.尺寸 $+\mathcal{U}(-\delta_{size},\ \delta_{size})$
    否则，如果 op是字体替换 **则**
      e_i.字体 $\leftarrow$ 示例（字体库） $\setminus \{e_i$.字体$\}$
    结束如果
  否则如果 op是颜色扰动 **那么**
    如果 op 是低对比度**那么**
      e_i.颜色 $\leftarrow$ SampleNear(M.背景色 ϵ)
    否则如果 op是高对比度／冲突 **那么**
      e_i.颜色 $\leftarrow$ Invert(M.主导色)+噪声
    结束如果
  否则如果 op是 图形/图像扰动 **则**
    如果 op是替换 **则**
      e_i.src $\leftarrow$ 采样(图像库)
    否则如果 op是分辨率降低 **则**
      e_i.质量 $\leftarrow$ 下采样(e_i.src,factor)
    否则如果 op是模糊 **则**
      e_i.效果 $\leftarrow$ 高斯模糊(e_i.src σ)
    结束如果
  结束如果
  更新： 在元数据 M'中更新元素 e_i。
结束循环
渲染： $D' \leftarrow$ 渲染(M') {使用更新后的 JSON 重新渲染设计}
返回 D' , M'

- **审美判断与区域选择：** 由于这些属于分类任务，我们采用麦克尼马尔检验来分析模型之间的不一致性。结果显示存在显著差异，p 值分别为 $p = 0.004$ 和 $p < 0.001$，证实了我们的微调流程有效。
- **精确定位：** 对于交并比得分，我们进行了配对 t 检验。我们的微调模型显著优于最强基线（GPT-5），t 统计量为 $t = 4.82$，p 值为 $p < 0.01$。

I 基准展示

本节通过视觉示例，进一步说明 AesEval 基准中各个审美维度的评估标准。如表 10、表 14、表 17、表 15、表 13、表 12 和表 16 所示，每个展示都通过并列对比设计，呈现某项标准所体现的正面与负面特征。这些示例有助于阐明评判布局对齐、色彩协调、图形质量和字体可读性等方面所依据的标准，为我们的基准方法提供直观参考。



图5：由专业设计师收集的真实世界中的缺陷案例代表性样本。与合成扰动基准不同，这些设计由专业设计师筛选而来，旨在呈现他们在真实工作流程中遇到的真实审美缺陷。这些样本展现了多种缺陷，例如：(a) 视觉杂乱和方向不一致（如“世界艾滋病日”），(b) 空间失衡和元素脱节（如“创业招聘会”），(c) 基于网格的对齐错误（如“学校杂志”），以及 (d) 字体排版遮挡（如“End Violence”）。该数据集可作为严格的分布外基准，用于评估模型在合成模式之外的泛化能力。

Task	原始提示	优化提示
审美判断	$指标$ +用一个词回答：“是”或“否”。	根据以下内容分析该设计$指标$。评估视觉元素，逐步判断是否它们符合标准。然后回答只需用一个词回答“是”或“否”。
区域选择	$指标$ +请仅提供给定边界框选项中不美观元素的索引。A. [] B.[]C.[] D.[]。	检查候选区域（A、B、C、D），并依据 $指标$ 进行判断。通过视觉细节进行推理，以确定哪个特定区域存在缺陷。然后输出该元素的索引。
精确定位	$指标$ +请仅提供不美观元素的边界框.....如果没有任何问题，请返回‘None’。	分析整个设计以准确找出任何违反$指标$的元素。如果发现缺陷，则逐步确定其精确的空间坐标。以.....格式输出边界框，或返回‘None’。

表9：推理模型的原始提示词与优化提示对比。

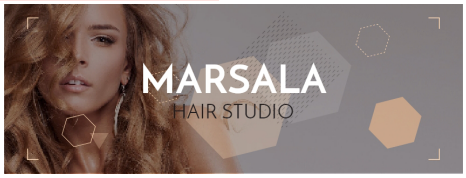
基准展示	
▷布局对齐、图形质量	
	
说明：背景清晰，文字居中对齐。	背景模糊，中间的文字 没有对齐。

表 10：AesEval 基准中布局对齐和图形质量的示例。

基准展示	
▷ 字体可读性	
	
说明：左下角的黄色字体 清晰可见。	左下角的黄色文字变得 模糊且无法辨认。

表 11：AesEval 基准中的字体可读性示例。

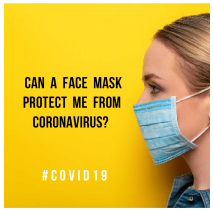
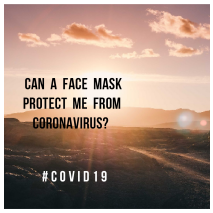
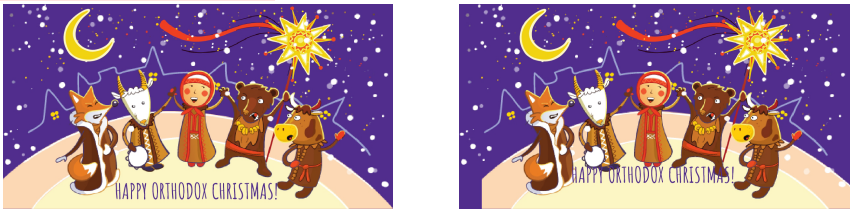
基准展示	
▷ 图形相关性	
	
说明：背景中有一位女性，穿着一件 mask, 这与主题相关。	背景是一张美丽的风景照片，但 与主题不符。

表 12：AesEval 基准中图形相关性的示例。

基准展示

▷布局-留白，布局-层叠



说明：文本位于空白区域，没有留白，也没有重叠。文本被移到了动物的脚边，因此下方留有空白，元素也出现了堆叠。

表 13：AesEval 基准中布局-留白和布局-层叠的示例。

基准展示

▷ 布局-平衡



说明：图片的整体布局平衡且协调。图片的背景向上移动，破坏了整幅图片的平衡。

表 14：AesEval 基准中的布局-平衡示例。

基准展示

▷ 布局-层级



说明：图片中的所有字体都保持一致，营造出层级感。图片中的字体受到干扰，看起来与前一个字体不在同一层级上。没有层级感。

表 15：AesEval 基准中的布局-层级示例。

基准展示	
▷ 图形质量、色彩对比	
	
说明：背景清晰，黑白色- 彩对比度良好。	背景被模糊，颜色从 从白色到黑色，与背景没有对比度。

表 16：AesEval 基准中的图形质量和色彩对比示例。

基准展示	
▷ 色彩和谐、色彩吸引力、色彩心理学	
	
解释：整幅图的色彩和谐且 美观。	中间的元素变成了绿色，这使得 整幅图看起来不协调，也缺乏吸引力。 绿色显得怪异，并会造成不良的心理影响。

表 17：AesEval 基准中的色彩和谐、色彩吸引力和色彩心理学示例。