arXiv:2602.01796v1 [cs.HC] 2 Feb 2026

CritiqueCrew：编排多视角对话式设计评议

Xiaojiao Chen 艺术与考古图像实验室，中华人民共和国教育部，浙江大学，中国浙江杭州
chenxiaojiao@zju.edu.cn

Jiahuan Zhou 艺术与考古实验室 图像，中华人民共和国教育部，浙江大学 中国浙江省杭州市
12021203@zju.edu.cn

Yunfeng Shu*中华人民共和国教育部艺术与考古图像实验室，浙江大学，中国浙江省杭州市
yfshu@zju.edu.cn

Ruihan Wang 中华人民共和国教育部艺术与考古图像实验室，浙江大学，中国浙江省杭州市
ruihanwang@zju.edu.cn

Qinghua Liu 中华人民共和国教育部艺术与考古图像实验室，浙江大学，中国浙江省杭州市
nico_yazawa@zju.edu.cn

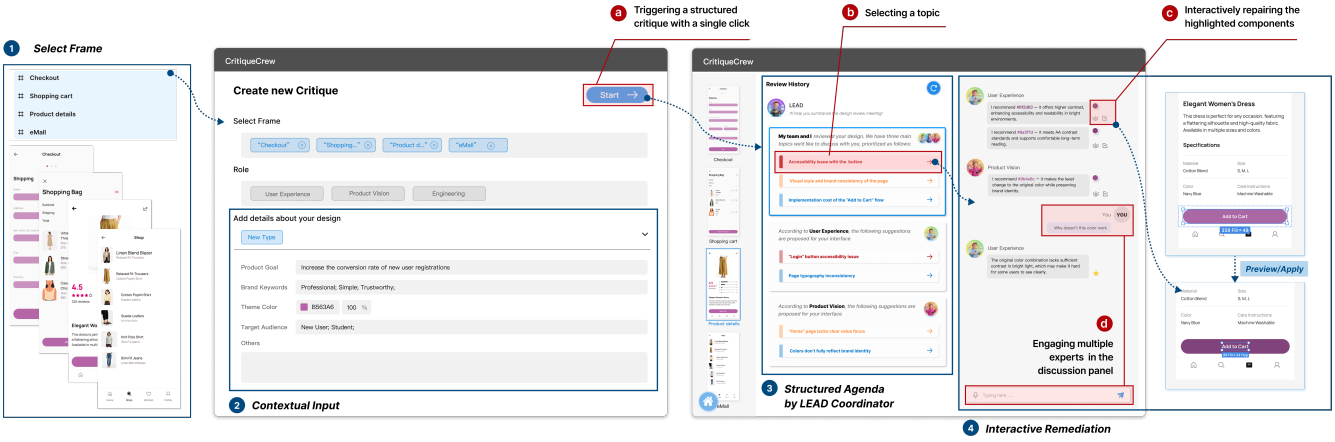


图 1：CritiqueCrew 将多视角评议流程直接嵌入 Figma。设计师首先选择目标画框，并提供产品目标和品牌关键词等上下文信息。随后，首席协调员将多个专家角色的评议综合为结构化议程。借助情境化修正工具和多视角对话，设计师可以顺畅地探索权衡并应用修改。这个流程将评议从静态检查清单转变为持续且可解释的协作，为设计师提供可操作的见解。

摘要

用户界面设计师在用户需求、业务目标和工程限制的交汇处，面临日益沉重的认知负荷和跨职能摩擦。现有自动化工具往往只提供静态的“问题列表”，缺乏可操作的修复方案，还会打断创作心流。我们推出 CritiqueCrew，这是一款通过对话式评议为设计师提供支持的 Figma 工具。CritiqueCrew 会生成通过实施对不同专家角色（用户体验、产品经理、工程师）的多视角编排，实现多维洞见。它借助上下文反馈与交互式修正，将抽象评议转化为具体行动建议。在两项独立对照研究（共计 $N = 48$）中，与传统静态检查器相比，CritiqueCrew 显著提升了设计质量与主观体验。此外，我们的结果证实，对专家角色进行结构化编排——而非采用统一模型——是培养信任与创造力支持的关键。我们的工作表明，人工智能如何通过将多视角对话与交互式修复相结合，从“问题审计者”转变为“解决方案共同创造者”，并为未来的创意工具提供设计启示。

*通讯作者。

本作品采用知识共享署名 4.0 国际许可协议。*CHI '26*，西班牙巴塞罗那© 2026 版权所有：著作权归所有者/作者所有。ACM ISBN
979-8-4007-2278-3/2026/04 https://doi.org/10.1145/3772318.3791391

CCS 概念
• **以人为中心的计算 → 交互系统与工具;人机交互实证研究**.

关键词
设计评议、对话式交互、人类–人工智能协作、交互式设计工具

ACM 参考文献格式：
Xiaojiao Chen、Jiahuan Zhou、Yunfeng Shu、Ruihan Wang 和 Qinghua Liu。2026。CritiqueCrew：统筹多视角对话式设计评审。载于《*2026*年*CHI*计算机系统人因大会论文集》（*CHI '26*），*2026*年*4*月*13*日至*17*日，西班牙巴塞罗那。ACM，美国纽约州纽约市，20页。
https://doi.org/10.1145/3772318.3791391

1 引言
当代设计实践日益体现为跨学科之间复杂的协商。设计师不再只是“视觉工匠”，而是逐渐转变为“系统协调者”，必须在用户需求、业务目标和工程限制之间取得平衡 [23, 36]。这种战略角色的提升让设计师在产品决策中拥有更大的影响力，但也带来了新的矛盾。例如，用户体验设计师和产品经理（PM）经常在产品愿景、用户研究优先事项和功能路线图方面存在分歧 [4, 14]。因此，设计师必须不断切换不同学科的思维方式——这一过程会带来沉重的认知负荷，削弱创意专注力，甚至可能导致职业倦怠 [27, 54]。

这种日益增长的复杂性促使学术界和业界纷纷探索自动化工具，以减轻设计师在重复性评估任务上的负担。早期工具主要侧重于单一维度的静态检查，例如可用性启发式评估或 WCAG 无障碍合规性 [12, 18, 37, 61]。然而，这些工具仅以“问题列表”的形式提供反馈，打断了设计师的创作流程 [10]，并割裂了问题与潜在解决方案之间的联系 [2, 11]。近年来，一些研究尝试利用大语言模型（LLM）改进评议流程，例如 UICrit [11] 能够生成结构化的设计问题报告。然而，总体而言，这些系统仍局限于静态、一次性的清单式反馈，未能与设计师的工作流程深度融合，也无法支持真实评议中所需的多维权衡与持续协商。

这凸显了一个严重的研究空白：在实际工作中，设计评议不仅关乎发现问题，更是一种充满协商、权衡与共同创造的社会过程 [31]。现有系统难以支持这一过程，主要有两个原因：（1）它们通常依赖单视角模型，难以呈现产品经理、工程师和设计师等内部利益相关者相互冲突的立场与目标；（2）它们提供的反馈大多是静态的，缺乏交互式的从评议到行动的闭环，也无法无缝融入设计师的实际工作流程。要克服这些局限，就需要重新界定人工智能的角色——从静态的问题审查者转变为动态的解决方案共同创造者 [39]——这也呼应了人机交互领域更广泛的呼吁，即重新构想人工智能作为创意工作中协作伙伴的角色 [8, 32, 34]。

为此，我们推出了集成于 Figma 的对话式评议工具 *CritiqueCrew* [15]。CritiqueCrew 将问题发现、解释和修复整合为一个无缝闭环，使人工智能能够贯穿整个过程，从识别问题到探索解决方案。它采用两项核心设计策略：（1）**多视角批评**，由不同的专家模块模拟跨职能团队，生成结构化反馈；（2）**赋能交互式修正**，通过画布高亮和实时预览等情境化交互，将抽象反馈转化为具体、可执行的建议，同时确保设计师对创意决策保有完全控制权。

我们通过两项受控研究（$N = 48$）评估了 CritiqueCrew 的有效性。研究1将我们的系统与一款先进工具进行基准比较，而研究2则单独考察多视角架构相对于统一专家基线的具体贡献。在本文中，我们从两个结果维度界定设计质量：**(1)问题覆盖度**和 **(2)解决方案有效性**（专家对视觉美学、可用性与合规性以及设计完整性的评分；见第 5.3 节）。这一界定有意与我们的多视角批评框架保持一致：角色专长视角旨在拓宽问题发现范围（从而提高**问题覆盖度**），而跨视角综合有助于形成更连贯且可执行的修正方案（从而提高**解决方案有效性**）。结果表明，CritiqueCrew 显著提高了问题覆盖度和解决方案有效性，同时也提升了设计师对工具的主观易用性评分与信任度。

本研究的主要贡献如下：

- CritiqueCrew，一个采用将多视角对话与上下文内交互式修复相结合的新颖方法的系统。该方法将设计评议转化为可执行工作流，为设计师提供多样化的跨职能视角。

- 两项用户研究（$N = 48$）的结果证明了多视角编排的益处。我们的研究结果表明，与统一专家模型相比，CritiqueCrew显著提升了设计质量与效率，同时降低了认知负荷、增强了信任，并提高了感知到的创意支持。

- 关于将人工智能的角色从“问题审查者”转变为“解决方案共同创造者”的关键洞见，为下一代赋能型创意支持工具提供了可执行的设计启示。

2 相关工作
2.1 人类与人工智能协作：从工具到伙伴
早期的人机交互愿景设想，计算机不应仅仅是工具，而应成为增强人类认知能力的协作者 [33]。数十年来，围绕交互系统和混合主动性交互 [21]开展的研究进一步确立了这样一种理念：人类与机器可以灵活地共享控制权，各自发挥互补优势，共同完成一项任务。随着大语言模型（大语言模型）的最新进展，这一愿景已变得更加具体：人工智能系统如今可以参与构思，生成不同的视角，并进行

来回推理，将自身角色从助手转变为共同创作的伙伴。

协同创作系统已在写作、视觉设计、编程及其他创意领域广泛应用，使人工智能从被动生成器转变为主动贡献者 [42, 46, 53]。这些系统有助于外化推理、为探索提供支架，并拓展用户的问题空间。在面向设计的工作流程中，人工智能也被用于弥合跨职能角色之间的沟通鸿沟，例如帮助设计师向开发者或利益相关者传达设计意图 [4]。这些研究共同呈现出一种转变：人工智能日益被视为能够提供领域相关视角的对话伙伴。

尽管此前的系统已探索人工智能如何协作完成创意任务，但对于人工智能如何参与设计评议——这一过程不仅涉及错误检测，还关乎阐释与设计方向的塑造——关注仍少得多。设计评议植根于艺术教育 [13], , 是一项基础实践，能帮助设计师明确创作意图，并评估他人如何解读其作品。由于评议通常通过小组对话互动 [38], 展开，因此它本质上具有社交性和阐释性，而非机械过程。这些特性使评议成为一个颇具潜力的场景；在此场景中，日益具备对话式多视角推理能力的新兴人工智能系统，不仅可以充当工具，还可以成为协作伙伴。

2.2 设计评议作为一种社会性与反思性实践

设计评议被广泛视为设计迭代的基础机制，与错误检测或性能优化截然不同 [3]。Schön 将设计描述为“与情境展开的反思性对话” [48],，在这一过程中，从业者在与不断演变的材料互动时揭示假设并重新界定问题。在此过程中，评议通过促使设计师阐明其选择背后的推理，并将默会判断转化为明确论证 [5, 16, 52]，支持行动中的反思。不同于旨在实现二元合规性的验证导向程序，评议有助于设计师应对模糊性并解读相互冲突的约束。

评议从根本上说也是一种社会性和协商性活动。设计认知研究表明，评价性讨论依赖于不同心智模型的对齐以及理由的交流 [43, 49]。在专业场景中，评议往往涉及跨职能参与者，例如设计师、工程师和产品经理，他们各自带来不同的评价标准。这些视角之间的相互作用会形成富有成效的张力，从而揭示盲点并促成战略性权衡 [7, 47, 58]。

这些理论视角表明，为评议提供计算支持必须超越静态规则检查。有效的系统不应只是被动的检查者，而应促进对话式参与，并近似模拟人类团队中的多视角推理。这一区别凸显了传统自动化评估工具与系统需求之间的差距：后者应能够支持高质量设计评议所必需的阐释性、往复互动过程。

2.3 自动化用户界面/用户体验评估

与上述设计评议所具有的阐释性和多视角特征相比，评估领域历来侧重于客观验证。传统方法，如启发式评估和可用性测试 [24], , 高度依赖专家检查来发现缺陷。尽管结构化评审有助于这一过程 [3], , 但它们仍然耗费大量资源且主观性很强，因此需要能够降低新手使用门槛的自动化解决方案 [30]。

为扩大评估工作规模，研究人员开发了利用计算机视觉和机器学习的工具，以严格按照规则验证设计。UIS-Hunter [61] 和 Nighthawk[35] 等系统运用深度学习，自动检测像素级的“设计异味”和 GUI 异常。另一些工具，例如 UIGuider [59] 和 UiAnalyzer [18], , 则通过构建知识图谱来评估不符合规范的风险。然而，这些方法将评估视为二元分类任务，缺乏处理行动中反思过程中特有的模糊约束所需的细腻度。

近年来，视觉语言模型（VLM）的出现，通过将视觉感知与语义推理相结合，为自动化评估开辟了新途径。Duan 等人 [12] 的研究表明，多模态模型可以结合用户界面截图与设计原则，生成详细建议；混合方法则将启发式方法的精准性与 VLM 的灵活性相结合 [37, 60]。除了静态分析，研究人员还采用智能体模拟虚拟用户行为，以动态评估可用性 [56]。与此同时，研究也致力于解决数据稀缺问题；UICrit [11] 等工作建立了大规模专业反馈数据集，以提升通用视觉语言模型的评议能力。尽管取得了这些进展，要超越表层视觉分析，实现对界面有效性的深入、行为驱动理解，仍是一项重大挑战 [25]。

归根结底，一个根本性的缺口依然存在：现有工具扮演的是问题审查者，而非评议伙伴。无论是采用启发式方法还是视觉语言模型（VLM），其输出都大多是静态的——问题清单或注释截图 [55]。这些系统缺乏**富有成效的张力**和**跨职能协商**能力，而这正是第 2.2 节所定义的；因此，设计师不得不手动弥合“分析”与“行动建议”之间的鸿沟。CritiqueCrew 通过将评估置于交互式多视角对话中，并在其中共同创造解决方案，来应对这一问题。

2.4 从自主模拟到多视角编排

大语言模型的最新进展重新激发了人们对智能体范式的兴趣。一个具有代表性的里程碑是 Generative Agents 框架[45], , 它融合了记忆与规划，以模拟社会行为。尽管这类方法在模拟方面效果显著，但专业设计工作流往往需要更有条理的协作机制，以确保可靠性。

因此，研究已转向功能角色编排，为大语言模型分配专门角色以分解任务。在生成式设计中，PrototypeFlow [62] 和 GUIDE[28]等系统使用中央控制器来协调模块，以实现结构化的

用户界面生成。同样，在评估中，多视角配置会模拟不同的利益相关者：PosterMate [50]生成受众反馈，而 PaperEval [22] 则分配验证角色以进行学术评估。这些研究表明，将问题分解为功能角色有助于改善结果。

然而，现有的大多数系统都采用单向流水线的方式生成静态输出。它们很少支持被认为是评议核心要素的协商式对话和富有成效的张力（第 2.2 节）。专业评议不只是模拟反应，还需要主动协调相互冲突的专家约束，并将其转化为可执行决策。

CritiqueCrew 通过在对话式框架内协调多视角协作来应对这一问题。我们没有依赖完全自主的智能体，而是实现了一种结构化的多视角架构，让不同功能角色通过中介式话语进行互动。这一设计选择优先考虑可解释性，而非不受约束的自主性，从而确保系统能够促进跨职能专家团队进行严谨推理。

3 形成性研究

为确保我们的系统设计真正植根于设计实践工作流程和核心痛点，我们开展了一项形成性研究。我们的目标是深入探究设计师、产品经理和工程师在实际产品开发流程中进行设计评议和反馈时面临的困难，并基于这些见解，探讨 AI 设计伙伴应发挥何种作用以及应具备哪些核心能力。

3.1 方法与参与者

我们共招募并访谈了 15 位在产品开发一线工作的专业人士。按照 Francis 等人提出的方法 [17], ，我们采用迭代式数据收集与分析流程，并以理论饱和作为停止准则。我们跟踪了每次访谈后新编码的出现情况。对前 12 位参与者进行详细分析后，我们确立了核心主题框架：第 12 次访谈之后，关于跨职能冲突不再出现新的一级编码。为确保研究严谨性，之后的三次访谈作为验证性案例。由于这些访谈验证了已有主题，却未带来额外的概念性见解，因此本研究的发现主要基于最初 12 位参与者提供的丰富且各具特色的数据。包括对全部 15 位参与者进行数据收集在内的完整研究方案，均经大学机构审查委员会（IRB）审查并批准。

参与者。在 12 位核心参与者中，我们纳入了用户界面（UI）设计师（P1、P2、P3）、交互设计师（P4、P5、P10）、产品经理（P6、P7、P11）以及前端工程师（P8、P9、P12）。他们的行业经验从 1 年到 10 年不等，任职机构规模各异，既有大型科技公司，也有初创团队。这种多样性确保收集的数据能够反映不同组织结构和产品阶段中的常见挑战。

访谈流程与数据分析。我们进行了半结构化访谈，每次持续 30–45 分钟。所有访谈均在线进行，并在征得参与者同意后录音并转录。访谈讨论围绕两个核心主题展开：

- **经验**：在端到端设计工作流中，参与者在设计评议过程中遇到了哪些真实的经验和痛点?
- **需求**：在了解“AI专家评审小组”的早期概念后，他们设想并期待理想的人工智能协作者具备哪些功能和形式?

我们按照 Braun 和 Clarke [6] 提出的六阶段指南进行了主题分析。我们采用混合归纳—演绎方法：首先从数据中归纳生成编码，再根据有关“经验”和“需求”的研究问题演绎完善主题。两位研究人员独立沉浸式阅读核心访谈记录，并生成初始编码。经过多轮讨论，这些编码得到反复完善并归并，最终形成稳定的编码手册。随后，我们将该编码手册应用于三次验证性访谈（P13–P15），以检验主题覆盖度。关于可靠性，我们优先采用协商共识，而非统计性的评定者间信度指标（例如 Cohen’s Kappa），这与人机交互领域的定性研究标准一致 [40]。所有分歧均通过讨论解决，最终归纳出贯穿整个数据集的三个核心主题。完整的编码手册，包括所有主题的定义和代表性引文，详见附录D。

3.2 研究结果

3.2.1 超越用户界面边界——设计师寻求主动的多维专家反馈，以应对跨职能冲突。我们的访谈一致表明，设计评议远不止于审美或可用性评估。相反，它是协商和调和来自跨职能利益相关者、尤其是产品管理与工程团队之间相互竞争的需求的主要场所。尽管这种协商至关重要，但也相当耗时。正如一位交互设计师（P5）所描述的那样，从内部评审到实现跨团队对齐，一轮完整的评议流程“可能需要大约五个小时。”如此巨大的时间投入，反映了弥合不同学科视角之间固有张力所需付出的劳动。处于这些相互竞争的约束交汇点上，设计师肩负着寻找最优解决方案的任务。这一位置凸显了在设计流程早期获得主动且整合的专家反馈的严重需求，以便预见并缓解此类冲突。以下两种冲突情境尤其凸显了这一需求：

- **权衡工程可行性与设计创新**：工程实用主义与设计创新之间出现了明显的张力。一位前端工程师（P8）讲述了一个案例：设计师提议使用非标准字体，以实现独特的视觉效果。从设计师的角度来看，这是对视觉创新的追求，旨在提升用户体验。然而，工程师回应称，尽管技术上可以实现，这种做法却会违反既定的编码规范，并造成大量技术债务，使后续维护更加复杂。这个案例表明，设计师职责不仅限于构思，还包括在新颖的设计

理念与工程架构的长期可扩展性和稳定性之间进行复杂权衡。

- **在业务需求与设计系统指南之间进行协商**：另一个反复出现的冲突，是如何在眼前的业务目标与长期的设计一致性之间取得平衡。一位UI设计师（P2）描述了自己与产品经理的一次协商，对方坚持将一项严重的业务功能放在“界面顶部最显眼的位置。”对这位设计师而言，这一要求不仅违背了既定的设计系统指南，还需要创建一个非标准组件，从而损害产品一致性。设计师发现自己不得不在短期业务目标与用户体验的长期完整性之间寻求平衡。正如P2所说，这种冲突迫使团队“花很多时间召开多次会议”以达成妥协。

总的来说，这些反复出现的张力表明，设计师在协作项目中扮演着关键的冲突调解者角色。他们需要的不只是零散、单向的反馈——他们需要一个虚拟专家顾问团，能够模拟真实的决策场景，并整合来自工程、产品愿景和用户体验等方面的观点。设计师希望及早预判、揭示并化解潜在冲突，将团队资源从代价高昂的后期返工转向价值更高的创意创新。

3.2.2 超越问题清单——设计师需要情境化的解决方案来推动迭代。 我们的研究结果表明，仅识别问题的反馈被认为是不够的。正如多位设计师所强调的，最有价值的评议不仅诊断问题，还提供通向解决方案的可执行路径。一位高级设计师（P3）清楚地阐述了这一点，她将理想反馈定义为“告诉我问题在哪里，并给我几个可行的修复选项。”她讲述了一个实例：对她的设计“太乱”的模糊评议，在主管识别出根本原因（图标颜色过多）并给出具体补救措施后变得具体。这体现了一种有效的反馈结构，我们将其概括为三部分流程：诊断（“是什么”）、定位（“在哪里”）和建议（“为什么”）。

对完整反馈闭环的这种期望，也延伸到设计师对人工智能工具的期待。一位初级设计师（P10）设想的理想工具，应超越标记对比度错误，并主动“推荐若干符合 *WCAG* 的安全颜色值。”其根本目标是卸下反复试错的繁琐工作，并加速创造性问题解决。因此，人工智能伙伴若要真正有效，就不能只是被动的评议者。它必须充当协作式问题解决者，通过交付的不仅是数据，还有果断、可实施的建议，提供一条从问题检测到创造性解决方案的支架式路径。

3.2.3 超越审计工具——设计师寻求赋能且值得信赖的伙伴关系，以支持创作自主权。关于人机协作的讨论主要围绕信任与控制这两个相互交织的主题展开。参与者强调，他们是否采用某个人工智能工具，取决于该工具能否建立一种助力而非支配其创作过程的伙伴关系。

建立这种信任有两大支柱。第一是透明度和可解释性 [1]。正如交互设计师 P5 所说，信任并非一蹴而就，而是逐步建立的。他强调，当人工智能能够为其建议给出理由时，信任便会逐渐形成。他表示：“如果人工智能能为它的建议给出理由，那就能建立起我的信任。”这凸显了设计师理解人工智能建议背后基本原理的严重必要性。因此，推理过程不透明的“黑箱”系统不太可能获得认可。第二大支柱是可靠性 [1]。P3 指出，当人工智能误解提示或生成无关输出（通常称为人工智能幻觉 [26], ）时，其可信度会立即受损。没有持续稳定且可靠的性能，信任便无法维持。

除信任之外，设计师也强调保有控制权和专业自主权。这一原则体现为对赋能而非取代的诉求，设计师坚持自己拥有最终决策权。正如 P3 恰如其分地所说，理想的人工智能是“一位不会干涉我做决定的顾问。”落实到实际中，这意味着人工智能应提供多样化选项、阐明权衡，并助力探索，同时始终将最终判断权留给人类设计师。从本质上说，设计师寻求的是可靠、透明且尊重其自主权、增强其能力的协作者，而非削弱其权威的自动化审计工具。

4 系统设计

我们的形成性研究揭示，设计师在与自动化工具协作时，深切渴望获得多维度洞见、情境化补救以及能够赋能的伙伴关系。为系统地回应这些期待，我们开发了 CritiqueCrew，这是一款集成于 Figma 的工具，旨在将人工智能从单纯的“问题审查者”转变为“解决方案共同创造者”。本节首先介绍系统的设计目标与整体架构，随后详细阐述其核心多视角评议模型，最后展示用户如何通过完整的交互工作流程与系统协作。

4.1 设计目标

基于形成性研究的三项核心发现，我们为 CritiqueCrew 确立了三个设计目标：

DG1：融入多视角、专家级评议。CritiqueCrew 不依赖单一视角，而是协调不同的功能角色，生成综合考虑用户体验、产品愿景和工程可行性约束的反馈。

DG2：将评议无缝转化为可行动的解决方案。CritiqueCrew 不提供抽象建议，而是提供交互式修正工具，让设计师能够在设计情境中直接预览并应用修改。

DG3：建立赋能且互信的合作关系。为建立信任，系统将交互设计为对话式且易于理解的形式。系统会明确阐述权衡与理由，同时确保设计师始终是最终决策者。

4.2 系统架构

CritiqueCrew 采用客户端—服务器架构，旨在将先进的分析能力与流畅的前端体验紧密集成。整体系统架构如图 2 所示。

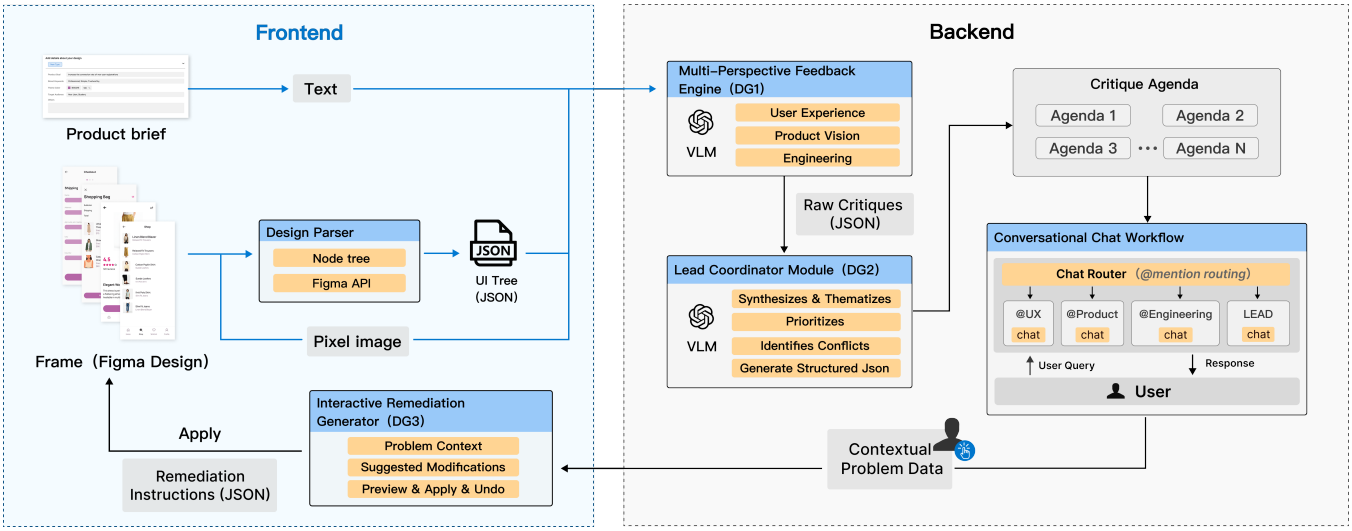


图 2：CritiqueCrew 的系统架构

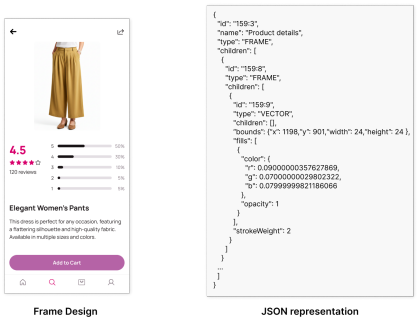


图 3：框架设计与 JSON 表示

前端是一个使用 React [41] 构建的轻量级 Figma 插件。它是用户交互的主要界面，负责呈现评审议程和一系列交互式修正工具。借助 Figma 插件 API，前端可以实时读取画布中的设计图层数据，并将用户行动建议（例如突出显示元素或更新属性）写回画布，从而确保工作流程流畅。

后端采用模块化设计，由四个核心组件构成，用于协调评议流程：

- **设计解析器**：将 Figma 复杂的原生节点树转换为适合大语言模型处理的结构化 JSON 格式。该 JSON 不仅保留界面的层级结构，还整合了语义和视觉属性（例如文本标签、位置、尺寸和色彩），使原始设计数据能够被下游分析解读。图 3 展示了一个简单的框架及其 JSON 表示。

- **多视角反馈引擎**：多维度评议的核心（DG1）。该引擎基于 Lang- Graph [29] 框架实现，以确保执行流程具有确定性；它实例化了由 OpenAI 的 GPT-4o [44]驱动的并行专家模块。每个模块都会评估解析后的设计数据，并从不同的功能视角（例如用户体验、产品愿景、工程可行性）进行分析，生成专项评议，随后将其汇总为一份全面的评审。

- **首席协调员模块**：作为系统的“编排中枢”，旨在为设计师提供连贯的指导（DG3）。它综合来自不同专家模块的原始反馈，将其归纳为各个主题，为问题排定优先级，并明确指出冲突或权衡。此过程将零散的见解转化为结构化、可追溯的决策支持。此外，该模块还负责管理一种轻量级路由机制，在交互式对话中将特定于上下文的查询（例如，@UX、@PM、@Engineer）发送给相应的专家模块，同时作为一般咨询的默认响应方。

- **交互式修正生成器**：这是连接评议与行动建议的关键桥梁（DG2）。根据每个已识别问题的上下文，此组件会动态生成修复建议，并生成前端所需的用户界面代码。设计师随后可以直接在画布上预览、应用或撤销修改，将抽象的反馈转化为即时解决方案，同时不打断其工作流程。

4.3 多视角评议模型

为了实现我们的设计目标，我们开发了一个**多视角评议模型**，它由专门的专家模块组成，这些模块近似于跨职能评审团队的推理过程，同时配备一名负责综合与赋能的首席协调员。为确保每个模块遵循其特定的功能视角，我们设计了严谨的系统提示词（详见附录 B）。

4.3.1 专家模块：近似跨职能视角。为实现 DG1，我们的模型编排了一个由三个核心专家模块组成的“虚拟评审小组”。这些模块旨在模拟现实世界产品开发中多样化的评价标准，提供兼顾深度与广度的反馈：

• **用户体验模块**：用户的守护者。该模块综合了用户研究员、UI设计师和无障碍专家的角色。它依据尼尔森可用性启发式和 WCAG 标准等行业基准，评估可用性、信息架构、一致性和无障碍性。其指导性问题是："该设计对所有用户而言是否清晰、高效且具备无障碍性？"

• **产品愿景模块**：目标导航者。该模块代表产品经理和业务利益相关者的战略视角。设计师可以提供产品目标（例如，提高新用户注册转化率）或品牌关键词（例如，专业、简约）。基于这些背景信息，该模块运用零样本推理，评估设计是否符合业务目标、产品战略和品牌形象。其指导性问题是："这项设计能否有效推动业务目标，并忠实传达我们的品牌价值观？"

• **工程模块**：务实的工程师。该模块体现了前端开发人员所面临的限制，充当技术把关者。它侧重于评估工程可行性、估算实施成本，并识别潜在的性能瓶颈。其指导性问题是："这项设计能否利用当前技术栈高效实现？会对成本和性能产生什么影响？"

为确保一致性和互操作性，每个模块的输出都遵循统一的结构化格式：issue_id、source_role、component_target（链接到 Figma 图层 ID）、severity、critique_text、rationale、suggestion。这种结构化表示使人工智能生成的见解可计算、可排序，并可直接转化为行动。

4.3.2 首席协调员：从审查者到赋能者。为实现 DG3，我们引入首席协调员模块来统筹整个评议流程。其首要任务并非做出最终判断，而是通过以下三个关键步骤，赋能设计师，使其成为最终决策者：

• **综合与归纳主题**：根据相关用户界面组件，汇总不同专家模块的原始反馈。这一过程将原本可能零散的评议意见转化为结构化、便于浏览的议程，有效降低设计师的认知负荷。• **确定优先级**：根据问题严重程度对议程进行排序。例如，严重的无障碍障碍（如违反 WCAG AA）会优先于轻微的美学建议，引导设计师优先处理影响最大的问题。

• **呈现冲突与权衡**：当建议出现分歧时——例如，产品模块建议使用品牌色来提高转化率，而用户体验模块则指出其对比度不足——首席协调员不会强行给出所谓的正确方案，而是明确呈现其中的权衡（例如，视觉突出度与无障碍性），并将决策交由设计师，确保这种张力能够促进思考，而非令人无所适从。

4.4 交互工作流程

CritiqueCrew 的核心交互工作流程被设计为一个流畅的"调用—分析—交互—解决"闭环，将设计评议从一次性事件转变为融入创作过程的持续对话。交互工作流程如图4所示。

(1) **调用评议并查看议程**：设计师在 Figma 中完成草稿或某个里程碑后，只需单击一次即可调用 CritiqueCrew，开启评审会话（图4）。系统会立即收集所选框架的图层数据，并发送给后端专家模块进行并行分析。几秒钟内，首席协调员便会综合结果，并在侧边栏中呈现结构化评审议程。该议程不仅为后续改进提供了清晰的路线图，也从一开始就建立起多维结构化对话（DG1），为建立信任（DG3）奠定基础。(2) **通过交互式修正工具结合上下文提供反馈**：随后，设计师可以从议程中选择任意问题，开始处理（图4）。选中问题后，系统会在画布与反馈之间建立紧密联系：Figma 画布会自动缩放并突出显示相应的用户界面组件，而侧边栏则切换为显示详细反馈、理论依据以及结合上下文的交互式修正工具（图4）。例如，如果检测到色彩对比问题，系统会提供符合 WCAG 标准的调色板。设计师可以即时预览更改，并通过单击一次应用更改，使问题定位、方案探索和设计执行无缝衔接，同时不打断创作流（DG2）。(3) **通过多视角对话解决冲突**：对于涉及权衡的复杂问题，设计师可以打开专门的主题对话面板（图4）。在此，系统不仅会呈现不同功能角色给出的最终建议，还会总结其不同观点，让设计师仿佛置身于一场透明的评审会议，各种不同视角都能得到充分展现。此外，设计师可以使用 @ 功能直接向特定角色提问，例如："@Engineer 为什么这种动画的实现成本很高？"这种机制将人工智能传统的"黑箱"推理转变为透明且人人参与的审议过程，揭示设计决策背后的矛盾，让设计师能够做出最终判断，同时强化彼此信任的合作关系（DG3）。

5 评估

我们采用了顺序评估策略。研究1评估CritiqueCrew相对于业界现行工具的整体有效性。受所观察到的显著性能提升的推动，我们随后设计并开展了研究2——使用一组独立的参与者以及新的设计任务——以解构这些改进的来源，具体将我们的多视角架构的贡献与交互

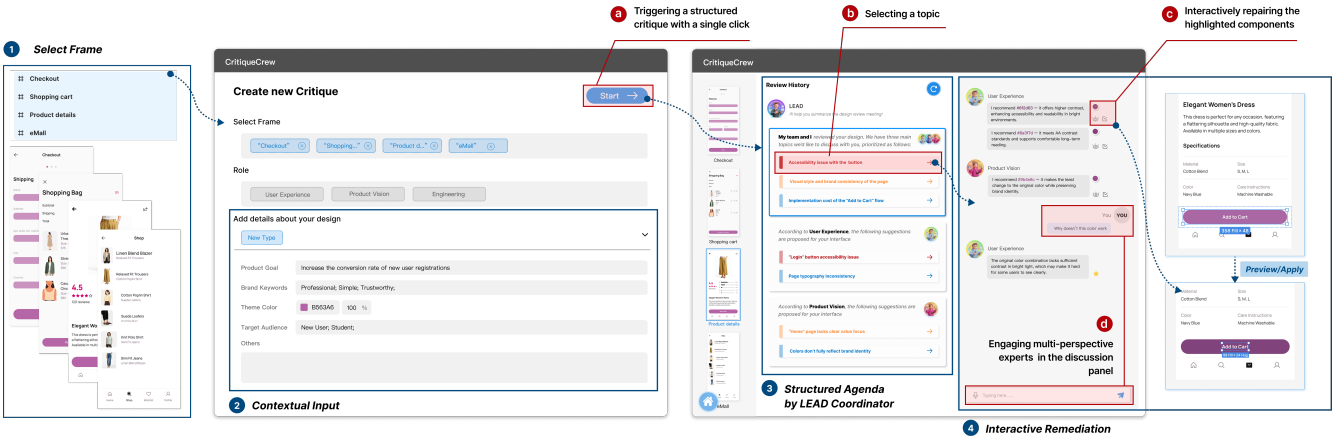


图4：交互工作流程

模态分离开来。这一两阶段方法旨在回答以下研究问题：

(1) **RQ1：设计质量**

完整的 CritiqueCrew 系统与当前最先进的工具相比表现如何？**专业专家角色的编排**是否确实比单一通才模型带来更好的结果？(2) **研究问题2：用户体验** CritiqueCrew 的对话式工作流如何影响设计师的主观体验，包括可用性、认知负荷、信任和感知到的创造力支持？(3) **研究问题三：交互模式**与静态分析工具相比，设计师如何看待并利用 CritiqueCrew 中不同功能角色提出的评议，尤其是在这些角色的观点相互冲突时？

5.1 研究1：CritiqueCrew 与基线（SpecAI）

研究1将 CritiqueCrew 与具有代表性的静态评估插件 SpecAI 进行比较。此比较旨在评估我们完整系统及其核心对话式范式的实际价值。

5.1.1 研究设计。我们采用被试内设计，每位参与者都在两种不同条件下完成一项设计优化任务。此设计有效控制了个体差异（例如设计经验）。

条件1：基线（SpecAI）。为了将 CritiqueCrew 与当前先进的自动化评估工具进行比较，我们选取了一款真实的 Figma 社区插件，名为 *SpecAI* [2]，作为基线。该插件利用人工智能提升设计质量、可访问性和文案撰写水平，并在侧边栏中以静态问题列表的形式呈现反馈。这种交互模式忠实地代表了我们希望改进的“审查员式”工具。在本研究开展时，SpecAI 约有 2,400 名用户，因而能够代表当前实践中先进的静态分析工具。

条件 2：CritiqueCrew。在此条件下，参与者使用完整的 CritiqueCrew 系统，并与多个专家模块进行对话，并通过画布高亮和实时预览等交互探索修改。两种工具均配置为检测相同类别的问题，以确保比较公平。

5.1.2 参与者。我们进行了先验功效分析（G*Power），假定效应量较大（$d_z = 0.8, \alpha = .05, 1 - \beta = .80$），这与以往对人工智能辅助设计工具的评估相一致。分析表明最低样本量为15。为确保估计结果稳定，我们的目标是每项研究至少招募24名参与者。

我们成功为研究1从在线设计社区招募了24名专业用户界面设计师（15名女性，9名男性）。参与者的平均年龄为26.8岁（$SD = 1.77$），平均拥有3.5年（$SD = 0.9$）的专业设计经验。招募通过专业人脉网络进行，两项研究之间没有参与者重叠。每位参与者获得一张礼品卡作为报酬。研究方案经大学机构审查委员会（IRB）审查并批准。

5.1.3 任务与材料。我们为实验准备了两份多帧Figma设计稿：一个在线教育平台的课程详情页，以及一个电子商务应用的结账流程。为呈现真实的设计产物，每份设计稿由连续的4至5个框架组成，并被分配到两种实验条件之一。两份设计稿均植入了数量相近的预定义问题（约16个）。这些问题有意覆盖本系统三个**视角**（用户体验、产品愿景和工程可行性）的核心关注领域，并包含若干潜在冲突，以评估设计师在面临权衡时如何做出决策。

为确保任务等效，我们邀请两位外部设计师开展了试点研究，他们确认完成两项任务所需的认知负荷和时间相当。研究期间，参与者在每种条件下各有 20 分钟，使用提供的工具，在指定的模型稿中尽可能多地发现并解决问题。我们采用拉丁方设计对任务和条件的顺序进行平衡，以减少顺序效应。

5.1.4 流程。 整个研究历时约 80 分钟，通过视频会议在线进行。流程分为四个部分：引言和知情同意（5 分钟）、两种工具各自的教程和正式任务（各 30 分钟），以及最后的半结构化访谈（15 分钟）。任务期间，我们鼓励参与者边思考边说出自己的想法。访谈重点了解他们使用两种工具的比较体验，以及他们的决策过程。

5.2 研究2：角色专业化分析

虽然研究1证明了CritiqueCrew优于静态分析工具，但这一比较将多个变量捆绑在一起（后端架构 + 对话界面）。为严谨地识别成功的关键因素，研究2通过控制所有交互可供性，单独考察角色编排机制。

5.2.1 研究设计。研究2采用了严格独立的参与者样本（无人参与过研究1）。为确保研究结果并非特定设计文件所致，同时保持领域复杂度的可比性，我们开发了两套**全新**多画板 Figma 原型。这些原型保留了与研究1相同的行业主题（在线教育和电子商务），但内容和布局结构完全不同。每套原型都植入了约16个难度相当的问题。本研究采用被试内设计，每位参与者均完成两项任务，一项使用统一专家系统，另一项使用多视角系统，且任务顺序经过平衡。

条件 1：CritiqueCrew（统一专家）。在此条件下，后端配置为运行一个 GPT-4o 实例，并将其设定为“全面设计专家”。模型接收了一份合并后的指南，该指南将用户体验、产品愿景和工程可行性的标准整合到一条系统提示词中。各条件下的前端交互保持一致。与多视角设置不同，评议生成未按角色区分；聚合和优先级排序由统一专家在单次响应中隐式完成，而非通过单独的显式协调层进行。

条件 2：CritiqueCrew（多视角）。在此条件下，参与者使用完整系统，其中用户体验设计师、产品经理和工程师角色由首席协调员显式编排。评议由不同的专家模块生成，从而能够显式识别冲突并进行权衡推理。首席协调员通过整合跨角色反馈并呈现冲突，使综合与优先级排序过程显式化。

5.2.2 参与者。研究2招募了另一组24名专业用户界面设计师（11名女性，13名男性）。如前所述，这组样本与研究1完全独立。他们的人口统计特征相近：平均年龄为27.5岁（$SD = 2.1$），专业经验为3.2年（$SD = 1.1$）。所有参与者均获得标准报酬，且研究方案经大学伦理委员会审查并批准。

5.3 数据收集与分析

我们采用混合方法，同时收集定量数据和定性数据。

定量数据。 我们采用多维评估框架对设计质量进行操作化定义：

- **问题覆盖度**：参与者成功解决的预定义问题数量。鉴于任务时长固定，该指标也可作为效率的替代指标，反映参与者在规定时间内的产出。
- **解决方案有效性**：三位不了解实验条件的资深设计专家，独立评估了所有匿名化最终设计。在近期多维度用户界面评估指标的基础上，专家依据结构化质量模型 [11, 57, 63], ，评估了三个维度：**视觉美感**、**可用性与合规性**和**设计完整性**。他们使用7点评分（1 = 非常差， 7 = 非常好）对解决方案的有效性进行单项评分。
- **主观问卷**：我们收集了系统可用性量表（SUS）（仅研究1）、NASA任务负荷指数认知工作负荷量表以及人工智能信任量表（TAI）的得分 [19]。对于研究2，我们还加入了创造力支持指数（CSI）[9] ，以支持创造性探索。

在统计分析中，我们首先进行 Shapiro–Wilk检验，以评估配对差值的正态性。满足正态性假设时，我们使用配对样本t检验；否则使用 Wilcoxon符号秩检验。我们报告检验统计量、p值以及相应的效应量（t检验采用 Cohen’s d；Wilcoxon检验采用秩双列相关r）。

定性数据。定性数据包括出声思考环节和访谈的文字记录。两名研究人员使用主题分析对文字记录进行编码。分析重点在于了解设计师如何解读并权衡不同功能角色提供的反馈，以及结构化对话如何影响他们的决策。

6 结果

本节介绍用户研究的发现，以回答我们的三个核心研究问题。我们首先报告关于设计质量（RQ1）和主观用户体验（RQ2）的定量比较结果，随后深入探讨设计师与 CritiqueCrew 交互模式的定性见解（RQ3）。

6.1 RQ1：CritiqueCrew 提升设计质量与效率

我们的第一个研究问题（RQ1）探讨 CritiqueCrew 是否能带来更好的设计成果。我们报告了两个关键指标的结果：**问题覆盖度**（效率）和**解决方案有效性**（质量）。统计结果的完整摘要见表 1，视觉对比见图 5 和图 6。

6.1.1 问题覆盖度。首先，就效率而言，与基线工具相比，参与者使用 CritiqueCrew 时识别出的问题显著更多（$M = 9.58$ $SD = 1.86$ $M = 7.96$ ，）。由于数据不符合正态性假设，我们进行了 Wilcoxon符号秩检验，结果证实这种提升具有统计显著性（$W = 49.0$， $p = .006$， $r = .59$）。

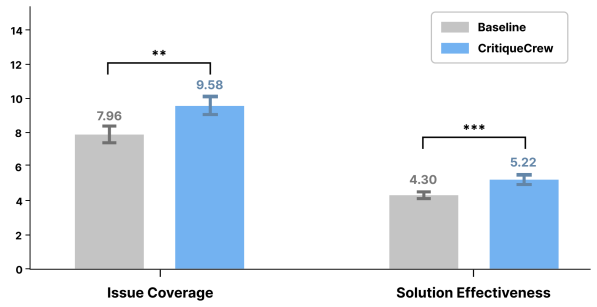


图5：在真实设计任务中，CritiqueCrew 与基线工具的设计成果质量比较。与基线工具相比，CritiqueCrew 带来了更广的问题覆盖度（左）和显著更高的解决方案有效性（右）。误差线表示标准误。$p < .01$，***$p < .001$。**

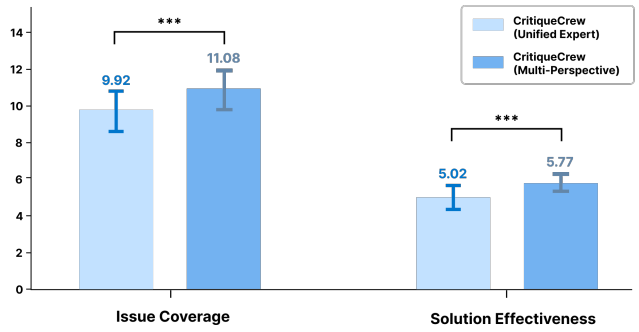


图6：两种 CritiqueCrew 模式的设计成果质量比较：统一专家与多视角。多视角模式在问题覆盖度（左）和解决方案有效性（右）上均显著更高，证实了基于角色的评议分解的益处。*$p < .001$。**

角色专门化效应（研究2）。为确定这一提升的来源，研究2比较了多视角模式与统一专家模式。结果显示，多视角模式的问题覆盖度显著更高（$M = 11.08$，$SD = 1.18$），而统一专家模式的问题覆盖度较低（$M = 9.92$，$SD = 1.18$）。这一差异具有统计显著性，且效应量较大（ $W =$ 9.0 $p < .001$ $r = .82$ ，，）。这表明，将评议分解为专业化角色有助于发现更广泛的问题。

6.1.2 解决方案有效性。 CritiqueCrew 的优势在解决方案有效性这一严重指标上更为显著。三名独立专家的盲评显示出极佳的评定者间信度（ICC(2,*k*) = 0.8），表明专家判断高度一致。

使用 CritiqueCrew 得出的解决方案 ($M = 5.22$，$SD = 0.98$) 的评分显著高于基线方案 (4.30 $SD = 0.47$ t t 23 ，)。配对 - 检验确认了这一改进 (( ) = 3.90 $p < .001$ $d = 1.19$ 95 [0 43, 1 40] ，，，% CI . .)。

角色专门化效应（研究2）。研究2进一步支持了这一发现：多视角模式的设计 ($M = 5.77$，0.49 SD) 显著优于统一专家模式的设计 ($M = 5.02$，$SD = 0.67$)。这一差异具有显著性 ($W =$ 5.0 $p < .001$ $r = .85$ ，，)，且评定者间信度同样较高（ICC(2,*k*) = 0.85）。专家指出，多视角反馈带来了更具战略性和整体性的设计优化。

总之，我们的结果表明，CritiqueCrew 的多视角范式全面提升了设计成果的数量（问题覆盖度）和质量（解决方案有效性）。至关重要的是，研究2证实，这些提升并非仅仅是界面造成的，而是由角色编排机制本身驱动的。通过将评议拆解为不同的功能视角，系统帮助设计师超越表层修正，转向更深入且具备权衡意识的优化。

6.2 研究问题2：CritiqueCrew 提升主观用户体验

研究问题2考察了 CritiqueCrew 的对话式工作流对设计师主观体验的影响。问卷数据显示，与静态基线相比，CritiqueCrew 在系统可用性、认知负荷和对人工智能的信任方面表现出显著优势（研究1；见图7）；此外，与统一专家变体相比，其多视角配置进一步优化了这些指标（研究2；见图8）。表2总结了关键统计数据。

6.2.1 系统可用性量表（*SUS*）。在研究1中，CritiqueCrew 的系统可用性量表平均分为 77.29（$SD = 15.20$），而基线工具的得分为 55.52（$SD = 9.92$）。配对 t 检验证实，这一差距具有高度显著性（$t(23) = 6.37$，$p < .001$，$d = 1.30$）。基线得分低于 68 分的行业基准，而 CritiqueCrew 则明显超过这一门槛，表明该系统可用性很高，且学习曲线较短。

6.2.2 认知负荷（*NASA*任务负荷指数）。在研究1中，与基线相比，参与者使用 CritiqueCrew 时报告的认知负荷显著更低（ $M =$ 40.57 $SD = 9.75$ $M = 51.99$ $SD = 8.24$ ，），降低了 **21.9%**（$t(23) = -5.24$，$p < .001$，$d = -1.07$）。

角色专业化效应（研究2）。研究2进一步阐明了该架构的作用。与统一专家模式（$M = 59.9$，$SD = 6.24$）相比，多视角模式产生的工作负荷更低（$M = 54.9$，$SD = 5.47$）。这一差异具有显著性（$t(23) = 2.74$，$p = .012$，$d = 0.56$）。这些结果表明，尽管引入了多个视角，结构化编排仍能有效管理复杂性，而非增加认知负担。

6.2.3 对人工智能的信任（*TAI*）。CritiqueCrew 的信任评分显著高于基线（$M = 3.76$，$SD = 0.68$），而基线的评分为（$M = 2.89$, $SD =$ 0.37 t 23 $= 6.13$ $p < .001$ $d = 1.25$ ；（），，）。

角色专业化效应（研究2）。多视角评议带来的信任度（$M = 2.93$，$SD = 0.36$）高于统一专家（$M = 2.71$，$SD = 0.28$）。配对 t-检验确认了这一效应（$t(23) = -3.81$，$p < .001$，$d = 0.78$），表明角色差异化的评议能够增强用户信心。

表 1：RQ1 中问题覆盖度和解决方案有效性的定量结果摘要。

研究	指标	描述性统计	Test	效应量
1	问题覆盖度	基线： $M = 7.96$， $SD = 1.71$ CC： $M = 9.58$， $SD = 1.86$	威尔科克森 W $= 49.0$, $p = .006$	$r = .59$
1	解决方案有效性	基线： $M = 4.30$， $SD = 0.47$ CC： $M = 5.22$， $SD = 0.98$	$t(23) = 3.90$, $p < .001$	$d = 1.19$
2	问题覆盖度	统一： $M = 9.92$， $SD = 1.18$ 多项： $M = 11.08$， $SD = 1.18$	威尔科克森 W $= 9.0$, $p < .001$	$r = .82$
2	解决方案有效性	统一： $M = 5.02$， $SD = 0.67$ 多项： $M = 5.77$， $SD = 0.49$	Wilcoxon W $= 5.0$, $p < .001$	$r = .85$

注： 基线 =基线（SpecAI）， CC =CritiqueCrew， 统一 =统一专家， 多项 =多视角。

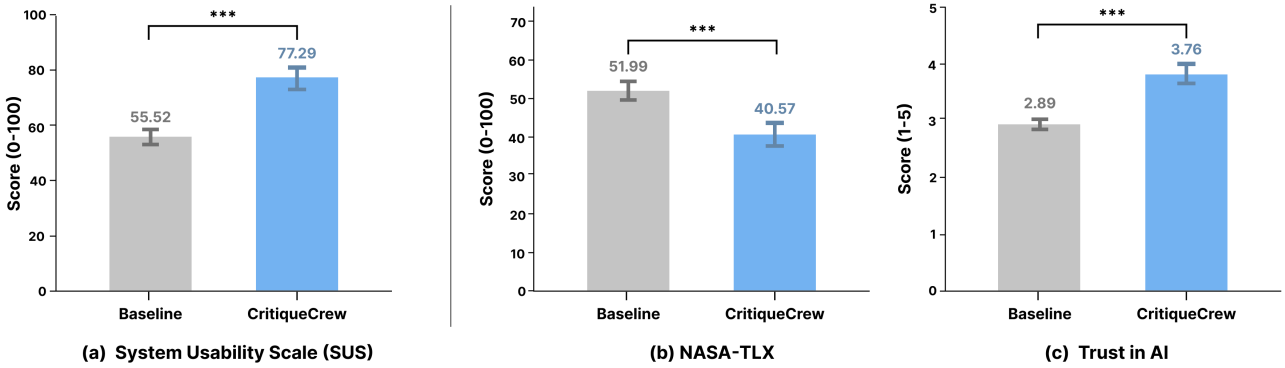


图7：CritiqueCrew 与基线工具（SpecAI）的主观体验指标比较。(a) 系统可用性量表（SUS，0–100；仅研究1）。(b) NASA任务负荷指数的心理工作负荷（0–100；数值越低越好）。(c) 对人工智能的信任（TAI，1–5）。由于仅在研究1中收集了 SUS 数据，而 NASA 任务负荷指数和 TAI 评分汇总了参与者在所有使用基线工具和 CritiqueCrew 的研究环节中的回答，因此面板 (a) 与面板 (b) 和 (c) 在视觉上分开呈现。在所有指标上，参与者都认为 CritiqueCrew 的可用性更高、心理工作负荷更低，也比基线工具更值得信任。* $p < .001$。**

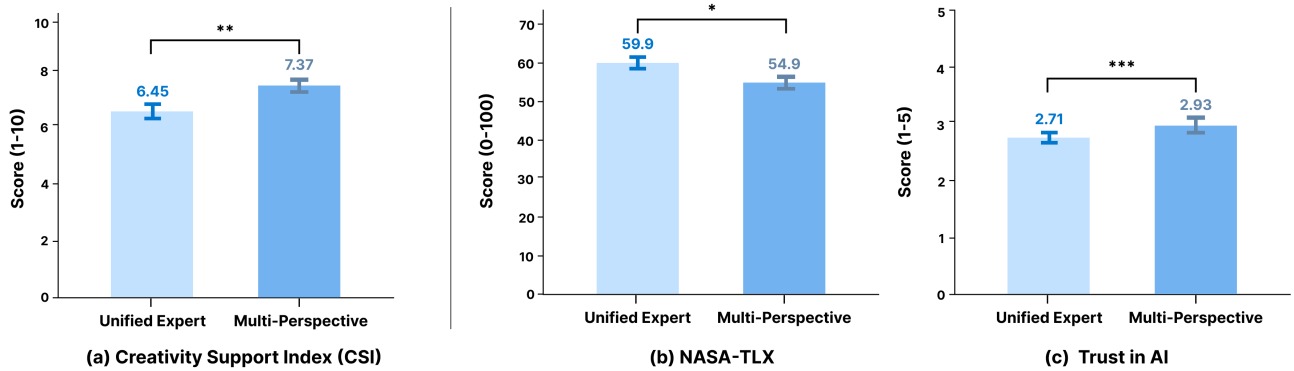


图8：研究2中 CritiqueCrew 的统一专家模式与多视角模式的主观体验比较。(a) 创造力支持指数（CSI，1–10；仅研究2）。(b) NASA任务负荷指数的心理工作负荷（0–100；越低越好）。(c) 对人工智能的信任（TAI，1–5）。与统一专家模式相比，多视角模式的 CSI 得分更高、NASA任务负荷指数得分略低，对人工智能的信任评分也更高，这表明有多个专家角色可供选择时，参与者感受到更强的创造力支持、更低的认知负荷，也更愿意依赖系统的建议。* $p < .05$， $p < .01$，*** $p < .001$。**

表2：研究问题2中关于主观体验的量化结果摘要。

研究	指标	描述性统计	Test	d	95% 置信区间
1	SUS	基线：$M = 55.5$，$SD = 9.9$；CC：$M = 77.3$，$SD = 15.2$	$t(23) = 6.37$，$p < .001$	1.30	[15.1, 28.5]
1	NASA任务负荷指数	基线：$M = 52.0$，$SD = 8.2$；CC：$M = 40.6$，$SD = 9.8$	$t(23) = -5.24$，$p < .001$	-1.07	—
1	TAI	基线：$M = 2.89$，$SD = 0.37$；CC：$M = 3.76$，$SD = 0.68$	t（23）$= 6.13$，$p < .001$	1.25	—
2	NASA任务负荷指数	统一：$M = 59.9$，$SD = 6.2$；多项：$M = 54.9$，$SD = 5.5$	t（23）$= 2.74$，$p = .012$	0.56	—
2	TAI	统一：$M = 2.71$，$SD = 0.28$；多项：$M = 2.93$，$SD = 0.36$	$t(23) = -3.81$，$p < .001$	0.78	—
2	CSI	统一专家：$M = 6.45$，$SD = 0.91$；多视角：$M = 7.37$，$SD = 0.69$	$t(23) = -3.11$，$p = .005$	0.64	—

注：基线 = 基线（SpecAI），CC = CritiqueCrew，统一专家 = 统一专家，多视角 = 多视角。SUS = 系统可用性量表；NASA-TLX = 任务负荷指数；TAI = 对人工智能的信任；CSI = 创造力支持指数（如论文中所定义）。效应量为 Cohen's d；负值 d 表示第二个列出的条件（根据行的不同，可能为 CC 或多视角）的得分较低。

6.2.4 创造力支持指数（*CSI*）。尽管 CSI 不属于基线比较的一部分，研究2显示，与统一专家相比，多视角批评显著提升了创造力支持（$SD = M = 7.37$，0.69 $M = 6.45$ $SD = 0.91$），。配对 t t 23 d =检验证实了这一改善（ ( ) = −3.11，p = .005，）。0.64

CritiqueCrew在所有可用性和信任指标上均显著优于静态基线。至关重要的是，研究2表明，多视角架构进一步放大了这些体验收益。结构化编排专家角色并未增加复杂性，反而降低了认知负荷并增强了创意支持，这表明设计师更偏好角色差异化指导所带来的清晰性，而非通用的统一响应。

6.3 研究问题三：CritiqueCrew促进战略设计合作

我们的定性分析揭示了设计师如何看待CritiqueCrew的多视角批评，以及这一对话范式如何从根本上改变了他们的设计优化流程。我们确定了三个核心主题。

6.3.1 从表层修正到跨职能策略。 CritiqueCrew促使设计师发生了认知转变，使他们的流程从执行表层修正转向开展战略思维。不同于会引发线性"清单式"工作流程的静态基线，CritiqueCrew通过呈现建设性冲突，促进了积极的决策。当产品愿景角色建议使用醒目的品牌色按钮，而用户体验角色指出其对比度不足时，设计师不得不权衡取舍，而不是被动地修正孤立的问题。正如P4（研究1）所指出的，这促使他重新评估整个设计：*"CritiqueCrew帮助我发现，'eMall'页面缺少关键的产品价格信息，这促使我从信息架构和业务逻辑完整性的角度重新评估整个设计。"*

研究2进一步印证了这种多视角方法的价值。参与者明确指出，与统一专家相比，不同的视角让人觉得"更有说服力，也更贴近现实"。正如 P1（研究2）所解释的，统一评论者给出了"更简洁但泛泛的建议"，而多视角系统则像一个"全面的安全网"。然而，这种深度也伴随着权衡。像 P17（研究2）这样的注重执行的设计师指出，相互冲突的观点有时会增加简单任务的认知负荷："有时候，多视角信息太多了……我现在不在乎工程师怎么想；我只想完成设计。"这表明，尽管多方面的视角有利于提升战略质量，但需要更多主动处理。

最终，这一工作流程清晰展现了人工智能角色的转变。正如 P3（研究1）所描述的："我通常会先让它评估界面，然后结合它从不同维度给出的反馈和我自己的设计经验来进行修改。"这使人工智能从规定性审计员转变为战略伙伴。

6.3.2 情境化修复保留创作流程。CritiqueCrew 的一项关键优势在于能够将反馈直接整合到设计画布中，解决了基线中"阅读—理解—定位—修改"这一碎片化循环。参与者一致称赞其实际效用。P2（研究1）指出"一键复制文本功能"省去了繁琐的解析过程，而 P6（研究1）则提到，基线要求他"手动选择并定位"组件，CritiqueCrew 则通过"一键预览"实现了自动化。

通过最大限度地降低执行过程中的认知成本，系统让设计师得以保持创作心流。P2（研究1）解释说，使用其他工具时，为了保持专注，她更愿意"先手动修改一次，再让工具帮我评估修改后的草稿"。CritiqueCrew 的无缝整合使这类防御性做法不再必要，将实时反馈从潜在干扰转变为高效的助推器。

6.3.3 结构化对话建立信任与能动性。 对 CritiqueCrew 的信任源于可解释性与能动性。与基线中不透明的建议不同，CritiqueCrew 基于角色的理由说明帮助设计师理解修改背后的"原因"。P7（研究1）指出，基线模糊的反馈常常迫使他浪费时间去验证问题："我不得不花时间思考是否真的存在问题。"

研究2表明，角色差异化是建立这种信任的关键。将多视角系统与统一专家进行比较时，参与者认为，不同角色人格有助于他们更清晰地理解人工智能的推理方式。研究2的日志分析显示，设计师会主动使用 @ 功能来询问特定角色（52% 咨询用户体验设计师，32% 咨询产品经理，17% 咨询工程师），并将

他们视为各有所长的队友。当角色意见不一致时，设计师会进行协商以寻找“折中方案”，或将冲突视为激发创意的契机。一些设计师（如研究2中的 P8）会直接拒绝相互矛盾的建议（“有些建议我不会采纳”），但仍认为辩论颇有价值：“这给了我一个新想法。”这证实，真正的信任并非源于人工智能从不出错，而是源于赋予设计师自主掌控协作过程的能力。

7 讨论

本研究评估了 CritiqueCrew，这是一款旨在将人工智能从被动审查者转变为战略性评议伙伴的对话工具。结果表明，与静态基线相比，CritiqueCrew 的多视角编排显著提升了设计质量（RQ1）和用户体验（研究问题2）。关键的是，研究2证实，这些益处源于专家角色之间有组织的协作，而非仅仅来自底层模型的能力。在本节中，我们将从设计理论的角度解读这些发现，将我们的方法与现有方法进行比较，并展望未来方向。

7.1 主要发现：从验证到反思性对话

我们的研究发现表明，CritiqueCrew 超越了静态工具的传统“验证”范式，迈向“反思性对话”。首先，该系统提升设计质量，不仅是因为发现并解决了更多问题，还因为它促进了更深入的战略思维。研究2表明，将评议拆分为专业化角色（用户体验、产品经理、工程师），比统一专家模型带来了更高的问题覆盖度和解决方案有效性。这说明，不同角色之间富有成效的张力 [7] 有助于设计师发现盲点，而这些盲点可能会被单一的“正确”答案所掩盖。

其次，尽管信息密度有所增加，CritiqueCrew 仍降低了认知负荷并提升了信任。参与者并未将多视角反馈视为“噪声”，而是将其视为一种用于权衡推理的结构化框架。这与分布式认知理论相符 [20], : 智能产生于不同视角之间的协商。通过将这种协商外化，CritiqueCrew 使设计权衡的“黑箱”变得可见且可操作，帮助设计师从表层修正转向整体优化。

7.2 对设计的启示

我们的研究为未来的人工智能赋能创意工具提出了三项关键启示：

(1) **超越问题检测，助力探索解决方案**。未来的工具应将重点从“问题出在哪里”转向“如何改进”。我们的研究结果表明，人工智能能够成为激发创造力的催化剂，这与 UIDEC [51] 的理念相符：约束可以充当创意脚手架。CritiqueCrew 将抽象规则转化为不同专家角色的“立场”，以此实现这一目标。研究2中更高的创造力支持度评分证实，这种多视角方法不仅有助于识别问题，也能支持生成兼顾权衡的策略性解决方案，使设计师能够超越单纯的合规要求。

(2) **善用基于角色的交互范式**。采用多视角系统能够拓宽评议维度，是一种有效的交互范式。将抽象规则（例如 WCAG）转化为具体角色（例如“无障碍专家”）提出的建议，更符合设计师的心智模型。然而，这种拟人化也带来了过度信任的风险。我们的研究结果提出了一种缓解方法：**有意呈现建设性冲突**。通过呈现不同视角（例如产品愿景和用户体验角色的观点），系统促使设计师担任最终裁定者。在我们的研究中，设计师积极使用 @提及功能来质疑和交叉核查建议，展现了角色差异化和对话控制如何帮助将人的判断置于核心位置，而非让其被自动化取代。

(3) **深度情境整合是建立信任的关键**。 工具必须无缝融入现有工作流。我们的研究表明，直接在画布上高亮内容并实时预览，可以最大限度地减少情境切换，使设计师能够将人工智能视为自身工作流程的自然延伸。然而，对大语言模型的依赖是一把双刃剑。“一键应用”功能带来的效率提升可能会放大幻觉造成的负面影响。CritiqueCrew 通过可解释性和对话式澄清来应对这一问题。通过允许设计师追问推理过程（例如，“@工程师，为什么成本这么高？”），我们赋予设计师审查人工智能逻辑的能力。这表明，未来的工具必须以可审查性为设计目标，将解释和行动建议直接关联到共享产物，从而最大限度地降低盲目采纳的风险。

7.3 局限性与未来工作

尽管我们的结果令人鼓舞，但我们也承认，研究设计和系统范围存在一些具体局限。

首先，尽管研究2主要旨在考察基于角色的编排所带来的作用，但多视角条件也通过明确设置一名首席协调者，将综合过程外显化。我们的统一专家基线同样会汇总并优先排序反馈，但这一整合过程是在单次模型响应中隐式完成的。因此，观察到的改进很可能反映了角色专门化推理与明确协调的综合效果，而非任何单一机制的作用。未来的工作可以通过消融研究或析因设计进一步厘清这些因素，分别改变角色专门化程度和协调结构。

其次，我们的评估采用了预定义原型图，而非设计师自己的创意作品。虽然这确保了严格的实验控制以及不同条件之间的可比性，但可能无法充分体现设计师对自己作品的情感依附，以及他们对作品所处情境的深入了解。在现实环境中，设计师在接受评议时可能会更具防御性，或考虑得更加周全。不过，参与者指出，这些原型图很贴近现实，而且任务限制有效模拟了常见的行业场景。

第三，我们的评估聚焦于单个设计师。尽管设计师通常是不同学科之间的主要桥梁，但我们并未直接观察真实的产品经理或工程师将如何与系统交互。这一重点是有意为之，旨在评估

该工具对其主要用户群体的价值，但未来的工作应扩大范围，纳入多样化的利益相关者，以验证我们的专家角色画像的准确性。

第四，尽管我们的样本量（两项研究共计$N = 48$）为检测大效应提供了足够的统计功效，但更广泛的推广仍需更大且更多样化的参与者群体。此外，我们对设计质量的操作化依赖于专家对最终静态模型稿的评分。虽然这是人机交互领域的标准做法，但它无法捕捉下游效应，例如终端用户行为转化率或长期代码可维护性，而这些需要通过纵向部署来评估。

最后，CritiqueCrew 目前评估单个画面。当前实现实例化了一组固定的三个角色，尚不能根据不同任务类型动态调整。未来的迭代将扩展以支持跨多个屏幕的用户流程分析。

未来工作将沿三个战略方向推进。首先，我们计划开展一项纵向实地研究，将 CritiqueCrew 部署到真实的设计团队中。这将使我们能够考察冲突频率与协商策略如何在长期项目中演变，以及人工智能生成的评议如何随时间影响设计所有权。其次，我们旨在将系统的情境感知从单个画面扩展到支持用户流程分析。通过使系统理解多个屏幕之间的过渡与交互逻辑，我们可以就信息架构和旅程一致性提供更全面的反馈。最后，我们设想一个可定制的专家小组，设计师可以配置角色画像，甚至创建针对其特定项目需求的自定义角色。这将使 CritiqueCrew 从固定工具转变为可适应、不断演进的设计伙伴。

8 结论

我们提出了 CritiqueCrew，重新构想人工智能在设计评议中的作用——从静态的“问题审查者”转变为动态的“解决方案共同创造者”。通过在结构化工作流程中编排多视角专家角色，该系统弥合了抽象反馈与创造性行动之间的鸿沟。CritiqueCrew 将评议直接融入设计流程，把孤立的问题清单转化为可执行、可解释且能顾及情境的解决方案。

在两项对照研究（$N = 48$）中，CritiqueCrew 的表现显著优于最先进的静态工具和统一专家基线。使用我们系统的设计师解决了更多问题，取得了更高质量的成果，同时认知负荷显著降低，信任度也更高。重要的是，我们的研究结果证实，这些优势源自角色编排机制本身：通过呈现不同视角（如用户体验与工程）之间富有成效的张力，该系统能够促发更深入的战略思维，而非仅仅纠正错误。

本研究为设计中的人机协作提出了一种新范式：超越自动化检查，转向经统筹的多视角对话。本研究为新一代创意支持工具提供了可操作的见解，并证明，结构化的角色差异化是构建可信且高效的人工智能伙伴的关键，尤其是在复杂的跨职能领域中。

致谢

本研究获得国家人因工程重点实验室基金会（编号：HFNKL2024W03）资助。

参考文献

[1] Saleh Afroogh、Ali Akbari、Emmie Malone、Mohammadali Kargar 和 Hananeh Alambeigi。2024年。《对人工智能的信任：进展、挑战与未来方向》。《人文与社会科学通讯》11，1（2024年），1–30。[2] Spec AI。2024年。Spec AI - 用户界面/用户体验反馈工具 | Figma AI 插件。https://www.specai.io/。访问日期：2025-09-10。[3] Lorans Alabood、Zahra Aminolroaya、Dianna Yim、Omar Addam 和 Frank Maurer。2023年。《设计评议方法的系统性文献综述》。《信息与软件技术》153（2023年），107081。[4] Mário Bexiga、Sergii Garbatov 和 João C. Seco。2020年。《缩小低代码生态系统中设计师与开发人员之间的差距》。载于《第23届ACM/IEEE模型驱动工程语言与系统国际会议论文集：配套论文集》。ACM，1–10。[5] Eli Blevis、Youn-kyung Lim、David Roedl 和 Erik Stolterman。2007年。《将设计评议作为研究方法，以建立可持续性与交互技术之间的联系》。载于《在线社区与社会计算国际会议》。Springer，22–31。[6] Virginia Braun 和 Victoria Clarke。2006年。《在心理学中运用主题分析》。《心理学定性研究》3，2（2006年），77–101。[7] Louis L Bucciarelli。1988年。《工程设计的民族志视角》。《设计研究》9，3（1988年），159–168。[8] Xiang 'Anthony Chen、Tiffany Knearem 和 Yang Li。2025年。《GenUI 研究：探索生成式 UI 工具的设计，以支持用户体验从业者及其他群体》。载于《2025年ACM交互系统设计会议论文集》。计算机协会，1179 1196 10 1145 3715336 3735780 –。[9] 2014 doi：。/。Erin Cherry 和 Celine Latulipe。《通过创造力支持指数量化数字工具对创造力的支持》。《ACM人机交互汇刊》（TOCHI）21，4（2014年），1–25。[10] Mihaly Csikszentmihalyi 和 Mihaly Csikzentmihaly。1990年。《心流：最优体验心理学》。第1990卷。纽约Harper & Row。[11] Peipei Duan、Chieh-Yang Chen、Gang Li、Björn Hartmann 和 Yang Li。2024年。《UICrit：借助 UI 评议数据集提升自动化设计评估》。载于《第37届ACM用户界面软件与技术年会论文集》（UIST '24）。ACM。[12] Peipei Duan、Jed Warner、Yang Li、Darren Gergle 和 Björn Hartmann。2024年。《利用大语言模型为 UI 线框图生成自动反馈》。载于《CHI计算机系统人为因素会议论文集》。ACM，1–20。[13] Edmund Burke Feldman。1994年。《实用艺术批评》。Prentice Hall，纽约。[14] K. J. Kevin Feng、Tony W Li 和 Amy X. Zhang。2023年。《理解软件组织中专业用户体验从业者的协作实践与工具》。载于《2023年CHI计算机系统人为因素会议论文集》（德国汉堡）（CHI '23）。计算机协会，美国纽约州纽约市，文章 764, 20 10 1145 3544548 3581273 页。doi：。/。[15] 2024 Figma。《Figma：协作式界面设计工具》。https://www.figma.com/。访问日期：2025-09-11。[16] Gerhard Fischer、Andreas C Lemke、Thomas Mastaglio 和 Andres I Morch。1991年。《评议在协作式问题解决中的作用》。《ACM信息系统汇刊》（TOIS）9，2（1991年），123–151。[17] Jill J Francis、Marie Johnston、Clare Robertson、Liz Glidewell、Vikki Entwistle、Martin P Eccles 和 Jeremy M Grimshaw。2010年。《多大的样本量才足够？为基于理论的访谈研究将数据饱和操作化》。《心理学与健康》25，10（2010年），1229–1245。[18] Shang Gao、Yaxin Wang 和 He Liu。2024年。《UiAnalyzer：评估应用程序的 UI 是否存在违反功能布局设计规范的风险》。《专家系统与应用》239（2024年），122408。[19] Robert R Hoffman、Shane T Mueller、Gary Klein 和 Jordan Litman。2023年。《可解释人工智能的衡量指标：解释质量、用户满意度、心智模型、好奇心、信任与人机协作表现》。《计算机科学前沿》5（2023年），1096257。[20] James Hollan、Edwin Hutchins 和 David Kirsh。2000年。《分布式认知：迈向人机交互研究的新基础》。《ACM人机交互汇刊》（TOCHI）7，2（2000年），174–196。[21] Eric Horvitz。1999年。《混合主动式用户界面的原则》。载于《SIGCHI计算机系统人为因素会议论文集》。ACM，159–166。[22] Shuai Huang、Qihang Wang、Weizhen Lu、Lei Liu、Zhiyong Xu 和 Yixuan Huang。2025年。《PaperEval：由多智能体系统驱动的通用、定量且可解释的论文评估方法》。《信息处理与管理》62, 6 2025 104225 10 1016 2025 104225（），。doi：。/j.ipm.。[23] 2018 Yun-Han Huang。《理解敏捷环境中用户体验设计师与开发人员之间的协作困难》。硕士论文。普渡大学

Robin Jeffries、James R. Miller、Cathleen Wharton 和 Kathy M. Uyeda。1991年。《真实世界中的用户界面评估：四种技术的比较》。载于《SIGCHI计算机系统人为因素会议论文集》，Scott P. Robertson、Gary M. Olson 和 Judith S. Olson（编）。ACM，119–124。[24] Jihyeon Jeon、Jaehong Yoon、Minji Kim、Sunggyun Shim、Yoonsik Choi、Hwaran Kim 和 Yoonsik Yu。2025年。《多模态大语言模型能否捕捉界面如何引导用户行为？多模态 UI/UX 设计理解基准》。arXiv:2505 05026 [ ]（），–。doi：。/ [25] . cs.HC Ziwei Ji、Nayeon Lee、Rita Frieske、Tiezheng Yu、Dan Su、Yan Xu、.....和 Pascale Fung。2023年。《自然语言生成中的幻觉综述》。《ACM计算调查》（CSUR） 55, 12 2023 1 38 10 1145 3571730 Paul A Kirschner。《认知负荷理论：认知负荷理论对学习设计的启示》。[26] 10页。2002 Kai Kolthoff、Fabian Kretzer、Christoph Bartelt、Alexander Maedche 和 Simone Paolo Ponzetto。2025年。《GUIDE：大语言模型驱动的图形用户界面生成分解，用于自动化原型设计》。arXiv:2502 21068 [ ] [27] . cs.HC LangChain。《LangGraph：借助基于图的控制构建稳健的智能体系统》。https://github.com/langchain-ai/langgraph。访问日期：2025-08-10。Chunggi Lee、Sanghoon Kim、Dongyun Han、Hongjun Yang、Young-Woo Park、Bum Chul Kwon 和 Sungahn Ko。2020年。《GUIComp：具有实时多方面反馈的图形用户界面设计助手》。载于《2020年CHI计算机系统人为因素会议论文集》。计算机协会，1–13。doi:10.1145/3313831.3376327[29] 2024 [30] Jung-Joo Lee、Miia Jaatinen、Anna Salmi、Tuuli Mattelmäki、Riitta Smeds 和 Mari Holopainen。2018年。《共同创造项目的设计选择框架》。《国际设计学刊》12，2（2018年），15–31。[31] Kyu Lee。2025年。《迈向生成式用户界面设计的工作定义》。arXiv:2505 15049 [ ] [32] . cs.HC J. C. R. Licklider。《人机共生》。《IRE电子领域人为因素汇刊》HFE-1，1（1960年），4–11。Zhangqi Liu。2025年。《人机共同创造：智能系统协作设计框架》。[33] 1960 arXiv预印本 arXiv:2507.17774（2025年）。[34] Zeyu Liu、Chong Chen、Jun Wang、Ziyi Zhang、Zhenyu Chen 和 Tao Chen。2023年。《Nighthawk：通过视觉理解全自动定位用户界面显示问题》。《IEEE软件工程汇刊》49，1（2023年），403–418。[35] Geetha LS、Yousef A Baker El-Ebiary、Bandla Srinivasa Rao、Revati Ramrao Rautrao、T Subha Mastan Rao、Janjhyam Venkata Naga Ramesh 和 Omaia Al-Omari。2025年。《敏捷软件开发的挑战与解决方案：实施实践的管理视角》。《国际高级计算机科学与应用杂志》16，3（2025年）。[36] Yachen Lu、Tricia Knearem、Suso Dutta、Amy Xu、Walter S Lasecki 和 Ranjitha Kumar。2024年。《仅有人工智能还不够：在基于启发式方法的用户界面静态检查中应用人工智能的混合技术方案》。载于《CHI计算机系统人为因素会议扩展摘要集》。ACM，1–7。[37] Kurt Luther、Amy Pavel、Wei Wu、Jari-lee Tolentino、Maneesh Agrawala、Björn Hartmann 和 Steven P Dow。2014年。《CrowdCrit：众包并汇总视觉设计评议》。载于《第17届ACM计算机支持的协同工作与社会计算会议配套出版物》。21–24。[38] Gloria Mark、Daniela Gudith 和 Ulrich Klocke。2008年。《工作中断的代价：更快的速度与更大的压力》。载于《SIGCHI计算机系统人为因素会议论文集》。107–110。[39] Nora McDonald、Sarita Schoenebeck 和 Andrea Forte。2019年。《定性研究中的可靠性与评定者间信度：计算机支持的协同工作与人机交互实践规范和指南》。《ACM人机交互会议录》3，CSCW（2019年），1–23。[40] Meta开源。2025年。React。https://github.com/facebook/react，用于构建用户界面的 JavaScript 库。访问日期：2025-09-11。[41] Caterina Moruzzi 和 Sandra Margarido。2024年。《人机共同创造力的以用户为中心框架》。载于《CHI计算机系统人为因素会议扩展摘要集》。ACM，1–9。[42] Yeonjoo Oh、Suguru Ishizaki、Mark D Gross 和 Ellen Yi-Luen Do。2013年。《建筑设计工作室中设计评议的理论框架》。《设计研究》34，3（2013年），302–325。[43] OpenAI。2024年。GPT-4o。https://openai.com/index/hello-gpt-4o/。访问日期：2025-09-11。[44] Joon Sung Park、Joseph C O'Brien、Carrie J Cai、Meredith Ringel Morris、Percy Liang 和 Michael S Bernstein。2023年。《生成式智能体：人类行为的交互式模拟》。载于《第36届ACM用户界面软件与技术年会论文集》（UIST '23）。ACM，1–22。[45] Moin Reza、Jordan Thomas-Mitchell、Pela Dushniku、Nicole Laundry、Joseph Jay Williams 和 Anatoliy Kuzminykh。2025年。《以人为本的人工智能协同写作：在整个写作过程中使研究与用户需求保持一致》。arXiv：预印本 [46] Jullia Saad、Suéllen Martinelli、Leticia S Machado、Cleidson RB De Souza、Alexandre Alvaro 和 Luciana Zaina。2021年。《软件初创企业中的用户体验工作：文献主题分析》。《信息与软件技术》140（2021年），106688。[47] Donald A Schon 和 Vincent DeSanctis。1986年。《反思型实践者：专业人士如何在行动中思考》。[49] David W Shaffer. 2003. 牛津设计工作室画像：设计教学法民族志. 技术报告。WCER 工作论文。 [50] Dayeon Shin, Dahyun Lee, Gary Hsieh, and Gloria Y.-Y. Chan. 2025. PosterMate：面向海报设计的受众驱动协作角色代理。载于第*38*届*ACM*用户界面软件与技术年会论文集（*UIST '25*）。ACM。[51] Atefeh Shokrizadeh, Boniface Bahati Tadjuidje, Shivam Kumar, Sohan Kamble, and Jinghui Cheng. 2025. Dancing With Chains：在约束条件下运用 UIDEC 进行 UI/UX 设计构思。载于*2025*年*CHI*计算机系统人为因素会议论文集。计算机协会，1–23。doi:10.1145/3706598.3713785[52] Erik Stolterman. 2008. 设计实践的本质及其对交互设计研究的启示。《国际设计学刊》2, 1（2008）。[53] Vivian Stray, Anamaria Barbala, and Vilde T. Wivestad. 2025. 软件开发中的人机协作：开发人员使用 GitHub Copilot 和 ChatGPT 的混合方法研究。载于第*33*届*ACM*软件工程基础国际会议论文集。ACM，1325–1332。[54] John Sweller. 1988. 问题解决过程中的认知负荷：对学习的影响。认知科学12, 2（1988），257–285。[55] G. Vera-Amaro and J. R. Rojano-Cáceres. 2025. 借助人工智能实现无障碍网站设计：系统性文献综述。《信息与软件技术》（2025），107821。[56] Wenyu Xiang, Haotian Zhu, Siwei Lou, Yechen Chen, Ruirui Zhang, Haining Wang, and Hong Mei. 2024. SimUser：通过模拟不同用户与移动应用交互来生成可用性反馈。载于*CHI*计算机系统人为因素会议论文集。ACM，1–17。[57] Liwenhan Xie, Yanna Lin, Can Liu, Huamin Qu, and Xinhuan Shu. 2025. DataWink：利用大型多模态模型复用和改编基于 SVG 的可视化示例。*arXiv* 预印本 *arXiv:2507.17734*（2025）。[58] Anbang Xu, Huaming Rao, Steven P Dow, and Brian P Bailey. 2015. 在课堂中使用群体反馈进行迭代设计过程的研究。载于第*18*届*ACM*计算机支持的协同工作与社会计算会议论文集。1637–1648。[59] Bohan Yang and Shuai Li. 2024. UIGuider：使用领域知识图谱方法检测隐式设计指南。《电子学》13, 7（2024），1210。[60] Bohan Yang and Shuai Li. 2024. UISGPT：利用大语言模型自动检测移动端界面设计异味。《电子学》13, 16（2024），3127。doi:10. 3390/electronics13163127[61] Bohan Yang, Zhenchang Xing, Xin Xia, Yupan Liu, Tiezhu Zhang, and Wei Zhao. 2021. 别这么做！依据设计指南排查复杂用户界面中的视觉设计异味。载于*2021*年*IEEE/ACM*第*43*届软件工程国际会议（*ICSE*）。IEEE，761–772。[62] Min Yuan, Jiahui Chen, Yuchong Hu, Shuai Feng, Mingliang Xie, Golnaz Mo-hammadi, 等。2024. 迈向用户界面设计中的人机协同：通过意图澄清和对齐增强基于多智能体的用户界面生成。arXiv:2412 20071 [ ] [63] . cs.HC Mathieu Zen, Nicolas Burny, and Jean Vanderdonckt. 2023. 一种基于质量模型的用户界面美学测量方法。*ACM*人机交互论文集7, EICS（2023），1–47。

表 3：参与者人口统计信息和职业角色。

PID	性别	Age	经验	职业角色
P1	女性	18至24岁	2年	UI设计师
P2	女性	32–38	7年	UI设计师
P3	女性	32–38	7年	UI设计师
P4	女性	25–31	2年	交互设计师
P5	女性	25–31	3年	交互设计师
P6	女性	25–31	5年	交互设计师
P7	Male	25–31	2年	产品经理
P8	女性	25–31	4年	产品经理
P9	Male	25–31	4年	产品经理
P10	Male	25–31	3年	软件工程师（前端）
P11	女性	25–31	5年	软件工程师（前端）
P12	Male	25–31	5年	软件工程师（前端）

A 形成性访谈研究

形成性研究参与者的详细信息（性别、年龄、角色和设计经验年限）见表3。所有参与者在日常工作中均有使用 Figma 的经验，也参与过设计评议，并熟悉从产品设计到实现的完整工作流程。

B 系统提示词

本节介绍了 CritiqueCrew 中每个专家模块和首席协调员所使用的详细系统提示词。

用户体验模块提示词

角色与目标

作为用户体验专家（用户体验研究员／UI设计师／无障碍专家），评估可用性、无障碍性和整体用户体验，并提供具体、可执行的改进建议。

输入

1) UI 截图 2) UI JSON 3) 产品背景（目标、品牌关键词、主题色、目标用户）

输出格式（JSON）

```
{ "反馈_id": "唯一_id", "来源_角色": "用户_经验", "问题": [{ "节点 ID": "元素_id", "节点名称": "元素_名称", "元素类型": "文本|框架|矢量图形|...", "问题类型": "可用性|无障碍性|导航|...", "严重程度": "严重|高|中|低", "描述": "问题描述", "修复": { "行动建议":"行动建议", "具体建议": "具体改进", "技术方案": "实现计划"
    } }],"摘要": {...}, "优先级": "严重|高|中|低
", "详细_分析": "整体分析" }
```

分析范围

1) 可用性：信息清晰度、任务流程、交互效率、一致性。2) 无障碍性 （WCAG 2.1 AA 和 AAA）： 文本对比度（普通文本为 ≥4.5:1，大号文本为 ≥3:1）、触控目标最小尺寸（44×44px）、可读字体大小（≥16px）、键盘导航、语义化标签。3) 移动端优化：响应式设计、触控/手势体验、加载性能。4) ...

执行规则

- 仅关注界面与交互问题。- 每个问题都必须包含 nodeId，并描述用户影响。- 数据缺失时，提供保守、通用的建议，并确保输出有效的 JSON。

产品愿景模块提示词

角色与目标

作为产品愿景专家（产品经理/业务策略师），评估设计是否符合业务目标、品牌形象和市场定位，并提供可执行的建议。

输入

1) UI 截图 2) UI JSON 3) 产品背景（目标、品牌关键词、主题色、目标用户）

输出格式（JSON）

```
{ "feedback_id": "唯一_id", "source_role": "产品_愿景", "issues": [ { "nodeId": "元素_id", "nodeName": "元素_名称", "ele mentType": "文本|框架|矢量图形|...", "issueType": "业务_目标|品牌|内容|市场|...", "severity": "严重|高|中|低", "description": "问题描述", "remediation": { "action": "行动建议", " specificSuggestion":"具体改进", "technicalSolution": "实现计划" } }],"summary": {...}, "priority": " 严重 | 高 | 中 | 低 ",
```

"detailed_analysis"："整体分析"
}

分析范围

1) 业务目标对齐：对产品目标的支持、转化流程优化、关键功能的优先级排序，以及对关键绩效指标的潜在影响。2) 品牌一致性：正确使用主题色，与品牌关键词（例如：专业、简洁、值得信赖）对齐，视觉层级，以及品牌价值观的情感表达。3) 内容策略：文案的清晰度与可读性、行动号召的有效性、信息优先级排序、用户流程引导。4) 市场定位：差异化、与目标市场中用户期望的对齐、竞争优势。5) ...

执行规则

- 每个问题都必须包含节点 ID，并说明其对业务/品牌的影响。- 提供基于产品战略和品牌形象的建议。- 重点关注设计选择如何影响用户转化和品牌传播。- 缺少数据时，给出保守、通用的建议，并确保输出有效的 JSON。

工程模块提示词

角色与目标

作为工程专家（从前端／软件工程师的角度），评估设计的 工程可行性、实施成本和性能影响 。提供具体且可执行的技术解决方案。

输入

1) UI 截图 2) UI JSON 3) 产品背景（目标、品牌关键词、主题色、目标用户）

输出格式（JSON）

{ "feedback_id": "唯一_id", "source_role": "工程", "问题": [{ "节点 ID": "元素_id", "节点名称": "元素_名称", "元素类型": "文本|框架|矢量图形|...", "问题类型": "可行性|性能|无障碍性|...", "严重程度": "严重|高|中|低", "描述": "问题描述", "修复": { "行动建议":"行动建议", "具体建议": "具体改进", "技术方案": "实现计划/代码级建议" }

}],"摘要": {...}, "优先级": "严重|高|中|低", "详细_分析": "总体技术分析" }

分析范围

1)工程可行性：CSS/JS 复杂度、第三方依赖、浏览器/设备兼容性、响应式实现挑战。2) 性能影响：图像优化、字体加载、CSS/JS 复杂度、渲染成本（首次内容绘制、最大内容绘制、累积布局偏移、首次输入延迟）。3)无障碍性实现支持：针对用户体验模块标记的问题提供技术修复方案（对比度、触控目标尺寸、键盘导航、屏幕阅读器兼容性）。4) 开发成本估算：按复杂度等级评估工作量（轻微样式调整与新增组件/架构级变更）。5) 设计缺陷检测：识别实现不一致、交互过于复杂或缺失的无障碍支持。6) ...

执行规则

- 重点关注可行性、成本和性能，而非纯粹的可用性。- 每个问题必须包含节点 ID、技术挑战、成本估算，以及适用时的替代方案。- 提供切实可行的实现细节（必要时提供 CSS/JS 示例）。- 与用户体验模块协作，为无障碍性问题提供具体修复方案。

首席协调员提示

角色与目标

作为首席协调员，你在对话式设计评议中担任引导者。你的目标是帮助设计师做出明智的决策，而不是替他们做决定。你综合来自三个专家模块（用户体验、产品愿景、工程）的输入。

输入

1) 三个专家模块识别出的反馈和问题 2) 产品背景和目标 3) 模块提供的结构化评审议程事项

输出格式（JSON）

{ "对话_开场":"首席协调员引言", "总体_得分": 1-10, "议程_事项": [{ " 优先级": "严重|高|中|低", "标题": "议程标题", "组件_组":"用户界面组件", " 受影响的 _ 角色 ": ["用户 _ 经验 ", " 产品 _ 愿景 ", " 工程 "],

```
        "issue_summary": "问题摘要",
        "建议": "建议的行动",
        "conflicts": [
          {
            "conflicting_roles": ["模块1", "模块2"],
            "冲突_描述": "冲突描述",
            "tradeoff_question": "开放式权衡问题"
          }
        ],
        "estimated_effort": "工作负荷估算"
      }
    ],
    "冲突_to_surface": [...],
    "组件_分析": [...],
    "积极_亮点": ["..."],
    "后续_对话_要点": ["..."]
}
```

分析范围

1) 综合并归纳主题：按组件归类反馈，突出跨领域主题，避免简单罗列。2) 确定优先级：根据严重程度排列问题（严重的无障碍性问题 >核心流程阻碍 > 业务不一致 > 技术债务 > 美学优化）。3)呈现冲突：当模块意见不一致时，呈现权衡，并通过开放式问题引导设计师。

执行规则

- 扮演引导者，而非决策者。- 使用自然、对话式且鼓励性的语言。- 提供结构化议程，避免认知过载。- 明确指出冲突，并将其表述为问题（例如："我们如何在此处兼顾无障碍标准与品牌一致性？"）。- 强调集体观点（"我们观察到……" "团队建议……"），而非给出绝对判断。

C 应用案例

为进一步说明该系统的工作流程与价值，本节介绍了取自研究中电子商务（图 9）和在线教育（图 10）场景的详细应用案例。每个案例都聚焦于优化一个具有代表性的单帧，展示 CritiqueCrew 如何将抽象的多视角评估转化为具体、可执行的设计修改，从而帮助设计师高效提升设计质量。

D 定性分析编码手册

表 4 详细介绍了我们在形成性研究期间制定的编码手册。该表概述了三个核心主题、其相关子代码、定义，以及参与者具有代表性的引述。

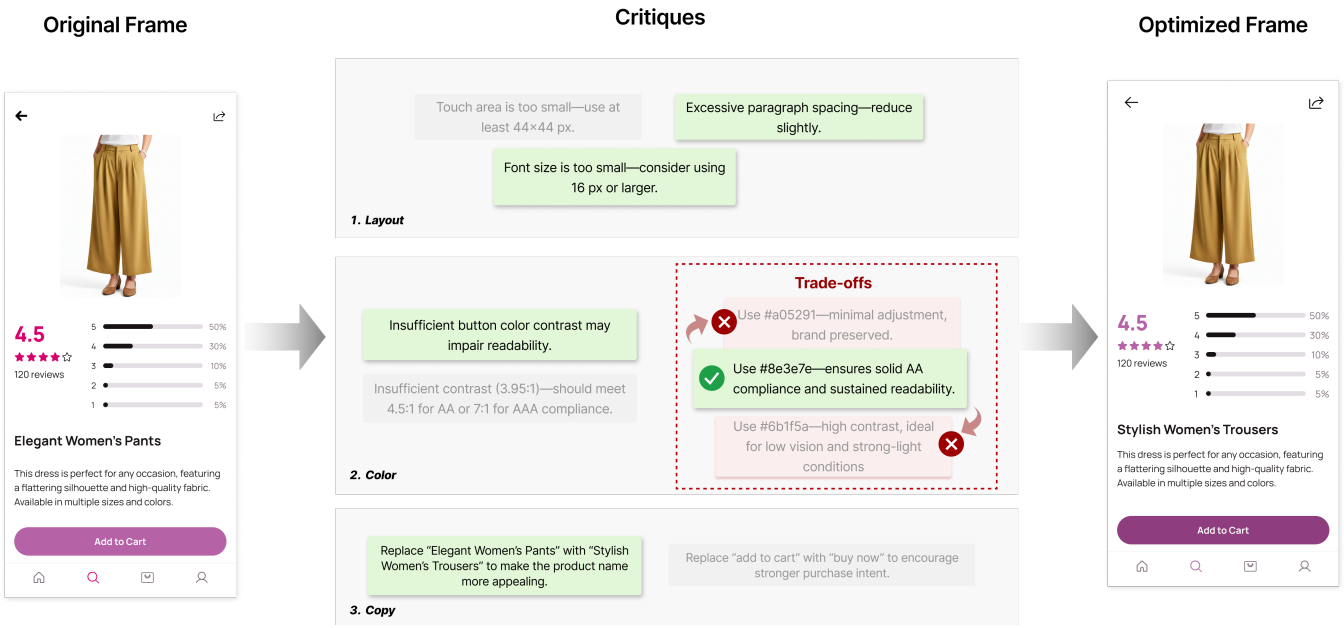


图 9：将 CritiqueCrew 应用于电商产品页面原型单个画面的案例研究。系统从三个维度提供评议：布局、色彩和文案。值得注意的是，色彩对比建议以"权衡"的形式呈现，使设计师能够在品牌一致性与无障碍合规之间做出明智决策，从而得到右侧的优化画面。

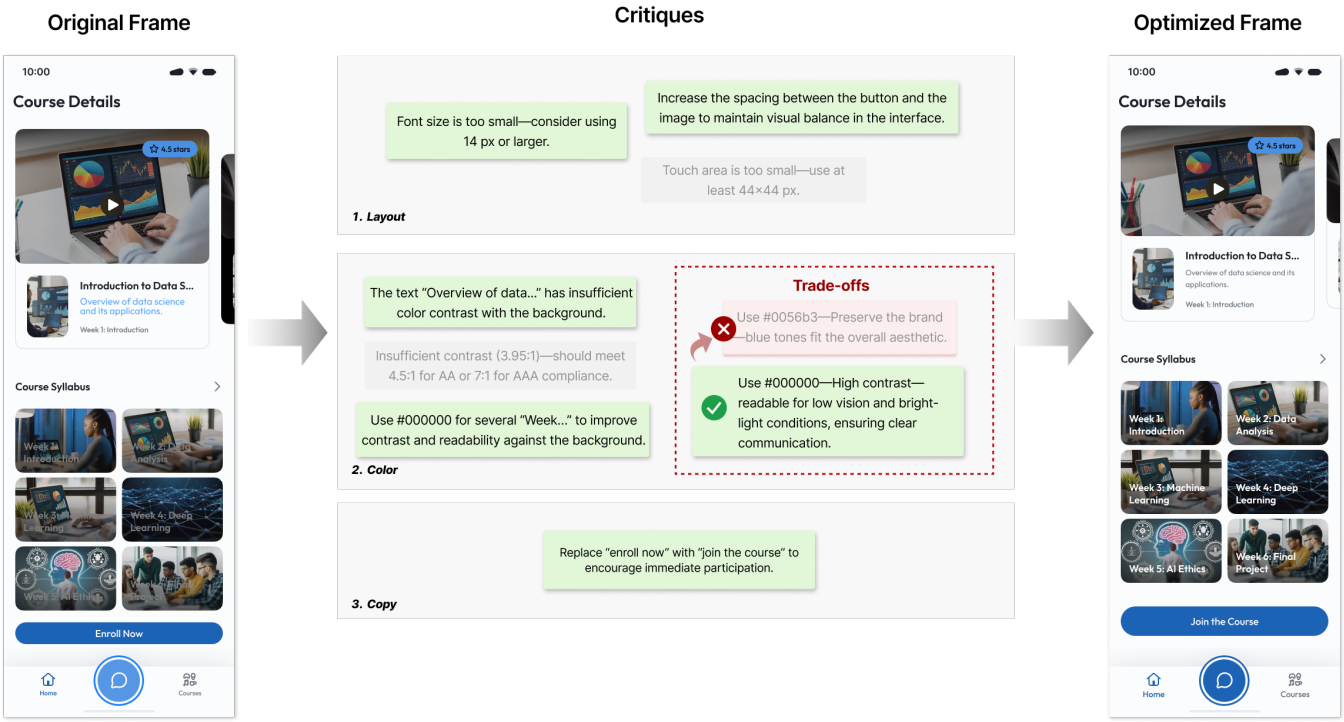


图 10：将 CritiqueCrew 应用于在线课程详情页原型单个画面的案例研究。系统从多个维度识别设计问题。针对文本对比度的色彩评议以"权衡"的形式呈现，清晰展示了品牌美学与可读性之间的冲突，引导设计师完成最终优化版本。

表 4：基于形成性研究访谈（N=15）得出的主题分析编码手册。

主题	子代码	定义	代表性引文
协商相互冲突约束	**技术债务与视觉新颖性**	独特的视觉设计需要非标准工程实现，从而增加维护成本时产生的矛盾。	“设计师想要一种用于弹窗标题的‘*Semibold*’字体……我告诉他们，我们没有这个字体包。要支持它就意味着要增加开发成本。这并非不可能，但我拒绝了以免为非标准需求大开方便之门。”（P10，前端工程师）
	业务关键绩效指标与设计一致性	激进的业务目标（例如，流量入口）迫使设计违反既定的用户界面组件指南。	“产品经理坚持把‘*Robot*’功能放在顶层以提高可见性……但我们的平台指南并不支持这个组件。我们争论了很久，并最终妥协，将其折叠进一个菜单。”（P2，用户界面设计师）
	边界情况下的逻辑漏洞	依靠下游角色（开发人员/人工智能）来识别产品需求中的逻辑漏洞而这些漏洞是产品经理容易忽略的。	“作为产品经理，我可能会忽略一些边界情况，例如账单日期恰逢月末的非工作日。人工智能可以帮助穷尽这些人类难以完全应对的情境涵盖。”（P7，产品经理）
信任与可解释性	**循证比率-nale**	只有当人工智能的反馈有引用来源作为依据（例如指南、心理学理论），而非不透明的观点时，用户才会信任它。	“当人工智能提供建议的依据时，我才会信任它……比如引用某个具体的心理学理论。如果有参考文献，我就信任它；否则，我会怀疑幻觉。”（P5，交互设计师）
	情境化依据（产品需求文档）	人工智能需摄取项目特定文档（产品需求文档），以避免泛泛而谈、毫无帮助的建议。	“如果人工智能不了解我们内部的‘黑话’，也不了解具体的业务背景……它就只会生成 *99%* 正确但毫无用处的废话。它必须理解产品需求文档。”(P8，产品经理)
	白盒推理	人工智能需要揭示其决策过程，尤其是在解决冲突时。	“它不应该只给出最终结果，而必须展示推理过程……如果它是一个黑箱，我就不会信任它对冲突意见的判断。”*(P6*，交互设计师)
工作流集成	**阶段依赖效用**	需求的明显转变：“开放式灵感-”在早期阶段，相较于“严格合规检查”在后期阶段。	“在概念阶段的早期，我希望获得开放式建议以激发灵感。但到了交付阶段的后期，我更信任人为设定的限制……人工智能最适合检查色彩对比度合规性等细节。”（P5，交互设计师）
	赋能与解释国家	人工智能需要解释冲突背后的“原因”，并提供权衡方案，将最终决策权留给人类。	“人工智能不应只给出最终结果，还需要展示推理过程……像调解者一样，解释了不同观点之间的权衡。”（P3，UI设计师）