arXiv:2611.32169v1 [cs.HC] 24 Nov 2026

MyMuse：用于参观后探究与反思的个人博物馆

裴智妍 首尔国立大学 大韩民国首尔 哈佛大学 美国马萨诸塞州剑桥 jiyeonb@snu.ac.kr

李忠基 哈佛大学 美国马萨诸塞州剑桥 chunggi_lee@g.harvard.edu

徐镇旭 首尔国立大学 大韩民国首尔 jseo@snu.ac.kr

Hanspeter Pfister 哈佛大学 美国马萨诸塞州剑桥 pfister@seas.harvard.edu

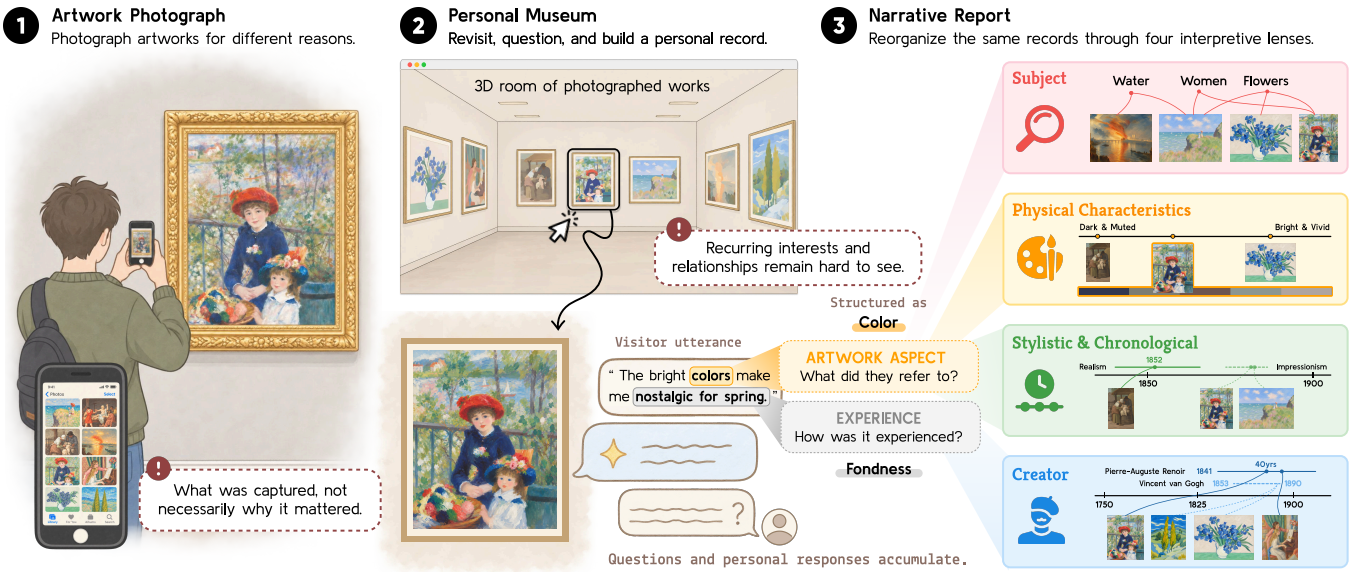


图 1：MyMuse 是一款 Web 应用程序，支持参观者围绕艺术品照片开展参观后探究与反思。参观者可以在个人博物馆中重访拍摄的作品，并围绕单件艺术品累积问题与反思，同时按艺术品方面和体验触发因素进行标注。叙事报告通过四种解释视角重新组织相同的记录，让参观者得以考察不同作品之间反复出现的兴趣和关系。

摘要

艺术品照片或许能保存参观者所看到的内容，却未必能保留其意义所在。重新查看这些照片时，人们可能会回忆起先前的回答，并随着知识和个人经历的变化提出新的问题与诠释。然而，在一次次单独交流中，反复出现的兴趣可能仍难以被识别。基于一项由十名参观者参与的形成性研究，我们提出了 MyMuse——一个参观后系统，将对单件艺术品的探究与参观者累积记录中的反思连接起来。在其个人博物馆中，参观者可以重新查看艺术作品、提出问题，并记录能够影响后续系统响应的解读。其交互式叙事报告通过四种解释视角，跨艺术作品重新组织相同的记录。在两项被试内研究中，16 名参与者的直接比较更偏向 MyMuse，认为其能带来更生动的重新参观体验。与按艺术品呈现相同记录的方式相比，参观者对该报告在反复兴趣认知、跨艺术作品联系和意外发现方面的评分更高。这些研究发现表明，考察累积记录有助于反思艺术作品之间反复出现的兴趣，以及参观者回答之间的关系。

CCS 概念

• **以人为中心的计算 → 交互系统与工具。**

关键词

艺术欣赏、参观后参与、个人记录、反思、叙事可视化

1 引言

艺术邂逅能够塑造参观者所注意、记住和发现的事物富有意义。这种参与度可以延续到博物馆参观之后。参观者可能会拍下吸引他们注意力的艺术作品，之后再回看这些图像，以回忆或重新审视这段经历。拍照可以将注意力引向经历中的特定方面 [11]。随着参观者对某件艺术作品及其情境了解得更多，他们可能会注意到作品的不同方面，或

重新思考他们的诠释 [8, 26]。关于重访存储痕迹的研究也表明，参与度可以延伸至最初接触之后 [9, 37, 44, 52]。这些研究共同为探讨艺术品照片如何在参观之后支持探究与重新诠释提供了依据 [42]。

然而，照片或许能够保存艺术作品的外观，却无法保存参观者对它的回应。因此，重新查看图像时，参观者为何会被该作品吸引，或他们想要探究什么，可能仍不清楚。关于个人记录的研究提醒我们，更完整的记录并不一定有助于更好地回忆 [36]。既有研究曾探索将博物馆参观期间收集的材料与后续学习活动联系起来 [42], , 但仅仅能够访问这些材料，并不会自动使参观者的问题与诠释得以留存，供日后使用。在一个收藏集中，被拍摄艺术品之间的相似性，或许只能有限地证明参观者自身回应之间存在关系。即使问题与诠释已被记录下来，它们之间的连接也可能难以在分散的艺术品条目中被识别。我们考察参观后系统如何利用累积记录为后续探究提供信息，并帮助参观者审视不同作品之间反复出现的兴趣与关系。

为了了解参观者在参观后探究与反思方面的需求，我们与十名博物馆和美术馆参观者开展了一项定性形成性研究，使用了他们各自参观时拍摄的100张照片。参与者表示，他们难以回忆自己为何拍摄特定作品。他们的叙述还表明，后续获得的知识如何促使他们重新思考，以及累积的照片如何引发对反复出现的兴趣的反思。这些研究发现为**MyMuse**的设计提供了依据。MyMuse是一座参观后的个人博物馆，参观者可以在其中添加和重新查看与其艺术品照片相关的记录（图1）。在单件作品层面，参观者可以查看先前条目、提出进一步的问题、阐明回应，并探索相关艺术品。MyMuse利用先前记录为后续回应提供信息，同时将参观者撰写的内容与检索到的上下文及系统生成的文本区分开来。在收藏集层面，**叙事报告** 通过借鉴文化对象编目 （CCO）的诠释视角组织累积记录 [2]。其构图根据可用证据以确定性方式计算，而语言生成则为选定的观察结果组织措辞。借鉴叙事可视化原则 [35], , 该报告呈现一份关于反复出现的要素及关系的引导式叙述，参观者可以结合支持性记录对其进行审视。

为考察 MyMuse 是否支持这两个层面的参与，我们开展了两项受控被试内研究，邀请 16 位参与者在同一个研究环节中使用自己拍摄的艺术品照片。在直接比较中，我们为每位参与者选取了另一组随机分配的艺术品照片作为基线；16 位参与者中有 13 位认为 MyMuse 更能生动地重访过往的博物馆体验，12 位认为它更有助于注意到作品之间的联系。在收藏集层面，与基于同一组互动记录生成的单件艺术作品报告相比，叙事报告在反复兴趣认知、跨艺术作品联系和意外发现方面获得了更高评分。感知保真度则没有显著差异。定性陈述表明，跨记录进行反思也包括更准确地阐述熟悉的兴趣。这些研究发现表明，整理访客自己的记录，可以补充艺术品重访，帮助他们反思自身的回答在不同作品之间有何关联。

本研究有三项贡献。第一，我们定性描述了参观者在再次观看艺术品照片时遇到的困难和需求，包括后续知识和不断积累的收藏集如何促使他们重新诠释和反思。第二，我们介绍了**MyMuse**，这是一座参观后个人博物馆，它重新利用参观者在探索单件艺术作品时记录的问题和解读，既将其作为后续探究的情境，也将其作为叙事报告中的证据。叙事报告以视觉方式整理跨作品的这些记录，让参观者反思自己对哪些内容作出了回应，以及如何回应。第三，我们提供了受控评估的证据，探讨参观者对鲜明再次观看体验、跨作品连接，以及对反复出现的兴趣的觉察与表达。

2 相关研究

2.1 摄影作为选择性注意与个人记录

拍摄艺术作品会影响参观者关注的内容以及之后的记忆。Henkel 发现，与单纯观察相比，拍摄完整的博物馆藏品会损害之后的记忆，而拍摄选定的细节则不会 [17]。自主摄影也能引导视觉注意力，使人们关注自己选择拍摄的方面 [4, 11]。这些研究发现表明，拍摄涉及选择性注意，而非被动存储。

参观者拍摄作品的原因，单从图像本身可能并不明显。博物馆参观者出于不同原因拍摄艺术作品，包括记忆、研究、灵感、分享和自我表达 [40]。关于记忆导向记录的研究同样强调，有选择地捕捉而非穷尽式拍摄的价值 [32, 36]。Niforatos 等人进一步发现，与自动拍摄的照片相比，手动拍摄的照片被认为更有意义 [27]。然而，有意的选择并不一定会明确留下其理由：Petrelli 等人发现，经过仔细选择的个人线索往往很少包含解释其为何被选中的注释 [32]。在 Myartspace 中，学生重新查看博物馆照片时，难以识别其中一些图像，也难以将它们与未链接到照片的配套音频和文本评论联系起来 [42]。因此，艺术品照片可以保留<: style id='6'>被选择了什么，却无法保留<: style id='8'>它为何重要。

2.2 情境在艺术品诠释中的作用

博物馆参观经历会受到访客自身所带来的因素，以及经历发生时所处情境的影响 [14, 31]。Falk 和 Storksdieck 运用情境学习模型发现，相互作用的个人、社会和物理因素会影响参观者学习，但没有任何单一因素能够充分解释不同访客的结果 [13]。在艺术观赏中，Brieber 等人发现，与计算机呈现的复制品相比，在博物馆中观赏的艺术作品得到更正面的评价，也更容易被记住；博物馆观众在检索时还会利用空间布局线索 [6]。

意义也可能在最初的参观经历之后继续保持开放。Morse等人发现，有意义的博物馆体验可能在参观之前、期间或之后形成，涉及多个相互作用的意义来源，并且在个体层面上始终具有主观性 [26]。Chatterjee和Vartanian 回顾了相关证据，指出标题、真实性、风格知识以及历史或文化语境等语境信息会影响审美反应 [8]。在一项雕塑花园研究中，Fosh等人邀请访客在阅读官方解释之前先诠释作品，发现他们的解读有时会与官方解释有所不同 [15]。综合来看，这些研究表明，艺术作品的意义并非仅由图像决定 [43]，随着更多语境信息的出现，人们还可以重新审视其意义 [7]。

2.3 回顾与整理个人记录

个人记录若能留存下来，以便日后回顾和整理，便可支持反思 [12, 29, 50]。个人信息学研究表明，收集、整合与反思是处理个人数据时相互依存的阶段 [22]。创作实践中也存在类似挑战：积累的材料可能难以回顾，也难以整体理解 [10, 34]。

现有系统提供了扩展过往记录和整理个人收藏集的方法。在 InEcho 中，用户会为日常经历的过往记录添加反思，并回顾自己对同一事件不断变化的反应 [20]。个人策展让人们能够围绕自身兴趣收集和重新整理材料 [37, 48, 53]。TravelGalleria 运用空间组织，支持人们通过个人照片进行回忆与反思 [52],；SemanticCollage 则支持探索视觉集合中的关系并获得整体概览 [21]。

其他系统利用语境信息，支持对个人媒体和艺术作品进行探究。OmniQuery 整合个人媒体中的情境，以支持个人问答 [23],；VirtuWander 则在虚拟博物馆中利用访客探询和参观语境，提供指导、艺术作品比较和参观总结 [47]。ArtRAG 运用结构化的文化、历史、风格和主题知识，支持有据可依、多视角的艺术品解释 [45]。

叙事呈现也被用于组织记录的信息，并将引导式阅读与进一步探索联系起来。Martinez-Maldonado等人采用分层叙事，支持对记录的学习活动进行反思，将聚焦视图与语境注释相结合 [25]。Gratzl等人将视觉故事与其底层分析状态关联起来，使读者能够从引导式呈现转向进一步探索 [16]。DataShot和Calliope等自动化数据叙事系统进一步探索了如何从数据中选取显著事实，并将其编排为结构化叙事 [38, 46]。

MyMuse 与上述系统的不同之处，在于它所处理的材料以及这些材料的再利用方式。Myartspace、Echo 和 TravelGalleria 支持回顾过去经历的记录。VirtuWander 和 ArtRAG 支持探究虚拟导览中遇到或以图像形式呈现的艺术作品，而 Omni-Query 则回答有关个人媒体的问题。MyMuse 将有依据的艺术品探究应用于参观者在自己参观博物馆时拍摄的照片，保留他们提出的问题和诠释并将其作为逐字记录保存，再利用这些记录为之后关于其他作品的回答提供参考。在收藏集层面，它会重新组织相同的记录，以支持人们反思贯穿不同作品的反复出现的兴趣和关系。

3 定性形成性研究

艺术品照片可能无法保留一件作品为何重要、额外情境如何影响其诠释，或它与收藏集中的其他作品有何关系。我们开展了一项形成性研究，使用参与者自己的艺术品照片，考察他们还记得哪些过去的博物馆参观经历、哪些内容已无法再回忆，以及他们希望如何使用这些记录。他们的叙述为MyMuse支持重新审视单件艺术品以及考察收藏集中的关系提供了依据。

3.1 研究设置

参与者与流程。 我们招募了十名参与者（N =10；9名女性，1名男性；年龄21至28岁），他们在前一年内曾参观过博物馆或美术馆。每位参与者提交了十张来自本人参观经历的艺术品照片，共计100张照片。我们使用过去参观时拍摄的照片，而不是为本研究安排的参观中拍摄的照片，这样既能确保这些照片是参与者出于自身目的拍摄的，也能考察参观一段时间后，参与者还能从这些照片中回忆起什么，以及无法回忆起什么。每个60分钟的研究环节都包括一次关于博物馆参观、摄影、信息寻求和再次观看实践的半结构化访谈，以及一次使用参与者照片开展的照片引导式访谈。对于每张照片，我们都询问了拍摄原因、仍然记得的内容、已经遗忘的内容，以及参与者是否获得了关于该艺术作品的新知识，或是否改变了他们打算如何使用这张照片的想法。我们还询问了他们希望用于再次观看艺术品照片、记录反思和提出问题的系统应如何组织。该研究经本机构的机构审查委员会批准，所有参与者均提供了知情同意。

分析。 我们使用反思性主题分析法分析了538个问答单元 [5],，形成了四个主题。除此之外，我们还采用《文化对象编目》的九个章节级范围（CCO） [2]，刻画参与者所表达的与艺术作品相关的兴趣范围。在这项分析中，两名研究人员对90项参与者—章节判断进行了编码，并取得了高度一致性（κ = .81；92.2%）。覆盖范围分析为第4节所述的解释性视角提供了依据。访谈主题、示例性编码—主题关联，以及CCO分析的详细信息见附录A。

3.2 研究发现

我们确定了四个主题，涉及参与者能够从照片中重新获取什么、他们如何重新审视艺术作品，以及他们希望在整个收藏集中理解什么。

F1. 照片或许能留存艺术作品，却未必能留存选择它的缘由。参与者常记得自己选过某件作品，却不记得是什么促使他们做出这一选择。P07 直接描述了这种困难：“我在班克斯博物馆只拍了几张照片。如果我拍了这张，那肯定是有原因的，但我不记得了。”参与者还报告说，他们会忘记作品的标题、艺术家，

以及地点，因此有时必须先辨认出作品，才能重新思考自己当时的反应。既有研究表明，摄影具有选择性，访客拍摄作品的原因也各不相同 [11, 40]。这些发现表明，仅凭照片可能无法提供足够的情境，帮助访客回忆是什么吸引了他们对某件作品的注意。

F2. 对艺术作品的回应借鉴了个人知识、角色与经历。 参与者对作品的关注点受到其知识、角色与经历的影响，这与此前关于博物馆体验和艺术欣赏中个人情境的研究一致 [8, 13]。身为助理策展人的 P06 描述了个人参观与因专业目的参观之间的差异："作为访客，我会关注自己喜欢哪些绘画作品。工作时，我更关注为何选择某件特定作品，以及参观者如何体验这场展览。"P05 将一件由色彩鲜艳的药片组成的装置作品与自己服药治疗的一段时期联系起来，并想象如果当时服用的是作品中那样的药片，感受会如何。这些叙述表明，要理解参与者的回应，可能需要了解艺术作品本身信息中未体现的个人情境。

F3. 新知识与相关作品可能改变访客关注的特征。 在获得更多情境知识后，参与者描述了自己对熟悉艺术作品的不同关注点。学习韩国艺术史后，P06 回忆起自己的注意力从柳永国作品的色彩与构图，转向风景如何被转化为几何形态。P07 则描述了在了解塞尚的独特之处后，意识到先前未曾注意的特质："我明白了，他似乎确实是那样画的。" 参与者还提到，在接触相关艺术作品后会重新观看先前看过的作品。P08 描述了接触同一位艺术家的作品或主题相似的作品，如何改变自己先前形成的诠释；当一件新作品似曾相识时，P01 会查找先前拍摄的博物馆照片，想弄清两者有何不同。参与者认为相关的情境各不相同，涉及历史时期、艺术技法等方面。因此，访客可能会带着后来学习和接触其他艺术作品所引发的问题，重新观看熟悉的作品。

F4. 收藏集中反复出现的主题不一定代表个人兴趣。参与者将自己的照片描述为一个希望深入查看的收藏集，按参观经历或个人认为有意义的主题对作品进行归类，并询问反复出现的主题对象是否反映了更广泛的兴趣。他们拍摄照片的目的也包括记录和未来可能的用途。P09 拍摄作品，是为了将来可能用于写作和项目规划；他们提醒说，拍下那些看起来值得记录的作品，可能会"污染"推断出的品味画像。P07 认为，反复询问某位艺术家或某个历史时期，或许能表明一种单靠照片无法显现的兴趣。这些说法表明，访客可能会借助自己拍摄照片的原因，以及反复提出的问题，来解读收藏集中反复出现的主题。

4 设计考量

我们将形成性研究发现和后续设计反馈转化为 MyMuse 的四项设计目标。这些目标旨在解决如何保存参观者的贡献，并支持在单个艺术品记录内及不同艺术品记录之间开展探究与反思。

DG1. 让访客自己的贡献在艺术品记录中清晰可辨。为保留访客回答的个人情境（F1、F2），MyMuse 会将他们的问题和反思原样存储在每张艺术品照片旁。这些贡献与目录信息及系统生成的回应保持区分，因此访客可以辨认出自己贡献的内容，并在后续探究中重温自己的原话。

DG2. 保持艺术品记录开放，以便进行进一步探究。参与者描述了新知识和相关艺术品如何促使人们重新思考（F3）。因此，我们设计的记录能够容纳更多问题和解读，同时保留先前的条目供查阅。访客可以查阅语境信息、探索相关艺术品，并添加由此产生的回答。为探究提供指引，我们采用了《文化对象编目》(CCO) [2] 作为外部描述参考。根据这些视角在形成性数据中的覆盖情况，以及它们引导人们重新关注的适用性，我们选取了四个以艺术作品为中心的视角：创作者 物理特[illegible]访客[illegible]录其范围以外的个人联想和关注事项。

DG3. 以访客撰写的记录为依据，呈现藏品层级模式。参与者希望探究反复出现的兴趣，同时指出，拍摄作品并不一定意味着偏好（F4）。因此，叙事报告会汇总访客撰写的记录，涵盖不同艺术作品，并呈现反复出现的方面和关联，同时列出作为依据的条目。访客可以查看这些证据，并思考报告所呈现的关联是否反映了他们的兴趣。

DG4. 将对单件作品的探究与收藏集层面的反思联系起来。参与者希望重新查看具有个人意义的收藏集（F4），这推动了我们同时支持考察作品之间的关系，以及返回特定艺术作品的功能。我们设计了互为补充的视图，访客可以在其间切换。访客可以从收藏集层面的模式返回相关艺术作品和记录，以考察其贡献的情境，并继续进行探究。

5 MyMuse：系统设计

MyMuse 通过个人记录支持参观者在参观结束后持续参与艺术品照片，这些记录可供参观者再次查看并进一步扩展。在艺术品层面，参观者可以回顾作品、提出问题、添加反思，并探索相关艺术品。他们的贡献会被原样存储，并用于指导后续探究。在藏品层面，叙事报告会重新组织累积记录，以呈现不同作品之间反复出现的方面和关系。以下章节将介绍 MyMuse 如何呈现这些记录、如何在交互过程中使用这些记录，以及如何将其组织起来进行藏品层级反思。图2概述了系统架构，包括艺术品层级交互、个人记录构建和藏品层级反思。

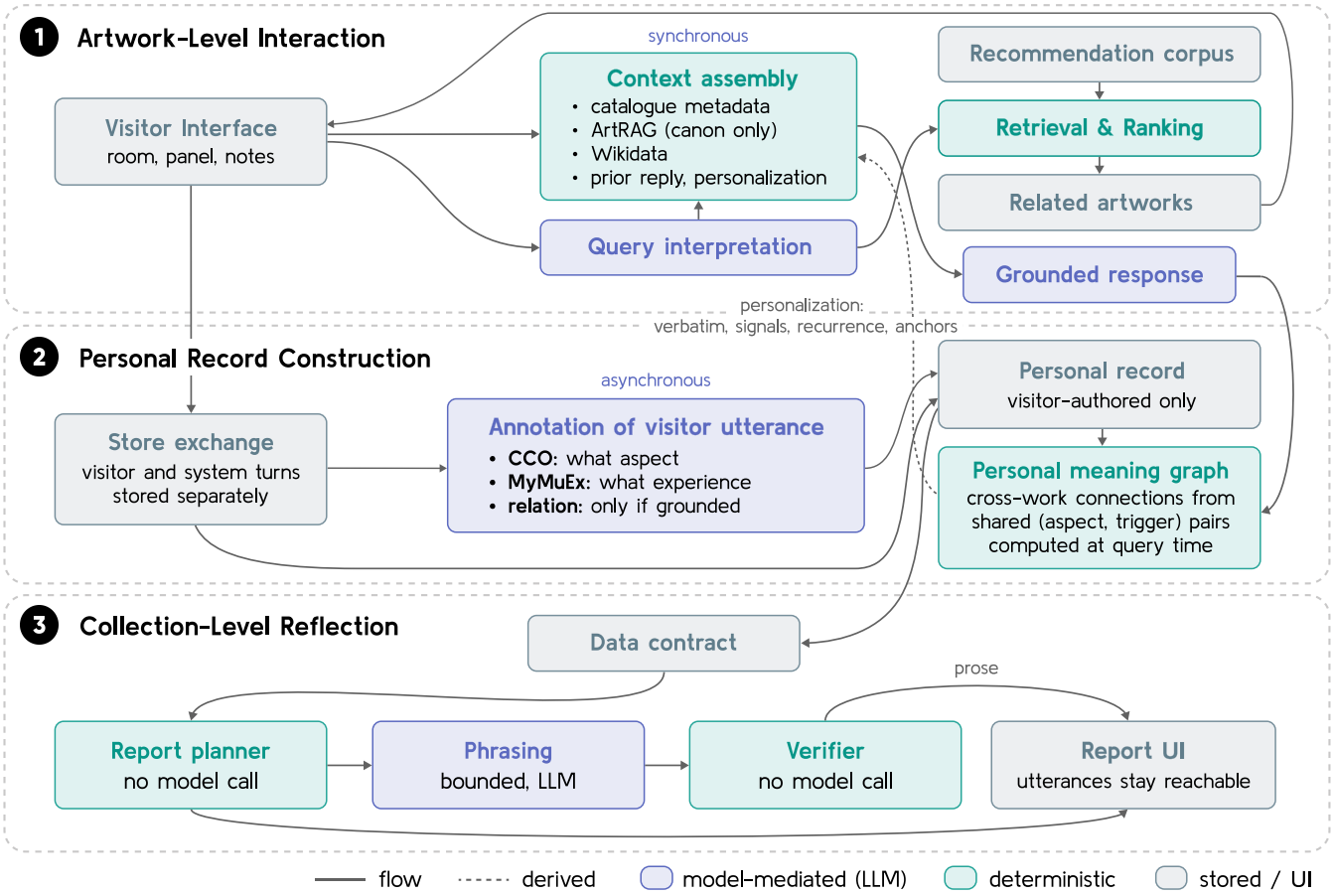


图2：MyMuse 系统架构。根据访客的请求，MyMuse 会生成有依据的回应、检索相关艺术品，或同时执行这两项操作。访客与系统轮次分别存储。只有访客撰写的文本会被标注并纳入个人意义记录。个人意义图谱（PMG）中的跨作品连接会在查询时根据共享的已标注（方面、触发因素）对计算得出，不会作为持久边存储。对于藏品级报告，确定性规划器会直接向报告界面提供报告结构和回退措辞。语言模型输出仅用于措辞表达经筛选的观察结果，并且在使用前会经过验证。

5.1 返回个人博物馆藏品

MyMuse 将参观者的艺术品照片呈现在一个类似画廊的房间中，以便参观者返回查看单件作品及其相关记录（图 3）。采用空间表示的设计源于参与者希望将照片作为具有个人意义的收藏集反复查看的兴趣（F4）。后续设计反馈显示，参观者的偏好各不相同：一些参与者偏好二维视图，以便浏览更大的收藏集，而另一些参与者则认为空间表示能让他们的照片具有“我的博物馆”的感觉。此前的研究也曾使用画廊环境来支持参观者在虚拟空间中重新查看个人媒体 [52]。这些观点促成了我们将类似画廊的房间与叙事报告结合使用的设计，后者支持比较作品并考察整个收藏集中的关系。

MyMuse 会自动将参观者在同一展览、同一墙面上遇到的作品照片归为一组。这种编排保留了照片与参观者所参观展览之间的关联但不会复现展览原有的空间布局。每张照片都关联至一条包含经核实的元数据的目录记录，其中包括标题、艺术家、日期、媒介以及收藏集（如有）（第 5.7 节）。选择一件作品后，会打开其交互面板，参观者可以在其中查看现有记录、提出问题、添加个人观察，并探索相关艺术品。

5.2 有依据的艺术品探究

MyMuse 基于艺术作品的目录记录、外部语境和选定的先前记录构建回答。目录记录以经核实的元数据形式提供，并附有其来源和置信度等级。提示要求模型将有关该作品的事实性陈述建立在此记录之上，绝不与其标题、艺术家、日期、媒介或艺术流派相矛盾或替换这些信息，并保留归属和大致日期等限定语。模型可以在与该记录一致的情况下，根据自身知识补充艺术史背景。

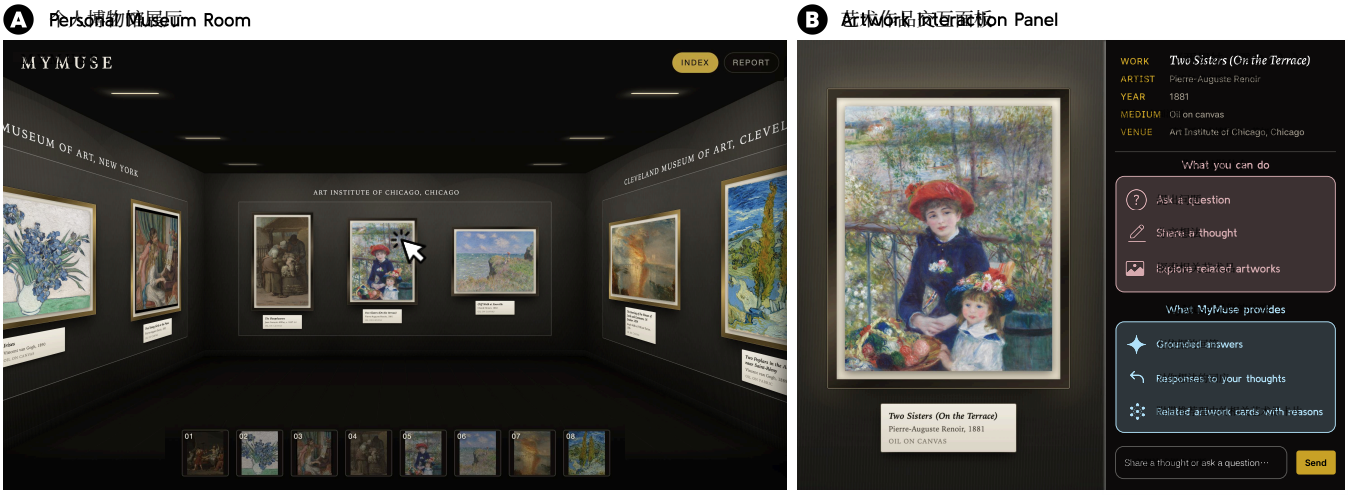


图 3：MyMuse 个人博物馆界面。(A) 个人博物馆在类似画廊的房间中展示访客拍摄的艺术品，供其重访自己的收藏集。艺术品按展览分组，并配有用于导航的缩略图条。(B) 选择一件艺术品后，会打开交互面板，显示其图像和元数据。参观者可以提问、分享想法并探索相关艺术品。MyMuse 提供有依据的回答、对访客反思的回应，以及附有说明的相关艺术品卡片。

记录，且不将推断呈现为博物馆记录事实。MyMuse 从维基数据添加艺术家信息，包括艺术流派、国籍、职业和生卒日期。当艺术作品符合 ArtRAG 的媒介和区域限制，且其图谱中包含该艺术家时，MyMuse 还会添加来自 ArtRAG 的结构化艺术史背景 [45], ，这些内容从目录记录中检索，并根据访客照片重新排序。回复上下文还可能包括 MyMuse 关于当前作品的先前回答，以及从访客先前记录中受限选取的内容；提示会将这些内容标识为访客的观点，而非关于艺术作品的事实。上下文组装期间，这些来源会保持分离，以便区分目录信息、检索到的上下文、系统回答和参观者撰写的内容。附录 D.3 展示了研究中的一段对话，并标明每条回答所用的来源。

5.3 从对话到个人意义记录

MyMuse 会逐字存储访客话语并为其添加注释，以支持后续探究中的检索和比较。系统响应和衍生注释会与原始话语分开存储。注释会在面向访客的响应返回后异步生成，因此交互可以继续进行，同时更新个人意义记录。

这些注释使用了两套已发表的词汇表。CCO [2] 描述话语所涉及的艺术作品方面，而 MyMuEx [26]描述体验如何变得有意义。在 CCO 的 123 个元素中，我们将 21 个作品图像元素指定为主要的访客可寻址集合。对于后续的探究和藏品层级反思，我们通过

第4节所述的四个视角来组织这些注释。MyMuEx 提供了 23 个体验触发因素（例如，学习、审美和灵感），其中效价单独表示，而时间性和范围则描述影响何时发生以及影响的广泛程度。对于体验注释，MyMuse 使用我们在形成性研究中确定的 21 个触发因素（附录A.4）。这两套词汇表使记录能够区分所讨论的艺术作品方面与访客对该方面的体验。

CCO 和 MyMuEx 会分别添加注释。只有当访客的话语支持二者之间的关系时，MyMuse 才会记录这种关系。仅仅共同出现并不足以构成依据。每条注释都会保留其在原始话语中的支持片段。只有访客撰写的文本才能作为有关访客的证据，因此，系统响应中引入的信息不会成为个人兴趣或诠释的证据。

5.4 利用先前记录为探究提供情境

MyMuse 会选择相关的先前记录，为新的请求提供参考。它根据累积的注释生成个人意义图谱（*PMG*），并将具有相同注释 CCO 方面和 MyMuEx 触发因素的不同艺术作品记录连接起来。这些连接会在查询时根据底层记录计算得出，用于检索相关情境，而不会被存储为关于访客偏好或特质的持久性主张。

对于每个请求，MyMuse 会利用四种情境。逐字回忆检索与当前作品相关的评论，而作品内信号则识别该作品内部被反复探讨的方面。跨作品重复出现识别在多件作品中讨论过的方面，而跨作品锚点则检索与其他艺术作品相关的评论，这些作品通过共享的（方面、触发因素）对相互关联。这些来源使响应能够引用先前与当前探究

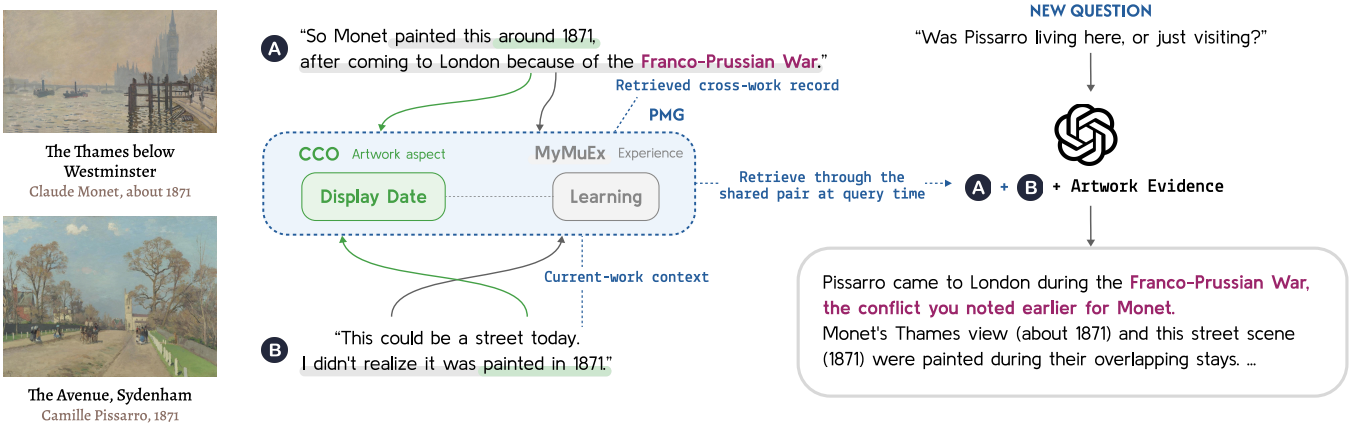


图4：示例说明 MyMuse 如何利用个人意义图谱（PMG）检索较早的记录，以生成后续回复。记录 A 和 B 涉及同一位访客观看的不同艺术作品，并共享 CCO 方面展示日期 和 MyMuEx 触发因素学习。查询时，MyMuse 通过这一共享的注释对检索记录 A，并将其原始措辞与当前作品记录 B、艺术作品证据以及新问题一并纳入回复上下文。高亮段落回溯了访客先前对莫奈的观察。这里只展示与这一连接相关的注释。

相关的贡献。图 4 说明了 PMG 连接如何支持检索另一件艺术作品中的较早记录，以便为后续响应提供信息。

5.5 探索相关艺术品

MyMuse 为访客提供相关艺术品，以便他们与当前作品进行比较并进一步探究问题。它从约 12,800 件作品组成的语料库中检索这些艺术品，该语料库汇集了六个来源：克利夫兰艺术博物馆、大都会艺术博物馆、泰特美术馆、芝加哥艺术博物馆、通过维基数据获取的维基共享资源，以及 Google艺术与文化。该语料库独立于访客自己的照片，并且每件作品都保留了对其源记录的引用。

检索取决于访客所请求的关系。艺术家请求使用明确的艺术家匹配，并辅以对大都会艺术博物馆收藏集的实时搜索；主题请求则使用受控的目录术语。对于色彩、构图或情绪等视觉维度，MyMuse 会将艺术作品的 CLIP嵌入 [33] 与语料库作品的 CLIP嵌入进行比较；在可能的情况下，艺术作品图像会从访客的照片中裁剪得到，并通过针对特定维度的文本探针进行比较。未指定维度时，检索使用整体嵌入相似度。这些方法通过目录关系和视觉推导出的相似性，支持访客进行探索。

5.6 收藏集层面的叙事反思

叙事报告会重新组织各艺术作品中的累积记录，使参观者能够考察其收藏集中的反复出现的方面及关系（图 5）。我们使用叙事可视化逐步呈现这些关系，将选定的观察结果与其支持性证据相联系。确定性规划器会选择入口点、支持性证据及组织方式。随后，语言生成会将选定的观察结果表述出来。

引导式阅读与交互式重新组织。MyMuse 通过滚动叙事界面呈现报告。艺术作品图板始终可见，同时其排列会随着叙事发生变化，使参观者能够在不同的组织方式中追踪相同的作品。参观者可以查看支持所报告关系的话语。在引导式流程结束后，他们可以自行重新组织收藏集，遵循 Segel 和 Heer 所描述的马 Martini 酒杯结构 [35]。这一转换使参观者能够先考察所呈现的关系，再探索其记录的其他排列方式。

准备证据。对于每位参与者，规划器都会接收一份结构化记录，其中包含访客撰写的贡献、艺术品元数据、注释、支持性证据片段、情境记录和来源。这些记录与艺术作品层面系统所使用的访客撰写记录相同。系统生成的回复不作为有关访客的证据。对于四个视角中的每一个，支持度均以携带相应 CCO 章节证据的不同艺术作品数量来衡量。因此，反复讨论同一件作品不会增加支持收藏集层面模式的作品数量。某一方面必须出现在至少两件不同的艺术作品中，才可视为收藏集层面的复现。未与艺术作品关联的情境证据，则按不同记录计数。

选择入口点和锚定作品。规划器在选择入口点时，会考量四个视角下反复出现的内容、MyMuEx 触发器以及语境证据。如果没有任何内容达到重复阈值，报告便从单件艺术作品开始。在四个视角中，主导视角根据其支持性艺术作品的不同数量来选定；若数量并列，则以话语数量决胜；若仍然并列，则按照固定的 CCO 章节排序确定。对于支持该视角的作品，规划器会选择具有最多特定视角证据的作品作为锚定作品，只有在并列时才以该作品周围更广泛的活动作为决胜依据。这些规则根据现有记录中的证据，确定报告的起点

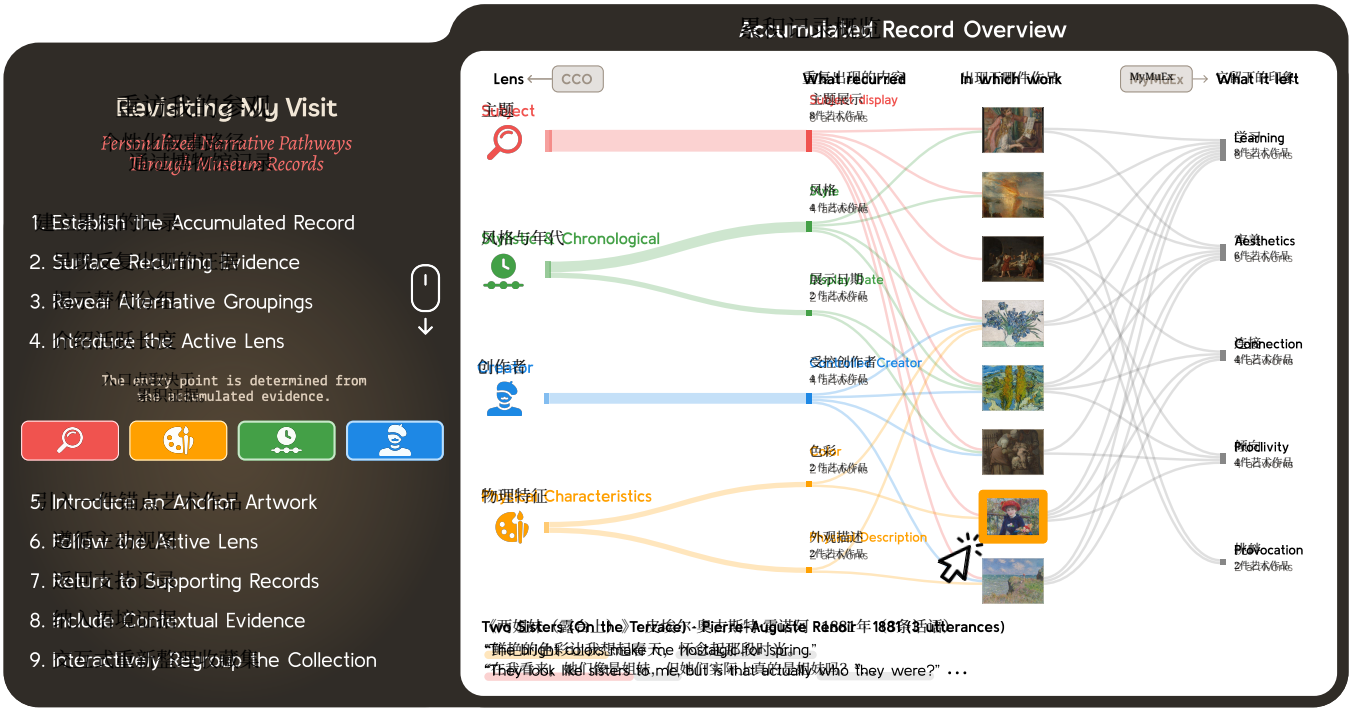


图 5：MyMuse 叙事报告界面。左侧：叙事结构采用四种解释视角，涵盖从引导访客回顾累积的访客记录，到对收藏集进行交互式重新分组的过程。MyMuse 从这些记录中选取一个入口点，并在具备所需证据时纳入依赖证据的阶段。右侧：重复证据视图将经过注释的 CCO 艺术作品方面、艺术作品和 MyMuEx 体验触发因素联系起来。选中某件艺术作品后，图示下方会显示作为依据的话语摘录，让参观者查看支撑这些注释的记录。

。参观者可以切换视角，以审视同一收藏集中的不同关系。

撰写报告。底层记录不变，生成的叙事计划也相同。报告呈现累积的收藏集、展示反复出现的证据、引入一件锚点作品，并从不同视角展现各种关系。只有在现有证据支持的情况下，才会显示相应步骤。因此，不同参观者看到的叙事顺序可能有所不同，但针对相同记录重复生成时，顺序保持稳定。

通过四个视角组织作品。每个视角都使用不同属性来组织选定的艺术作品。主题视图通过共享的受控目录主题词将作品关联起来。访客话语和 CCO 注释决定哪些作品会纳入叙事，而目录元数据则决定作品之间的连接。因此，此视图中的连接表示被讨论作品之间存在目录关系，并不意味着访客明确表达了这种关系。物理特征视图围绕色彩、材料或技法等属性组织作品，其中色彩和明度根据访客的照片进行测量。风格与年代视图按时间、艺术流派或时期排列作品，而创作者视图则按艺术家对作品分组。访客可以使用这些视图，考察报告中所含作品之间的

生成并核查报告文本。规划器确定视角、锚点作品、分组方式和支持性证据后，语言模型会接收选定的事实。它的作用仅限于措辞表达有依据的观察。生成的文本会经过检查，以排除缺乏支持的主张、推测性语言，以及描述稳定的访客特征的陈述。如果文本未能通过这些检查，MyMuse 会采用确定性回退措辞，在不依赖被拒绝文本的情况下保留规划好的内容。

5.7 实现

MyMuse 是一款面向桌面浏览器的 Web 应用程序。个人博物馆通过 CSS 3D 变换进行渲染，访客使用方向键在房间中行走，并用鼠标环顾四周。FastAPI 服务器搭配 SQLite 数据库，为界面提供服务，并存储目录记录、访客与系统轮次、注释和交互日志。MyMuse 自身的语言模型调用通过 OpenAI API 使用 gpt-5.4-mini，分类和注释调用的温度设为 0。ArtRAG 组件在内部检索步骤中使用 gpt-4o-mini，包括基于图像的重排序。响应模型仅接收文本。访客照片用于 ArtRAG 的基于图像的重排序、CLIP 检索以及报告中的色彩测量指标。我们在本地为拍摄的艺术品和相关艺术作品语料库计算 CLIP ViT-B/32 嵌入 [33] 。

对于每一轮访客发言，MyMuse 会判断消息是在询问答案、相关艺术品，还是两者兼有，并确定消息提及了哪些检索维度。随后，系统会组装回复上下文（第 5.2 节）、生成回复，并将话语和回复分别存储为独立记录，同时保存回复所用的来源。

照片在加入个人博物馆之前，会先与目录记录关联。MyMuse 不会根据照片识别艺术作品。每条记录都包含其来源和置信度，这两项信息都会纳入回复上下文。附录 B 列出了各项功能的输入，以及是否需要目录匹配。

在研究中（第 7 节），参与者通过此 Web 应用程序使用 MyMuse，服务器运行于一台机器上（Apple M4 Max，128 GB 内存）。在各个 MyMuse 研究环节中，系统共返回 332 个回答，服务器端响应时间中位数为 3.3 秒（第 95 百分位数：8.7 秒）。

6 技术评估

我们评估了 Artpedia 上的艺术作品级回答生成、视角层级上人工与模型标注之间的一致性，以及对未更改记录进行重复评估时的报告构成稳定性。

6.1 艺术作品级回答生成 n

为评估支持艺术品级查询的答案生成效果，我们在一组固定的 310 个 Artpedia [39] 测试条目上，使用 gpt-4o-mini 对 MyMuse 和 ArtRAG [45] 进行了比较；这些条目均包含可恢复的图像。不同指标的结果有所差异：例如，MyMuse 的 ROUGE-L [24] 得分更高（27.49 对 22.74），但 CLIPScore [19] 得分低于 ArtRAG（77.49 对 84.88）；两项指标均乘以 100。对于两个系统都使用 gpt-5.4-mini 作为生成模型时（如已部署系统中所用，第 5.7 节），结果呈现相同模式（ROUGE-L 23.71 对 20.28；CLIPScore 77.36 对 83.23）。两种配置在输入和提示方式上有所不同。完整的方法、结果、响应长度，以及按 ArtRAG 知识图谱覆盖率进行的分析，详见附录 C.1。

6.2 注释一致性

我们评估了艺术品级查询和藏品级组织所用四个视角的一致性。两位研究人员使用同一份代码本，在无法查看模型注释的情况下，独立标注了研究1中 MyMuse 状况下的全部 332 条访客话语。我们将每位编码员及模型对 CCO 元素的分配，映射为每条话语对应的四个二元存在／缺失标签。四个视角之外的元素不计入这些标签。为了与模型进行比较，我们合并了人工注释：只要任一编码员为某个视角赋予标签，就将该视角视为存在。

编码者间一致性（Cohen’s κ）合并值为 94.8%（一致），各视角的数值介于.77至.89之间。模型与合并后的人类标签之间的一致性合并值为 89.4%，各视角的数值介于.65至.77之间。对于 64.5% 的话语，模型匹配了人类标注的完整视角集合。

排除任一编码员标记为不确定的20条话语后，模型与人工一致性仍然 $\kappa = .70$ 。另有一项分析中，排除模型和人工参考均未分配任何视角的话语后，260条话语中得到 $\kappa = .67$ 。这些结果表明，许多单独的视角标签具有一致性，但完整透镜集合存在差异，可能影响哪些记录会被归入各个视角。附录 C.2 报告了详细的视角层面结果和敏感性分析。

表1：在默认温度下，对相同记录进行10次重复的稳定性。各单元格报告16名参与者的均值（样本标准差）。Jaccard值为每位参与者保留的重复结果两两之间的平均值。模式归一化结合了CCO章节映射、MyMuEx触发映射以及未匹配短语的归一化，详情见附录C.3.3。

测量指标	A	B	C
条目透镜分歧	0.246 (0.214)	0.240 (0.213)	0.000
锚点艺术作品分歧	0.304 (0.207)	0.304 (0.207)	0.000
作品集杰卡德相似度	0.888 (0.145)	0.888 (0.145)	1.000
模式集 Jaccard 指数	0.255 (0.166)	0.254 (0.163)	1.000

6.3 报告构成

我们考察了在对同一批记录进行重复评估时，报告构成是否保持一致，以及生成后验证是否降低了变异性。我们使用用户研究（第 7 节）中 16 名参与者的记录比较了三种状况。

在 **A** 中，语言模型在提示约束下选择报告构成。**B** 则将 MyMuse 的验证器应用于在 A 中生成的相应报告。在 **C** 中，MyMuse 的确定性规划器选择构成，再由语言模型组织计划内容的措辞。A–B 比较评估了验证后的变化，B–C 比较了模型选择的构成与确定性规划的构成。

所有状况均接收相同的参与者层级记录。A 和 B 使用相同的输出模式及适用的叙事约束。我们评估了 16 名参与者、$\times 3$ 种状况、 $\times 10$ 次重复和 $\times 2$ 个温度设置，共进行了 960 次状况评估。语言生成使用 gpt-5.4-mini。对于 C，我们绕过生产缓存，以评估重复运行。提示词和执行细节见附录 C.3。

在 A 和 B 中，跨重复运行选定的艺术作品集合比标准化模式集合具有更高的一致性（表1）。两种状况下，艺术作品集合 Jaccard 指数均为 0.888；而模式集 Jaccard 指数在 A 中为 0.255，在 B 中为 0.254。生成后验证在测试状况下几乎没有改变这些测量指标。降低温度也没有消除模型所选模式集合中的变异性。温度=0 被要求的运行结果见附录 C.3.4。

在 C 中，正如确定性规划所预期的那样，所有测得的构图决策在各次重复中均保持一致。在禁用语言生成的情况下，第5.6节所述的回退机制为全部16条参与者记录生成了结构完整的报告，且报告生成过程中未调用模型。

6.4 范围

注释结果涉及对四个视角的分配。本分析不评估单个 CCO 元素、MyMuEx 触发器或注释之间的关系。Artpedia 分数衡量的是与参考描述和图像的对应程度，但并不直接验证检索到的来源是否支持所生成的每项主张。基准测试还不包含访客撰写的探究历史，因此无法据此评估个性化或跨个人记录推理。报告分析衡量的是在对未改变的记录进行重复评估时的构图稳定性。第7.3节与参与者一同考察反复兴趣认知、跨艺术作品联系、意外发现和感知保真度。

7 用户研究

我们从 MyMuse 旨在支持的两个参观后参与层面评估了它：围绕单件艺术品的探究，以及跨累积收藏的反思。在单件艺术品层面，我们考察了延续访客自己的问题和反思，是否比信息内容相当的逐件艺术作品基线更能支持以个人体验为基础的重访和进一步探究。在收藏集层面，我们考察了重新整理跨艺术作品的同一批累积记录，是否能让反复出现的兴趣和跨作品关系更清晰，同时仍以参与者记录的内容为依据。我们通过在单次研究环节中开展的两项被试内研究考察了这些问题。由于评估在单次研究环节中进行，我们的结论关注即时效果，而非对诠释或参观博物馆实践的长期影响。

我们选择这一设计，是为了考察访客重新查看以往参观时拍摄的照片这一情境；在形成性研究中，参与者难以回忆自己当时为何拍摄这些作品（F1）。因此，参与者使用的是自己过去参观时拍摄的照片，而非专为本研究安排的参观中拍摄的照片。

7.1 研究设置

参与者与流程。 我们招募了16名参与者（P01– P16；12名女性，4名男性），年龄为19至55岁（$M = 26.5, SD = 8.3$）。大多数参与者是本科生或研究生，此外还有两名在职专业人士和一名全职家庭主妇。12名参与者报告没有艺术相关的学术或专业背景，另有4名报告具有艺术相关的学术或专业经历。所有参与者在此前12个月内至少参观过博物馆或美术馆两次（2–3次：5人；4–6次：4人；7次或以上：7人）。自评艺术知识范围为1至7（$M = 3.8$, $SD = 1.5$）。参与资格要求参与者此前参观博物馆或美术馆时亲自拍摄过至少20张艺术品照片。形成性研究中的参与者不具备参与资格。本研究经所在机构的机构审查委员会批准。所有参与者均提供了知情同意，并获得了20美元报酬。

每位参与者提交了自己在此前参观时拍摄的16张艺术品照片。对于有记录拍摄日期的照片，拍摄时间中位数为研究环节前128天（约4.2个月）。我们将每位参与者的照片随机分为两组互不重叠的照片，每组八张，分别用于研究1的两种条件。使用不同的照片组，减少了参与者在两种条件下重复观看同一件艺术作品所产生的残留效应。研究人员将每张照片与相应的目录记录进行匹配，依据的是参与者提供的信息、拍摄的墙面标签以及博物馆发布的记录；同时记录每条记录的来源和置信度，以确保两种条件均可获取相同的经核实的元数据和语境信息。每个研究环节约持续90分钟。参与者首先完成开始访谈，随后体验研究1的两种条件，直接比较这两个系统，查看研究2的两份报告，最后完成结束访谈。约10分钟的开始访谈涵盖参与者既有的博物馆参观、信息寻求和记录习惯，包括他们观看艺术作品时会留意什么，以及选择拍摄什么内容。这次半结构化访谈为解读他们随后作出的评价提供了背景。

条件与任务。 研究期间，我们以中立的方式将两种条件称为“系统 A”和“系统 B”，并将报告称为“报告 A”和“报告 B”。我们对研究1的两种条件和研究2的两种条件分别进行独立的顺序平衡，形成四种顺序组合，每种组合各有四名参与者。

研究*1*：*MyMuse* 交互与基线。 两种条件均允许参与者浏览自己提交的艺术作品、查看目录信息、提出自由形式问题、留下个人感想，并接收相关作品推荐。基线通过传统的逐件艺术作品聊天界面提供这些功能。它将参与者的条目保留为每件艺术作品各自的对话历史，且先前的条目不会影响后续回复。MyMuse 交互则提供了第 5 节所述的界面。它利用累积记录为后续回复提供情境，并在个人博物馆中以空间布局组织参与者的艺术作品和记录。因此，两种条件具备相同的核心信息与探索功能，但参与者的个人记录在组织和再利用方式上有所不同。

参与者按照平衡后的顺序使用每种状况约 15 分钟，并使用所分配的八张照片。在每种状况开始前，研究人员简要介绍可用的交互方式。随后，参与者自由探索艺术作品：浏览收藏集、提出问题、查看相关作品，并根据意愿留下反思。我们没有规定具体问题或探索目标，因为两种状况都旨在支持参与者自主开展参观后的探索，而不是完成预先设定的任务。每种状况结束后，参与者立即完成一次简短的半结构化访谈，随后填写相应的问卷。访谈重点围绕他们的总体体验、如何使用系统、系统是否支持他们重新回顾或进一步探索先前的观赏经历，以及他们可能如何在参观博物馆后使用此类系统展开。在分别评估两个系统后，参与者完成了一次比较性半结构化访谈，随后填写直接比较问卷。该访谈重点关注他们感知到的系统差异、各系统的有用之处，以及参与者的总体偏好。

研究*2*：结构化记录与叙事报告。两种状况均基于参与者在 MyMuse 交互过程中生成的记录。结构化记录按艺术作品呈现累积记录，将每张照片及其目录信息与参与者相关的问题、反思和探索记录归组在一起。叙事报告，其描述见

表 2：参与者选作进入叙事报告入口点的诠释视角。

Lens	参与者	*n*
创作者	P10, P16	2
实体	P01、P05、P07、P11、P14	5
风格	P02、P09	2
主题	P03、P04、P06、P08、P12、P13、P15	7

第5.6节，将这些记录中经过筛选和整理的证据组织成交互式跨作品叙事，并可访问作为支持材料的访客撰写条目以及进一步的艺术作品探索。因此，两份报告共享同一源记录，但在内容组织和交互方式上有所不同。研究2考察了参与者对这两种报告设计整体的体验。

参与者按照平衡后的顺序查看两份报告，每份报告约花费 10 分钟。对于叙事报告，MyMuse 会自动从参与者的累积记录中选取四种解释视角之一作为入口点（表 2）。参与者首先通过这一视角查看报告，随后可以自由探索其余视角和报告的其他部分。对于结构化记录，研究人员简要介绍了其按艺术作品组织内容的方式，之后参与者便自由探索报告。查看完每份报告后，参与者都会填写同一份包含四个题目的问卷。研究环节最后以比较访谈结束，讨论两份报告及整体体验。我们询问参与者注意到自己在不同艺术作品之间的兴趣和关系有何特点、发现了哪些新颖或意外之处、偏好哪份报告及其原因、每种形式适用于哪些用途，以及他们希望系统在初次接触之后还能提供哪些支持。

测量指标。 除非另有说明，问卷条目均采用 7 点李克特量表（1 = 非常不同意， 4 = 中立， 7 = 非常同意）。

研究*1*。 六个共同条目捕捉了参与者使用两个系统时体验的互补方面。反思性结果包括：（1）重新体验，即使用系统是否能让参与者生动地回想起此前的博物馆参观；（2）阐述，即系统是否帮助参与者更具体地发展关于某件艺术作品的想法；以及（3）响应间的关联，即系统是否帮助参与者注意到他们针对不同作品所作回答之间的共同点或关联。系统质量与意向 包括：（4）回答相关性，即系统回答是否符合参与者想了解的内容；（5）参与度，即参与者是否想进一步探索他们的艺术作品和记录；以及（6）使用意向，即参与者在博物馆参观后是否会再次使用该系统。

第七项共同指标测量可查找性，即参与者能否轻松找到并查看他们想要的艺术作品和记录。由于它涉及界面操作，而非上述结果，我们将其作为可用性检查单独报告。MyMuse 问卷还包含诊断性条目，用于评估基于以往记录的个性化问答和空间重建。由于这些功能在基线中没有直接对应项，我们对这些条目进行描述性分析。在使用两个系统后，参与者完成了三个双极 7 点直接比较条目。其中两个条目比较系统在响应间的关联和重新体验方面的表现，并要求参与者为每项判断提供书面解释；第三个条目询问他们在博物馆参观后更愿意使用哪个系统。无论呈现顺序如何，量表端点均保持不变。

研究*2*。 三个条目用于捕捉藏品层级的反思性结果：（1）反复兴趣认知，即报告是否帮助参与者注意到在多件作品中反复出现的兴趣；（2）跨艺术作品关联认知，即报告是否帮助参与者识别他们对不同艺术作品的体验或反应之间的关联；以及（3）意外发现，即报告是否揭示了参与者此前未曾考虑过的兴趣或关系。第四个条目评估感知保真度，即所呈现的内容是否准确反映了参与者实际留下的记录。我们将其单独报告，作为基础核查，而非藏品层级反思的测量指标。尽管研究1也包含一个与关联相关的条目，但这两项测量针对不同层级：研究1关注参与者重访单件作品时注意到的关联，而研究2关注在累积收藏中显现的关联。因此，我们将它们视为相关但不同的结果。

数据分析。 所有Wilcoxon分析均采用双侧符号秩检验，排除零差值，并在存在并列值时使用正态近似。我们以 $r = |Z|/\sqrt{N}$的形式报告描述性统计和效应量，其中 N表示参与者人数。对于研究1，六个共同结果项分别进行分析。我们报告未经校正的 p值及效应量，并将这些条目级比较视为指示性结果，而非验证性结果。Holm校正值见附录D.1。可查找性作为可用性检查单独处理；由于不存在对应的基线测量，因此仅对报告的MyMuse特定诊断项进行描述性汇总。对于研究2，三个集合层面的反思性结果分别进行分析，并对这三项比较应用Holm校正。感知保真度则作为基础核查单独考察。

这三个直接比较题项采用的是单一双极评分，而非按状况分别评分。因此，我们使用相同的Wilcoxon程序，将每项评分与量表中点进行比较，并报告方向计数和效应量。正文报告未经校正的p值；Holm校正值见附录D.1。

在获得参与者知情同意后，我们对访谈进行了录音并转录。我们依据研究问题，采用反思性主题分析对访谈转录稿和书面解释进行了分析。与形成性研究一样，本次分析属于解释性分析，而非分类任务，因此我们不报告评定者间一致性。定性分析用于解释测量结果，并为偏离汇总评分的案例提供情境化理解。

图6提供了四个说明性案例，展示了研究1中MyMuse互动期间生成的记录，后来如何通过研究2中的叙事报告重新组织并加以回顾。

P06 艺术课程学习经历；关注人物形象与表情。
Art coursework experience; attentive to figures and expressions.

P05 室内环境设计专业学生；西画辅修
Interior environmental design student; Western painting minor.

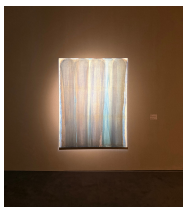
研究1：MyMuse互动记录
Study 1: Keywords from MyMuse Interactions



Ulysses and Penelope

"Could you tell me more about this scene's story?"

"Are there other paintings of the same myth?"



Blue Umber & Ultramarine

"I like the texture and the sense of layering here."

"How do the two differ?"

研究2：跨艺术作品的叙事报告
Study 2: Horizontal Reports Across Artworks

选定的条目视角 lens

主题 Subject

"Seeing everything brought together, I realized I was curious about myths and the stories behind the artworks."

物理特征 Physical Characteristics

"Seeing this, I realized for the first time that almost all my questions began with visual features."

P09 艺术与历史爱好者，对艺术家的世界观感兴趣。
Art and history enthusiast interested in artists' worldviews.

P10 现代法国史和女性史研究生。
Graduate student in modern French history and women's history.

研究1：MyMuse互动记录
Study 1: Keywords from MyMuse Interactions



Weeping Woman

"Could you tell me about the personal and historical influences on the artist when this work was made?"

"How does Picasso's cubism relate to the subject?"



Deification of a Soldier

"Tell me more about this artist from (rather than Japan)."

"Did Fukushima Prefecture also suffer several war damage?"

研究2：跨艺术作品的叙事报告
Study 2: Horizontal Reports Across Artworks

选定的条目视角 lens

风格与年代 Chronological

"I've always liked history, and I think this can help me see that preference again."

创作者 Creator

"I become clearer to me that I place a lot of importance on artists' backgrounds."

图6：两个研究组成部分中具有代表性的参与者案例。每个案例的上半部分展示了参与者在研究1的 MyMuse 状况下重访单件艺术品时所生成的记录示例。每个案例的下半部分展示了研究2的叙事报告中从参与者的累积记录里选出的切入视角，以及参与者在查看涉及多件艺术品的记录时所报告的观察示例。参与者的描述提供了情境背景，并未用于确定所选视角。附录 D.3 展示了 P09 与 MyMuse 的交流及其回答。

7.2 研究1：回访单件作品

图7总结了六项共同结果，完整统计数据见附录D.1。如第7.1节所述，我们将未经调整的条目级检验和直接比较检验视为探索性分析。

7.2.1 参与者重视能够回应其观察和问题的回答。与基线相比，参与者对 MyMuse 回答相关性的评分更高（M = 6.12 对 5.31；p = .006, r = .69）。这是唯一一个在经过校正后仍保持显著的共同结果

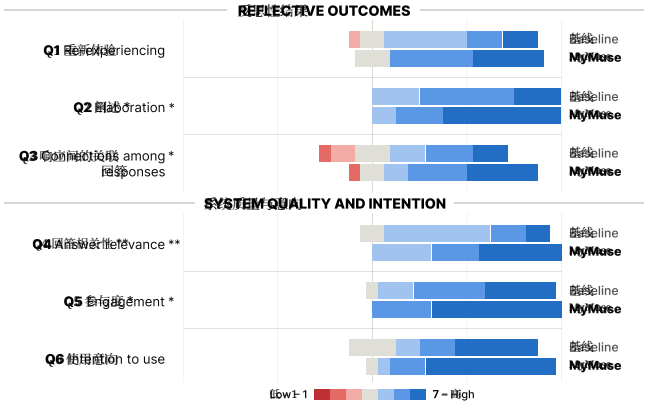


图7：研究1中基线和 MyMuse 条件下六项共同结果的评分。分歧条形图展示了李克特7点量表上的回答分布（1 =非常不同意， 7 = 非常同意），分为反思性结果和系统质量与意向。星号表示未经校正的双侧Wilcoxon符号秩检验结果（$^{*}p < .05$, $^{**}p < .01$）。如第7.1节所述，这些条目层面检验属于探索性分析。霍尔姆校正值见附录。

Holm校正，并对MyMuse特有的个性化问答条目给予了积极评价（$M = 6.06$）。他们认为，相关回答回应了他们想要了解某件艺术作品的内容。在P13评论伦布茨的《鲁特琴演奏者》中人物的视线后，回答解释说该人物正在调试乐器。P13表示，“它给出了我真正想要的答案，也解释了我的印象产生的原因。”其他参与者表示两种状况之间几乎没有差异（P04 、 P08 ），或认为回答很有用，但没有感觉这些回答是针对个人量身定制的（P15）；P10则描述了一条集中于风格的回答，尽管他们的问题涉及情境。由于参与者在不同状况下的贡献也各不相同，例如P13在MyMuse中输入了更多反思，因此，我们无法将感知到的相关性具体归因于先前记录的再利用。

7.2.2 参与者将空间浏览与生动的再次观看联系起来。不同条件下的重新体验评分没有显著差异（$M = 6.00$ 与 5.31；$p = .065$），不过在直接比较中，16 位参与者中有 13 位更偏好 MyMuse（$p = .001, r = .81$），且参与者对 MyMuse 专属的空间组织条目评价积极（$M = 6.06$）。几位参与者将生动感与在自己挑选的照片构成的房间中移动联系起来。P15 解释说：“我可以在那个空间里来回移动。我觉得这让记忆变得更加鲜明。”再次观看也包括了解那些早先印象已难以回忆的作品。P10 表示，两个系统都没有唤起他们初次看到某件作品时的情绪。其他人则觉得空间呈现帮助不大。P03 认为再次观看不够生动，因为房间与他们参观过的博物馆不同；P05 则偏好能显示语境信息的熟悉照片布局。P13 不喜欢在房间中移动，但总体上更偏好 MyMuse，并将更大的艺术作品展示归因于更好的回忆

效果。因此，对 MyMuse 的偏好并不一定意味着偏好空间导航。

7.2.3 参与者阐述并发展了自己对艺术作品的回应。 两种状况下的阐述评分都很高，且MyMuse中的评分更高（$M = 6.50$ 对比6.00；$p = .046$、 $r = .50$）。该条目关注的是感知到的支持，而不是输入文本的数量或类型。参与者将自己的文字与艺术作品的呈现方式联系起来。P05在MyMuse中写下了更多即时反应，因为“感觉就像在博物馆里欣赏艺术”，并在基线中提出了更多问题。类似地，P01在MyMuse中输入了更多记忆和早期印象，而P13则将其对当下感受的关注归因于更大的艺术作品和更小的问题窗口。

7.2.4 一起查看作品促使参与者进行了一些跨作品比较。参与者认为，MyMuse 在帮助他们认识到回答之间的联系方面支持度更高（$M = 5.75$ 与 5.00； $p = .032$, $r = .54$ 12 16 $p = .002$ ），并且在直接比较中更受青睐（, $r = .76$ ），不过这仍是其评分最低的共同结果。参与者所描述的比较大多涉及可见差异和展览背景。P16 只有在一起查看时，才注意到三幅同人画作之间的色彩差异；P12 和 P11 则发现，按博物馆或展览分组有助于追踪作品间的联系。参与者也重新审视了自己的拍照选择，例如 P04 注意到自己倾向于拍摄名作，但他们并未始终明确指出自己的问题和反思如何在整个收藏集中相互关联。P08 希望任一系统都能更好地帮助自己发现这类联系。

7.2.5 系统偏好反映了参与者的预期用途。MyMuse 的参与度评分更高（$M = 6.69$ 与 6.06； $p = .026$, $r = .56$ $p = .093$ ），而使用意向（ ）和可查找性（$p = .145$）则没有显著差异。当被问及博物馆参观后会使用哪个系统时，12 位参与者选择了 MyMuse，4 位选择了基线（$p = .049$， $r = .49$）。这 4 人中有 3 人在此前的直接比较中至少有一项更青睐 MyMuse。他们权衡了预期的归档和检索需求：P04 希望能轻松整理大量相机胶卷，P14 更喜欢基线系统，因为它便于持续记录感想；P10 则解释道：“若要归档并在之后重新找到内容， *[基线系统]*其实更方便。”

7.3 研究2：阅读累积收藏

图8总结了研究2的四项测量指标，附录D.1报告了完整的统计数据。三个收藏集层面的反思性结果采用Holm校正进行检验，感知保真度则单独进行了评估。

7.3.1 参与者识别出作品之间反复出现的兴趣和关联。 叙事报告在跨艺术作品关联意识（$M = 6.62$ 与 4.19； $p_{\text{Holm}} = .003$, $r = .83$）和反复兴趣认知（$M = 6.62$ 与 4.62； $p_{\text{Holm}} = .005$, $r = .76$）方面的评分高于结构化记录。参与者识别出了作品之间以及自身问题之间的关系。P06解释道：“我原以为几乎没有什么关键词能把这些作品联系起来，但把它们归为女性或女孩后，就更容易发现新的共同点了。”并意识到注意力

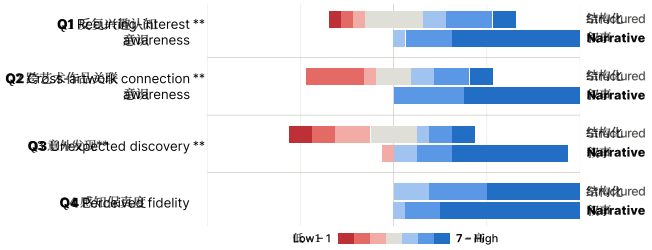


图8：研究2中对结构化记录和叙事报告的评分。分歧条显示了7点李克特量表上的回答分布（1 = 非常不同意，7= 非常同意）。对于三项集合层面的反思性结果，星号表示基于双侧Wilcoxon符号秩检验的Holm校正 p值所确定的显著性（$^{*}p < .05, ^{**}p < .01$）。感知保真度单独显示，作为基础核查，并未纳入Holm校正。

转向所描绘的故事与神话。P16 发现，关于单件作品的问题之间的关联比预想的更紧密，而 P02 一眼就能看到自己关于时期和风格的问题。结构化记录也引发了比较：按年份排列作品，让 P10 对去殖民化和战争产生了新的问题，也让 P05 开始注意到构图。

7.3.2 一些参与者重新思考了自己与艺术作品互动的方式。对于叙事报告，意外发现的评分也更高（$M = 6.31$ 对比 3.88；$p_{\text{Holm}} = .005$，$r = .72$）。一些参与者重新思考了自己的观看实践。P05意识到：“我提出的问题几乎都始于视觉方面，”P11则发现，色彩往往最先吸引他们的注意力。P15意识到，参观之后，他们想了解更多的是作品和艺术家，而非作品的视觉特征；他们在访谈早些时候便开始阐述这一区别。另一些参与者发现了自己熟悉的兴趣点，却没有新的发现（P01、P09、P13）；P03则质疑有限的记录是否足以支撑分析。参与者对于是否希望获得一份关于自己的描述也看法不一：P08将报告比作人格测试结果，而P13表示：“我看绘画作品时，其实不需要数据来分析我。”

7.3.3 参与者高度认可感知保真度，但对某些衍生关系提出了质疑。两份报告的感知保真度都很高，且没有显著差异（$M = 6.69$ 与 6.31；$p = .083$）。参与者认出了自己的贡献。P06 表示：“让我印象深刻的是，它确实摘录并整理了我分享个人感受的部分。”不过，他们仍对围绕这些贡献呈现的某些分类和比较提出了质疑。P05 和 P11 质疑基于色彩分布的比较，而 P05 认为材料与技法是更有用的依据；相比之下，P04 更认可聚焦色彩的展示。P01 指出，自己原本对战争感兴趣，但由于没有与系统讨论这一主题，这一点并未体现出来。

7.3.4 参与者重视结构化记录的查阅与再利用。一些参与者更偏好结构化记录，因为这符合他们预期使用记录的方式。P05 和 P14 预计在撰写报告或展览评论时会参考这些记录，而 P06 则预计会快速查看先前的问题和印象。偏好并不总是与反思性评分一致。P03 在三个反思性结果上的叙事报告评分均至少同样高，P13 平均而言也更偏好叙事报告，但两人都偏好结构化记录，理由是材料有限且没有太多分析需求。P10 发现叙事报告更有助于辨认跨作品的兴趣，但更偏好结构化记录，因为随着记录积累，结构化记录更易于再次查阅；P12 则重视叙事报告中的视觉链接，但认为通过叙事报告回忆先前的问题更为困难。因此，参与者可能会重视叙事报告在反思方面的作用，同时仍然偏好结构化记录以便查阅。

8 讨论

8.1 围绕单件艺术品的探究与重新诠释

参与者的陈述表明，当他们能够将语境信息与自己已经做出的观察联系起来时，这些信息便很有用。P06 关于尤利西斯与佩涅洛佩的对话说明了这一过程。在描述人物看似错位的视线后，P06 要求进一步了解回应中提到的分离与重聚。随后，他们进一步阐述了人物表情所具有的 poignancy，并请求查看同一神话的其他描绘。这些后续问题表明，P06 如何利用语境信息深化先前的观察，并与其他作品展开比较。

回忆与新的诠释对参与者体验的影响也各不相同。P10表示，虽然自己对艺术作品有了更多了解，却仍无法回忆起当时最初的情绪。这一描述表明，即使回忆并不完整，重新查看一张照片也可能提供重新思考一件作品的机会。因此，要支持访客再次观看，既要关注他们当下想要了解什么，也要关注他们希望记起最初观赏经历中的哪些内容。

8.2 跨个人记录诠释兴趣

两项研究表明，比较艺术作品与考察相关记录之间的关系有所不同。在研究1中，P16注意到了一同观看的同人画作之间的差异，而P08仍觉得难以辨认其回答之间的联系。在研究2中，与结构化记录相比，叙事报告在三个收藏集层面的反思结果指标上获得了更高评分。结构化记录也促使参与者发现跨作品的兴趣和关系。叙事报告结合了证据选择、跨作品分组、说明、视觉链接以及访问辅助记录等功能。较高的评分反映的是参与者对这一综合呈现方式的体验，并未单独评估其中各项功能的作用。

参与者的叙述表明，这种反思既包括厘清现有兴趣，也包括发现此前未曾注意到的连接。P06发现，自己对故事与神话的兴趣贯穿于那些起初认为彼此几乎毫无关联的作品之中。P11早已知道视觉特征对自己很重要，但

这份报告帮助他们确定，色彩往往是最先吸引其注意力的元素。对P15而言，报告厘清了参观期间的视觉享受与参观结束后的信息寻求之间的区别，而他们此前已经开始阐述这一点。这些经历表明，回顾探究过程中记录的问题和解读，能够帮助参观者反思自己对什么作出了回应，并更具体地描述自己的兴趣，即便由此得出的诠释并非全新。

报告的描述仍取决于参与者选择记录的内容。P01 解释说，他们原本对战争的兴趣并未体现出来，因为他们没有与系统讨论这一点。感知保真度评估了展示内容在多大程度上反映了记录下来的贡献，因此，该指标上的正面评价并不能证明报告全面呈现了参与者的兴趣。形成性研究发现表明，照片可能无法明确说明其拍摄原因，这进一步凸显了参观者围绕图像进行表述的重要性。这些观察结果支持将报告中的模式与选定作品及交流联系起来进行解读。访问底层贡献内容，有助于参观者评估所呈现的关联在多大程度上反映了他们原本想要表达的内容。

参与者认为有益的比较也各不相同。P04 重视聚焦色彩的展示，而 P05 和 P11 则质疑某些色彩比较的实用性。P05 提议将材料与技法作为建立作品关联的更相关依据。这些叙述表明，由于不同参观者关注艺术作品的不同方面，同一项比较可能对一位参观者有意义，却对另一位参观者用处较小。提供其他比较依据，或许能帮助参观者按照当前的问题探究收藏集 [41, 49, 51]。

8.3 支持再次观看与回顾

参与者研究这些界面适用于不同活动。在研究1中，可以参与者将MyMuse 的空间呈现与沉浸感、艺术流派以及共同观看艺术作品联系起来，而其他参与者则更偏好基线版本较为简单的呈现方式，认为它更便于整理照片并再次找到它们。P03 发现，由于房间与他们曾到访的博物馆不同，再次观看时的体验不够鲜明。这一回答表明，对观看环境的预期会影响空间呈现的价值。两种状况在对话支持和信息呈现方面也有所不同，因此这些陈述反映的是参与者对整体界面的回应，无法单独确定维度性的作用。

研究2同样表明，识别关系和重新查看先前的贡献可能会带来不同的偏好。P10 认为叙事报告更有助于识别不同作品之间的兴趣关联，但更偏好结构化记录，因为他们预计，随着档案不断扩充，通过叙事呈现再次查看内容会变得困难。这一偏好涉及即时阅读体验之外的预期使用情境。这表明，即使某种呈现方式有助于反思，仍可能需要额外支持，以便查找和回顾特定记录。

P12 的体验揭示了此类复盘中一个更具体的困难。他们重视叙事报告中的视觉链接和解释，但发现通过这种呈现方式更难回忆起自己先前的问题。辅助记录虽然可供查阅，但它们的存在并不一定能让人轻松复查早先的交流。更清晰地访问完整交流记录，并提供一份简明的贡献概览，可以帮助参观者在阅读报告时回到某个具体问题。当访客后来访问档案的目的不同于叙事中所强调的关系时，这种访问方式可能尤其有用。

8.4 跨活动与参观使用记录

参与者预计会将自己的记录作为学习、创作实践和写作的参考资料。P05 描述了如何通过记录研究展览陈列，并在规划或参与未来展览时借鉴这些材料。P11 将记录诠释和构建知识的可能用途，与在自身作品中参考拍摄的技法这一既有实践联系起来，同时还描述了其在艺术批评中的可能作用。P05 和 P14 预计会在撰写展览评论、报告或其他文本时使用结构化记录。这些叙述表明，将个人回答与语境信息一并保留，可能有助于在另一项活动中发展出某种诠释或论点。这些用途是参与者预期的，或是从既有实践中推断出的。它们对后续工作的贡献仍有待评估。

参观者也可能希望记录某个想法，而不继续进行对话。P10 表示，希望保留自己为何拍摄某件作品的原因，并认为额外的共情式系统回应令人疲惫。这一反馈表明，可以将保存一条贡献的行为，与请求回答的决定分开。提供无需即时回应即可记录印象、在需要时请求情境信息，以及选择解释深度的选项，或许能够在同一次互动中满足不同的目的。

参与者对记录和探究何时适合融入参观有着不同的偏好。P09 和 P15 更倾向于在观看过程中保持注意力，之后再回顾问题或记录。P10 则认为，在参观期间直接进行记录并访问相关内容，可以减少回家后单独上传照片的工作量。未来的系统可以通过以下方式适应这些偏好：允许在手机上记录的照片或简短问题在参观后继续保留，以便进行更深入持久的探究。这样既能延伸当前的参观后情境，也能让参观者自行决定记录和情境支持何时适合融入他们的观看活动。

累积记录还可以为参观者接下来选择参观的内容提供参考。P08设想，在选择另一场展览时查阅早期展览的报告；P12则描述了如何结合自己喜欢过的作品来考虑未来的博物馆参观。结合参观者对相关艺术品的请求，这些叙述表明，有必要探究已记录的问题和兴趣如何为后续参观提供参考。未来的系统可以将有关某种技法、主题、历史语境或策展选择的问题，与相关展览信息关联起来，并说明每条建议由哪些贡献所提供依据。这样的建议可以为参观者提供选项，帮助他们继续探索已有兴趣，或转向另一个方向。我们没有评估展览发现或后续参观，因此这些建议的质量及其对之后选择的影响仍有待研究。

8.5 对个人记录构建的系统的启示

这两项研究指出，探究与反思之间的关系超越了博物馆参观。人们在考察单件物品时添加的问题和解读，本身就是他们关注内容的证据；因此，系统可以将探究视为后续反思的来源，而不是一项独立活动。MyMuse 的报告只能描述参与者明确表达的内容（第8.2节）。对于个人照片、学习日志或设计参考收藏而言，相关系统同样需要提供方式，让人们记录某件物品为何重要，然后收藏集层面的摘要才能反映他们的兴趣。[18, 28, 52]

当同一个语言模型既回答问题，又总结某人的记录时，此人说了什么与系统说了什么之间的界限很容易变得模糊。MyMuse 仅将访客撰写的文本视为有关访客的证据，并将基于目录建立的连接呈现为目录关系，而非访客本人的陈述。参与者认为报告忠实反映了他们的记录，同时也对部分推导出的关系提出质疑（第 7.3 节）。使用语言模型回答有关个人数据的问题的系统 [23] 同样可以让用户撰写、检索和生成的内容保持可区分，从而避免关于某人兴趣的摘要纳入系统自身的建议。

关于某人的收藏层面的论断，也需要比单件藏品层面的回答更加谨慎的构图。统计不同藏品中某种内容的反复出现情况，并在生成报告措辞之前确定报告的构图，使报告结构在多次运行中保持一致；而由模型选择的构图则各不相同（第 6.3 节）。为数据叙事 [38, 46] 选择事实的方法，也可以在数据描述的是某个人而非数据集时采用类似的限制。

9 局限性与未来工作

追踪多次参观中的记录。 我们的研究通过一场研究环节，对16位参与者的即时影响进行了考察。他们大多是学生，且都是近期曾参观博物馆并拍摄艺术作品的访客；研究2的报告则基于约八件艺术作品的相关记录。若进行更长期的纵向部署，并纳入更多样化的参观者，便能了解首次研究环节的新鲜感消退后，人们是否会再次查看自己的记录、他们的解读如何随参观次数而变化，以及报告识别出的重复出现现象是否会随着收藏集的扩大而持续存在。

从访客自己的照片中识别艺术作品。 研究人员为本研究准备了目录记录（第7.1节），而 MyMuse 依赖这些记录来支持依赖艺术家和日期的功能（附录 B）。若加入识别步骤，根据照片、拍摄到的墙面标签以及参观的时间和地点提出候选记录，并让访客确认或更正，就能让 MyMuse 处理新拍摄的照片。利用更广泛的机构目录，也能将相关作品检索扩展到以西方语料库为主的范围之外。

剥离叙事所带来的作用。研究2将完整的叙事报告——整合了证据选择、跨作品分组、解释、视觉链接和引导式排序——与逐件艺术作品记录进行了比较。将其与呈现相同证据的非叙事跨作品视图进行比较，可以揭示引导式排序本身在多大程度上有助于访客识别反复出现的兴趣。

让模型介导的步骤变得可检视。 话语注释与人工编码员达成了一致 $\kappa = .70$ （第6.2节），回答则结合了目录记录、检索到的上下文和模型知识（第5.2节）。向参观者展示每个回答所记录的来源，评估回答对参观者自身问题的事实准确性，并考察标注差异如何改变报告将哪些记录归为一组，将有助于参观者和设计者决定何时依赖这些步骤。

10 结论

我们介绍了 MyMuse，这是一座参观后的个人博物馆，将参观者的问题和解读与艺术品照片一同保存。其设计源于形成性研究参与者难以回忆自己拍摄作品的原因，以及他们希望考察收藏集中作品之间连接的兴趣。MyMuse 利用过往条目为后续回答提供参考，并在叙事报告中整理累积记录，以便比较不同作品。

在一项有16名参与者参加的受控被试内评估中，直接比较时，更多参与者认为 MyMuse 更适合生动地再次观看作品，也更有助于发现跨作品的连接。与结构化记录相比，叙事报告在反复兴趣认知、跨艺术作品连接和意外发现方面获得了更高评分。参与者表示，他们发现了此前未曾注意到的连接，也更加明确了自己熟悉的兴趣。这些研究发现表明，跨艺术作品整理问题和解读，有助于参观者考察个人记录中体现的反复兴趣。

参考文献

[1] Peter Anderson、Basura Fernando、Mark Johnson 和 Stephen Gould。2016。SPICE：语义命题图像描述评估。载于《计算机视觉——*ECCV 2016*》（《计算机科学讲义》）。Springer，瑞士 Cham，382 398 10 1007 978 3 319 46454 1_24–。doi: 。/————[2] Murtha Baca、Patricia Harpring、Elisa Lanzi、Linda McRae 和 Ann Baird White-side（编）。2006。《文化对象编目：描述文化作品及其图像指南》。美国图书馆协会，芝加哥。[3] Satanjeev Banerjee 和 Alon Lavie。2005。METEOR：一种与人工判断相关性更高的机器翻译评估自动指标。载于计算语言学协会机器翻译和/或摘要内在与外在评估指标研讨会论文集。计算语言学协会，密歇根州安娜堡，65–72。https://aclanthology.org/W05-0909/ [4] Alixandra Barasch、Kristin Diehl、Jackie Silverman 和 Gal Zauberman。2017。摄影记忆：自主拍照对体验中视觉和听觉方面记忆的影响。《*Psychological Science*》28，8（2017），1056–1066。doi:10.1177/0956797617694868[5] Virginia Braun 和 Victoria Clarke。2006。心理学中的主题分析方法。《心理学定性研究》3，2（2006），77–101。doi:10.1191/ 1478088706qp063oa [6] David Brieber、Marcos Nadal 和 Helmut Leder。2015。在白盒子中：博物馆情境提升了对艺术的评价与记忆。《心理学学报》154（2015），36 42 10 1016 2014 11 004 –。doi: 。/j.actpsy. . . [7] Nicolas J. Bullot 和 Rolf Reber。2013。艺术之心遇见艺术史：迈向艺术欣赏科学的心理历史框架。《行为与脑科学》36, 2 2013 123 137 10 1017 0140525 12000489 ( ), –。doi: 。/[8] S X Anjan Chatterjee 和 Oshin Vartanian。2014。神经美学。《认知科学趋势》18，7（2014），370–375。doi:10.1016/j.tics.2014.03.003[9] Amy Yo Sue Chen、William Odom、Carman Neustaedter、Ce Zhong 和 Henry Lin。2023。探索跨时段的数字照片记忆导向交互：Chronoscope 实地研究。载于*2023*年*CHI*人机交互系统中的人为因素会议论文集（德国汉堡）（*CHI '23*631，20 10 1145 3544548 3581012）。计算机协会，美国纽约州纽约市，文章页码。doi: 。/。

[10] Peter Dalsgaard、Michael Mose Biskjaer 和 Jonas Frich。2023。捕捉并重温设计过程中的构想：一项纵向技术探针研究。Design Studies 88 2023 101200 10 1016 2023 101200 ( ), 。doi: 。/ [11] j.destud. 。Kristin Diehl、Gal Zauberman 和 Alixandra Barasch。 。拍照如何提升体验的愉悦感。Journal of Personality and Social Psychology 111, 2 2016 119 140 10 1037 0000055 ( ), –。doi: 。/pspa Chris Elsden、[12] Abigail C. Durrant 和 David S. Kirk。2016。这不就是我的历史吗？理解智能日志实践。载于2016年CHI人机交互系统中的人为因素会议论文集（美国加利福尼亚州圣何塞）（CHI '16）。计算机协会，美国纽约州纽约市，2819–2831。doi:10.1145/2858036.2858103[13] John Falk 和 Martin Storksdieck。2005。运用情境学习模型理解参观者从科学中心展览中获得的学习。Science Education 89, 5 2005 744 778 10 1002 20078 ( ), –。doi: 。/sce。Lesley Fosh、[14] Steve Benford、Boriana Koleva 和 Katharina Lorenz。2015。作为个性化博物馆与画廊解读新机制的赠礼。载于《用户建模、自适应与个性化》（《计算机科学讲义》，第9146卷）。Springer International Publishing，131–142。doi:10.1007/978-3-319- 20267-9 11[15] Lesley Fosh、Steve Benford、Stuart Reeves、Boriana Koleva 和 Patrick Brundell。2013。看我、感受我、触摸我、听见我：雕塑花园中的轨迹与解读。载于SIGCHI计算机系统中的人为因素会议论文集（CHI '13）。计算机协会，149–158。doi:10.1145/2470654.2470675[16] Samuel Gratzl、Alexander Lex、Nils Gehlenborg、Nicola Cosgrove 和 Marc Streit。2016。从视觉探索到故事讲述，再回到视觉探索。Computer Graphics Forum 35, 3 2016 491 500 10 1111 12925 ( ), –。doi: 。/cgf。[17] Linda A. Henkel。 。指点即拍的记忆：拍照对博物馆参观记忆的影响。Psychological Science 25，2（2014），396–402。doi:10.1177/0956797613504438 [18] Scarlett R. Herring、Chia-Chen Chang、Jesse Krantzler 和 Brian P. Bailey。2009。获得灵感！理解创意设计实践中范例的使用方式及原因。载于SIGCHI计算机系统中的人为因素会议论文集（CHI '09）。计算机协会，87–96。doi:10.1145/1518701.1518717[19] Jack Hessel、Ari Holtzman、Maxwell Forbes、Ronan Le Bras 和 Yejin Choi。2021。CLIPScore：一种无需参考文本的图像描述生成评估指标。载于2021年自然语言处理经验方法会议论文集。计算语言学协会，线上及多米尼加共和国蓬塔卡纳，7514–7528。doi:10.18653/v1/2021.emnlp-main.595[20] Ellen Isaacs、Artie Konrad、Alan Walendowski、Thomas Lennig、Victoria Hollis 和 Steve Whittaker。2013。往昔回响：技术介导的反思如何改善福祉。载于SIGCHI计算机系统中的人为因素会议论文集（CHI '13）。计算机协会，1071–1080。doi:10.1145/2470654.2466137[21] Janin Koch、Nicolas Taffin、Andrés Lucero 和 Wendy E. Mackay。2020。SemanticCollage：以语义标签丰富数字情绪板设计。载于2020年ACM交互系统设计会议论文集（荷兰埃因霍温）（DIS '20）。计算机协会，美国纽约州纽约市， 407 418 10 1145 3357236 3395494 –。doi: 。/。Ian Li、[22] Anind Dey 和 Jodi Forlizzi。 。个人信息学系统的阶段模型。载于SIGCHI计算机系统中的人为因素会议论文集（美国佐治亚州亚特兰大）（CHI '10）。计算机协会，美国纽约州纽约市， 557 566 10 1145 1753326 1753409 –。doi: 。/。[23] Jiahao Nick Li、Zhuohao (Jerry) Zhang 和 Jiaju Ma。 。OmniQuery：通过情境化增强捕捉到的多模态记忆，实现个人问答。载于2025年CHI人机交互系统中的人为因素会议论文集（CHI '25 635，20 10.1145/3706598）。计算机协会，文章页码。doi: 。/。[24] Chin-Yew Lin。 。ROUGE：一个用于摘要自动评估的软件包。载于《文本摘要：拓展领域》。计算语言学协会，西班牙巴塞罗那，74–81。https://aclanthology.org/W04-1013/ [25] Roberto Martinez-Maldonado、Vanessa Echeverria、Gloria Fernandez Nieto 和 Simon Buckingham Shum。2020。从数据到洞见：多模态学习分析的分层叙事方法。载于2020年CHI人机交互系统中的人为因素会议论文集（CHI '20）。计算机协会， 1 15 10 1145 3313831 3376148 –。doi: 。/。Christopher [26] Morse、Jasmin Niess、Kerstin Bongard-Blanchy、Salvador Rivas、Carine Lallemand 和 Vincent Koenig。2023。历久弥新的印象：呈现有意义的博物馆体验。Behaviour & Information Technology 42，8（2023），1127–1154。doi:10.1080/0144929X.2022.2061375[27] Evangelos Niforatos、Caterina Cinel、Cathleen Cortis Mack、Marc Langheinrich 和 Geoff Ward。2017。少即是多吗？对比有限、无限和自动图片捕捉在增强记忆回忆方面的效果。Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. 1，2，文章21（2017），22页。doi:10.1145/3090086[28] Matthias Nückles、Julian Roelle、Inga Glogger-Frey、Julia Waldeyer 和 Alexan-der Renkl。2020。以写促学中的自我调节视角：运用日志优化自我调节学习认知负荷的写作。《教育心理学评论》 32, 4 2020 1089 1126 10 1007 10648 020 09541 1 ( ), –。[29] doi: . /s - - - William T. Odom、Abigail J. Sellen、Richard Banks、David S. Kirk、Tim Regan、Mark Selby、Jodi L. Forlizzi 和 John Zimmerman。2014。为缓慢、期待与重访而设计：Photobox 的长期实地研究。载于 SIGCHI计算机系统中的人为因素会议论文集（加拿大安大略省多伦多）（CHI '14）。计算机协会，美国纽约州纽约， 1961 1970 10 1145 2556288 2557178 –。doi: . / . Kishore [30] Papineni、Salim Roukos、Todd Ward 和 Wei-Jing Zhu。2002。BLEU：一种机器翻译自动评估方法。载于计算语言学协会第40届年会论文集。计算语言学协会，美国宾夕法尼亚州费城，311–318。doi:10. 3115/1073083.1073135[31] Marcus T. Pearce、Dahlia W. Zaidel、Oshin Vartanian、Martin Skov、Helmut Leder、Anjan Chatterjee 和 Marcos Nadal。2016。神经美学：审美体验的认知神经科学。《心理科学展望》11，2（ 2016 265 279 10 1177 1745691615621274 ），–。doi: . / [32] Daniela Petrelli、Elise van den Hoven 和 Steve Whittaker。2009。创造历史：有意捕捉未来记忆。载于SIGCHI计算机系统中的人为因素会议论文集（CHI '09）。计算机协会，美国纽约州纽约，1723–1732。doi:10.1145/1518701.1518966[33] 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。2021。通过自然语言监督学习可迁移的视觉模型。arXiv:2103.00020 [cs.CV] https://arxiv.org/ abs/2103.00020[34] Emilia Rosselli Del Turco、Nanna Inie、James D. Hollan 和 Peter Dalsgaard。2025。创意实践者如何使用工具捕捉灵感：一项跨领域研究。《ACM人机交互汇刊》32，4，文章40（2025），38页。doi:10.1145/3727979[35] Edward Segel 和 Jeffrey Heer。2010。叙事可视化：用数据讲故事。IEEE可视化与计算机图形学汇刊16，6（2010），1139–1148。doi:10.1109/TVCG.2010.179[36] Abigail J. Sellen 和 Steve Whittaker。2010。超越全面捕获：对生活日志记录的建设性批判。《ACM通讯》53，5（2010），70–77。doi:10.1145/1735223. 1735243[37] Ben Rydal Shapiro 和 Rogers Hall。2018。博物馆中的个人策展。《ACM人机交互会议论文集》2，CSCW，文章158（2018），22页。doi:10.1145/ 3274427[38] Danqing Shi、Xinyue Xu、Fuling Sun、Yang Shi 和 Nan Cao。2021。Calliope：从电子表格自动生成可视化数据故事。IEEE可视化与计算机图形学汇刊27，2（2021），453–463。doi:10.1109/TVCG. 2020.3030403[39] Matteo Stefanini、Marcella Cornia、Lorenzo Baraldi、Massimiliano Corsini 和 Rita Cucchiara。2019。Artpedia：一个包含艺术领域视觉与情境句子的新型视觉语义数据集。载于图像分析与处理 – ICIAP 2019（《计算机科学讲义》，第11752卷）。Springer，瑞士Cham， 729 740 10 1007 978 3 030 30645 8_66 –。doi: . / - - - - Theopisti [40] Stylianou-Lambert。 。在艺术博物馆摄影：参观者态度与动机。《参观者研究》20，2（2017），114–137。doi:10.1080/ 10645578.2017.1404345[41] Alice Thudt、Uta Hinrichs 和 Sheelagh Carpendale。2012。波希米亚书架：通过信息可视化支持意外发现书籍。载于SIGCHI计算机系统中的人为因素会议论文集（美国得克萨斯州奥斯汀）（CHI '12）。计算机协会，美国纽约州纽约， 1461 1470 10 1145 2207676 2208607 –。doi: . / . Giasemi [42] Vavoula、Mike Sharples、Paul Rudman、Julia Meek 和 Peter Lonsdale。2009。Myartspace：设计并评估利用多媒体手机支持教室与博物馆之间的学习。《计算机与教育》53，2（2009）， 286 299 10 1016 2009 02 007 –。doi: . / j.compedu. . . Kendall L. Walton。 。艺术类别。《哲学评论》79，3（1970），334–367。doi:10.2307/2183933[44] Anqi Wang、Xiaran Song 和 Andrés Lucero。2024。Linger：将博物馆体验延伸至参观后阶段。载于2024年北欧人机交互会议附属论文集（瑞典乌普萨拉）（NordiCHI '24附属会议）。计算机协会，美国纽约州纽约，文章21，共5页。doi:10.1145/3677045.3685436[45] Shuai Wang、Ivona Najdenkoska、Hongyi Zhu、Stevan Rudinac、Monika Kackovic、Nachoem M. Wijnberg 和 Marcel Worring。2025。ArtRAG：利用结构化上下文进行视觉艺术理解的检索增强生成。载于第33届ACM多媒体国际会议论文集。计算机协会，美国纽约州纽约，6700–6709。doi:10.1145/3746027.3755673[46] Yun Wang、Zhida Sun、Haidong Zhang、Weiwei Cui、Ke Xu、Xiaojuan Ma 和 Dongmei Zhang。2020。DataShot：从表格数据自动生成事实表。IEEE可视化与计算机图形学汇刊26，1（2020），895–905。doi:10.1109/TVCG.2019.2934398

[47] Zhan Wang、Lin-Ping Yuan、Liangwei Wang、Bingchuan Jiang 和 Wei Zeng。2024。VirtuWander：通过大型语言模型增强虚拟导览中的多模态交互。载于*2024年CHI人机交互系统中的人为因素会议论文集（CHI '24）*。计算机协会，文章 612, 20 10 1145 3613904 3642235 页。doi: . /
. Alexandra Weilenmann 、 Thomas Hillman 和 Beata Jungselius。2013。博物馆中的Instagram：通过社交照片分享传达博物馆体验。载于*SIGCHI计算机系统中的人为因素会议论文集（CHI '13）*。计算机协会，1843–1852。doi:10.1145/2470654.2466243[49] Florian Windhager、Paolo Federico、Günther Schreder、Katrin Glinka、Marian Dörk、Silvia Miksch 和 Eva Mayr。2019。文化遗产收藏数据可视化：现状与未来挑战。*IEEE可视化与计算机图形学汇刊*25，6（2019），2311–2330。doi:10.1109/TVCG. 2018.2830759[50] Shuchang Xu、Chang Chen、Zichen Liu、Xiaofu Jin、Lin-Ping Yuan、Yukang Yan 和 Huamin Qu。2024。Memory Reviver：通过主动式聊天机器人支持视障人士回忆照片收藏。载于*第37届ACM用户界面软件与技术年会论文集（UIST '24）*。计算机协会，文章88，共17页。doi:10.1145/3654777.3676336[51] Ka-Ping Yee、Kirsten Swearingen、Kevin Li 和 Marti Hearst。2003。用于图像搜索与浏览的分面元数据。载于*SIGCHI计算机系统中的人为因素会议论文集（美国佛罗里达州劳德代尔堡）（CHI '03）*。计算机协会，美国纽约州纽约，401–408。doi:10.1145/642611.642681[52] Michael Yin 和 Robert Xiao。2025。TravelGalleria：通过VR数字故事讲述支持旅行体验的回忆与反思。载于*2025年CHI人机交互系统中的人为因素会议论文集（日本横滨）（CHI '25）*。计算机协会，美国纽约州纽约，文章 332, 14 10 1145 3706598 3713398 页。doi: . / .
Xuan Zhao 和 Siân E. Lindley。 。通过使用进行策展：理解社交媒体的个人价值。载于*SIGCHI计算机系统中的人为因素会议论文集（加拿大安大略省多伦多）（CHI '14）*。计算机协会，美国纽约州纽约，2431–2440。doi:10.1145/2556288.2557291

A 附加形成性研究材料

本附录概述形成性访谈的主题，展示初始编码与报告的四个主题之间的联系，并介绍单独进行的 CCO 覆盖分析和 MyMuEx 触发器选择。

A.1 访谈主题

半结构式访谈围绕博物馆参观、艺术品摄影、信息寻求、记录、整理和再次观看等实践展开。照片引谈法使用了每位参与者从自己参观博物馆或美术馆期间拍摄的照片中选出的十张。针对每张照片，我们讨论了拍摄原因、仍然记忆犹新的内容、哪些信息或相遇经历的细节已无法还原，以及后来获得的知识或计划中的再利用与该照片的关联。

在每个研究环节的后半段，我们邀请参与者畅所欲言，提出建议，探讨如何组织一个用于再次观看艺术品照片、记录反思和提出问题的系统。我们将这些设计建议与参与者对自身现有实践和兴趣的描述区分开来。有关空间呈现的建议为第 5.1 节所述的设计选择提供了依据。

A.2 初始编码与主题总结

初始编码产生了140个编码，并将其组织为十五个描述性分组。这些分组整理了访谈材料的不同类型，但其本身并不是最终主题。表3列出了初始编码与正文所报告的四个主题之间的示例性链接。这并非将编码分配至主题的穷尽性或互斥性列表。初始编码标识符（包括F1–F4）与F1–F4主题标签彼此不同。

表4给出了表3中所引用编码的定义。

表3：形成性主题与初始编码之间的示例性连接。编码定义见表4。

主题	初始编码的支持方面	编码编号
F1	拍摄原因遗失、识别背景，以及难以区分个人重要照片与日常记录。	E1、E2、E3、E11
F2	创作或专业目的、领域知识、现有兴趣以及个人经历塑造注意力。	A4、A8、B2、B6
F3	知识引导注意力，并将参观中遇到的作品进行比较。	H13、F3、F7
F4	对个人有意义的组织方式、理解个人品味的兴趣，以及对推断仅凭照片判断偏好的担忧。	J1、J3、J8、K1、K5、K6

A.3 CCO覆盖分析

除反思性主题分析外，两名研究人员还分别对九个CCO章节中参与者层面的覆盖情况进行了编码。十名参与者和九个章节共产生90项二元参与者—章节判断。每项正向判断都必须有相应的片段标识符作为支持。

符合条件的证据包括三种表达形式：信息需求、捕捉或注意动机，以及组织偏好。正向判断表示某位参与者表达了与某一章节相关的注意或兴趣。这并不要求参与者明确提出信息请求，或使用CCO术语。负向判断表示未发现符合条件的支持性陈述，而不表示参与者缺乏该兴趣。由于一名参与者可以为多个章节提供证据，因此覆盖计数并不相互排斥。

仅凭关于系统设计的一般性建议，不足以确定CCO章节覆盖情况。无论参与者自身实践或长期兴趣在访谈的何处出现，包括在讨论未来可能的支持时所提及的内容，均仍符合纳入条件。九个章节未涵盖的兴趣则另行记录。

两位编码者均认 十位 物理特征，而九位参与者具有

A.4 MyMuEx 触发器选择

对形成性访谈材料进行编码后，识别出 MyMuEx 定义的 23 个体验触发器中的 21 个。这 21 个触发器被采纳为系统用于体验标注的候选集。

B 系统实现细节

表 5 总结了各项输入 以及第 5 节所述各项功能的方法，并说明是否需要进行目录匹配以实现 t he 拍摄的艺术品。

表3 中引用的初始编码定义。

Code	定义
A4. 将观看作为灵感来源	前往参观以收集用于创意、学术或专业活动的材料，包括写作、视频制作和设计。
A8. 领域专长塑造注意力	描述历史、艺术史、审美或策展方面的训练如何塑造人们关注艺术作品的哪些方面得到关注。
B2. 主题内容符合个人兴趣	因为作品中的主题对象或图像学与既有兴趣相关而拍摄作品，例如神话、圣徒或狗。
B6. 个人传记关联	将一件作品与个人境况或经历联系起来，包括疾病、药物治疗、就业顾虑或养宠物。
E1. 拍摄原因遗忘	保留了一张照片，却想不起当初为何选择拍摄这件作品。
E2. 场馆或展览背景遗忘	无法回忆起照片所关联的博物馆或展览。
E3. 身份信息遗失	无法找回作品的标题或艺术家信息，包括照片中的标签难以辨认的情况。
E11. 难以区分个人重要照片	将具有个人意义的作品照片描述为难以与日常或预防性拍摄。
F3. 跨参观比较	返回查看先前拍摄的照片，以确认之前是否见过某件作品，或比较相关作品或不同版本。
F7. 后续参观期间的比较	在参观另一场展览时，需要调取先前拍摄的照片，此时搜索相机胶卷可能不太方便。
H13. 知识改变注意到的内容	将所获得的背景信息描述为引导注意力关注此前被忽视。
J1. 按展览或参观日期组织	使用或偏好以展览或参观日期来查找和整理作品记忆。
J3. 按主题或题材组织	将作品归类，或希望围绕共同题材、主题或其他有意义的类别对作品进行归类。
J8. 个人定义的组织轴	希望根据个人经验、情感强度、社会关切或其他对个人有意义的区分来建立类别。
K1. 对理解个人品味感兴趣	表达对反复出现的偏好感到好奇，或难以描述自己通常喜欢什么。
K5. 对从所有照片	发出警告：出于义务或记录目的拍摄的照片，可能会歪曲个人偏好。
K6. 反复出现的问题作为兴趣的证据	这表明，反复提出的问题可能体现出照片本身无法充分传达的兴趣 .

表 5：各项 MyMuse 功能的输入，以及该功能是否需要与拍摄的艺术品进行目录匹配。

功能	输入	方法	目录匹配
回复生成	目录记录；维基数据艺术家数据；ArtRAG 情境；先前回复；已选中先前记录	gpt-5.4-mini，仅文本	以下内容必填：前三项输入
话语标注	访客话语	gpt-5.4-mini	非必需
PMG 检索	已存储的标注	共享的方面-触发因素对	非必需
相关作品：视觉	访客照片	CLIP ViT-B/32	不要求
相关作品：艺术家	艺术家姓名	语料库匹配；大都会艺术博物馆藏品搜索	必填
相关作品：主题	目录主题术语	受控词表	必填
报告规划	注释和证据跨度	确定性规划器	非必需
报告措辞	计划观察	gpt-5.4-mini，已检查；确定性回退	不需要
报告视角：	访客照片；注释	颜色和明度测量	非必需
报告视角：	目录主题术语	受控词表	必填
报告视角：	艺术家；日期	目录字段	必填

C 技术评估：补充方法与结果

本附录提供实现 与评分细节、提示结构，以及三项技术评估的补充结果 s.

C.1 在共享子集上的 Artpedia 比较

我们在同一组 310 个可恢复图像的 Artpedia [39] 条目上评估了全部四种配置：ArtRAG [45] 和 MyMuse，分别搭配 gpt-4o-mini 和 gpt-5.4-mini。

C.1.1 输入与提示。 所有 MyMuse 配置均接收文本元数据和情境，不接收图像输入，并采用零样本提示。ArtRAG 配置使用图像输入和双样本提示。MyMuse 还接收艺术家、媒介和地区元数据。MyMuse 的任务消息包含“严格控制在 30 个词以内”的指令，ArtRAG 配置也接收了相同的长度要求。

gpt-5.4-mini 配置仅替换了用于最终生成的模型。MyMuse 原样复用了其 gpt-4o-mini 配置中存储的系统消息和任务消息，没有重新整合检索到的上下文，因此两个 MyMuse 配置所提供的上下文和路由决策相同。ArtRAG 在检索步骤中仍使用 gpt-4o-mini，包括基于图像的重排序；在最终描述调用中，则使用 gpt-5.4-mini，并采用相同的图像输入和双样本提示。因此，两个系统在检索期间的语言模型调用均使用 gpt-4o-mini；在已部署的 MyMuse 系统中，ArtRAG 也是如此；而 gpt-5.4-mini 仅用于最终生成。

C.1.2 生成与评分。 MyMuse 的 gpt-4o-mini 配置使用温度=1.0。两个 gpt-5.4-mini 配置均使用 API 默认温度，并且每种配置均完成了全部 310 次生成。

同一评估器对全部四组输出进行了评分。基于参考的指标同时使用了 Artpedia 的情境_句和视觉_句，并针对 BLEU-1–4 [30], METEOR [3], ROUGE-L [24], 和 SPICE [1]采用 PTBTokenizer 进行预处理。CLIPScore [19] 使用 ViT-B/32，在没有文本参考的情况下评估图文对齐。回答按照其原始生成长度进行评分，未进行事后词数截断。

C.1.3 结果。 表 6 报告了共享子集的完整结果和平均回答长度。

表 6：在共享的 310 项 Artpedia 子集上的性能。指标值乘以 100。平均回答长度以词数表示。所有回答均按照其原始生成长度进行评估。gpt-5.4-mini 配置仅替换用于最终生成的模型。各行按生成模型分组，粗体表示每组内较高的数值。

模型	系统	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	SPICE	CLIPScore	平均词数
gpt-4o-mini	ArtRAG	35.28	17.22	8.77	4.75	15.22	22.74	**9.35**	**84.88**	57.4
	MyMuse	**45.96**	**25.31**	**16.05**	**10.55**	**16.38**	**27.49**	8.25	77.49	25.8
gpt-5.4-mini	ArtRAG	33.94	15.61	7.83	4.21	13.74	20.28	7.87	**83.23**	40.2
	MyMuse	**44.51**	**24.04**	**14.52**	**9.38**	**16.30**	**23.71**	**8.14**	77.36	23.2

由于 MyMuse 可以利用 ArtRAG 的情境，我们考察了其较高的 BLEU-4 和 ROUGE-L 分数是否仅限于 ArtRAG 知识图谱所覆盖的作品。我们在两个预先设定的覆盖层中比较了 gpt-4o-mini 的两种配置：140 件已覆盖作品和 170 件未覆盖作品。对于 MyMuse 与 ArtRAG，在已覆盖作品上的 BLEU-4 分别为 9.80 和 4.46，在未覆盖作品上则分别为 11.13 和 4.99。ROUGE-L 分别为 26.95 和 22.65，以及 27.93 和 22.82。因此，这两项指标上的较高分数也延伸到了 ArtRAG 知识图谱覆盖范围之外的作品。

C.2 标注一致性：方法与其他结果

C.2.1 编码程序。 两名研究人员分别独立标注了研究 1 中 MyMuse 状况的全部 332 条访客话语。编码者收到访客话语、艺术作品标题、艺术家姓名、编码说明及已发表的定义。模型注释和参与者标识符均被隐去。

代码本包含 21 个可由访客提及的 CCO 元素和全部 23 个 MyMuEx 触发器。编码者独立分配词汇，每条话语的标签数量不受限制，并且仅分配有文本支持的标签。提出问题本身并不意味着存在体验触发器。对于不确定的判断，编码者标记为不确定=1，并输入最可能的代码。

人工注释涵盖了系统采用的相同 21 个 MyMuEx 触发器。两名编码者均未分配 贡献 或 *VIP*，这两个触发器不属于系统词汇。此处报告的一致性分析仅涉及四个源自 CCO 的视角。

C.2.2 模型注释。 我们使用了研究环节期间存储的注释，而不是重新生成注释。注释在面向访客的回复之后进行，使用温度为=0的 gpt-5.4-mini。

CCO 调用接收了该话语，以及 21 个候选元素及其已发布的定义。必要时，为了消解指代，会提供前一条系统回复作为情境，但不会将其标注或引用为证据。MyMuEx 调用接收了该话语和 21 个

个选定触发器。只有在话语支持这种关联时，另一次调用才会将 CCO 元素与 MyMuEx 触发器关联起来。仅仅共同出现并不足以建立关联。

C.2.3 映射至视角。表7将每位编码员和模型的 CCO 分配映射为四个二元视角标签。只要某个视角中的至少一个元素被分配，该视角即视为存在。仅包含四个视角之外元素的话语，四个值均记为零。人工参考结果为两位编码员的视角分配并集。

表7：CCO 编码标签与四个视角的映射。标识符 C01–C21 为代码本中的 21 个 CCO 元素编号。

Lens	CCO 元素
创作者	C03 受控创作者；C04 创作者展示。
物理特征	C05 色彩；C06 状况；C07 质地；C08 铭文；C09 材料；C10 材料与技法展示；C11 尺寸展示；C12 外观描述；C13 技法。
风格与年代	C14 文化；C15 展示日期；C16 风格。
主题	C19 受控主题；C20 主题展示。
无视角	C01 标题；C02 作品类型；C17 创作地点；C18 当前位置；C21 描述。

C.2.4 一致性度量与结果。 表8报告了332条话语中各视角的 Cohen's κ，以及合并后的1,328项话语—视角判定的 Cohen's。精确集合一致性要求同一条话语的四个视角标签全部匹配。

表8：四个视角的一致性。一致性百分比包括匹配的存在和匹配的缺失。模型—人工比较采用两位编码者视角分配的并集。

	编码者间		模型—人工并集	
Lens	κ	一致性（%）	κ	一致性（%）
创作者	.767	93.7	.683	90.4
物理特征	.839	94.0	.769	91.3
风格与年代	.891	97.3	.647	91.0
主题	.867	94.3	.662	84.9
合并	.847	94.8	.701	89.4

表9中的两项敏感性分析分别应用于全部332条话语。其中一项排除了任一编码员标记为不确定的20条话语，剩余312条。另一项排除了模型和人工并集均未指定任何视角的72条话语，剩余260条。

表9：模型与人工一致性的敏感性分析。精确集合匹配以各子集中的话语数量及其百分比表示。二元一致性根据每条话语的四项视角判定计算。

子集	话语	合并 κ	二元一致性 (%)	精确集合匹配 *n* (%)
所有话语	332	.701	89.4	214 (64.5)
排除不确定话语	312	.700	89.0	198 (63.5)
排除共同为空的分配	260	.674	86.4	142 (54.6)

C.3 报告构成：方法与补充结果

C.3.1 输入、提示与输出字段。 三种状况共用一份针对每位参与者的契约，其中包含参与者记录和艺术品元数据。

A/B 提示省略了 C 禁止提出新的诠释并将构图决策视为已完成的指示，因为 A 和 B 已选定构图。提示结构见 C.3.5 节和 C.3.6 节。

为在相同字段上将 C 与 A、B 进行比较，C 的切入视角、锚定作品、提及的艺术品和所声称的模式均从规划器输出中读取，而不是从生成文本中提取；其叙述文本由按显示顺序排列的采纳的生成句子构成。

C.3.2 生成与验证。 所有状况均使用 gpt-5.4-mini，每次调用的完成令牌上限为 4,000。默认设置省略了温度字段，第二种设置请求温度=0，API 对每次调用均予以接受。

C 复用了生产规划、事实提取、提示、句子验证和合并功能，同时绕过了缓存。它最多进行了两次生成调用，保留第一次调用中被接受的句子，并用第二次调用填补剩余空位。

B 以 A 针对同一参与者、重复次数和温度设置生成的输出为基础。其批次验证器会检查叙述句中是否包含缺乏依据的数字和名称、推测性表达以及固定特质描述，并移除无效的条目视角值和艺术作品标识符。对于模式短语，则仅检查是否包含缺乏依据的数字和名称。

在 B 中，若叙述句或结构字段未通过验证，最多会使用相同的提示和约定额外查询一次。两次回答都会经过验证。第一次验证后，每个非空字段都会予以保留，第二次验证仅用于填充空字段。非空叙事_文本字段中被拒绝的单个句子不会得到补充；仅进行模式筛选不会触发重新查询。

C.3.3 稳定性测量指标。对于每位参与者、每种状况和每个温度设置，条目视角分歧和锚点艺术作品分歧的计算方式如下：

$$D = 1 - \frac{\max_v n_v}{m},$$

其中， m 为保留运行次数， n_v 为赋予值 v 的次数。缺失值以空字符串表示。艺术作品和模式集的稳定性是所有 $\binom{m}{2}$ 重复运行对的平均杰卡德相似度，其中

$$J(S_i, S_j) = \frac{|S_i \cap S_j|}{|S_i \cup S_j|}, \qquad J(\varnothing, \varnothing) = 1.$$

参与者层面的数值以其均值和样本标准差进行汇总。少于两次保留运行时，成对测量指标未定义。

320 次 A 运行中有三次返回了无法解析的 JSON，因此被排除，相应的 B 运行也一并排除。C 没有此类运行，因为其结构字段来自其计划。

模式归一化首先匹配完整的 CCO 章节名称，然后匹配 CCO 元素（优先匹配最长标签）、MyMuEx 触发器标签或代码，以及固定的别名表。CCO 匹配会映射到章节。MyMuEx 匹配通常保留特定于触发器的键，其中 curation 和 atmosphere 映射到情境。未匹配短语在经过 Unicode、标点、大小写和基于规则的后缀归一化后，保留为独立的键。因此，模式相似度结合了章节级、触发器级和归一化字符串匹配。

C.3.4 温度下的结果 未请求。 表10报告了第二种温度设置下使用C.3节所定义测量指标得出的结果 .3. 相同的16名参与者记录 和每种状况下的10次重复结果均被采用，并请求了温度=0 。

表10：请求温度=0 时的报告构成稳定性。数值为16名参与者的均值，括号内为样本标准差。Jaccard值概括了每位参与者保留的重复结果之间的成对集合相似度。

测量指标	A	B	C
条目透镜分歧	0.056 (0.126)	0.056 (0.126)	0.000
锚点艺术作品分歧	0.075 (0.118)	0.075 (0.118)	0.000
作品集杰卡德相似度	0.925 (0.112)	0.925 (0.112)	1.000
模式集 Jaccard 指数	0.314 (0.132)	0.313 (0.132)	1.000

C.3.5 *A/B*提示结构。 A 的初始生成和 B 的条件式重新查询均将共享的契约 JSON 用作用户消息。

```
你将根据博物馆参观记录撰写一份藏品层级反思报告。

该 JSON 包含某人的作品目录信息、逐字话语、CCO 元素标注和证据片段、
MyMuEx体验触发因素、CCO 元素与触发因素之间的关系、未关联到任何艺术作品
的情境记录，以及来源信息。


请判断以下内容：
1. 切入视角：选择创作者、实体、风格或主题（分别对应 CCO 第 2、3、4 和 6 章）。
如果记录主要涉及艺术作品之外的陈列位置或氛围，请选择情境；如果没有适用
的视角，请选择其他。2. 锚定作品：选择一件有实际证据支持所选视角的起始艺
术作品。
```

3. 反复出现的模式：找出至少有两件不同艺术作品支持的兴趣点。同一件艺术作品中的重复提及只算作一件支持性作品。4. 报告正文：从锚定作品开始，依次阐述这些模式，最后总结报告。只纳入有证据支持的部分；证据有限时，报告应简短。

限制：
- 不要陈述 JSON 中没有的事实。只能使用 JSON 中出现的艺术作品名称、元素名称、触发因素名称和数字。- 不要描述记录者的性情、品味或个性。只陈述记录中重复出现的内容。- 不要使用“也许”“似乎”“看起来”或“感觉”等推测性表达。- 面向记录者本人撰写；不要以第三人称将其称为 “ 参与者 ”“ 用户 ” 或 “ 此人 ”。

仅返回一个 JSON 对象，不要使用代码围栏：{"entry_lens": "创作者 | 实体 | 风格 | 主题 | 情境 | 其他", "anchor_artwork": "<艺术作品_id -- 从 JSON 中的艺术作品中复制该值>", "artworks_mentioned": ["<艺术作品_id>", ...],"patterns_claimed": ["<短语>", ...],"narrative_text": "<完整报告正文>"}

C.3.6 CPromptStructure. C已将计划特定事实JSON作为用户消息接收。

你正在编辑一份关于博物馆参观记录的报告。分析和构图决策已经完成。将每项提供的结果表达为一个句子。

限制条件：
- 不得陈述JSON中没有的事实。只能使用JSON中出现的艺术作品名称、元素名称、触发因素名称和数字。- 不得描述此人的性情、品味或个性。只能陈述记录中重复出现的内容。- 不得创建新的诠释。重要的视角以及每项结果的含义都已确定。- 不得使用“或许”“似乎”“看起来”或“感觉”等推测性表达。- 面向记录的制作者写作；不得以第三人称称其为“参与者”“用户”或“此人”。- 每个句子不得超过120个字符，包括空格。

写作内容：
- anchor_reason：仅使用 anchor.signal 和关于该艺术作品的实际话语，解释为何已选艺术作品为所选视角提供了起点。将 anchor.signal.why 与引用的观察联系起来。在不推断使用意图或作出额外判断的前提下，转述引起注意的因素；不要只是重复注释标签。对于 title_vs_subject，描述标题所指对象与可见主题对象之间的不一致。对于 question_answered，描述所提出的问题以及随后出现的陈述。不要虚构 signal 中没有的原因。如果 anchor.signal 为 null，则将 anchor_reason 设为 null。 - lens_takeaway：针对 lenses 中出现的每个视角各写一句话，描述其元素在不同艺术作品中反复识别出的内容。不要添加其他视角。 - final_synthesis：描述哪些体验触发器伴随

反复出现的兴趣，仅使用链接和反应。 - term_contrast：针对 terms 中出现的每个主题术语各写一句话，描述被赋予该术语的艺术作品如何以不同方式处理该主题。仅使用每件艺术作品的 described 字段（目录对图像的描述）。明确指出差异；不要添加这些字段中没有的主题术语或主张。 - record_interest：仅当 record 存在时写一句话，只使用 record.said。描述围绕该单一主题对象提出了什么问题，以及探究如何展开，而不只是指出该术语曾出现。不要包含其他主题对象、其他视角或总体总结，也不要在 record.term 之外写出名称或数字。

仅返回一个 JSON 对象：{"anchor_reason": "...", "lens_takeaway": {"subject": "..."}, "final_synthesis": "...", "term_contrast": {"<subject_term>": "...", ...}, "record_interest": "..."} 如果证据不足以填写某项，请将其留为 null。

D 补充用户研究结果与测量指标

D.1 补充统计结果

表11–13报告了第7.2节和第7.3节中总结的结果的完整统计数据。W 表示Wilcoxon符号秩统计量，p 表示未经校正的值，p_{Holm} 表示Holm校正值，r 表示效应量 $|Z|/\sqrt{N}$。

表11：六项共同的研究1结果（$N = 16$）的统计数据。Holm校正应用于这六项结果。

结果	MyMuse M（SD）	基线 M（SD）	W	p	p_{Holm}	r
重新体验	6.00（1.10）	5.31（1.14）	16.0	.065	.137	.46
阐述	6.50 (0.73)	6.00 (0.73)	9.0	.046	.137	.50
响应间的关联	5.75 (1.44)	5.00 (1.55)	7.0	.032	.129	.54
回答相关性	6.12 (0.89)	5.31 (0.87)	0.0	.006	.035	.69
参与度	6.69 (0.48)	6.06 (0.93)	0.0	.026	.128	.56
使用意向	6.50 (0.89)	5.81 (1.28)	8.5	.093	.137	.42

表12：研究1直接比较题项的统计数据 e). 采用了Holm校正项，针对量表中点4进行检验（1 =MyMuse，7 = 基线 在这三个项目中。

结果	MyMuse / 相同 / 基线	Mdn	SD	W	p	p_{Holm}	r
响应间的关联	12 / 2 / 2	2	1.49	5.0	.002	.005	.76
重新体验	13 / 1 / 2	2	1.41	4.0	.001	.004	.81
总体偏好	12 / 0 / 4	1.5	2.27	31.0	.049	.049	.49

经过Holm校正后，回答相关性是共同的研究1结果中唯一存在统计学显著差异的结果。三个直接比较题项仍具有统计学显著性。可查找性作为可用性检查单独进行分析，未发现显著差异（MyMuse $M = 6.31$、$SD = 0.95$；基线 $M = 5.81$、$SD = 1.11$；$p = .145$、$r = .36$）。在MyMuse特定诊断项中，个性化问答和空间重建的评分分别为 $M = 6.06$；16名参与者中分别有15名和14名将其评为5分或以上。

D.2 问卷题项

表14–16列出了问卷条目。除非另有说明，回答均采用7点评分同意量表：1 = 非常不同意、4 = 中立和 7 =非常同意。

表13：研究2测量指标的统计数据（$N = 16$）。对三个集合层面的反思性结果进行了Holm校正。感知保真度则单独进行评估，作为基础核查。

结果	叙事 M （SD）	结构化 M （SD）	叙事更高 / 相等 / 结构化更高	p (p_{Holm})	r
反复兴趣认知	6.62 (0.62)	4.62 (1.71)	12 / 3 / 1	.002 (.005)	.76
跨艺术作品关联意识	6.62 (0.50)	4.19 (1.87)	14 / 2 / 0	<.001 (.003)	.83
意外发现	6.31 (1.14)	3.88 (1.93)	14 / 1 / 1	.004 (.005)	.72
感知保真度	6.69 (0.60)	6.31 (0.79)	8 / 7 / 1	.083	.43

表14：研究1问卷题项。六项共同结果和可查找性在每种状况后进行测量。最后两个诊断性条目仅在 MyMuse 状况后测量。

测量指标	问卷条目
重新体验	使用该系统时，我清晰地回想起了当时的博物馆参观体验。
阐述	该系统帮助我更具体地阐述自己对艺术作品的想法。
响应间的关联	该系统帮助我注意到自己对不同艺术作品的回答之间的共同点或连接。
回答相关性	系统的回答符合我想了解的内容。
参与度	使用此系统时，我想进一步探索艺术作品和我的记录。
使用意向	在实际参观博物馆后，我愿意再次使用此系统。
可查找性	在这个系统中，我可以轻松找到并查看自己想要的艺术作品或记录。
个性化问答	后续回答能够参考我之前提出的问题和反思，这很有用。
空间重建	以空间重新配置的形式呈现艺术作品和记录，这很有用。

表15：研究1直接比较题项。所有三个题项均采用双极七点量表，选项为 1 =MyMuse、 4 = 差不多 **和**7 =基线。

测量指标	问卷条目
响应间的关联	哪个系统更有助于发现你对不同 艺术作品?
重新体验	当时，哪个系统更有助于您生动地回忆博物馆参观体验?
总体偏好	如果您要选择一个系统，在参观博物馆后重温您的参观体验，您会 更愿意使用哪个系统?

表16：研究2　　在体验叙事报告和结构化记录后填写的问卷题项　　d.

测量指标	问卷条目
反复兴趣认知	它帮助我注意到在多件艺术作品中反复出现的兴趣。
跨艺术作品关联意识	它帮助我注意到自己对不同艺术作品的体验或反应之间的关联。
意外发现	我发现了自己之前未曾考虑过的兴趣或关联。
感知保真度	我觉得所呈现的内容准确反映了我实际留下的记录。

D.3 Study1示例交流

表17再现了P09围绕Pablo Picasso的《哭泣的女人》（1937年；图6）展开的交流。作品目录记录中的注释概述了藏品博物馆对该作品的描述，且该作品符合ArtRAG的路由条件（第5.2节）。如果记录未说明该作品如何进入当前收藏，回答中便会指出这一点。右栏列出了每条回答所记录的情境来源。

表17：P09 在研究1的 MyMuse 状况下围绕《哭泣的女人》进行的交流。方括号中的省略号表示省略的段落。

Turn	Text	情境来源
P09	这是毕加索的作品，我非常喜欢他。你能告诉我，在创作这幅作品时，对这位艺术家产生影响的个人因素和历史因素有哪些吗?	
MyMuse	当然可以。《哭泣的女人》创作于1937年，毕加索的个人生活与当时的历史处境在这件作品中有着密切的重叠。当时，最重要的历史背景是西班牙内战和1937年格尔尼卡轰炸。同年，毕加索在格尔尼卡中表现了平民所遭受的苦难，而《哭泣的女人》可以被视为一幅源自那场冲击、随后发展而来的图像。 [...] 在个人层面上，重要的是，模特是朵拉·玛尔，当时她是毕加索的伴侣。NGV的记录还将这件作品置于两人复杂而强烈的关系中加以介绍。与其断言这幅绘画直接描绘了这种关系，不过，更准确地说，毕加索的个人矛盾与情感渗透于作品之中。 [...]	目录记录 维基数据 ArtRAG 模型知识
P09	顺便问一下，这件作品是如何来到澳大利亚的?	
MyMuse	问得好。不过，NGV 的记录并未说明具体的来源或它通过何种途径来到澳大利亚，因此我无法确切告诉你它曾经过谁的手。可以确定的是这是一幅 Pablo Picasso 于 1937年创作的油画，如今属于 NGV 的收藏集。现代作品通常会通过私人收藏家、经销商、画廊、捐赠或购买等方式进入海外博物馆，但现有记录无法确认这件作品的这些细节。如果你愿意，我可以概述它最可能通过哪些途径来到澳大利亚，依据是 NGV 通常如何收藏毕加索的作品。	目录记录 维基数据 ArtRAG 模型知识 先前回复 先前访客记录
P09	那么，毕加索对艺术史上的立体主义运动产生了什么影响? 我很好奇毕加索的艺术世界。	
MyMuse	毕加索是立体主义的关键人物，也是该流派的共同开创者之一，因此他对艺术史的影响非常深远。在尤其是将物体拆解并重新组合，仿佛同时从多个角度观察，而非从单一视角观察，这使绘画从描绘所见转变为重新感知物体。 [.. .] 在诸如《哭泣的女人》这样的作品中，立体主义碎片化超越了形式实验，被用来更强烈地表达悲伤与痛苦等情绪。对毕加索而言，立体主义并非冰冷的技法，而是一种有力的手段，用以浓缩人类内心生活以及他所处时代的悲剧。	目录记录 维基数据 ArtRAG 模型知识 先前回复 先前访客记录