arXiv:2603.02056v1 [cs.HC] 2 Mar 2026

视觉注意控制建模的资源理性原则

Yunpeng Bai 新加坡国立
大学 新加坡

摘要

理解人们如何分配视觉注意，是人机交互（HCI）的核心议题；然而，现有注意计算模型往往要么仅作描述、针对特定任务，要么难以解释。我的学位论文提出了一个基于模拟的资源理性框架，将视觉注意建模为在感知、记忆和时间约束下进行的序贯决策过程。我利用部分可观测马尔可夫决策过程，将阅读和多任务处理等视觉任务形式化为有界最优控制问题，从而使注视和注意切换等眼动行为能够由理性适应自然产生，而非通过手工编码或纯粹依赖数据驱动得出。这些模型在模拟环境中实例化，涵盖传统文本阅读以及佩戴智能眼镜边走边阅读的情境；它们既再现了经典的实证效应，也解释了观察到的理解与安全之间的权衡，并针对时间压力和界面变化提出了新的预测。总之，这项工作为视觉注意提供了统一的计算解释，并为理论驱动、资源高效的人机交互设计提供了新工具。

CCS 概念

• **以人为中心的计算 → 人机交互理论、概念与模型**。

ACM 参考格式：
Yunpeng Bai. 2026. 用于建模视觉注意控制的资源理性原则。载于《*2026* 年 *CHI* 人机交互系统会议扩展摘要集（*CHI EA '26*）》，*2026*年*4*月*13*日至*17*日，西班牙巴塞罗那。ACM，纽约州纽约，美国，6 页。
https://doi.org/10.1145/3772363.3799190

1 引言

视觉注意是获取信息的主要渠道 [7, 10, 27]。它塑造人们在日常环境中的感知与行动，包括在交互系统中的感知与行动 [12, 21]。除了采集视觉输入，注意分配还会影响用户在设备交互过程中如何协调感知与行动 [18, 32–34]，并支持多任务处理中的情境意识与安全 [2, 23]。尽管视觉注意至关重要，目前仍缺乏系统的计算理论来解释人类如何在不同任务和交互情境中控制视觉注意 [2]。

模拟为构建此类解释提供了一条有原则的途径，使关于注意控制的机制假设得以形式化并接受检验 [2, 3, 24, 25]。然而，经典模型往往依赖手工设计的策略（例如 E-Z Reader、SWIFT） [11, 30],，这限制了它们对不同任务、环境和界面设计的适应能力 [3, 25]。近期的数据驱动方法取得了出色的预测性能（例如 ScanDL、EyeTtention、基于 VQA 的模型） [5, 6, 8, 9, 15–17], , 但通常需要大规模训练数据，并优先考虑行为模仿，而非解释注意力为何 会如观察结果所示那样分配 [3]。

我的学位论文通过将视觉注意建模为资源理性控制，回应了这些局限。我将用户形式化为有限理性的智能体，它们会在感知和认知资源有限的情况下，权衡认知成本与环境成本，选择注意行动以最大化任务效用 [14, 22, 25]。这一原则不预先规定固定的启发式方法，而是让注意力策略根据任务需求和约束，自行形成有限最优策略。具体而言，我使用部分可观测马尔可夫决策过程（POMDPs） [35], , 将注意控制表述为不确定性下的序贯决策，从而以状态、观测、转移和奖励构成明确且可解释的结构。策略通过深度强化学习（近端策略优化） [31],进行学习，使智能体能够通过与模拟环境交互来习得注意力策略，而非依靠大规模人类注视数据进行监督。尽管核心建模框架已初步建立，学位论文中的若干关键方面仍有待完善，尤其是将基于模拟的建模构想为一种人机交互设计方法，以及对个体差异和群体差异的处理；这些也是目前持续改进的重点。

2 研究目标与方法

在攻读博士期间，我沿着一条连贯的研究路径不断推进，逐步从粗粒度模型转向细粒度模型，以研究视觉注意控制。我首先在一个立足现实情境的多任务处理场景中研究注意切换——即行走时通过智能眼镜阅读——在这一场景中，用户必须决定何时以及在数字内容与物理环境之间的何处分配注意力 [2]。随着智能设备日益无缝地融入日常生活，这一场景呈现了普遍存在且日益重要的日常挑战，也凸显了在现实条件下平衡信息摄取与安全的必要性。

随后，我将阅读作为典型的细粒度注意任务加以聚焦。阅读是获取信息的基本途径，但眼动并非线性采样过程：它们与不断演变的理解和记忆状态紧密耦合 [3, 28]。在这一尺度上，注意控制作用于单个字母和单词层面，需要在低层眼动动作与高层认知过程之间进行精确协调。我开发了一个计算模型，共同解释眼动

本作品依据知识共享署名 4.0 国际许可协议授权。
CHI EA '26，西班牙巴塞罗那 © 2026 版权所有者/作者保留版权。ACM ISBN 979-8-4007-2281-3/2026/04
https://doi.org/10.1145/3772363.3799190

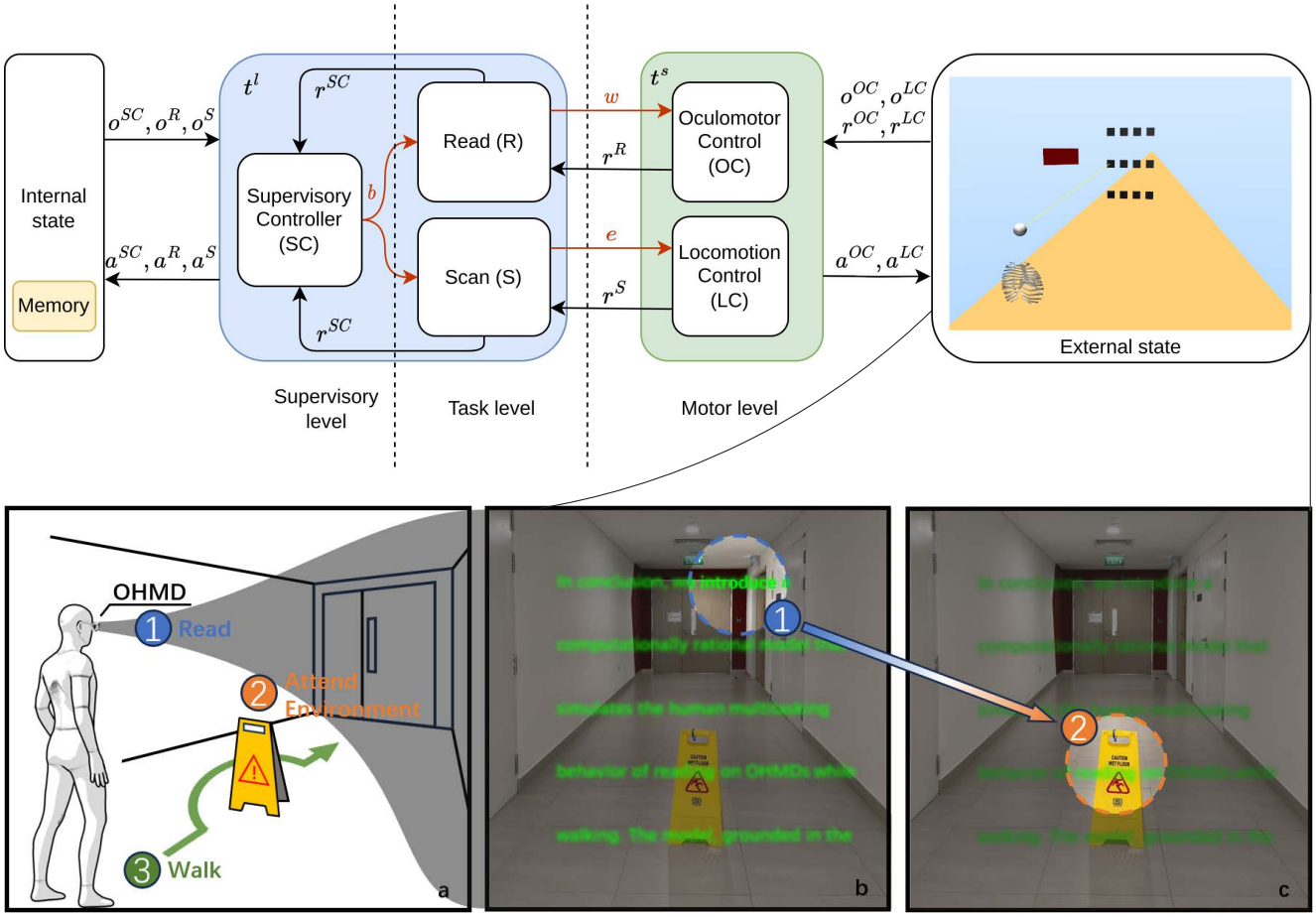

图 1：抬头多任务模型：一种用于建模边走边阅读时注意分配的分层资源理性模型。该模型通过在相互竞争的任务需求下分配视觉注意，使模拟智能体能够协调在光学头戴式显示器（OHMDs）上阅读与安全行走。上图展示了分层强化学习架构，该架构将注意控制分解为监督层、任务层和运动层级，使有界最优控制所带来的阅读与安全之间的权衡具有可解释性。下方各图展示了外部环境中的模拟多任务场景：(a) 智能体在有危险因素的环境中行走的第三人称视角；(b) 智能体阅读光学头戴式显示器（OHMDs）上呈现文本的第一人称视角；(c) 智能体关注环境标识以支持安全导航的第一人称视角。

控制以及统一的资源理性原则下动态的、基于记忆的理解 [3]。

在这一发展脉络的指引下，我的研究围绕三个相互关联的目标展开：

- **将多任务处理建模为资源理性控制过程**。我将人类的多任务处理形式化为不确定性下的序贯决策过程：用户在行走时，在智能眼镜与周围环境之间切换视觉注意力。智能体旨在有限的视觉感知资源下，最大限度地提高文本理解度，同时尽量降低因忽视环境信息而引发的安全风险（图 1）。关键问题包括：如何指定能够体现理解度与安全之间权衡的部分可观测马尔可夫决策过程组成部分，以及注意力策略如何适应步行速度和界面设计（例如文本间距和布局）等情境因素。

- **将文本阅读建模为资源理性的控制问题**。我将阅读形式化为一个精细控制过程：读者在字母和单词层面决定注视位置，旨在视觉敏锐度、记忆容量和时间有限的约束下，最大限度地提高理解水平，同时尽量降低眼动和时间成本（图2）。关键问题包括：部分可观测马尔可夫决策过程形式化如何捕捉眼动行为对词汇属性和时间压力的适应，以及这些适应如何影响理解结果。

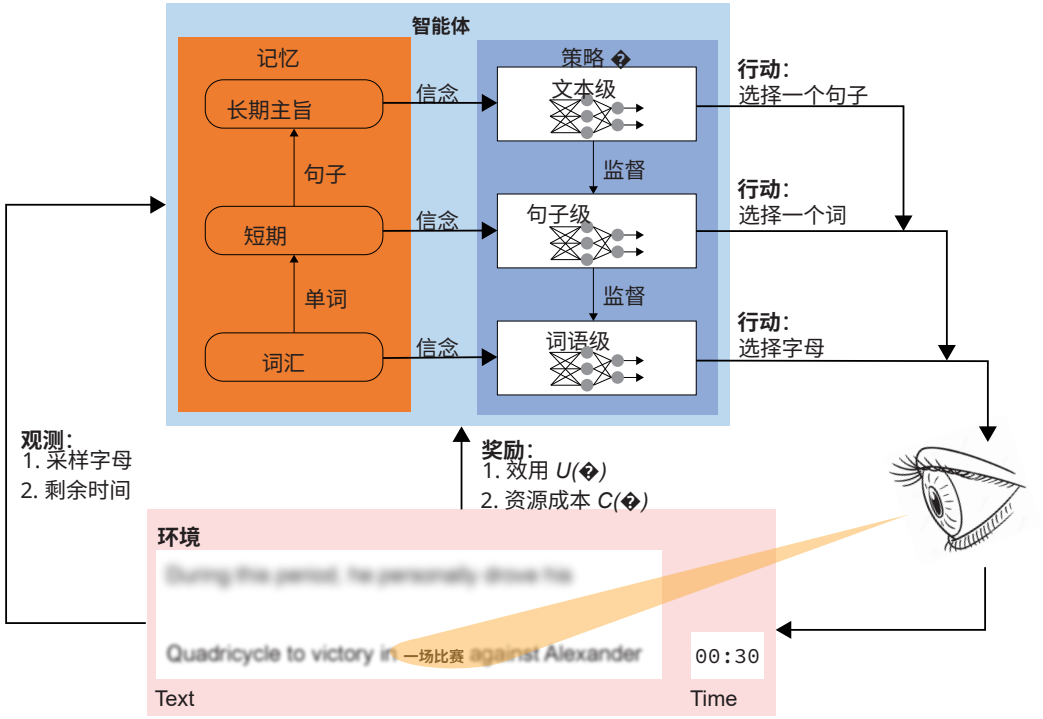

图2：资源理性阅读模型的架构。该模型将读者表示为一个序贯决策智能体，在有限记忆、视敏度和时间的约束下分配视觉注意，以最大化文本理解。由于对文本的观测是局部且带有噪声的，智能体会在三个表征层级上维持概率信念：词汇、句子和文本，并由相应的记忆系统（词汇存储、短期记忆和长期概要记忆）提供支持。决策按层级组织：文本层级控制器指导句子选择，句子级控制决定词语层级优先级，而词语级控制则支配对可见字母的眼动决策。奖励将理解效用与资源成本相结合，使注视模式、跳读和回视成为有界最优策略。这一架构将眼动控制与记忆驱动的理解联系起来，为人类阅读行为提供了一种可解释的计算说明，对预测型和基于模拟的人机交互研究具有参考意义。

- **对阅读中的注意分配进行建模，探究个体差异与无障碍性。**基于阅读框架，我研究注意控制策略如何因人而异。这包括捕捉母语与非母语读者之间的差异，如 SB-SAT [1] 和 MECO [19, 20], 中所报告的差异；也包括利用 CopCo 语料库 [4, 13, 29]对无障碍相关效应进行建模，例如阅读障碍读者所表现出的效应。目标是将个体差异解释为资源约束和习得控制策略的差异，而非临时拼凑的行为规则。

在所有项目中，我都采用一种一致的方法，基于分层深度强化学习 [26]。分层控制将复杂的注意分配问题分解为多个相互作用的时间尺度，使大规模且符合认知现实的任务变得可处理，同时保留可解释性。尽管当前的实现聚焦于以阅读为中心的任务，但底层的部分可观测马尔可夫决策过程形式化并不假设存在文本结构；状态表示和奖励函数可以针对视觉搜索、仪表盘监控或用户界面探索等非文本任务重新定义。

在多任务模型（图 1）中，注意控制被划分为监督层级、任务层级和运动层级控制器：高层策略决定何时在阅读与环境监控之间切换注意力，而低层策略则执行眼动和行走动作，以获取与任务相关的信息。这种分解使安全与理解之间的权衡能够自然地体现为有界最优解。

在阅读模型（图2）中，层级结构反映了语言与记忆过程的结构 [3]。文本层级控制器负责句子选择，句子级控制器根据不断变化的理解情况确定单词的优先级，而词级控制器则从可见字母中选择注视目标。由于观测是不完整且存在噪声的，智能体会在词汇、句子和文本表征上维持概率信念状态，并由相应的记忆系统提供支持。策略通过深度强化学习（近端策略优化）进行学习，无需人类注视数据的直接监督，使注视模式、跳读和回视能够在

认知约束下的任务优化过程中自然涌现。这种统一的方法既支持解释性建模，也支持基于模拟的交互系统中人类注意行为评估。

3 结果与贡献

我的第一条研究工作提出了*Heads-Up Multitasker*，这是首个模拟模型，用以刻画用户在行走时于光学头戴式显示器（OHMDs）上阅读过程中如何分配视觉注意。在四项实证与模拟研究中，该模型复现了关键的多任务现象，包括在不同行走速度与任务优先级下系统性的注意切换、由运动诱发的扰动导致的阅读表现下降、在不同光学头戴式显示器文本布局间准确恢复阅读，以及阅读效率与行走安全之间的适应性权衡。这些结果表明，有界最优控制为用户如何在安全关键的日常多任务场景中协调感知与行动提供了原则性解释，并为设计更安全、更可用的光学头戴式显示器界面提供了可预测的洞见。

在此基础上，我的第二条研究工作构建了一个分层资源理性阅读模型，将眼动控制与记忆驱动的理解统一起来。该模型将阅读解释为序贯决策，在单一计算模型中关联词级注视选择、句子级整合与文本级记忆形成。在未经人类眼动追踪数据训练的情况下，它复现了已确立的实证效应，包括注视时长、跳读和回视随词汇难度、可预测性与语境约束而发生的系统性变化，以及文本连贯性与先验知识对理解的影响。为扩展该模型，我收集了一个新数据集，其中参与者在明确的时间限制下阅读，并在人类与模型中均表现出趋同的适应：更紧的时间约束导致阅读更快、跳读增加、回视减少以及理解下降（图3）。这项工作提供了一种整合性的计算解释，共同说明阅读过程中的注视行为与不断演化的记忆表征。

除了总体指标外，我们还在个体扫描路径轨迹层面评估了模型。具体而言，我们使用扫描路径相似度的标准指标——归一化莱文斯坦距离（NLD），逐试次比较模拟生成的扫描路径与人类参与者的扫描路径。值得注意的是，这项比较针对的是一次模拟运行与一个人类试次，而非汇总后的人类热力图。我们的资源理性模型在各个时间条件（30 s、60 s、90 s）下均取得最低的 NLD 值，优于先前的阅读模型。除了轨迹相似度外，我们的方法还通过控制知觉敏锐度、记忆约束和成本敏感度的自由参数，支持对个体差异进行建模。通过调整这些参数，模型能够再现各项指标的分布，包括阅读速度、跳读概率和回视概率，使其与人类参与者中观察到的变异范围相符。这使得模型不仅能够模拟平均阅读行为，还能通过将参数配置拟合到个体数据，模拟特定的读者特征。

较简单的启发式模型或基于规则的模型或许能够捕捉孤立效应，而资源理性框架则提供了一种统一的生成式解释，在单一的决策论结构中将感知、记忆与任务目标联系起来。这使得人们能够系统地调整任务约束和设计参数，而纯描述性模型难以轻易做到这一点。

这些项目共同确立了视觉注意是一种分层控制形式，涵盖了跨任务和环境决定关注什么，以及在精细的知觉尺度上如何采样信息。它们展示了计算理性和分层强化学习如何在不依赖大规模行为模仿的情况下，生成可解释且具有预测能力的复杂人类行为模型。重要的是，这种结构使设计师能够在开展成本高昂的实证研究 [24]之前，模拟注意策略在采用不同界面布局、时间约束或资源假设时将如何适应。在实践层面，这两条研究路线都提供了开源模拟环境和数据集，支持系统性的“假设情景”分析以及对交互设计的可扩展评估。总的来说，本学位论文将基于模拟的资源理性建模定位为人机交互领域以理论为导向、数据高效的界面设计基础。

4 未来方向

我的研究将把视觉注意的资源理性建模拓展至适应性和交互性日益增强的场景。从理论角度看，下一项关键工作是将计算理性控制从单智能体行为推广到多智能体及社会情境场景。许多现实任务，例如课堂中的协作阅读，或 AR/VR 环境中的共享注意，都需要多个目标相互重叠但又各不相同的智能体彼此协调。对这类场景进行建模，将使我们能够研究注意如何不仅在个体内部、也在群体之间分配，从而为理解联合认知和协作式人机交互开辟新的途径。

在应用层面，我计划利用这些模型，为更广泛的人机交互场景中的自适应界面设计提供依据。阅读提供了一个受控且认知内容丰富的测试平台，而同一仿真框架也可以扩展到与可视化内容、图像和复杂用户界面的交互场景中；在这些场景下，用户必须在时间和资源约束下策略性地分配注意力。通过模拟用户在多任务处理、时间压力或高认知负荷下如何调整注意力，这些模型有助于在开展大规模用户研究之前，系统地分析各种设计方案，并指导开发时间感知阅读界面、AR系统中的预期性安全提示，以及能够根据预测的用户状态作出响应的自适应视觉界面和教育界面 [2, 24]。

具体而言，我旨在将深度生成模型整合到仿真流水线中，以提升可扩展性和真实感。大型语言模型可提供丰富的上下文预测，以支持语义推断和记忆构建 [32]，而扩散模型——

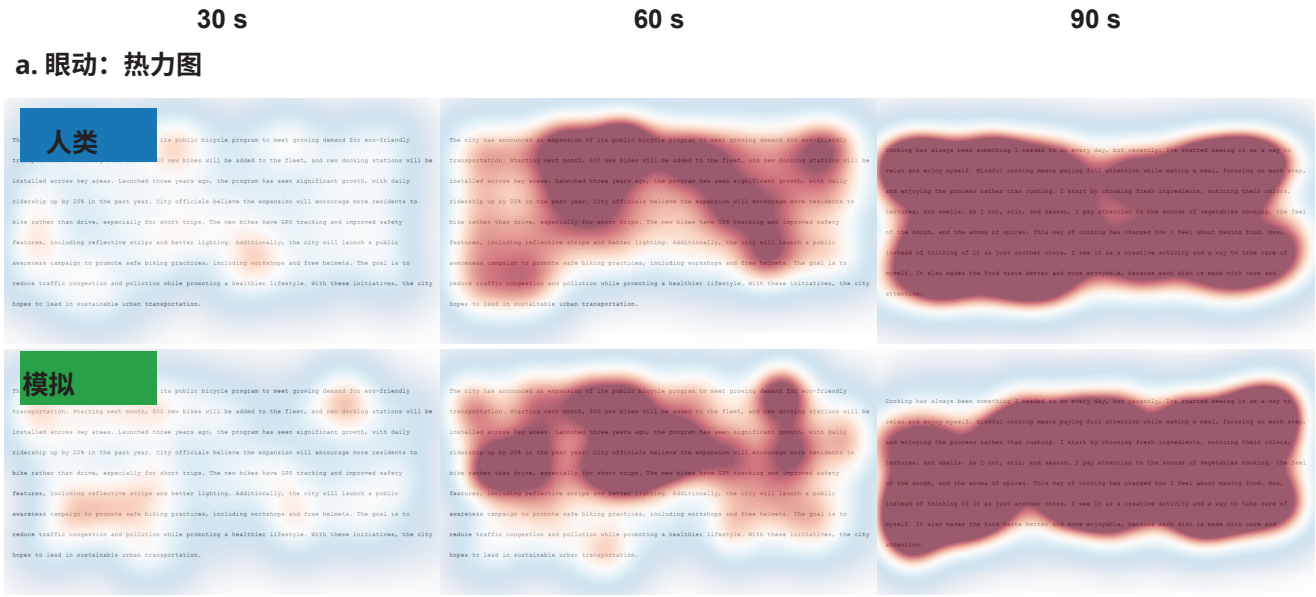

图3：不同时间约束（30 秒、60 秒、90 秒）下的阅读行为。随着可用阅读时间的变化，模型和人类读者会调整视觉注意策略，体现出对理解的资源理性控制。(a) 热力图。随着时间增加（从左到右），人类和模拟都会将注意力更广泛地分配到文本各处，并投入更多注视时间；而在截止期限紧迫时，注意力则集中于高覆盖区域。(b) 回忆。自由回忆反应呈现出相同的权衡：较长的阅读时间带来更丰富的要旨层面回忆，以及更完整、详尽的记忆。

或者可以使用基于Transformer的生成模型来合成逼真的眼动模式或受控交互语料库 [5, 6, 9]。整合这些组件将有助于开展基于模拟的分析，深入研究细微的理解和交互过程，并支持能够根据推断出的用户状态动态调整的界面。

最后，在我的学位论文路线图中，剩余的里程碑重点是扩展阅读模型，以捕捉个体差异和不同读者群体的特点。这包括对母语与非母语读者之间的差异以及无障碍相关差异进行建模，计划投稿至 UIST 和 CHI。这些方向共同旨在构建一个通用且可扩展的框架，用于对注意力感知交互系统进行建模与设计。

在学位论文研究的这一阶段，我尤其希望获得以下方面的指导：(i) 如何更好地将资源理性模拟定位为对人机交互设计实践的贡献，而不仅仅是认知理论；(ii) 为了确保可解释性，相对于设计师的可用性，需要多少机制细节；以及 (iii) 如何在一篇学位论文中界定个体差异建模的范围。这些问题促使我参加学位论文研究圆桌论坛。

参考文献

[1] Seoyoung Ahn、Conor Kelton、Aruna Balasubramanian 和 Greg Zelinsky。2020。基于注视行为预测阅读理解。载于 ACM眼动追踪研究与应用研讨会论文集。1–5。[2] Yunpeng Bai、Aleksi Ikkala、Antti Oulasvirta、Shengdong Zhao、Lucia J Wang、Pengzhi Yang 和 Peisen Xu。2024。Heads-Up Multitasker：模拟光学头戴式显示器上的注意切换。载于 2024年CHI人机交互大会论文集。1–18。[3] Yunpeng Bai、Xiaofu Jin、Shengdong Zhao 和 Antti Oulasvirta。2025。阅读的资源理性机制。（2025）。[4] Marina Björnsdóttir、Nora Hollenstein 和 Maria Barrett。2023。基于自然阅读丹麦语文本预测阅读障碍。载于第24届北欧计算语言学会议（NoDaLiDa）论文集。60–70。[5] Lena S Bolliger、David R Reich、Patrick Haller、Deborah N Jakobi、Paul Prasse 和 Lena A Jäger。2023。ScanDL：一种用于生成文本合成扫描路径的扩散模型。arXiv预印本 arXiv:2310.15587（2023）。[6] Luca S Bolliger、Daniel R Reich 和 Lukas A Jäger。2025。Scandl 2.0：一种生成式阅读眼动模型，可合成扫描路径和注视时长。《ACM人机交互学报》9，3（2025），1–29。[7] Marisa Carrasco。2011。视觉注意：过去25年。《视觉研究》51，13（2011），1484–1525。[8] Xianyu Chen、Ming Jiang 和 Qi Zhao。2021。预测视觉问答中的人类扫描路径。载于 IEEE/CVF计算机视觉与模式识别会议论文集。10876–10885。[9] Shuwen Deng、David R Reich、Paul Prasse、Patrick Haller、Tobias Scheffer 和 Lena A Jäger。2023。Eyettention：一种基于注意力的双序列模型，用于预测人类阅读时的扫描路径。《ACM人机交互学报》7，ETRA（2023），1–24。[10] Robert Desimone、John Duncan 等。1995。选择性视觉注意的神经机制。《神经科学年度评论》18，1（1995），193–222。[11] Ralf Engbert、Antje Nuthmann、Eike M Richter 和 Reinhold Kliegl。2005。SWIFT：一种阅读过程中扫视生成的动力学模型。《心理学评论》112，4（2005），777。[12] Mary Hayhoe 和 Dana Ballard。2005。自然行为中的眼动。《认知科学趋势》9，4（2005），188–194。[13] Nora Hollenstein、Maria Barrett 和 Marina Björnsdóttir。2022。哥本哈根语料库：自然阅读丹麦语文本的眼动追踪记录。arXiv预印本 arXiv:2204.13311（2022）。[14] Andrew Howes、Richard L Lewis 和 Alonso Vera。2009。任务和处理约束下的理性适应：对检验认知与行动理论的启示。《心理学评论》116，4（2009），717。[15] Yue Jiang、Zixin Guo、Hamed Rezazadegan Tavakoli、Luis A Leiva 和 Antti Oulasvirta。2024。EyeFormer：借助变换器引导的强化学习预测个性化扫描路径。载于第37届ACM用户界面软件与技术研讨会年会论文集。1–15。[16] Yue Jiang、Luis A Leiva、Hamed Rezazadegan Tavakoli、Paul RB Houssel、Julia Kylmälä 和 Antti Oulasvirta。2023。Ueyes：理解不同用户界面类型中的视觉显著性。载于2023年CHI人机交互大会论文集。1–21。[17] Yue Jiang、Changkong Zhou、Vikas Garg 和 Antti Oulasvirta。2024。Graph4gui：用于表示图形用户界面的图神经网络。载于2024年CHI人机交互大会论文集。1–18。[18] Jussi Jokinen、Aditya Acharya、Mohammad Uzair、Xinhui Jiang 和 Antti Oulasvirta。2021。触摸屏打字作为最优监督控制。载于2021年CHI人机交互大会论文集。1–14。[19] Victor Kuperman、Sascha Schroeder、Cengiz Acartürk、Niket Agrawal、Dominick M Alexandre、Lena S Bolliger、Jan Brasser、César Campos-Rojas、Denis Drieghe、Dušica Filipović Đurđević 等。2025。英语作为第二语言的文本阅读新数据：多语种眼动语料库（MECO）的第二波扩展。《第二语言习得研究》47，2（2025），677–695。[20] Victor Kuperman、Noam Siegelman、Sascha Schroeder、Cengiz Acartürk、Svetlana Alexeeva、Simona Amenta、Raymond Bertram、Rolando Bonandrini、Marc Brysbaert、Daria Chernova 等。2023。英语作为第二语言的文本阅读：来自多语种眼动语料库的证据。《第二语言习得研究》45，1（2023），3–37。[21] Michael F Land 和 Mary Hayhoe。2001。眼动在日常活动中有哪些作用？《视觉研究》41，25-26（2001），3559–3565。[22] Falk Lieder 和 Thomas L Griffiths。2020。资源理性分析：将人类认知理解为对有限计算资源的最优利用。《行为与脑科学》43（2020），e1。[23] Alexander Lingler、Dinara Talypova、Jussi PP Jokinen、Antti Oulasvirta 和 Philipp Wintersberger。2024。利用强化学习支持任务切换。载于2024年CHI人机交互大会论文集。1–18。[24] Roderick Murray-Smith、Antti Oulasvirta、Andrew Howes、Jörg Müller、Aleksi Ikkala、Miroslav Bachinski、Arthur Fleig、Florian Fischer 和 Markus Klar。2022。模拟能为人机交互研究做些什么。*Interactions* 29, 6 (2022), 48–53. [25] Antti Oulasvirta, Jussi PP Jokinen, and Andrew Howes. 2022. 计算理性作为交互理论。载于*2022*年*CHI*人机交互大会论文集。1–14。[26] Shubham Pateria, Budhitama Subagdja, Ah-hwee Tan, and Chai Quek. 2021. 分层强化学习：全面综述。*ACM*计算调查 *(CSUR)* 54, 5 (2021), 1–35. [27] Michael I Posner. 1980. 注意定向。实验心理学季刊 32, 1 (1980), 3–25. [28] Keith Rayner. 1998. 阅读和信息处理中的眼动：20年的研究。心理学通报 124, 3 (1998), 372. [29] David R Reich, Shuwen Deng, Marina Björnsdóttir, Lena Jäger, and Nora Hol- lenstein. 2024. 阅读不等于阅读：比较、模拟和利用不同群体的阅读行为。载于*2024*年计算语言学、语言资源与评估联合国际会议论文集（*LREC-COLING 2024*）。13586–13594。[30] Erik D Reichle, Keith Rayner, and Alexander Pollatsek. 2003. 阅读中眼动控制的EZ Reader模型：与其他模型的比较。行为与脑科学 26, 4 (2003), 445–476. [31] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. 近端策略优化算法。*arXiv*预印本 *arXiv:1707.06347* (2017). [32] Danqing Shi, Yao Wang, Yunpeng Bai, Andreas Bulling, and Antti Oulasvirta. 2025. Chartist：面向图表阅读的任务驱动眼动控制。载于*2025*年*CHI*人机交互大会论文集。1–14.[33] Danqing Shi, Yujun Zhu, Francisco Erivaldo Fernandes Junior, Shumin Zhai, and Antti Oulasvirta. 2025. 模拟触摸屏打字中的错误。载于*2025*年*CHI*人机交互大会论文集。1–13.[34] Danqing Shi, Yujun Zhu, Jussi PP Jokinen, Aditya Acharya, Aini Putkonen, Shumin Zhai, and Antti Oulasvirta. 2024. CRTypist：通过计算理性模拟触摸屏打字行为。载于*2024*年*CHI*人机交互大会论文集。1–17。[35] Richard S Sutton, Andrew G Barto, et al. 1998. 强化学习：导论。第1卷。MIT出版社，剑桥。