恰到好处的多样性：使用指标选择方法改进表达范围分析

Oliver Withington
o.withington@qmul.ac.uk 伦敦玛丽女王大学 伦敦，英国

Laurissa Tokarchuk
karchuklaurissa.tokarchuk@qmul.ac.uk
伦敦玛丽女王大学 伦敦，英国

arXiv:2304.02366v1 [cs.HC] 5 Apr 2023

ACM 参考文献格式：
OliverWithington和Laurissa Tokarchuk。2023年。《恰到好处的多样性：通过指标选择方法改进表达范围分析》。载于 数字游戏基础 *2023*（*FDG 2023*），*2023*年 *4* 月*12–14* 日，葡萄牙里斯本。ACM，美国纽约州纽约市，11 页。https://doi.org/10.1145/3582437.3582453

1 摘要

表达范围分析（ERA）是一种用于可视化程序化内容生成（PCG）系统输出的方法，在程序化内容生成研究中被广泛用于评估和比较生成器，通常用于就其在输出多样性和搜索空间探索方面的相对性 能作出比较性陈述。生成标准的ERA可视化需要选择两个指标，这两个指标必须能够针对所有待可视化的生成制品进行计算。然而，据我们所知，目前没有任何方法论或启发式方法能够证明选择某一特定指标对优于其他备选方案的合理性。以往的工作通常要么基于已确立但缺乏论证的规范或设计师直觉进行选择，要么针对所有可能的指标对生成多个可视化结果。本研究旨在通过识别指标配对的有价值特征，并证明具有这些特征的配对更有可能生成关于底层生成器的有信息量的ERA投影，为这一领域作出贡献。我们提出并研究了三个可量化的选择标准，用于评估指标对，并展示了如何将这些标准操作化，以对可用选项进行排名。尽管这只是对量化ERA指标对效用这一概念的初步探索，但我们认为，本文所探讨的方法能够使ERA对研究人员和游戏设计师都更加有用且易于使用。

2 引言

游戏的程序化内容生成（PCG）系统往往能生成海量、多样的制品和体验。无论是 Minecraft 中数十亿种可能的地图 [18]，还是《无人深空》中的数千万亿颗行星 [6]，PCG 系统可能生成的内容总量通常称为其“生成空间”，其规模可能极其庞大且复杂。理解并有效比较生成空间，是游戏研究人员和游戏设计师都至关重要的挑战。研究人员希望能够令人信服地证明，与以往的替代方案相比，他们的新系统在输出的多样性或质量方面有所提升。同样，商业游戏设计师也希望确信自己选对了生成方法，并能正确调整所选方法，从而生成最理想的生成空间。

理解生成空间的一种常见方法是将其可视化，把高度复杂的 n 维产物可能性空间浓缩成便于人类理解的二维表示。早期程序化内容生成研究中，将生成空间可视化为二维图的一种常见技术是表达范围分析（ERA），该技术由 Smith 等人开创 [21]。ERA 允许我们使用两个或更多感兴趣的可量化指标（例如衡量难度或美学特质的启发式指标）来可视化游戏关卡集合。这些指标用于在二维图中定位每个产物，进而帮助我们了解生成的内容如何分布在由这些指标定义的空间中。这有助于设计师了解其生成器的生成空间在该领域可能生成的更大关卡可能性空间中的位置，并将其与其他生成空间进行比较。

可以说，应用 ERA 时最重要的选择，是用于生成该图的指标对。然而，在以往研究中部署 ERA 时，这一选择通常并未得到论证，其影响通常也未被讨论。据我们所知，尚无先前工作旨在论证特定指标对相对于其他可选方案的价值，也没有方法帮助研究者或设计者决定应使用的最佳指标对。最常见的情况是，使用 ERA 的研究要么沿用先前工作中的指标配对（通常来自 Smith 等人的原始论文），要么为所有指标对生成大量 ERA 图。

本文的贡献正是在于为应用 ERA 选择指标这一领域。我们认为，指标对实际上具有若干特征，这些特征可能使其在应用 ERA 时比其他备选方案更可取；并且，在候选指标对上计算这些特征，可以构成有用的选择标准的基础。具体而言，我们考察的选择标准包括：

- 适应度独立性，即适应内容在指标对所定义的空间中均匀分布
- 相互独立性，即所选指标对之间的相关性最小
- 最小替代指标独立性，意指存在未被选定指标对计入的最小指标多样性

特此授予许可：可免费为个人或课堂使用而制作本作品的全部或部分数字副本或纸质副本，前提是不得为营利或商业利益而制作或分发副本，且副本首页须附有本声明及完整引文。对于本作品中由作者以外的其他方拥有版权的组成部分，必须尊重其版权。允许在注明出处的情况下进行摘要。若要以其他方式复制、重新发表、发布到服务器或重新分发到列表，则必须事先获得具体许可和/或支付费用。请通过permissions@acm.org申请许可。*FDG '23，2022*年*4*月*11*日至*14*日，葡萄牙里斯本© 2023 版权所有归所有者/作者所有。出版权获授权给 ACM。ACMISBN 978-1-4503-9855-8/23/04...$15.00https://doi.org/10.1145/3582437.3582453

本研究论证了：指标对越符合这些标准，使用它们的ERA可视化就越有用。综合来看，这些标准有助于确保生成器能够访问ERA图中尽可能多的区域，图中尽可能多的区域对生成器的探索具有价值，并且对于使用该指标对的生成空间而言，看似多样性的内容确实代表了跨越众多可能指标的多样性。

本文介绍了这些用于 ERA 指标对的选择标准，并探讨了它们对于希望了解或比较自身系统生成空间的游戏设计师和研究人员的价值。我们展示了如何计算这些特征，以区分并排列其他指标配对，从而在难以判断时，能够更明智地选择要进行可视化的指标对。我们还在热门研究领域——超级马力欧兄弟生成关卡——开展了一系列试点实验，运用这一方法评估并排列了 150 多种可能的指标配对，以便使用 ERA 对生成结果进行可视化。实验结果表明，尽管这些理念仍处于早期探索阶段，将这些选择标准应用于指标选择，仍能为如何在特定游戏领域中开展 ERA 提供宝贵见解，也可能有助于更广泛地理解如何设计优良指标。我们总结认为，这一方法能够在保留 ERA 现有优势的同时，显著提升其当前应用方式的可用性和效用，使其成为一种更稳健的方法，用于理解和比较程序化内容生成系统的生成空间。

3 相关工作

本节将介绍并讨论对本项目影响最直接、并为其提供灵感的研究，既突出支撑本项目的核心概念，也说明本项目旨在如何改进当前最先进的成果。

3.1 表现力范围分析

ERA 最初由 Gillian Smith 和 Jim Whitehead 在2010 [21], 中提出，作为一种可视化并理解生成空间的新方法。它的运行方式简单而直观，这无疑促成了它的广泛应用。要进行基础的 ERA 分析，设计者首先从生成器生成大量关卡，然后为每个关卡计算两个或更多指标。随后，可将这些关卡绘制在二维图中，最常见的是热力图。这些二维图可用于理解单个生成空间的特征与多样性，也可用于比较不同生成器生成的空间。ERA 最常见的应用领域是其最初的方向——游戏关卡的程序化内容生成，但它也被用于程序化内容生成研究的其他领域，例如可视化诗歌生成器的输出 [15], 、桌面游戏结果 [10] 以及涌现叙事系统 [16]。

在此后的数年中，ERA已被用于程序化内容生成研究中的若干其他用途。在Smith 等人的原始论文 [21]中，它主要用于定性地理解其系统输出倾向，重点关注诸如可能生成哪些类型的关卡，以及哪些关卡难以或不可能生成等特征。不过，他们也强调，ERA可以构成不同生成器之间定量比较的基础，这一用例此后在研究中非常流行。它常被用来借助启发式指标评估替代系统所能产生的输出多样性的相对程度。输出多样性常被视为程序化内容生成系统固有的理想质量，而ERA提供了一种比较替代系统输出多样性的途径，即比较哪个生成空间在ERA图中分布得更广。因此，它经常出现在新颖的程序化内容生成研究中，用以支持关于程序化内容生成系统在其生成空间中所呈现的多样性方面如何比较的论述 [2, 12, 13, 27]。

虽然 ERA 目前仍以其原始形式用于当代研究，但围绕这一概念也出现了若干演进，以提升其适用性和功能性。Adam Summerville 提出了使用角图同时可视化多个指标对的概念，从而克服了此前一次只能查看两个指标的局限 [22]。他们还引入了 e-距离度量，用于比较 n 维空间中生成空间的分布。Cook 等人开发了 Danesh 系统，用于在 Unity 中执行自动化 ERA，并利用他们称为随机 ERA（Randomised ERA）或 RERA 的方法，识别同一生成器在不同参数化方式下 ERA 的趋势 [5]。Kremenski 等人引入了“表现范围覆盖分析”，用于可视化设计师随时间推移与生成系统交互的趋势 [15]。我们还探索了使用降维和卷积神经网络嵌入等替代技术，以生成 ERA 风格的生成空间可视化，而无需定义或重新计算我们所关注的多样性类型[25, 26]。

3.2 生成空间、可能性空间与 ERA

生成空间和可能性空间是本研究以及更广泛的 ERA 领域中的重要概念。这些术语由 Mike Cook 在其研究 [4] 中推广，有助于凸显 ERA 的优势和不足，以及本研究方法的潜在益处。遗憾的是，这些术语的定义尚未得到普遍认可，一些颇具影响力的研究也会将生成空间和可能性空间互换使用[10]。不过，这一区分在概念上仍然十分有用，我们希望这些术语今后能得到更广泛的使用。

正如我们之前所讨论的，程序化内容生成系统的生成空间指的是包含特定系统能够生成的每一个产物的体量。相比之下，可能性空间指的是包含该领域中任何生成器能够表示的每一个产物的体量。通常，生成空间是包含在更大可能性空间内的较小体量。当游戏设计师在为某项任务选择不同的生成方法，或调整所选生成方法的参数时，这可以被概念化为：针对给定的可能性空间，选择一个最优的生成空间。同样，当设计师根据输出比较两个生成器时，他们比较的是同一可能性空间内生成空间的两种不同局部化位置。

这一术语为我们理解ERA 提供了一个视角。ERA图同时生成了程序化内容生成系统的生成空间以及构成其可能性空间的各个域的二维投影。正如生成空间的n 维体积被压缩为由指标决定的图一样，其背后潜在的、可能复杂得多的可能性空间也会被压缩。生成空间的ERA可视化的效用，很大程度上取决于所选指标对对可能性空间投影的影响。在开展ERA 时，很容易选定某些指标对，导致图中的大部分区域不包含优质内容，甚至由于大多数指标值彼此不兼容，图中的大部分区域根本不属于内容的可能性空间。本工作的这一方法旨在帮助设计师利用 ERA，更有效地投影底层可能性空间，从而更有效地投影特定的生成空间，同时保持 ERA 的可解释性与简洁性。

3.3 ERA 的指标选择

设计师在将常规 ERA 应用于生成空间时必须作出的主要决策，是选择用于可视化该空间的指标对。这一选择决定了可视化将捕捉哪些类型的多样性，以及可能忽略哪些类型的多样性。任何受到关注但未被纳入图表的指标，其分布都是未知的。以三个受到关注的指标为例，图表可能显示指标 1 和 2 具有很高的多样性，同时却掩盖了指标 3 的同质性。如果部署 ERA是为了让设计师根据他们已知的两个受到关注的指标，对生成空间的分布进行定性探索，那么就无需作出选择。然而，如果部署ERA 是为了论证某个生成器生成的内容总体上比另一种方案更加多样，或者设计师关注的指标很多，那么这一选择就不那么显而易见了。

一般而言，选择指标有两种方法。首先，研究人员可以重复使用既有文献中采用过的指标，最常见的是 Smith等人原始论文中的两个指标。在该论文中，Smith 等人提出了“线性度”指标，用作衡量关卡有多平坦的启发式指标，以及“宽容度”指标，用作衡量关卡难度的启发式指标。尽管这两个指标是为在其特定关卡生成系统的背景下提供有意义的洞见而设计的，但此后它们已在程序化内容生成研究中被广泛重复使用，既用于开展 ERA [2, 12, 13],，也作为指标，用来有意在基于质量-多样性搜索的程序化内容生成中寻找多样性 [1, 19]。以这种方式重复使用指标本身并无不妥，但缺乏对在新领域中重复使用这些指标的理由说明，意味着研究人员可能会使用并不适合其目的的指标。

我们还注意到，尽管“线性度”和“宽容度”这两个名称常用于描述指标选择，但用于计算它们的底层方法往往会根据所关注的具体游戏领域进行调整。宽容度最初是根据关卡中的跳跃、移动平台和敌人数量计算得出的一个得分 [21],，但后来也被计算为关卡中每个点平均难度的归一化评分 [12],，或者更简单地用奖励数量减去敌人数量来计算 [13]。线性度同样最初是通过线性回归计算关卡中每个水平位置对应的最高点与中心线之间的间隙 [20, 21],，但此后在不同情境中采用了截然不同的计算方式，例如 Alvarez 等人在其关于《进化地牢设计器》（EDD）的研究中，使用门与门之间路径数量的归一化计数 [2]。这些替代计算方式通常符合实际需要。例如，EDD 创建的关卡采用俯视视角，场景中没有致命间隙，因此原始公式并不适用。然而，在我们看来，即使是这些表面上相似的 ERA 指标对，其计算方式也存在如此大的差异，这凸显了启发式方法在帮助人们做出更明智选择方面的潜在价值。

指标选择的另一种方法是完全不进行选择，而是将所有可能感兴趣的指标对的图都进行可视化。这份更长的指标列表同样可以全面地从类似游戏领域的既有文献中获取 [2, 11],，或者由游戏领域专家设计一份列表 [15]。随后，可以使用 ERA 将这 n 个指标的所有可能配对进行可视化，通常采用Summerville的角落图概念 [22]。然而，虽然呈现所有可能的可视化结果可以减小其中某一个具有误导性或不理想而造成的影响，但并不能避免这些结果的存在。这也可以说会限制呈现单个图时的可读性，尽管情况不一定如此。

人们一直在努力为 ERA 中选择单个指标提供更充分的依据，最常见的形式是验证常用指标是否与人类对质量和多样性的感知具有相关性。Summerville 等人探究了常用指标在预测玩家体验方面的效力，并发现了很强的相关性 [23]。Marino 等人开展了类似的研究，但发现相关性要有限得多 [17]。Herve 与 Salge 探究了常用指标与《我的世界》地图专家评估之间的关系 [11]。这些研究对于确认或反驳特定指标是否捕捉到了设计师真正关心的质量或多样性形式而言，一直具有不可估量的价值。然而，由于它们关注的是单个指标的信息提供能力，因此在帮助选择指标配对方面的实用性有限。正如我们将在第 4 节看到的那样，使用两个与人类感知高度相关的指标应用 ERA，并不能确保绘图结果具有实用价值。两个这样的指标仍有可能对可能性空间产生无益的投影，例如空间中的大部分区域都不理想或无法访问。

3.4 Mario AI Benchmark

本研究的初步实验使用 Mario AI Benchmark，这是一个广泛使用的 AI研究平台，用于开发和评估进行关卡游玩与关卡生成的 AI 智能体。尽管我们赞同 Mike Cook 的观点，即超级马里奥在程序化内容生成文献中作为测试平台受到了过度关注，但在本研究中使用它是合理的，因为已有大量工作专门将 ERA 应用于马里奥关卡生成器。这既让我们能够借鉴种类繁多的既有指标，也使我们对最优指标选择所获得的任何见解都与研究社区更加相关。

我们使用Horn等人在2014年工作中生成的关卡语料库 [12], , 这些语料库由九种多样的马里奥关卡生成器生成，并被纳入最新版本的基准中；该基准还包含来自原始马里奥游戏的 15个关卡，总计 9015个关卡。这些关卡集合出于多种原因非常适合本研究。它们全部使用相同的编码系统，其中关卡以字符的二维矩阵表示，每个字符代表一种图块类型。这意味着我们可以在语料库中的所有关卡上使用相同的指标计算方法。此外，先前的研究已经证实，来自不同生成器的集合在具有实际意义的层面上存在多样性，而这种多样性可以通过替代的可视化方法检测出来[12, 25]。

为自动评估关卡，我们使用 Robin Baumgarten 基于 A* 的马里奥关卡游玩智能体。该智能体是首届马里奥AI竞赛的获胜智能体 [24], , 已纳入基准测试，并常用于程序化内容生成研究，以确认关卡可通关 [3, 7]。

4 指标对选择标准

正如我们所看到的，ERA 的应用在所选指标方面缺乏一致性，而且通常很少说明为何选择某些指标及其配对，而非其他选项。然而，我们认为，可以从 ERA 图的设计中推断出实用 ERA 图所具备的一些特征，并以此为基础制定启发式方法，指导指标对的选择。本节将介绍并讨论这些特征的相关性，以及如何将其操作化，从而形成指标对选择标准。我们提出的所有选择标准，都是针对从较大的候选指标池中选出的某一对指标计算的。对于一组可能值得关注的 n 个指标，共有 n(n-1)/2 种潜在指标对可供选择，以生成 ERA 图。这些选择标准旨在为从这些潜在指标对中选择某一对而非其他备选项提供可量化的依据。

4.1 标准 1：适应度独立性

我们认为，对于指标对而言，第一个有价值的选择标准是适应度独立性。所谓适应度独立性，是指在底层可能性空间的二维 ERA 投影中，适应度高或理想的关卡尽可能均匀地分布（示意图见图 1）。适应度一词由基于搜索的算法领域推广开来，用于指称可量化的指标，以衡量某个产物的优劣或强弱。需要注意的是，这项标准自然只适用于存在适应度概念的领域。不过，在生成游戏关卡的领域中，这一概念极为常见，因为程序化内容生成系统很容易生成无法通关或只能通关到某个阶段的游戏关卡。游戏关卡程序化内容生成中的适应度通常被视为较为二元的概念 [8, 14],，我们认为，无论领域中的适应度概念是连续的还是二元的，这项适应度独立性选择标准都会有所助益。无论如何，要使 ERA 中的探索具有价值，就需要有分布良好的可接受内容。

我们认为，对于 ERA 可视化而言，具备这一标准能够使其尽可能提供丰富的信息，这一点是理所当然的。有了这一标准，替代方案

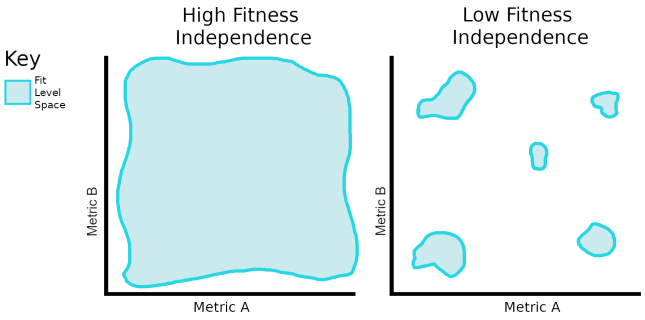


图1：以单一替代指标说明 ERA空间中高适应度独立性与低适应度独立性的存在

就可以直接比较其对生成空间的探索情况；或者，设计师可以调整某个生成器，使其生成空间位于空间中设计师希望的位置，同时确信空间中的每个位置包含适应内容的可能性都相同。相比之下，在适应度独立性较低的可视化中，探索可能并不可取，而且设计师也不会觉得自己可以随心所欲地针对空间中的任意位置进行优化，因为所有适应度较高的关卡都会紧密聚集在一起。对于确信这两个指标能够体现自己最关注的多样性的设计师而言，适应度独立性较低的 ERA图仍可能有用，但他们仍需更加留意适应内容在空间中的位置。适应度独立性较高则能减少对适应内容位置的担忧，让研究人员和设计师更专注于可视化本身。

本文所探讨的计算适应度独立性的方法，是将ERA图离散化为大小均匀的网格，其中网格中的每个单元由一定范围的指标值定义。随后，每个待可视化的关卡被归入该网格。然后，我们可以根据放置在其中的关卡计算每个网格单元的平均适应度，并进一步计算整个网格的平均单元适应度。因此，由某一指标对所定义的网格若取值较低，则表明适应度较高的关卡高度集中在网格的特定单元中，网格的更多部分由空单元或仅含低适应度得分的单元所占据；而较高的取值则表明分布更为均匀。

4.2 标准 2：互相关

指标对的第二项选择标准是互相关。实际上，这意味着我们希望一项指标的值与其指标对中另一项指标的值之间不存在或仅存在极小的预测关系（该标准的示例请参见图2）。这一标准之所以重要有两个原因。首先，要确保尽可能多的图确实可以被访问，换言之，尽可能多的图位于该领域底层的可能性空间之内。其次，要尽量减少ERA图中的冗余信息。如果一项指标的值能够高度预测另一项指标的值，那么在保持相同的

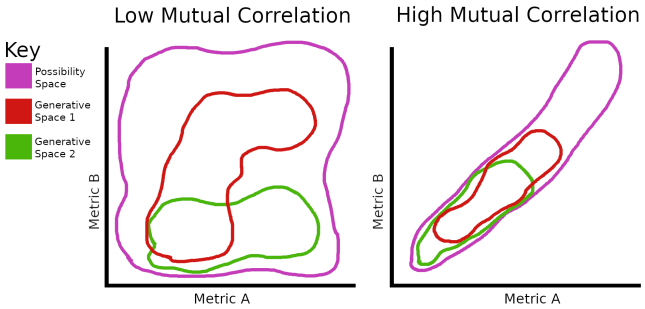


图2：低互相关与高互相关对可能性空间在ERA图中位置的影响，以及对生成空间位置的影响的示意图

关于生成器关卡分布的信息量。在另一种理想情况下，所选一对中的指标值之间的相关性最小，此时我们实际上是在最大化图表中获得的信息，同时也确保理论上可以访问图表中的更多部分。

符合此标准会使使用该指标对绘制的 ERA图更有意义。满足该标准后，我们便可以确信，ERA图中呈现的低多样性意味着生成器探索的可能性空间较小，而不是可能性空间本身的投影受到严重限制。这也更有可能使设计师能够区分其他生成空间的属性。指标之间的互相关越强，投影后的生成空间就越可能在表示受限可能性空间的图中区域发生重叠。正如我们将在下一节讨论第三项也是最后一项标准时所述，指标对之间较低的互相关也会提高其中一个指标与给定其他指标相关的可能性；我们很快将介绍这一特性为何值得追求。

为了计算我们在试点研究中各指标的相互独立性，我们使用斯皮尔曼等级相关系数计算每个指标所得值之间的线性相关，通常称为斯皮尔曼的 ρ。两个指标之间的相关性越强，一个指标的值对另一个指标的决定作用就越大，完整的ERA图也就越缺乏可探索性。我们首先存储斯皮尔曼的 ρ的绝对值，因为无论相关性是强正相关还是强负相关，我们都认为它同样不可取。然后我们用1减去该值，得到一个介于0和1之间的最终得分，其中1表示不存在相关性，使其与另外两个指标保持一致：低值不如高值理想。使用斯皮尔曼的 ρ 的主要缺点是，它无法帮助我们发现指标之间的非线性相关；如果存在这种相关性，仍会损害图表的可探索性，我们将在第7节中进一步讨论这一点。然而，它的优势在于，与用于研究非线性相关的方法相比，计算起来要简单得多，所需的计算量也小得多。

4.3 标准 3：替代指标相关性

我们在本研究中考察的第三个选择标准最难以概念化和计算，但也可以说是最重要的。它指的是与替代度量的相关性关卡。实际而言，这意味着我们希望尽量减少这样一种情况：相较于用于创建该图的指标组合，还存在其他有价值但与图中指标不相关的指标（关于单个指标的高低替代度量相关性示例，见图3）。其根本目标是，使 ERA 图对潜在可能性空间的投影尽可能信息丰富，并且其中呈现出的多样性或缺乏多样性，能够以比仅由这两个指标所决定的方式更全面的方式，代表多样性或其缺失。

为了理解为什么这一标准很有价值，我们将使用一个简短的示例。如果我们假设只对一组名为 A、B 和 C 的生成关卡中的三个指标感兴趣，那么，如果我们使用 A和B 作为指标对来进行ERA，即使所有内容在指标 C 上的取值都高度同质化，生成空间仍有可能在图中呈现出广泛分布。然而，如果发现指标 A 或 B 中任一个与指标 C 之间存在强相关性，那么我们就可以确信不会出现这种风险，并且 A 和 B 中呈现出的多样性也代表了 C 中的多样性。这正是我们希望通过这一选择标准评估的内容，不过我们的目标是找出所选指标对与其他所有感兴趣的指标之间的最低独立性，而不是与某个特定的其他指标之间的独立性。该标准的取值越高，我们就越能确信ERA图中的多样性代表了多个指标上的多样性，因此，相较于标准取值较低时，它代表的是更广义上的多样性。

为了使这一标准可操作化，我们采用与相互独立性标准类似的方法，不同之处在于，此时斯皮尔曼 ρ 的高相关性值是可取的。对于候选指标对M1和M2，我们分别计算每个其他指标与M1和M2之间相关性的 ρ，从而为每个其他指标给出两个值。我们只存储在M1或M2中找到的最高绝对值，因为该指标对中哪个指标具有相关性并不重要，相关性为正还是为负也不重要。这正是相互相关性标准支持这一标准的原因，因为拥有两个独立的指标会提高在此处找到一个对替代度量表现良好的指标的可能性。然后，可将每个其他指标的这组 ρ 值除以替代度量的数量，从而得到平均相关性得分。该得分越高，ERA图对其他形式的多样性的包容性就越强。

5 实验设计

在本节中，我们将介绍如何利用这一方法设计用于选择 ERA 指标对的初步实验，并解释关键设计决策背后的考量。这些实验旨在展示本研究所提出的新颖的选择标准的潜在实用价值，同时突出我们当前计算这些标准的方法的优势与不足。

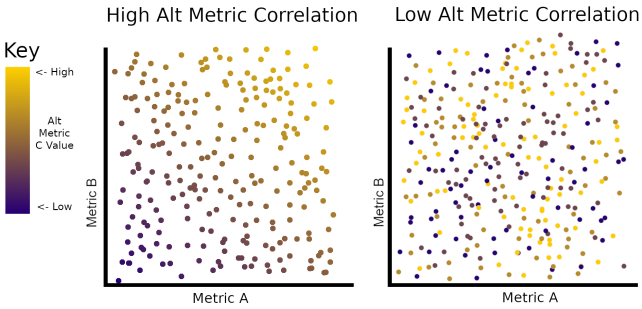


图 3：以单个替代指标为例，说明高、低替代指标相关性的影响，以及这对替代指标值在 ERA图中的分布可能产生的影响

5.1 评估的游戏领域

所分析的程序化内容生成空间全部来自 Horn等人的语料库中的生成关卡 [12]，我们在第3 节中对此进行了更详细的讨论。这个包含来自 9 个不同生成器的 9000 个生成关卡以及原版超级马力欧兄弟中的 15 个关卡的集合，已包含在当前版本的马里奥 AI 框架中。

在计算我们的选择标准并评估哪一组指标对最适合用于执行 ERA 时，我们聚焦于完整汇总的 9015 个关卡，因为我们认为，与集合中的任何单个生成器相比，这能更具代表性地呈现《马里奥》关卡的底层可能性空间。我们将在第 7 节中更详细地讨论并批判这一决定：虽然这种做法能更细致地呈现可能性空间，但实现起来也不太现实。在实践中，游戏设计师和研究人员很少能接触到某个游戏领域的多个不同生成器，以获得这类视角。

5.2 评估的指标

对于数据集中 9015 个《马里奥》关卡中的每一个，我们都计算了 18 项指标。指标的完整列表及其计算方法见表格 1。这 18 项指标及其计算方法主要借鉴了先前在 ERA 或质量多样性搜索中使用这些指标或类似变体的研究。与方块数量或重复模式（例如智能体可以站立的位置）相关的指标十分常见；在涉及模拟游玩以进行制品评估时，也常使用智能体动作的计数或比例 [11, 23]。我们提出的最大跳跃滞空时间和 Y值平均值指标则较为新颖。最大跳跃滞空时间记录智能体完成的最大跳跃距离，因为大幅跳跃往往会成为关卡中令人印象深刻的时刻，或造成难度陡增。Y值平均值更多地反映关卡的结构，通过记录智能体运行期间的平均高度，估算可站立地形的平均高度。由于本研究重点在于评估指标对，因此指标本身的设计并非研究重点。不过，我们希望结构性指标与智能体评估指标的数量相当：前者依据关卡结构本身计算，后者则依据 AI 智能体的游玩轨迹计算，以便比较这两类指标在高性能指标对中的表现。

我们还从智能体模拟中提取了智能体所能取得的水平进度占关卡总宽度的比例。在先前使用 Mario AI Benchmark 的研究中，这一特征通常称为“可玩性”或“适应度”[3, 7], ; 不过，我们采用“可玩性”这一说法，因为它更为具体。可玩性以 0 到 1 的得分表示，其中 1 表示智能体能够到达关卡终点。该指标未被纳入候选指标对，因为它被保留用于计算 FI 选择标准中的适应度启发式指标。

5.3 指标选择标准

主要结果计算了第 4节中解释的三个选择标准，即适应度独立性（FI）、互相关（MC）和替代度量相关性（AMC）。每个标准的得分介于 0 和 1 之间，得分越高越理想。这三个标准分别针对集合中的 153 个唯一指标对进行计算。

此外，我们还根据三个标准分别计算每个指标对的排名，然后可以利用这些排名确定每个标准下最佳的n 个指标对。我们还计算每个指标对在三个标准下的平均排名。这为评估指标对的整体表现提供了一个启发式指标，其中数值越低，表示其在三个指标上的总体表现越好。

5.4 智能体模拟

我们用于计算智能体评估指标以及关卡可玩性的智能体，是我们在第 3 节中介绍的 Robin Baumgarten 的 A*《马里奥》智能体。迄今为止，这一智能体已被广泛用于评估《马里奥》关卡的质量与特征，不过它也存在一些局限性，例如无法回溯。

由于智能体决策可能存在差异，我们在每个关卡上运行智能体五次，然后针对所有智能体提取特征，计算这五次运行的平均结果。

5.5 热图生成和分布分析

为了计算适应度独立性选择标准，以及部分结果可视化，需要将表现范围图划分为网格，以便计算每个网格位置中关卡的平均适应度。我们的方法是将每个指标对ERA图划分为20×20的网格，每个单元格的取值范围通过从该指标的最大值减去所找到的最小值，再除以20来确定。这种方法确保所有关卡都能被放置在图中，且无需选择特定于指标的范围。然而，这也意味着图内的分布会受到离群指标值存在的严重影响，我们将在第7节中进一步讨论这一点。

5.6 Data andVisualisationsPresented

在以下部分中，我们将展示多个结果表格和可视化，以说明这些实验中重要且有趣的发现。首先，我们展示 ERA可视化的】【。

表格 1：所研究的指标及其计算方式

Type	Name	如何计算
结构性	连续性	+1 每个与实心方块相邻的实心方块
	线性度	每个与实心方块水平相邻的实心方块+1
	区块计数	+1 每个实心方块
	敌人数量	+1 每个敌人
	奖励数量	+1 每个包含奖励的瓦片
	空瓦片数量	+1 每个空瓦片
	管道数量	+1 每个含有管道方块的瓦片
	密度	马里奥可站立的位置总数除以关卡长度
	清除列数	+1 每个仅含空瓦片的列
智能体提取	可玩性	智能体能够完成的关卡比例
	跳跃次数	+1 智能体使用的每次跳跃动作
	跳跃熵	跳跃次数除以总计操作次数
	速度	平均每秒导航的关卡比例
	耗时	智能体运行所用的时间（秒）
	敌人死亡总数	+1 每个在运行过程中死亡的敌人
	踩踏击杀数	+1 智能体通过跳到敌人身上击杀的每个敌人
	最大跳跃滞空时间	智能体处于未着地状态时，按帧数计算的最长跳跃
	着地时间占比	智能体停留在地面上的时间比例
	Y值平均值	运行期间马里奥的Y轴位置平均值

根据三个选择标准的平均排名，排名最高和最低的指标配对。这里既展示了可玩性热图，以显示可玩关卡在空间中的分布，也展示了完整关卡语料库的ERA可视化，其中不同生成器以不同颜色突出显示。这些内容旨在突出使用我们的选择标准时，强指标配对和弱指标配对的一些特征。

其次，我们展示了每种选择标准下按其数值排名的前5名指标配对，以及按其余三个排名的平均排名排序的前5名配对。这些内容主要用于展示应用选择标准后所得最终数据的样貌，并突出高性能指标对在三项标准下所获排名中的模式。这组20个配对也被用作我们的高性能指标对集合，用于最后两幅图，以说明至少在一项标准中获得高排名的指标配对中发现的模式。

最后，我们既展示了顶级高性能配对列表中每个单独指标的出现频率直方图，也展示了一张表格，用以枚举该列表中配对的计数：这些配对要么两者同为智能体所选指标，要么两者同为结构指标，或者由各一种组成。呈现这些内容是为了突出高性能配对构成中的有价值趋势，既为未来工作提供方向，也指出所呈现实现中的弱点。

5.7 硬件 & 软件使用

为了提取并保存实验中所包含的每个《马里奥》关卡的指标，使用了自定义 Java 代码来与 Mario AI Benchmark 交互。代码可在 github.com/ KrellFace/mario 指标提取 获取。选择_ _标准本身的计算以及可视化的生成使用 Python 完成，代码可在 github.com/KrellFace/ERA_ Metric Evaluation._ 获取。

所有软件均在一台配备i5-10310U 处理器和 16.0 GB 随机存取存储器的戴尔笔记本电脑上运行。从 Mario AI Benchmark中进行指标提取大约耗时 12 小时，而选择标准的计算大约耗时 1 小时 30 分钟。

6 结果与讨论

在本节中，我们将讨论所提出的方法，以及从初步实验中获得的相关见解。总体而言，我们发现有很多理由对所提出的方法充满热情。计算指标对的分节标准的过程简单且相对快速，而使用该方法所能获得的见解既可能有用，也可能出人意料。我们相信，相比于在缺乏信息的情况下做出选择，采用这些标准很可能会生成信息量更丰富、更公平的 ERA 图。为了了解这一点，我们可以查看图4和图5中按平均排名衡量表现最佳和最差的指标对所对应的可视化结果。在图4中，表现最佳的指标对不仅更容易区分不同生成器的生成空间，而且得益于 FI 和 AMC 选择标准，我们也能更有把握地认为，看起来是多样性的表现既在适应度方面有用，也体现了多个替代度量之间的多样性。

我们的试验性结果中一个有趣的模式是，在三个标准下，指标对的表现之间存在关联，或缺乏关联。正如我们在表 2 中所见，在一个选择标准中排名靠前的指标对，在其他选择标准中的表现往往非常差。我们预期 FI 标准和 MC 标准在一定程度上会具有关联性，因为二者都优先考虑空间中关卡的广泛分布。我们还预期 MC 标准和 AMC 标准会具有关联性，因为两个指标越不相关，就越

表 2：每种选择标准的前5名指标对，以及所有三项标准的平均排名。所有数值均四舍五入至3位有效数字

		FI		MC		AMC		
	指标对	**得分**	**Rank**	**得分**	**Rank**	**得分**	**Rank**	**平均排名**
前5名 FI Rank	密度-跳跃次数	0.517	1	0.885	30	0.518	77	36
	敌人数量-跳跃次数	0.514	2	0.494	101	0.521	74	59
	密度-平均Y值	0.496	3	0.985	3	0.551	39	15
	跳跃次数-敌人死亡总数	0.479	4	0.345	130	0.541	47	60.3
	PipeCount-TotalEnemyDeaths	0.462	5	0.406	122	0.5831	15	47.3
前5名 MC Rank	空白空间计数-最大跳跃滞空时间	0.0564	153	0.997	1	0.583	17	57
	速度-最大跳跃滞空时间	0.124	132	0.985	2	0.4521	128	87.3
	密度-平均Y值	0.496	3	0.985	3	0.551	39	15
	密度-清除列数	0.296	67	0.983	4	0.286	148	73
	踩踏击杀数-平均Y值	0.321	53	0.983	5	0.536	54	37.3
前5名 AMC Rank	空白空间数量-跳跃次数	0.122	133	0.465	106	0.610	1	80
	方块数量-敌人总死亡数	0.336	44	0.621	83	0.609	2	43
	方块数量-平均Y值	0.321	52	0.372	125	0.606	3	60
	连续性-敌人死亡总数	0.334	45	0.737	54	0.606	4	34.3
	方块数量-跳跃次数	0.359	30	0.522	99	0.604	5	44.7
前5名 平均值 Rank	密度-平均Y值	0.496	3	0.985	3	0.551	39	15
	PipeCount-Density	0.365	26	0.947	20	0.574	22	22.7
	密度-跳跃熵	0.456	6	0.896	29	0.530	60	31.7
	清空列-平均Y值	0.396	17	0.966	12	0.522	71	33.3
	连续性-敌人总死亡数	0.334	45	0.737	54	0.606	4	34.3

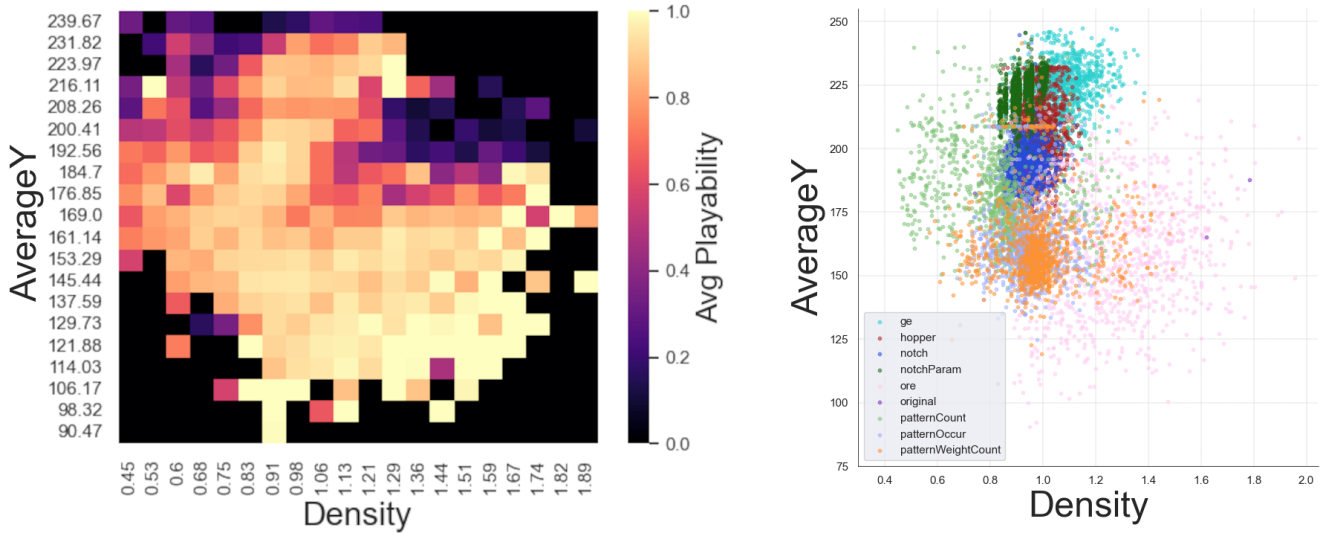


图4：就平均排名而言，排名最高的指标对"密度&Y值平均值"的热力图。展示了适应度热力图，以及使用这些指标对关卡语料库进行投影时生成的ERA图，其中每个生成器以不同颜色突出显示

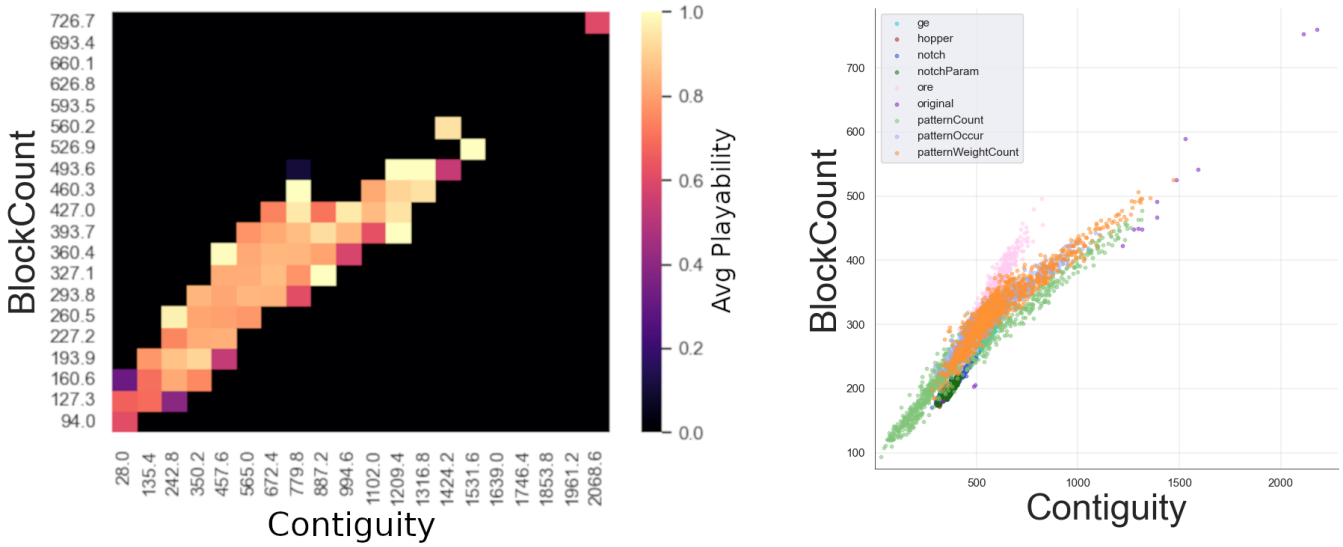


图5：排名最低的指标对“连续性& 区块计数”的热力图，以平均排名表示。图中同时展示了适应度热力图，以及使用这些指标投影关卡语料库时生成的ERA图，其中每个生成器以不同颜色突出显示

表3：对于每项标准及平均排名，在排名最高的5个指标对中，结构性特征与智能体提取特征的替代组合频率

	计数				
指标对类别	**平均排名**	**FI**	**MC**	**AMC**	**总计**
结构-结构	1	0	1	0	**2**
结构-智能体	4	4	2	5	**15**
智能体-智能体	0	1	2	0	**3**

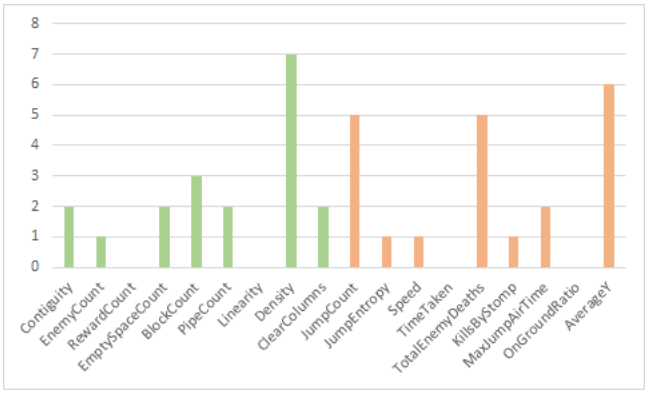


图6：每种选择标准下各指标在前5对中的出现频率及平均排名。结构指标为绿色，智能体提取为橙色。

其中一个与给定的其他指标具有强相关性的可能性。这些相关效应在这一领域似乎并不够强，无法确保指标对的存在。

它们在所有标准上均表现出色。例如，在平均排名方面位列第二的指标对 PipeCount- Density，在任何单项标准中的排名都未能进入前 20。不过，其他未经测试的指标或经过改进的已测试指标，仍有可能在三项标准上都取得良好表现。

如图6所示，我们所研究的各项指标在其出现在排名靠前的指标对中的频率方面分布广泛。我们针对每项标准考察了排名前5的指标对以及平均排名前5的指标对，总计得到20个高性能指标对和40个构成指标。处于高端的是密度指标，它出现在20个可能的指标对中的7个。在低端，有4个指标一次都未出现。尽管这一样本规模过小且具有特定领域性，无法据此作出任何普遍性结论，但这种偏斜程度颇为有趣，也表明我们作为研究人员，在选择所使用的指标时，应更加谨慎，避免未经充分考虑就作出选择。

对个别指标表现不佳的根本原因进行彻底调查超出了本研究的范围。不过，有些原因显而易见，例如离群值对我们计算 FI 选择标准的方式造成了损害。极端离群值会使非离群关卡在热力图中集中分布，因为热力图的维度由指标值的总范围决定，这会损害其分布程度以及随后的得分。这一点在最大跳跃滞空时间指标中最为明显：在其中一个关卡中，马里奥能够达到 148 帧的数值，远超出其余数百个关卡的数值范围（0 到 62）。这一高值源于该关卡的独特设计——敌人密度极高，以至于 A*智能体最高效的路线是在关卡的大部分时间里在敌人身上弹跳，从而导致长时间

不在地面上度过。这直接导致任何使用该指标的指标对在 FI 分数上排名较低，如表格2所示：原本表现强劲的“空隙数量-最大跳跃滞空时间”指标对在 MC 选择标准中排名第一。尽管这一点很有趣，但也提醒我们，在未进一步考虑如何处理离群值之前，不应再次使用这种计算适应度独立性的方法。

一个有趣的发现是，高排名指标对中，结构性特征与智能体评估特征相结合的指标对非常普遍。如表3所示，得分最高的20个指标对中，有15个符合这种形式。从统计上看，结构-智能体指标对占优，这是因为我们没有测试由相同指标组成的指标对，而总计153个指标对中，结构-智能体指标对略占多数，共有81个。不过，这种偏斜程度超过了这一偏差所能解释的范围；而且从直觉上看也合情合理，因为游戏智能体将关卡视为一系列必须执行的动作并与之交互，这与将其作为图块矩阵进行结构性分析截然不同。我们认为，这意味着二者相互关联的可能性较低，因此更有可能在 AMC 标准上取得良好得分。这一结果表明，ERA 实现可以通过采用不同的方法计算各指标对中的每个指标，来提高这些指标符合我们选择标准的可能性。不过，我们应谨慎避免过度解读这一单一数据点，因为它很可能只是我们所选特定指标的偶然结果。

7 局限性与未来工作

尽管我们对所提出的概念和高层次方法感到振奋，但它仍存在若干不足和局限，有待进一步改进。此外，还有大量后续工作可以开展，既能拓展我们提出的思路，也能推动对程序化内容生成系统生成空间的理解与可视化这一研究领域整体向前发展。

未来工作的一项重要方向，是探索计算本文所提出选择标准的其他方法。使用斯皮尔曼秩相关系数计算 MC 和 AMC 标准的秩相关既快速又高效，但其局限在于只能检测指标之间的单调关系，无法检测更复杂的非线性关系。未来工作可以进一步探索更复杂但计算量更大的替代方法，例如使用平滑样条。计算 FI 指标也有其他方法。一种简单的替代方法是计算网格单元中发现的最大关卡适应度，而非平均值；这种做法可能更加公平、有效，因为该标准关注的是能否在 ERA 图中广泛找到关卡，而不是其中是否已有很大比例被发现。

这一领域同样可以从进一步的探索与反思中获益良多，以思考哪些额外的选择标准对指标对具有价值。其中一项可能有价值、但超出本文范围的是涌现，即指标的取值并不由生成器的参数化方式决定。Smith 等人在首次引入 ERA 的论文中主张这一点很重要，我们认为它对于使 ERA 更具信息量、并使生成空间之间的比较更加公平仍然有价值。使这一标准可操作化将会很复杂，因为需要由同一生成器的多种参数化方式生成的多组关卡，但未来工作可以探索一种类似于 Cook 等人针对 Danesh 平台开发的 RERA 的方法 [5]。

我们的方法以及更广泛意义上的ERA都存在一个概念上的弱点，即无法知晓其底层的可能性空间。正如我们在第3.2节中所讨论的，调优程序化内容生成系统的过程，可以被视为试图在更广泛的可能内容可能性空间内，优化该系统生成空间的规模与位置。当使用 ERA来量化生成器的搜索空间探索时，研究人员想知道的是底层可能性空间中有多少部分正在被探索。然而，此前几乎没有工作旨在理解或量化那些包含生成空间的可能性空间。未来，探索如何获得与程序化内容生成系统无关的底层可能性空间视角，可能不仅会对稳健的ERA大有裨益，也会对总体上的程序化内容生成研究大有裨益。基于质量-多样性的程序化内容生成方法 [9] 为此提供了一条可能的途径，因为它们针对最大化探索进行了良好优化。由于它们也会使用指标以与 ERA 完全相同的方式创建搜索空间的投影，因此需要谨慎考虑应选择哪些指标。本工作的选择标准或许可以用于选择指标，以增加质量-多样性搜索的可能性空间探索，这也是我们计划在未来研究的方向。

我们所提出方法的一个实际缺点，单纯在于确定最佳指标对所需的时间。虽然计算时间相对较短，但显然远远长于开展 ERA 本身所需的时间。这使得该方法在实际ERA 应用中的吸引力大幅降低；随着候选指标数量的增加，其吸引力还会进一步下降。不过，增加候选指标的数量也可以说会通过帮助设计师对其进行筛选而提高我们的选择标准的价值。在未来的工作中，我们希望探究是否能够找到在多个相似类型的游戏领域中都表现优异的指标。这可能会与前文讨论的、探究指标与玩家感知之间关系的论文形成良好协同 [11, 17, 23]。根据我们的标准，在多个领域中得分较高的指标对，如果同时与人类感知中的质量或多样性相关，那么它们有望在无需评估的情况下，在新颖的程序化内容生成领域中发挥稳健的效用。

8 结论

本文介绍了我们通过采用指标选择标准来改进表达范围分析（ERA）实现的方法，使设计师能够更有依据地选择最适合其领域的一对指标。我们还介绍了这些标准的实现，以及将其应用于现有生成关卡语料库后获得的结果与见解。我们发现，采用这些标准有助于深入了解如何为某个领域制作最佳的表达范围分析可视化，以及如何设计指标来开展表达范围分析。我们希望这项工作能为未来探索新方法奠定基础，以便在程序化内容生成中选择最适合表达性可视化的指标对。

参考文献

[1] Alberto Alvarez、Steve Dahlskog、Jose Font 和 Julian Togelius。2019年。通过交互式约束 MAP-Elites 增强地牢设计中的质量多样性。载于*IEEE 2019*年游戏大会，*CoG 2019*（*IEEE*计算智能与游戏大会，*CIG*）。IEEE计算机学会。https: //doi.org/ 10 1109 2019 8848022 . /CIG. 。[2] Alberto Alvarez、Steve Dahlskog、Jose Font 和 Julian Togelius。2022年。交互式约束 MAP-Elites：特征维度表达能力的分析与评估。*IEEE*游戏汇刊14卷，第2期（2022年6月），202–211。https://doi.org/ 10 1109 2020 3046133 . /TG. 。[3] Michael Cerny Green、Luvneesh Mugrai、Ahmed Khalifa 和 Julian Togelius。2020年。运用场景拼接，根据游戏机制生成马里奥关卡。载于*IEEE 2020*年游戏大会（*CoG*）。IEEE，日本大阪，49–56。https://doi.org/10. 1109/CoG47356.2020.9231692[4] Michael Cook。2022年。心怀乐观：我们为何研究游戏人工智能？载于*IEEE 2022*年游戏大会（*CoG*）。IEEE，中国北京，560–567。https: //doi.org/ 10 1109 51982 2022 9893607 . /CoG 。[5] Michael Cook、Jeremy Gow、Gillian Smith 和 Simon Colton。2021年。Danesh：用于理解程序化内容生成器的交互式工具。*IEEE*游戏汇刊（ 2021 9426419 ）。[6] https://ieeexplore.ieee.org/document/ Hello Games（公司）和索尼电脑娱乐。2016。《无人深空》。[7] Matthew C. Fontaine、Ruilin Liu、Ahmed Khalifa、Jignesh Modi、Julian Togelius、Amy K. Hoover 和 Stefanos Nikolaidis。2021年。在生成对抗网络的潜在空间中呈现丰富多样的马里奥场景。载于第*35*届*AAAI*人工智能大会，*AAAI 2021*（第*35*届*AAAI*人工智能大会，*AAAI 2021*）。人工智能促进协会，5922–5930。[8] Roberto Gallotta、Kai Arulkumaran 和 L. B. Soros。2022年。用于程序化内容生成的代理不可行适应度获取 FI-2Pop。载于*IEEE 2022*年游戏大会（*CoG*）。IEEE，中国北京，500–503。https://doi.org/10. 1109/CoG51982.2022.9893592[9] Daniele Gravina、Ahmed Khalifa、Antonios Liapis、Julian Togelius 和 Georgios N. Yannakakis。2019年。通过质量多样性进行程序化内容生成。载于*IEEE 2019*年游戏大会，*CoG 2019*（*IEEE*计算智能与游戏大会，*CIG*）。IEEE计算机学会。https://doi.org/10.1109/ CIG.2019.8848053[10] Matthew Guzdial、Devi Acharya、Max Kreminski、Michael Cook、Mirjam Eladhari、Antonios Liapis 和 Anne Sullivan。2020年。作为程序化内容生成器的桌面角色扮演游戏。载于国际数字游戏基础大会（*FDG '20*）。计算机协会，美国纽约州纽约市。https://doi.org/10.1145/3402942.3409605 举办地：马耳他布吉巴。[11] J ean-Baptiste Hervé 和 Christoph Salge。2021年。比较《我的世界》定居点生成中的程序化内容生成指标与人工评估。载于第*16*届国际数字游戏基础大会（*FDG*）*2021*（*FDG'21*）。计算机协会，美国纽约州纽约市。https://doi.org/10.1145/3472538. 3472590 举办地：加拿大魁北克省蒙特利尔。[12] Britton Horn、Steve Dahlskog、Noor Shaker、Gillian Smith 和 Julian Togelius。2014年。《马里奥 AI 框架》中程序化关卡生成器的比较评估。载于第九届国际数字游戏基础大会论文集，*FDG 2014*，海洋自由号，加勒比海，*2014*年*4*月*3*日至*7*日，Michael Mateas、Tiffany Barnes 和 Ian Bogost（编）。数字游戏科学促进学会。http://www.fdg2014.org/papers/fdg 2014_paper_14.pdf[13] Mrunal Jadhav 和 Matthew Guzdial。2021年。图块嵌入：一种用于关卡生成的通用表示。*AAAI*人工智能与互动数字娱乐大会论文集17卷，第1期（2021年10月），34–41。https://doi.org/ 10 1609 17 1 18888 . / aiide.v i 。[14] Ahmed Khalifa、Scott Lee、Andy Nealen 和 Julian Togelius。2018年。Talakat：通过约束 MAP-Elites 进行弹幕射击生成。载于遗传与进化计算大会论文集。ACM，日本京都，1047–1054。https: //doi.org/ 10 1145 3205455 3205470 . / 。[15] Max Kreminski、Isaac Karth、Michael Mateas 和 Noah Wardrip-Fruin。2022年。通过表现范围覆盖分析评估混合主动式创意界面。载于第三届生成模型人机共创研讨会，第3124卷。[16] Ben Kybartas、Clark Verbrugge 和 Jonathan Lessard。2018年。对一种由可能世界驱动的涌现叙事系统进行表达范围分析。载于互动叙事，Rebecca Rouse、Hartmut Koenitz 和 Mads Haahr（编）。第11318卷。施普林格国际出版，尚，473–477。https://doi.org/10.1007/978- 3-030-04028-4_54 丛书名：《计算机科学讲义》。[17] Julian Mariño、Willian Reis 和 Levi Lelis。2021年。对程序化生成的马里奥关卡评估指标的实证评估。*AAAI*人工智能与互动数字娱乐大会论文集 11, 1 2021 44 50 10 1609 11 1 12785 （6月），– 。[18] https://doi.org/ 。/aiide.v i 。Marcus Persson。。《我的世界》。。[19] Anurag Sarkar 和 Seth Cooper。2021年。通过变分自编码器潜在空间中的质量多样性生成与混合游戏关卡。载于第*16*届国际数字游戏基础大会（*FDG*）*2021 (FDG'21)*。计算机协会，美国纽约州纽约市。https: // doi.org/10.1145/3472538.3472545 活动地点：加拿大魁北克省蒙特利尔。[20] Noor Shaker、Miguel Nicolau、Georgios N. Yannakakis、Julian Togelius 和 Michael O'Neill。2012年。使用语法进化为超级马力欧兄弟演化关卡。载于*2012*年 *IEEE*计算智能与游戏会议（*CIG*）。IEEE，西班牙格拉纳达， 304 311 10 1109 2012 6374170 –。[21] https://doi.org/。/CIG.。Gillian Smith 和 Jim Whitehead。2010。分析关卡生成器的表达范围。载于*2010*年游戏程序化内容生成研讨会论文集 - *PCGames '10*。ACM出版社，加利福尼亚州蒙特雷，1–7。https: //doi.org/ 10 1145 1814256 1814260 。/。[22] Adam Summerville。。拓展表达范围：程序化内容生成的评估方法。载于第十四届人工智能与互动数字娱乐会议。https://www.aaai.org/ocs/index. php/AIIDE/AIIDE 18 18085 /paper/view Adam Summerville、Julian 2018[23] R. H. Mariño、Sam Snodgrass、Santiago Ontañón 和 Levi H. S. Lelis。2017年。理解马里奥：平台游戏设计指标评估。载于第十二届国际数字游戏基础大会论文集。ACM，马萨诸塞州海恩尼斯，1–10。https://doi.org/10.1145/ 3102071.3102080[24] Julian Togelius、Sergey Karakovskiy 和 Robin Baumgarten。2010年。2009年马力欧人工智能竞赛。载于*IEEE*进化计算大会。IEEE，西班牙巴塞罗那， 1 8 10 1109 2010 5586133 –。https://doi.org/。/[25] CEC.。Oliver Withington 和 Laurissa Tokarchuk。2022。压缩并比较程序化内容生成器的生成空间。载于*2022*年 *IEEE*游戏会议（*CoG*）。IEEE，中国北京，143–150。https://doi.org/10.1109/ CoG51982.2022.9893615[26] Oliver Withington 和 Laurissa Tokarchuk。2022年。使用卷积神经网络嵌入可视化生成空间。http://arxiv.org/abs/2210. 17464 arXiv:2210.17464 [cs]。[27] Yahia Zakaria、Magda Fayek 和 Mayada Hadhoud。2022年。通过深度学习进行推箱子关卡的程序化生成：一项实验研究。*IEEE*游戏汇刊（2022年），1–1。https://doi.org/10.1109/TG.2022.3175795