arXiv:2602.10996v1 [cs.MA] 11 Feb 2026

交际中的人工智能体数值表征的涌现

Daniela Mihai*[1]、Lucas Weber[2]和Francesca Franzon[2]

*通讯作者：a-d.mihai@soton.ac.uk[1]南安普顿大学电子与计算机科学学院，英国[2]庞培法布拉大学翻译与语言科学系，西班牙巴塞罗那

人类语言提供了表达数量的高效系统，但仅仅是交际压力是否足以促使人工智能体涌现出数值表征，以及涌现的编码是否与人类数字有任何相似之处，仍是一个悬而未决的问题。我们研究了两个基于神经网络的智能体，它们必须在指称游戏中通过离散标记或连续草图来交流数量，从而探索符号式和图像式表征。在没有任何预定义数概念的情况下，智能体在两种通信渠道中都实现了较高的分布内通信准确率，并形成了精确性较高的符号—意义映射。然而，涌现代码不具组合性：对于未见数量，智能体无法生成系统化的消息，通常会重复使用训练中最大数量对应的符号（离散情形），或将外推值归并到同一幅草图上（连续情形）。我们得出结论，仅凭交际压力就足以精确传递已学习的数量，但若要产生组合式代码和泛化能力，还需要额外的压力。

1. 数量的交流与表征

数词，例如 *two*、*ten*、*sixteen*，使人类能够借助有限记忆对无限大的数量进行编码。它们的编码效率体现在若干属性上。首先，它们具有**精确性**：每个词都与一个数量一一对应；通过外推，这种映射可以无限延伸，为每个新增元素指定一个新词。数词还遵循透明的**组合规则**，从而减少需要记忆的形式数量：例如，掌握个位数和十位数的词语后，说话者便能推断出中间的数量。这些属性使其具有可学习性，让人们能够轻易推断概念—词语映射，并将其**推广**到新的词语和新的数量：在人类中，理解数词的结构便能解锁对任何数值概念的精确表征（Barner，2017）；而有人认为，在没有数词的情况下，对数量的表征只能是近似的。最后，数词的形式具有**任意性**，因为词语的任何属性都无法提示其所指的数量；不过，与语言中的普遍情况一样，最常出现的数词通常更短（Pajot、Sabl´e-Meyer 和 Dehaene，2025）。然而，数字的书面表征在任意性方面存在差异，并且可能起初是图像式表征，

其中数字符号的视觉属性直接再现其所编码的数量（Overmann，2025）。

尽管数字系统在不同文化中独立产生（Coolidge & Overmann, 2012），不同语言中的数词在很大程度上具有这些特性（Comrie, 2013），这表明它们反映了对相似编码压力的优化（Xu, Liu, & Regier, 2020; Deni´c & Szymanik, 2024）。

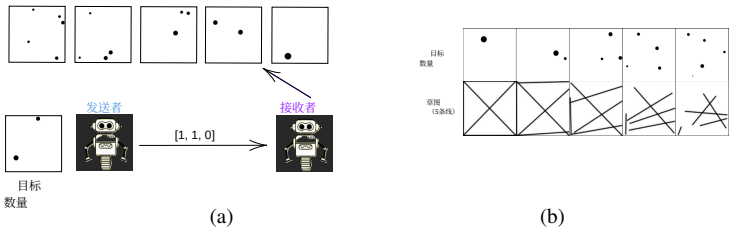
语言演化理论认为，结构化系统是在传递、经济性和泛化的压力下形成的（Kirby, Cornish, & Smith, 2008年）。在不同文化中，数字最初是为实现无歧义交际而产生的一种文化发明，用于记录储存的货物以及追踪交易中的数量（Menninger, 2013）。因此，我们探讨交际压力是否会塑造数词，并进而促成数概念的表征。

具体而言，我们探讨：（i）自然语言数词中观察到的特性是否也会出现在通过交际自我优化的编码中；（ii）交际的哪些方面促成了这些特性；以及（iii）由此形成的编码是否支持对未见数量的泛化，从而自举其概念表征。为此，我们实现了一种指称游戏（Lewis, 1969; Boldt & Mortensen, 2024; Brandizzi, 2023），其中通信协议完全通过两个神经网络智能体之间的协作互动形成，且它们不具备关于数词的先验知识。我们检验交际压力是否足以促使智能体发展出一种编码，以及这种编码是否呈现出精确性、组合性、泛化和任意性等涌现特性。我们还评估频率和学习策略（插值与外推）是否会影响通信成功率和编码结构。

2. 环境设置

通信任务。 我们采用指称游戏设置（图 1a），让两个神经网络智能体围绕数量进行交际。在每个试次中，发送方智能体会看到一张显示一定数量圆点的图片，并向接收方智能体发送一条消息。接收者根据消息，必须从数量不同的干扰项中找出显示相同圆点数量的目标 图片。选定目标后，两个智能体都会收到反馈；通过这种方式，智能体在完成任务的过程中形成通信协议。该设置包含一个训练 阶段，智能体在此阶段接收反馈并由此发展出自己的编码；以及一个泛化 阶段，智能体在此阶段围绕新数量进行交际。

刺激材料。 我们设计刺激材料，确保智能体只依赖数量，从而避免形状、颜色、圆点所占面积或周长等其他视觉属性造成的混淆因素。刺激图像的尺寸为 256 x 256 像素，每张图像都在白色背景上呈现数量不同的黑色圆点（例如图 1b）。为使数量信息独立于其他视觉变量（如黑色像素所占面积），我们将所有数量对应的圆点总黑色面积限制在画布面积的 5% 至 10%。



(a) (b)

图1。（a）指称游戏：接收者必须根据发送者发出的离散消息，从5种可能的数量中猜出正确的目标。请注意，虽然目标也是2，但它与发送者看到的并非同一个实例。（b）目标数量示例 $\{1, 2, 3, 4, 5\}$ 及其对应的5条线草图。

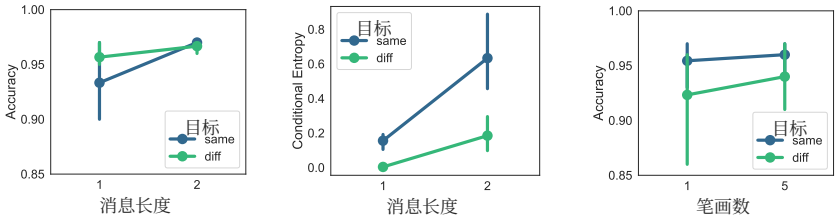
条件。 本研究的主要训练集大多仅包含 1 至 5 的数量（尽管我们也试验了更大的训练集，例如 1 至 20，并观察到相似结果）。外推／插值评估可包含另一种数量。每项实验的具体训练和测试数量均有说明。为检验数量表征能力是否独立于低层视觉属性而产生，我们测试了两种目标设置：相同和不同，分别表示智能体看到的是完全相同 的目标图像，还是来自同一数量类别的两个不同 实例。

智能体。 两个智能体均使用共享的预训练ViT编码器处理点图像（Radford等人，2021）。我们对两种类型的通信渠道进行建模。在**离散** 设置中，发送者配备了一个LSTM编码器（Hochreiter与Schmidhuber，1997），该编码器会输出一条消息，由从预定义词汇表中选出的离散标记序列组成；接收者的LSTM读取该消息，并据此选择一幅图像作为可能的目标。在**连续** 设置中，发送者在空白画布上绘制固定数量的直线，即一幅草图（Mi-hai与Hare，2021）。接收者通过与候选目标图像相同的ViT编码器读取草图。我们采用多类合页损失，以最大化选中正确目标图像的概率。训练使用每项实验指定的一组数量，测试则同时涵盖范围内数值和未见过的数值。

指标。 我们通过任务准确率评估通信成功率，其定义为接收者选出发送者所呈现数量的试次比例。此外，我们还评估逐渐形成的通信编码的属性：效率通过消息长度（词元）或草图笔画数/跨度来衡量；精确性以给定消息时数量的条件熵来衡量（草图通过聚类进行离散化）；泛化 通过未见数量上的准确率，以及这些数量对应的消息映射到哪些已见类别来衡量；任意性 则通过调整训练频率分布并检查更常见的数字是否获得更短或更简单的编码来衡量。

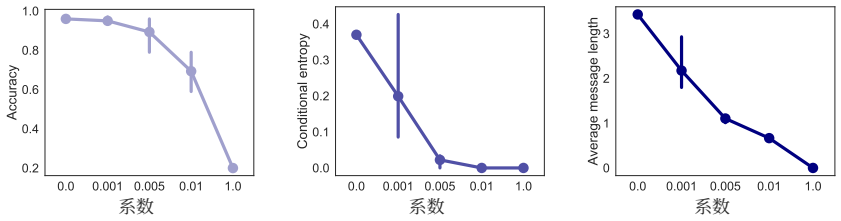
3. 结果

通信成功率与编码精度 首先，我们发现智能体能够可靠地交流训练过的数量，即使在不同 条件下也是如此；在该条件中，发送者和接收者看到的是同一数量的不同实例（图 2a 和 2c）。在不同设置下，系统的精确性有所提升，表现为熵降低（图 2b 和 4a）。在长度正则化条件下（图 3），我们观察到消息变得更短，同时准确率损失极小且熵更低，这表明系统受到压力，倾向于形成高效的双射编码。这说明，即使没有语言内部的约定，协调与信息瓶颈的共同作用也会促使精确性涌现。



(a) (b) (c)

图 2. (a) 离散通信准确率和 (b) 消息熵，以及 (c) 使用相同目标图片（same - 蓝线）和展示相同数量的不同图片（diff - 绿线）时的连续（即基于草图的）通信准确率。模型在数量 1 到 5 上训练和测试。结果取 3 个随机种子的平均值。

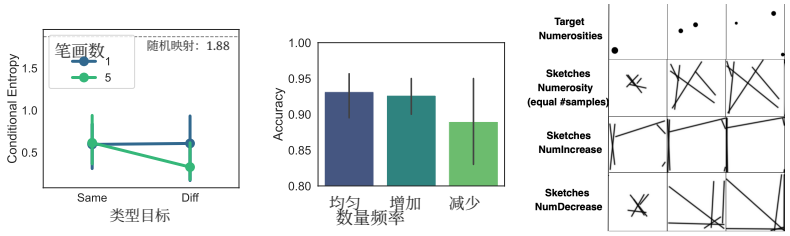


(a) (b) (c)

图 3. 离散通信——当通过对博弈损失施加正则化来惩罚较长消息时，智能体的消息行为。模型在数量 $\{1, 2, 3, 4, 5\}$ 上训练，允许可变消息长度，最长为 **5** 个词元，并且 $vocabulary = 3$。(a) 准确率，(b) 映射的条件熵，以及 (c) 不同正则化系数下的平均消息长度。

任意性胜过频率接下来，我们改变训练上下文，以评估与出现频率最低的数量相比，出现频率较高的数量是否会以更经济的方式编码。我们发现，调整类别频率（补充材料中详述的 Uniform/Increase/Decrease 设置）几乎不会改变准确率（图4b），也不会让高频数量采用更短的编码。

草图通道往往随输入图像元素的空间范围而变化，而非随频率变化（图4c）。人类数字常兼具任意性与一定程度的高频词汇化。在这一系列实验中，除准确率之外不存在传输成本，这可能削弱了类似齐普夫定律的简化倾向；而任意性得以保留，是因为它对协调来说已经足够好。



(a) (b) (c)

图4。（a）使用1条和5条线的草图进行交际时的条件熵，目标上下文有两种（两个智能体之间相同或不同）。（b）在两个经过调整的数量数据集上的通信准确率，这两个数据集的训练样本数量不相等：样本数从类别1递增至类别5（类别1样本最少，类别5最多），反之亦然。（c）分别在3种不同数据集（类别样本数量相等、递增和递减）上训练后得到的测试草图。

泛化与组合性 我们进一步考察智能体开发的编码能否泛化到未见数量（离散设置见图5上方）。我们发现，在分布外测试中，智能体通常会重复使用表示最大训练数量的消息或符号。对范围内数量进行测试的插值分析也表明，缺乏系统性结构。基于草图的交际表现相似，不过，平均而言，使用连续通道进行外推的表现优于随机猜测。然而，智能体并未发展出组合式草图，而是将所有大于训练中所见数量的情况都映射到同一张草图，这张草图通常与所见的最大数量相似（例如图5下方及补充材料）。涌现出的信号会稳定为整体标签：每个训练过的数量都对应一个独立的消息或草图，且不会重用子部分来组合新消息。

4. 对语言演化的启示

协调带来精确性和数值表征。 智能体能够快速习得低熵编码，精确地传达经过训练的数量。数量的传达不依赖于视觉特征，这表明智能体习得了抽象数值表征。

象似性有所帮助，但任意性占据上风。草图通信通道展现出一些大小和跨度线索，其中草图大小反映的是点的分布范围，而非数量；但离散和连续通信通道最终都形成了以任意的类别特定信号为主的形式。

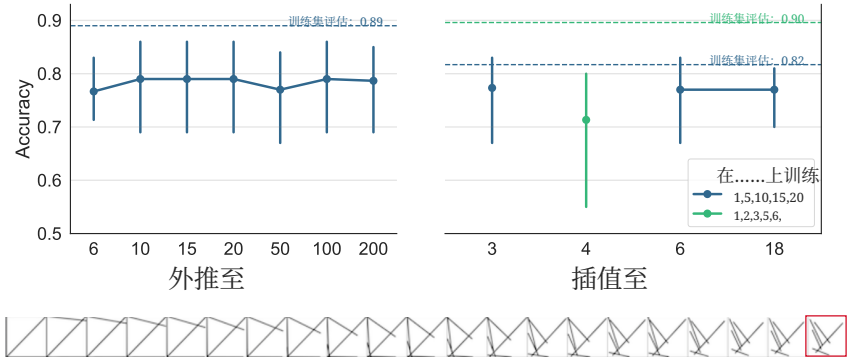


图5。（左）**离散外推**：模型在 $\{1, 2, 3, 4, 5\}$上训练。对于每个测试集，报告交际准确率（训练集为 + 个分布外数量）。对于每一次新的更高数量外推，模型都会使用此前为最高预训练数量（例如5）选择的相同消息。（右）来自两个不同训练集的**插值**结果。（下）从 $\{1 - 20\}$到*25* 的训练中进行草图外推的示例，以红色突出显示。

仅凭频率并不能产生高效编码。人类语言通常会让高频词更短，但在这里，由于没有为更长的消息设置明确成本，偏斜的输入分布并不会产生更高效的编码。加入产生和感知成本可能有助于涌现编码的简洁性。

缺乏组合性，缺乏泛化。面对更大的、未经训练的数量时，智能体会复用与最大训练数量相关联的消息。这一局限可能与涌现出的交际系统缺乏组合性有关：智能体使用整体式消息，既不复用可解释的消息部分，也不发展基于规则的组合。尽管在指称游戏中，编码不可分解特征组合的整体式消息经常涌现（LaCroix，2025；Freeborn，2025），但人类学习的证据表明，组合性在自举数值表征方面发挥着关键作用。在这一情境中诱导组合性仍是一项重要挑战。与由多个特征定义的指称对象（例如颜色和形状）不同，数字沿单一有序维度变化，没有天然可分解的特征可供编码。因此，要实现成功的泛化，就需要涌现出对该维度进行任意但系统性划分的方式，而非对感知特征进行透明映射。如何诱导形成这种结构仍是一个开放问题。其他学习设置或许有所帮助，例如群体训练（如 Mahaut 等人，2025年），或其他能够促进人类交际和神经网络组合性的机制，例如迭代学习、词汇约束，以及对符号复用的偏好（Lake 和 Baroni，2023年；Kirby 等人，2008年）。因此，探索此类机制或许是实现系统性数值泛化的关键。

致谢

FF 获得了欧洲研究委员会（ERC）通过欧盟地平线2020研究与创新计划（资助协议编号：101019291）提供的资助，以及加泰罗尼亚政府提供的资助（AGAUR 资助项目 SGR 2021 00470）。

参考文献

Barner, D. (2017). 语言、程序与数词意义的非感知起源。《儿童语言杂志》, *44*(3), 553–590. Boldt, B., & Mortensen, D. R. (2024). 基于深度学习的涌现式交流应用综述。《机器学习研究汇刊》. Brandizzi, N. (2023). 迈向更像人类的人工智能交流：涌现式交流研究综述。《*IEEE*开放获取》,*11*, 142317–142340. Comrie, B. (2013). 数词基数 (v2020.3) [数据集]. 收录于 M. S. Dryer & M. Haspelmath (编), 《世界语言结构地图集在线版》。Zenodo. Coolidge, F. L., & Overmann, K. A. (2012). 数量、抽象与符号思维的出现。《当代人类学》,*53*(2), 204–225. Deni´c, M., & Szymanik, J. (2024). 递归数词系统优化了词汇表规模与平均形态句法复杂度之间的权衡。认知科学,*48*(3), e13424. Freeborn, D. P. W. (2025). 信号博弈中的组合性理解。《综合》,*206*(3), 116. Hochreiter, S., & Schmidhuber, J. (1997). 长短期记忆。《神经计算》,*9*(8), 1735–1780. Kirby, S., Cornish, H., & Smith, K. (2008). 实验室中的累积文化演化：探究人类语言结构起源的实验方法。《美国国家科学院院刊》, *105*(31), 10681–10686. LaCroix, T. (2025). 组合信号演化中的信息与意义：T. lacroix。《逻辑、语言与信息杂志》, 1–21. Lake, B. M., & Baroni, M. (2023). 通过元学习神经网络实现类似人类的系统性泛化。《自然》, *623*(7985), 115–121. Lewis, D. K. (1969). 《惯例：一项哲学研究》。Mahaut, M., Dessi, R., Franzon, F., & Baroni, M. (2025). 预训练视觉深度网络异质群体中的指称性交际。《机器学习研究汇刊》. Menninger, K. (2013). 《数词与数字符号：数字的文化史》。Courier Corporation. Mihai, D., & Hare, J. (2021). 学习绘画：涌现式交流

通过素描实现。收录于 M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, & J. W. Vaughan (编), 《神经信息处理系统进展》(第34卷, 第7153–7166页)。Curran Associates, Inc. Overmann, K. A. (2025). 《文化数字系统：资料集》。Springer Nature. Pajot, M., Sabl´e-Meyer, M., & Dehaene, S. (2025). 数概念的组合性：来自数字频率的启示。《认知》, *263*, 106213. Radford, A., Kim, J. W., Hallacy, C., Ramesh, D., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). 在自然语言监督下学习可迁移视觉模型。收录于 M. Meila & T. Zhang (编), 《第*38*届国际机器学习会议论文集》(第139卷, 第8748–8763页)。PMLR. Xu, Y., Liu, E., & Regier, T. (2020). 跨语言数词系统支持高效沟通：从近似数量到递归。《开放思维》, *4*, 57–70.

补充材料

A. 操纵类别频率——实验设置

我们改变训练上下文，以评估更高频、因而更多样的数量是否会促成更高效的编码，而最不多样的数量是否对应更复杂的编码。在大多数实验中，我们使用均匀 分布，并为每个数量设置相等数量的训练样本（700 个）；此外，我们还定义了增加 和减少 设置。因此，我们将增加 设置定义为每个类别具有以下样本分布：从数量 1 到数量 5 依次递增的 100、 200、300、 400、 700；而减少 设置则相反。

该实验使用的模型具有可变的消息长度，最多为 5 个词元，词汇表大小为 3，并采用系数为 0.005 的消息正则化损失。图 6 展示了不同训练数据分布下离散通道的测试准确率，图 7 则说明了增加 和减少设置在 3 个随机种子下的精确映射。

主论文的图 4b 报告了不同随机种子下连续通道的准确率，其趋势与离散通道（图 6）相似。图 10 展示了测试时为训练集中各数量类别生成的更多草图示例。不同的行表示智能体在使用多样性较高或较低的训练样本进行训练时的结果，其中某些类别包含的实例较多或较少。图 8 展示了三种不同频率设置下草图的两两差异度。

结果表明，消息是任意的，消息复杂度与训练样本数量之间没有明确的相关性。唯一一致的趋势体现在不同训练设置下的准确率上，递减是离散通道和连续通道中最难解决的设置。

从这项实验中可以观察到，符号的复杂度不会随频率而变化，它们也不具有组合性（即不会组合使用高频数量的编码），但它们似乎都与元素所占面积相关。例如，数量 1 通常用一个符号表示，其线条集中在草图中心，且不会占据太多可用空间。随着数量增加，元素开始散布在目标图像空间中，线条也随之变长，并以一种尽可能覆盖绘图空间的方式排列。

B. 扩展结果

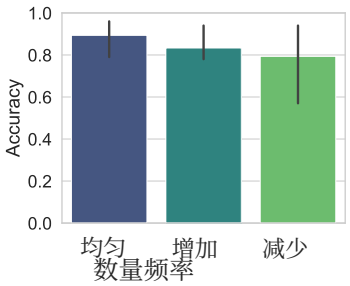
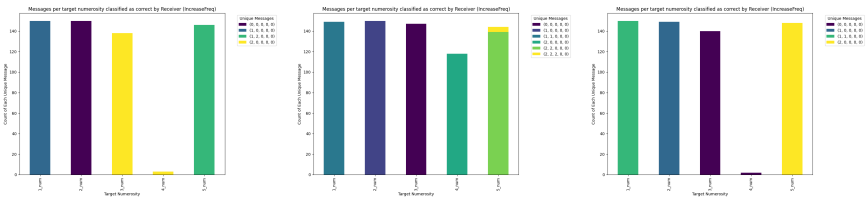
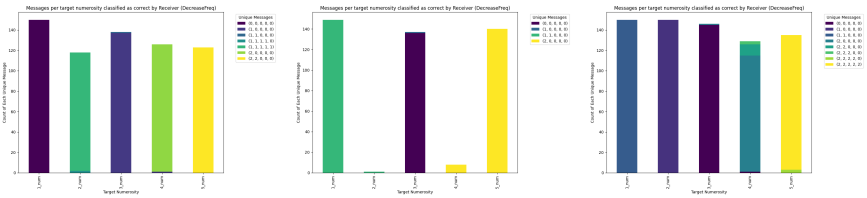


图 6. 在训练数值样本频率不同的设置下执行交际任务的准确率和标准差（基于 3 个随机种子）：均匀、增加 和减少 频率设置。

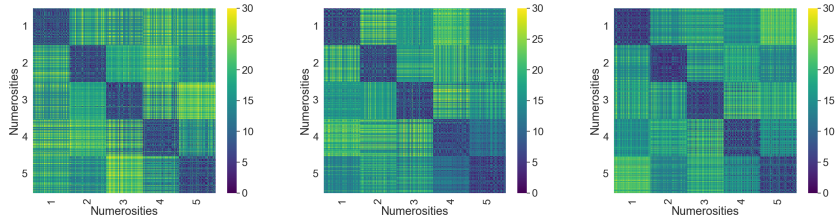


(a) 种子0　(b) 种子1　(c) 种子2



(d) 种子0　(e) 种子1　(f) 种子2

图 7. 在增加 （a、b 和 c）及减少 频率设置（d、e、f）中，针对正确分类的数量，在 3 个不同随机种子下消息到数值的精确映射。

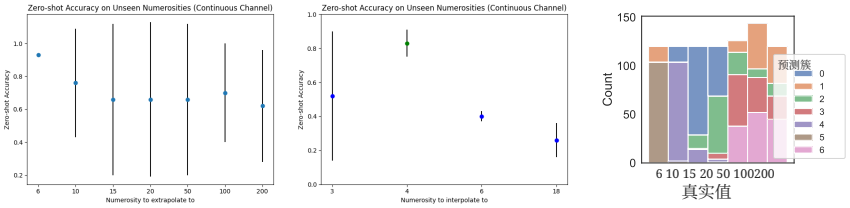


(a) 均匀 (b) 增加 (c) 减少

图 8. 不同数量频率下，t-SNE 降维后的像素空间中草图的两两不相似度：(a) 相等，(b) 从 1 递增至 5，(c) 从 1 递减至 5。

表1. 在数量 $\{1, 2, 3, 4, 5\}$ 上训练的固定消息长度模型，采用不同的$vocabulary$ 大小：$\{3, 5, 10, 100\}$。测试通信准确率和熵均为 3 个随机种子的结果。

设置	最大消息长度	词汇量大小	准确率	熵
差异目标	5	3	。95±.2	1。14±.34
	5	5	。96±0	1。15±.31
	5	10	.94±.02	1.13±.42
	5	100	.96±.008	0.76±.15



(a) 外推 (b) 插值 (c) 草图映射

图9。(a, b) 连续（即草图）通道在分布外测试数据上的零样本准确率。(a) 外推：在$train$ $=\{1, 2, 3, 4, 5\}$ 上训练的模型，并在$train$ $+$ 一个分布外数据集上进行测试。(b) 插值：插值中不同的颜色（绿色与蓝色）对应不同的训练数据——请参阅主论文中的图5。(c) 根据与外推数量的相似度进行聚类后的草图映射（条件熵：1.348）。

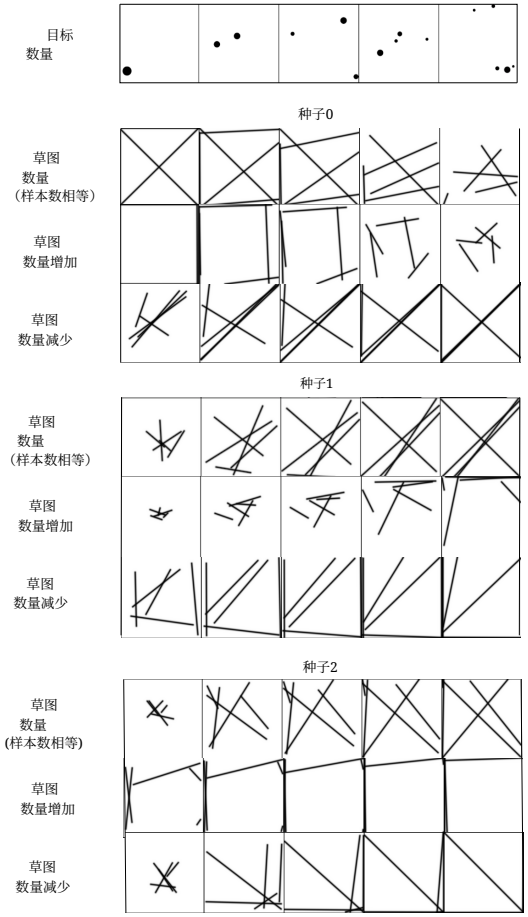


图 10。测试目标图像（第一行中的每个条目分别对应 1 至 5 之间的一个不同数量）以及在不同数量数据集上训练的智能体生成的草图，这些数据集包含数量各异的类别样本。结果基于 3 个训练随机种子进行报告。