arXiv:2602.02117v1 [cs.LG] 2 Feb 2026

最大冯·诺依曼熵原理：理论及其在机器学习中的应用

Youqi Wu*，Farzan Farnia†

摘要

冯·诺依曼熵（VNE）是量子信息论中的一个基本量，近年来也被机器学习采用，作为衡量核矩阵和核协方差算子多样性的谱度量。尽管在量子领域，关于在约束下最大化 VNE 的研究已十分成熟，但在数据驱动场景中，尚未明确提出经典最大熵框架在 VNE 上的原则性对应形式，尤其是其决策论和博弈论解释。本文将 Grünwald 和 Dawid 提出的最大熵原理的极小极大表述扩展到冯·诺依曼熵的情形，为在密度矩阵和迹归一化半正定算子上最大化 VNE 提供博弈论依据。这一视角为部分信息条件下的最大 VNE 解提供了稳健解释，并阐明了它们作为谱域中最少承诺推断的作用。随后，我们通过考察两个具有代表性的应用，展示由此得到的最大 VNE 原理如何应用于现代机器学习问题：通过基于核的 VNE 最大化，从多个归一化嵌入中选择一种核表示，以及根据部分观测到的条目补全核矩阵。这些示例表明，所提出的框架为基于 VNE 的核学习方法提供了统一的信息论基础。

1 引言

熵是信息论中的一个基本量，用于量化随机变量的不确定性，并在数据压缩、通信和统计推断中发挥核心作用。在经典情形下，香农熵 [1] 为概率分布的不确定性提供了数学刻画。在量子信息论 [2, 3], 中，这一概念被扩展到量子态，而量子态由密度矩阵表示。密度矩阵 ρ 是一个具有单位迹的半正定（PSD）算子，其关联的熵称为冯·诺依曼熵（*VNE*），定义为其特征值的香农熵：

$$S(\rho) \;=\; -\mathrm{Tr}(\rho \log \rho) \;=\; -\sum_i \lambda_i(\rho) \log \lambda_i(\rho), \tag{1}$$

其中，特征值 $\{\lambda_i(\rho)\}$构成一个概率分布。

近年来，VNE 已被机器学习采用，用于衡量数据表示的多样性和有效维数 [4, 5]。特别是，经过归一化而具有单位迹的核矩阵和核协方差算子，自然定义了类密度矩阵对象，其谱编码了全局相似性结构 [6, 7]。这一观察促使研究者将基于 VNE 的量（例如 Vendi 分数 [5], ）用作数据集和生成模型的谱多样性度量。然而，在若干机器学习应用中，底层密度矩阵并未被完整观测到，例如目标核矩阵存在缺失条目、测量结果含有噪声，或核协方差算子是根据有限数据估计得到的。这引出了一个根本问题：当密度矩阵仅被部分指定时，应如何开展基于 VNE 的分析?

在经典信息论中，可以通过最大熵原理，在信息不完整的情况下进行基于熵的统计推断。Jaynes 于 [8], 提出了这一原理，主张在所有符合给定约束的概率分布中，选择香农熵值最大的分布

*香港中文大学计算机科学与工程学系，yqwu24@cse.cuhk.edu.hk
†香港中文大学计算机科学与工程学系，farnia@cse.cuhk.edu.hk

。一个典型例子是，在均值和协方差矩阵固定的所有分布中，高斯分布的（微分）熵最大 [9]。除了变分形式外，最大熵原理还有一种操作性解释，正如 Grünwald 和 Dawid 在 [10],所展示的那样：最大熵解可视为博弈论框架中的极小极大最优策略；在该框架下，决策者与对手围绕一个分布模糊集展开竞争。

本文将这一框架扩展到冯·诺依曼熵，重点关注密度矩阵信息不完整的机器学习情形。给定与现有信息相符的密度矩阵模糊集 $\mathcal{C}$ ，我们考虑最大 *VNE* 原则，即选择 VNE 最大的密度矩阵 ρ^* ，并以此为基础进行推断：

$$\rho^\star \in \arg\max_{\rho\in\mathcal{C}} S(\rho). \tag{2}$$

我们的主要贡献是正式证明，Grünwald 和 Dawid 在 [10] 中提出的最大熵原理的博弈论解释可自然地扩展到矩阵值情形。这为在密度矩阵集合上最大化 VNE 提供了有原则的极小极大论证，与概率分布的经典情形直接对应。

我们进一步证明，这一极小极大框架不仅适用于 VNE，也适用于其他值得关注的谱熵泛函。具体而言，我们考察基于矩阵的 Rényi 熵，包括二阶二次 Rényi 熵和由最大特征值 [11, 12] 决定的无穷阶 Rényi 熵。我们还借鉴监督学习的极小极大表述，将这一框架扩展到条件情形 [13], ，从而使其能够应用于核对象由输入和输出共同产生的情形。

最后，我们通过机器学习中的两个具体应用来阐释所提出的最大 VNE 原则。第一个应用涉及从多个归一化嵌入中选择核表示。通过对相应的核矩阵进行凸组合，并针对混合权重应用 Max-VNE 原理，我们得到一种有原则的方法，用于选择最少承诺核混合，从而促进谱多样性。第二个应用针对基于部分观测的核矩阵补全：在所有与已观测条目一致的可行半正定补全中，我们选择使 VNE 最大化的补全。我们还通过标准图像数据集上的数值实验补充理论研究，展示最大 VNE 原则在数据驱动场景中的实际价值。

2 相关工作

最大熵原理源于香农熵 [1] 以及 Jaynes 提出的在一组约束 [8]下选择最少承诺模型的思想。在较新的文献中，一些密切相关的原理以在部分约束下最小化依赖性取代最大化熵。典型例子包括最小互信息 原理，该原理将最大熵推广到判别式学习，通过在边际约束和特征约束 [14], 下最小化 $I(X;Y)$；以及最小最大相关 原理，该原理在低阶约束 [15, 16]下选择使 Hirschfeld–Gebelein–Rényi（HGR）相关系数最小的联合分布。这些依赖性最小化原理具有极小极大解释，并在仅有部分信息时导出鲁棒学习规则。另一方面，Grünwald 和 Dawid [10]对最大熵本身的极小极大与博弈论基础进行了形式化。这些研究推动我们将基于熵和依赖性的极小极大原理从概率分布拓展至矩阵值熵目标。

在量子信息论中，不确定性由密度矩阵描述。密度矩阵是半正定、迹为单位的算子；冯·诺依曼熵则通过特征值分布 [2, 3], 将香农熵推广到算子值情形。量子系统的最大熵推理已得到广泛研究，尤其是在量子统计力学和推断领域，常用于选择与部分信息相一致的密度矩阵。这些研究促使人们直接在密度矩阵层面探讨最大熵原理，但通常以物理约束为基础，而非数据驱动的歧义集。

冯·诺依曼熵及相关谱熵泛函近年来在机器学习中日益受到重视，被用作衡量核矩阵和核协方差算子的多样性与有效维数的指标。Vendi 分数将多样性定义为归一化

相似度矩阵的冯·诺依曼熵的指数，并已被用作数据集和生成模型的可解释多样性度量 [5]„ 随后还发展出了相关的基于相似度的多样性族 [17]。与此同时，基于核的矩阵信息度量可直接根据归一化格拉姆矩阵估算类熵量，包括基于矩阵的 Rényi 型熵以及算子层面信息论视角 [4, 12]。这些理念已应用于生成模型的信息论评估和多模态学习 [6, 7, 18]。我们的工作有别于这些方法：我们为最大冯·诺依曼熵原理提供了博弈论基础，并将其应用于部分观测下的核诱导密度矩阵。

3 预备知识

3.1 熵与最大熵原理

设 $\mathcal{X}$ 为有限字母表，并令 $\Delta(\mathcal{X})$ 表示定义在 $\mathcal{X}$ 上的概率单纯形。对于分布 $P \in \Delta(\mathcal{X})$，香农熵定义为

$$H(P) \;:=\; -\sum_{x\in\mathcal{X}} P(x)\log P(x). \tag{3}$$

给定由约束（例如矩约束或边缘约束）指定的可行族 $\mathcal{P} \subseteq \Delta(\mathcal{X})$，经典最大熵表述 [8] 会寻找熵最大化分布 P^*，并据此采取行动：

$$P^\star \;\in\; \arg\max_{P\in\mathcal{P}} H(P). \tag{4}$$

3.2 密度矩阵与矩阵熵

令 $\mathbb{S}_+^n$ 表示实对称半正定矩阵构成的锥。密度矩阵 是满足以下条件的矩阵 $\rho \in \mathbb{S}_+^n$

$$\rho \succeq 0, \qquad \operatorname{Tr}\big(\rho\big) = 1. \tag{5}$$

令 $\lambda(\rho) = \big(\lambda_1(\rho)\dots,\lambda_n(\rho)\big)$ 表示 ρ 的特征值（按重数计）。由于 $\rho \succeq 0$ 且 $\operatorname{Tr}\big(\rho\big) = \sum_i \lambda_i(\rho) = 1$，我们有 $\lambda(\rho) \in \Delta_n$，其中 $\Delta_n := \big\{p \in \mathbb{R}_+^n : \sum_{i=1}^n p_i = 1\big\}$。 ρ 的 冯·诺依曼熵 （VNE）定义为

$$S(\rho) \;:=\; -\operatorname{Tr}\big(\rho\log\rho\big) \;=\; -\sum_{i=1}^{n} \lambda_i(\rho)\log\lambda_i(\rho), \tag{6}$$

其中 $\log\rho$ 是通过对特征值进行函数演算来定义的。

M更一般地，对于满足 $\alpha \neq 1$ 的 $\alpha > 0$，二阶为 α 的（量子）Rényi 熵定义为

$$S_\alpha(\rho) \;:=\; \frac{1}{1-\alpha}\log\operatorname{Tr}\big(\rho^\alpha\big) \;=\; \frac{1}{1-\alpha}\log\sum_{i=1}^{n}\lambda_i(\rho)^\alpha. \tag{7}$$

本文将用到一个特殊情形。二次（二阶）Rényi 熵为

$$S_2(\rho) \;=\; -\log\operatorname{Tr}\big(\rho^2\big), \tag{8}$$

3.3 核诱导密度矩阵与核协方差算子

设 $x_1,\dots,x_n \in \mathcal{X}$ 为数据点，并设 $k\colon \mathcal{X}\times\mathcal{X} \to \mathbb{R}$ 为与再生核希尔伯特空间（RKHS） $\mathcal{H}$ 及特征映射 $\phi\colon \mathcal{X} \to \mathcal{H}$ 相关联的半正定核。其对应的核（ Gram ）矩阵为 $K \in \mathbb{R}^{n\times n}$，其元素为

$$K_{ij} \;:=\; k\big(x_i, x_j\big). \tag{9}$$

根据 k 的半正定性，我们有 $K \succeq 0$。如果核函数 k 满足对于每个 $x \in \mathcal{X}$， $k(x,x) = 1$，则称其为归一化核函数。请注意，对于任意核函数 k，无论其是否归一化，归一化函数

$\widetilde{k}(x,y) = k(x,y)/\sqrt{k(x,x)k(y,y)}$ 将是一个有效的归一化核函数。根据上述定义，我们将迹归一化的核诱导密度矩阵与 K 相关联

$$\rho_K := \frac{K}{\mathrm{Tr}(K)}, \tag{10}$$

其中，$\mathrm{Tr}(K) = n$ 表示所假设的归一化核函数。因此，默认有 $\rho_K \succeq 0$ 和 $\mathrm{Tr}(\rho_K) = 1$，从而 ρ_K 是一个密度矩阵，公式 6–公式 8 中定义的矩阵熵可直接适用。特别是，$S(\rho_K)$ 提供了一种由核相似性结构诱导的谱多样性概念或有效维数。

等价地，也可以在算子层面进行处理。与 $\{x_i\}_{i=1}^n$ 的经验分布相关联的（未中心化）核协方差算子 $\mathcal{C}: \mathcal{H} \to \mathcal{H}$定义为

$$\mathcal{C} := \frac{1}{n}\sum_{i=1}^{n} \phi(x_i) \otimes \phi(x_i), \tag{11}$$

其中，$\phi(x) \otimes \phi(x)$表示秩一算子 $h \mapsto \langle h, \phi(x)\rangle_{\mathcal{H}}\phi(x)$。经过迹归一化后，$\rho_{\mathcal{C}} := \mathcal{C}/\mathrm{Tr}(\mathcal{C})$ 在 $\mathcal{H}$上定义了一个密度算子。$\rho_{\mathcal{C}}$ 的非零特征值与 ρ_K 的非零特征值一致（仅差一个缩放因子），因此可以在矩阵层面或算子层面一致地定义 von Neumann 熵和 Rényi熵。

4 最大冯·诺依曼熵原理

4.1 博弈论框架与广义熵

我们采用一种博弈论表述，该表述遵循最大熵原理及其在经典熵情形下发展出的极小极大解释 [10]。根据问题的表述，自然从可行集 Γ中选择一个未知状态 ρ，而决策者（DM）从行动集 $\mathcal{Q}$ 中选择一个行动 q，并承受损失 $L(\rho, q)$。

在本节中，我们假设 Γ 是有限维实欧几里得空间（E $\langle\cdot,\ \cdot\rangle$）的一个非空、凸且紧的子集，而 $\mathcal{Q}$是有限维欧几里得空间（F $\langle\cdot,\ \cdot\rangle$）的一个非空、凸且紧的子集。我们考虑损失函数 $L: \Gamma \times \mathcal{Q} \to \mathbb{R}$，它们是连续的，在$\rho$ 上仿射 Γ，并且在 q 上凸 $\mathcal{Q}$。特别地，该设定包括以下形式的期望损失模型

$$L(\rho, q) = \big\langle \rho,\, \ell(q)\big\rangle + c(q), \tag{12}$$

对于连续映射 $\ell: \mathcal{Q} \to$E 和 $c: \mathcal{Q} \to \mathbb{R}$。

给定状态 ρ，可达到的最小损失定义了广义熵 （贝叶斯风险 k)

$$H(\rho) := \inf_{q\in\mathcal{Q}} L(\rho, q). \tag{13}$$

请注意，在这种博弈论设定中，相应的极小极大值为

$$V_\Gamma := \inf_{q\in\mathcal{Q}} \sup_{\rho\in\Gamma} L(\rho, q). \tag{14}$$

定义1 (最大熵原理)。 给定可行集 Γ，最大（基于矩阵的）熵原理选择

$$\rho^\star \in \arg\max_{\rho\in\Gamma} H(\rho), \tag{15}$$

其中 H（ρ）是方程*13*中的广义熵。

该定义不依赖于熵泛函，因为不同的损失会通过公式 13 导出不同的广义熵。

4.2 最大熵原理的极小极大解释 e

现在我们将证明，在满足上一小节所讨论的假设时，最大熵原理具有精确的极小极大解释。具体而言，最大广义熵状态对应于自然在极小极大博弈中的均衡策略，而与之相关的行动则是稳健贝叶斯决策规则。

定理1 (状态线性极小极大定理)。 设 $\Gamma \subseteq \mathsf{E}$ 和 $\mathcal{Q} \subseteq \mathsf{F}$ 非空、凸且紧，并设 $L : \Gamma \times \mathcal{Q} \to \mathbb{R}$ 连续，在 Γ上关于 ρ 仿射，在 $\mathcal{Q}$上关于 q 凸。按方程*13*–方程*14*定义 $H(\rho)$ 和 V_Γ 。则：

1. 极小极大等式成立：

$$\inf_{q\in\mathcal{Q}} \sup_{\rho\in\Gamma} L(\rho, q) \;=\; \sup_{\rho\in\Gamma} \inf_{q\in\mathcal{Q}} L(\rho, q). \tag{16}$$

2. 存在一个鞍点 $(\rho^\star, q^\star) \in \Gamma \times \mathcal{Q}$ ，使得

$$L(\rho, q^\star) \;\le\; L(\rho^\star, q^\star) \;\le\; L(\rho^\star, q), \qquad \forall\, \rho \in \Gamma,\ \forall\, q \in \mathcal{Q}. \tag{17}$$

3. 鞍点处的状态使广义熵最大化，而该行动是极小极大稳健的：

$$H(\rho^\star) \;=\; \sup_{\rho\in\Gamma} H(\rho) \;=\; V_\Gamma \;=\; L(\rho^\star, q^\star). \tag{18}$$

证明。 证明推迟至附录。 □

4.3 冯·诺依曼熵的特例

现在，我们将这一框架特化到量子态。令 $\mathcal{D}_n := \{\rho \succeq 0 : \mathrm{Tr}(\rho) = 1\}$ 表示$n \times n$ 密度矩阵的集合。为便于正文中的技术处理，我们使用严格正定密度矩阵构成的紧子集

$$\mathcal{D}_{n,\varepsilon} \;:=\; \Big\{\rho \in \mathcal{D}_n :\ \rho \succeq \varepsilon I\Big\}, \tag{19}$$

其中 $\varepsilon \in$ (0,1)。我们假设 $\Gamma \subseteq \mathcal{D}_{n,\varepsilon}$ 是凸且紧的，并且决策者行动集为 $\mathcal{Q} = \mathcal{D}_{n,\varepsilon}$。考虑量子对数损失

$$L_{\log}(\rho, \sigma) \;:=\; -\mathrm{Tr}\big(\rho \log \sigma\big), \qquad \rho \in \Gamma,\ \sigma \in \mathcal{D}_{n,\varepsilon}, \tag{20}$$

它在 $\Gamma \times \mathcal{D}_{n,\varepsilon}$ 上定义良好且连续。在附录 B 中，我们指出，由$L_{\log}$ 诱导的贝叶斯风险与每个 $\rho \in \Gamma$的冯·诺依曼熵相同

$$\inf_{\sigma\in\mathcal{D}_{n,\varepsilon}} L_{\log}(\rho, \sigma) \;=\; S(\rho) \;:=\; -\mathrm{Tr}\big(\rho \log \rho\big). \tag{21}$$

推论1 (Max-VNE 作为极小极大均衡)。 将定理*1*应用于对数损失 $L_{\log}$，最大冯·诺依曼熵原理

$$\rho^\star \in \arg\max_{\rho\in\Gamma} S(\rho) \tag{22}$$

选出极小极大博弈中自然的均衡态。此外，相应的最优行动 $\sigma^\star$ 是一条极小极大鲁棒贝叶斯决策规则。

证明。证明见附录。 □

4.4 超越冯·诺依曼熵：迹熵与布雷格曼型分解。

公式 20 中的对数损失可分解为经典形式

$$L_{\log}(\rho,\sigma) = S(\rho) + D\big(\rho\|\sigma\big), \tag{23}$$

其中 $D(\rho\|\sigma)$ 是量子相对熵。这种“熵加散度”的形式并非 $S(\rho)$ 所独有：一大类迹熵 都可以作为某些损失的贝叶斯风险获得，而这些损失的后悔值是矩阵布雷格曼散度。

具体而言，设 $f:(0,\infty) \to \mathbb{R}$ 为凸且可微函数，并定义迹熵泛函

$$H_f(\rho) \;:=\; -\mathrm{Tr}\big(f(\rho)\big), \qquad \rho \in \mathcal{D}_{n,\varepsilon}. \tag{24}$$

对于附录 C 中定义的相应“f-损失” $L_f(\rho,\sigma)$，有如下精确的后悔值分解

$$L_f(\rho,\sigma) = H_f(\rho) + D_f\big(\rho\|\sigma\big), \tag{25}$$

其中 $D_f(\rho\|\sigma) \geq 0$是由 f生成的矩阵布雷格曼散度。特别地， $f(t) = t\log t$可还原公式 23，而 $f(t) = t^2$给出 $H_f(\rho) = -\mathrm{Tr}(\rho^2)$，这通过单调变换 $S_2(\rho) = -\log \mathrm{Tr}(\rho^2)$ 与 Rényi-2 熵等价。在附录中，我们讨论这些特殊情形。

5 机器学习中的应用

现在，我们通过两个具有代表性的机器学习问题，说明第 4 节提出的最大冯·诺依曼熵（Max-VNE）原理如何应用于涉及部分信息的基于核的表示。在本节中，我们始终假设归一化正半定核 k 满足$k(x,\ x) = 1$，且对所有输入 $x \in \mathcal{X}$均成立。在这一假设下，核协方差算子和核矩阵可以自然地解释为密度矩阵，无需额外进行迹归一化。

5.1 通过 Max-VNE 选择核混合

在许多学习场景中，数据可以通过多个归一化嵌入来表示，每个嵌入都会生成一个核矩阵。我们不固定单一表示，而是考虑一类受约束的凸核混合，并应用最大冯·诺依曼熵（Max-VNE）原理，选择一个满足先验约束且最少预设的混合。

为此，考虑 M 个归一化嵌入映射 $\phi_i : \mathcal{X} \to \mathbb{R}^{d_i}$，每个映射都会生成一个核（即内积）：

$$k_i(x,x') \;:=\; \big\langle \phi_i(x), \phi_i(x') \big\rangle, \;\; k_i(x,x) = 1, \;\forall x \in \mathcal{X}. \tag{26}$$

给定样本 $x_1, \ldots, x_n$, 令 $K_i \in \mathbb{R}^{n\times n}$ 表示相应的核矩阵 , 使得 $\mathrm{Tr}(K_i) = n$ 。我们通过加权拼接构造一个组合表示。对于权重 $\alpha = [\alpha_1,\ldots,\alpha_M] \in \Delta_M$，定义

$$\phi_\alpha(x) \;:=\; \big[\sqrt{\alpha_1}\phi_1(x), \ldots, \sqrt{\alpha_M}\phi_M(x)\big], \tag{27}$$

它满足对于所有 x， $\|\phi_\alpha(x)\|_2 = 1$ 。诱导核为

$$k_\alpha(x,x') \;=\; \big\langle \phi_\alpha(x), \phi_\alpha(x') \big\rangle \;=\; \sum_{i=1}^{M} \alpha_i\, k_i(x,x'). \tag{28}$$

因此，经验核矩阵是权重为 $[\alpha_1,\ldots,\alpha_M]$的凸组合

$$K(\alpha) \;:=\; \sum_{i=1}^{M} \alpha_i\, K_i, \qquad \alpha \in \mathcal{A} \subseteq \Delta_M, \tag{29}$$

其中 $\mathcal{A}$ 是凸集，且是概率单纯形 Δ_M 的子集。由于 $\mathrm{Tr}(K(\alpha)) = n$，归一化矩阵将为

$$\rho(\alpha) := \frac{1}{n} K(\alpha) \tag{30}$$

它是半正定且具有单位迹，因此可以视为密度矩阵。混合类公式 29 导出模糊集

$$\Gamma_{\mathrm{mix}} := \{\rho(\alpha) : \alpha \in \mathcal{A}\}. \tag{31}$$

仅已知数据表示属于此类时，我们应用 Max-VNE 原理并选择

$$\alpha^\star \in \arg\max_{\alpha \in \mathcal{A}} S\big(\rho(\alpha)\big), \qquad \rho^\star := \rho(\alpha^\star), \tag{32}$$

其中， $S(\cdot)$ 表示冯·诺依曼熵。根据第 4 节中的极小极大解释， $\rho^\star$ 对应于自然在 Γ_{mix} 中的均衡密度矩阵。

5.2 通过 Max-VNE 进行核矩阵与协方差补全

接下来，我们考虑根据部分观测补全核矩阵的问题。设$x_1, \ldots, x_n \in \mathcal{X}$为用于定义核矩阵 $K \in \mathbb{R}^{n \times n}$ 的数据点，并令

$$K_{ij} := k(x_i, x_j) \tag{33}$$

表示相应的核矩阵。在归一化条件 $k(x, x) = 1$ 下，有 $\mathrm{Tr}(K) = n$，而缩放矩阵

$$\rho := \frac{1}{n} K \tag{34}$$

是 $\mathcal{D}_n$ 中的密度矩阵。

假设仅观测到部分核矩阵元素 $\{K_{ij} : (i, j) \in \Omega\}$ ，对应于对 ρ的线性约束。与观测结果一致的密度矩阵集合为

$$\Gamma_\Omega := \Big\{\rho \in \mathcal{D}_n : \rho_{ij} = \tfrac{1}{n} K_{ij} \text{ for all } (i, j) \in \Omega\Big\}, \tag{35}$$

该集合是凸且紧的。应用 Max-VNE 原理，我们将补全核矩阵定义为

$$\rho^\star \in \arg\max_{\rho \in \Gamma_\Omega} S(\rho). \tag{36}$$

解 $\rho^\star$ 是与观测元素一致的最大熵补全，除约束所强制的内容外，不赋予任何额外的谱结构。根据第 4 节， $\rho^\star$ 在 Γ_Ω内可作极小极大解释，即为稳健均衡密度矩阵。随后，补全核矩阵可用于下游核方法，例如核主成分分析或谱聚类。

5-理论

6 数值结果

6.1 通过最大 V 选择预训练嵌入的混合 NE

我们在四个标准图像分类基准上，通过最大 VNE 原则评估了学习到的表示：ImageNet [19], CIFAR-100 [20], 可描述纹理数据集（DTD） [21], 和 FGVC Aircraft [22]。

模型与实验设置。 遵循表征学习领域的标准做法，我们考虑 DINOv2 [23], SigLIP [24], UNICOM [25], 和 OpenCLIP [26, 27] 模型。对于每个预训练嵌入，我们构造一个径向基函数核，并应用 Max-VNE 原理，获得相应核矩阵的最优混合权重 $\alpha^\star$ 。将每个嵌入乘以其对应权重后进行拼接，即可利用学得的权重构造混合特征表征。

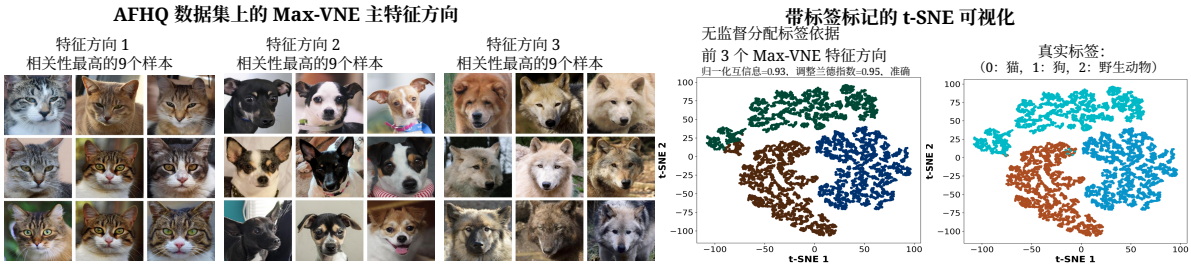


图1：在 AFHQ 数据集上，从 10% 观测条目进行 Max-VNE 核补全的实验结果。使用归一化互信息（归一化互信息）、调整兰德指数（调整兰德指数）和准确率（准确率）对恢复的核矩阵进行聚类性能评估。

表1：嵌入与 MAX-VNE 选定的嵌入混合的线性探测准确率（%，5 次运行的均值 ±标准差）。

模型	ImageNet	CIFAR100	DTD	Aircraft
OpenCLIP	78.9 ± 0.14	86.3 ± 0.02	77.7 ± 0.23	49.2 ± 0.39
SigLIP	81.8 ± 0.16	83.3 ± 0.04	80.3 ± 0.05	60.7 ± 0.41
DINOv2	80.0 ± 0.19	89.9 ± 0.13	79.9 ± 0.22	58.9 ± 0.63
UNICOM	77.4 ± 0.08	85.0 ± 0.04	71.6 ± 0.53	55.4 ± 0.85
Max-VNE	85.1 ± 0.12	91.5 ± 0.06	82.5 ± 0.17	65.1 ± 0.48

为评估 Max-VNE 混合方案能否为下游任务提供信息量更丰富的表征，我们采用标准的线性探测流程。具体而言，在嵌入之上训练一个线性分类器。使用 Adam 进行优化，学习率为 1×10^{-3}，批量大小为 256，训练 10 个周期。所有实验均在两张 NVIDIA RTX 4090 GPU 上进行。

结果。 表 1 汇报了所有数据集上的线性探测准确率。在四个基准测试数据集上，Max-VNE 混合嵌入的表现优于每一种单独的嵌入。

6.2 通过 Max-VNE 原理进行核矩阵补全

我们在三个视觉复杂度各异的标准图像数据集上进行了核补全实验：AFHQ [28]（动物面部）、MNIST [29]（手写数字）和 ImageNet-Dogs [30]（来自 ImageNet 的犬种）。

实验设置。 对于每个数据集，我们根据 OpenCLIP [26] 图像嵌入构建余弦相似度核矩阵。为模拟部分观测，我们随机屏蔽 90% 的非对角线条目。给定部分观测的核矩阵后，我们应用 Max-VNE 原理恢复补全核矩阵。随后，在恢复的核矩阵上执行谱聚类。

结果。 图 1、图 2 展示了在 AFHQ 和 MNIST 数据集上有意义的聚类结果。定量而言，我们在图中报告了归一化互信息 [31], 调整兰德指数 [32], 和聚类准确率（准确率） [33] 。总体而言，Max-VNE 在各数据集上表现良好，体现了其有效性。ImageNet-Dogs 上的结果见附录。

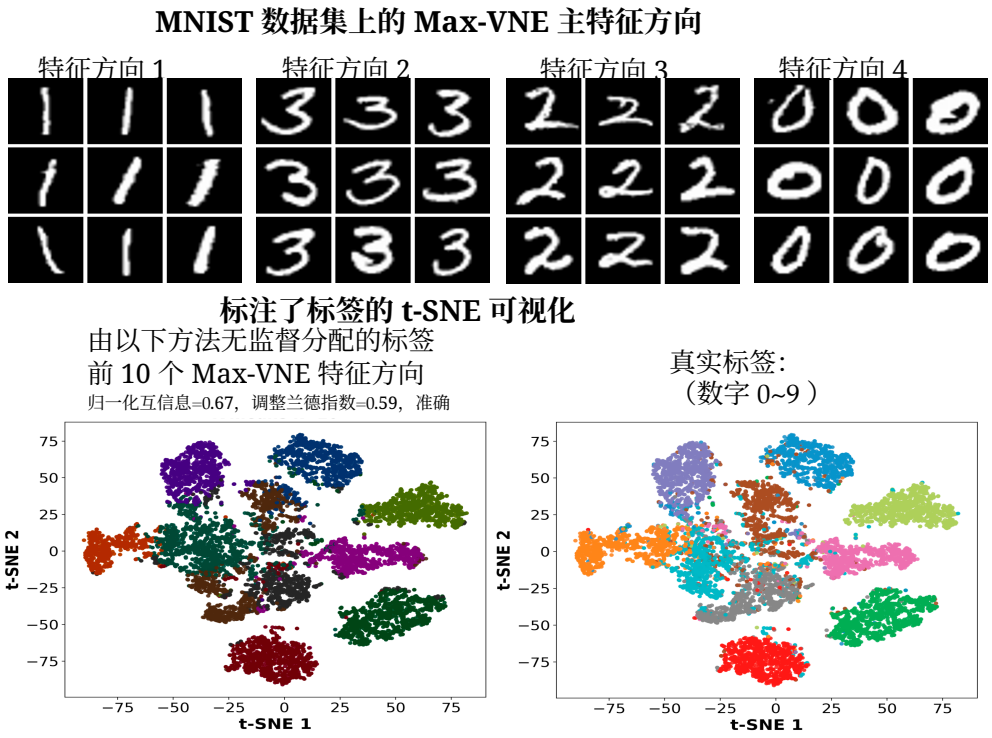


图 2：MNIST 数据集上图 1 的实验结果。

参考文献

[1] Claude E. Shannon。通信的数学理论。贝尔系统技术期刊，27(3):379– 423，1948年。[2] John von Neumann。量子力学的数学基础。普林斯顿大学出版社，1955年。译自1932年德文版。[3] Michael A. Nielsen 和 Isaac L. Chuang。量子计算与量子信息。剑桥大学出版社，2000年。[4] Francis Bach。核方法与信息论，2022年。[5] Dan Friedman 和 Adji Bousso Dieng。Vendi 分数：一种用于机器学习的多样性评估指标。机器学习研究汇刊，2023年。[6] Mohammad Jalali、Cheuk Ting Li 和 Farzan Farnia。对学习多模态分布的生成模型进行信息论评估。神经信息处理系统进展，36:9931–9943，2023年。[7] Azim Ospanov、Jingwei Zhang、Mohammad Jalali、Xuenan Cao、Andrej Bogdanov 和 Farzan Farnia。迈向可扩展的生成模型无参考评估。神经信息处理系统进展，37:120892–120927，2024年。

[8] E. T. Jaynes. 信息论与统计力学。《物理评论》，106(4):620–630，1957。

[9] Thomas M. Cover 和 Joy A. Thomas. 信息论基础。威立，第二版，2006。

[10] Peter D. Grünwald 和 A. Philip Dawid。博弈论、最大熵、最小差异与稳健贝叶斯决策理论。《统计年刊》，32(4):1367–1433，2004年。

[11] Alfréd Rényi。论熵与信息度量。载于 第四届伯克利数学统计与概率研讨会论文集，第1卷，第547–561页。加利福尼亚大学出版社，1961年。[12] Luis G. Sánchez Giraldo、Murali Rao 和 José C. Príncipe。使用无限可分核从数据中获取熵度量。*IEEE* 信息论汇刊，61(1):535–548，2015年。[13] Farzan Farnia 和 David Tse。监督学习的极小极大方法。载于 神经信息处理系统进展（*NeurIPS*），第29卷，第4240–4248页，2016年。[14] Amir Globerson 和 Naftali Tishby。判别式学习的最小信息原理。载于第*20*届人工智能不确定性会议（*UAI*）论文集，第193–200页。AUAI出版社，2004年。[15] Farzan Farnia、Meisam Razaviyayn、Sreeram Kannan 和 David Tse。最小HGR相关性原理：从边缘分布到联合分布。载于 *2015*年*IEEE*国际信息论研讨会（*ISIT*），第1377–1381页，2015年。[16] Meisam Razaviyayn、Farzan Farnia 和 David Tse。离散Rényi分类器。载于 神经信息处理系统进展，第28卷，第3276–3284页，2015年。[17] Amey P Pasarkar 和 Adji Bousso Dieng。Vendi 分数的近亲：面向科学与机器学习的一系列基于相似性的多样性度量。载于 国际人工智能与统计会议，第3808–3816页。PMLR，2024年。[18] Jingwei Zhang、Cheuk Ting Li 和 Farzan Farnia。对生成模型基于熵的新颖性进行可解释评估。载于 国际机器学习会议，第59148–59172页。PMLR，2024年。[19] Jia Deng、Wei Dong、Richard Socher、Li-Jia Li、Kai Li 和 Li Fei-Fei。ImageNet：大规模层次化图像数据库。载于*2009*年*IEEE*计算机视觉与模式识别会议，第248–255页。IEEE，2009年。[20] Alex Krizhevsky。学习来自微小图像的多层特征。技术报告TR-2009，多伦多大学，2009年。[21] Mircea Cimpoi、Subhransu Maji、Iasonas Kokkinos、Sammy Mohamed 和 Andrea Vedaldi。描述自然场景中的纹理。载于*IEEE*计算机视觉与模式识别会议论文集，第3606–3613页，2014年。[22] Subhransu Maji、Esa Rahtu、Juho Kannala、Matthew Blaschko 和 Andrea Vedaldi。Aircraft的细粒度视觉分类。*arXiv*预印本 *arXiv:1306.5151*，2013年。[23] Maxime Oquab、Timothée Darcet、Theo Moutakanni、Huy Vo、Marc Szafraniec、Vasil Khalidov、Pierre Fernandez、Daniel Haziza、Francisco Massa、Alaaeldin El-Nouby 等。DINOv2：无需监督学习鲁棒视觉特征。*arXiv*预印本 *arXiv:2304.07193*，2023年。[24] Xiaohua Zhai、Basil Mustafa、Alexander Kolesnikov 和 Lucas Beyer。用于语言图像预训练的Sigmoid损失。载于*IEEE/CVF*国际计算机视觉会议论文集，第11975–11986页，2023年。[25] Xiang An、Jiankang Deng、Kaicheng Yang、Jaiwei Li、Ziyong Feng、Jia Guo、Jing Yang 和 Tongliang Liu。UNICOM：用于图像检索的通用紧凑表征学习。*arXiv*预印本 *arXiv:2304.05884*，2023年。[26] Gabriel Ilharco、Mitchell Wortsman、Ross Wightman、Cade Gordon、Nicholas Carlini、Rohan Taori、Achal Dave、Vaishaal Shankar、Hongseok Namkoong、John Miller、Hannaneh Hajishirzi、Ali Farhadi 和 Ludwig Schmidt。OpenCLIP，2021年7月。

[27] Alec Radford、Jong Wook Kim、Chris Hallacy、Aditya Ramesh、Gabriel Goh、Sandhini Agarwal、Girish Sastry、Amanda Askell、Pamela Mishkin、Jack Clark 等。通过自然语言监督学习可迁移视觉模型。国际机器学习会议，第8748–8763页，2021年。[28] Yunjey Choi、Youngjung Uh、Jaejun Yoo 和 Jung-Woo Ha。StarGAN v2：面向多个领域的多样化图像合成。载于*IEEE/CVF*计算机视觉与模式识别会议（*CVPR*）论文集，2020年。[29] Yann LeCun、Léon Bottou、Yoshua Bengio 和 Patrick Haffner。应用于文档识别的基于梯度的学习。*IEEE*会刊，86(11):2278–2324，1998年。[30] Jeremy Howard。Imagewoof：ImageNet中由10个类别构成的子集，这些类别并不容易分类。https://github.com/fastai/imagenette#imagewoof，2019年3月。GitHub代码库。[31] Aaron F. McDaid、Derek Greene 和 Neil Hurley。用于评估重叠社区发现算法的归一化互信息，2013年。[32] Nguyen Xuan Vinh、Julien Epps 和 James Bailey。用于聚类比较的信息论度量：是否有必要进行机会校正？载于第*26*届国际机器学习年会论文集，第1073–1080页，2009年。[33] Harold W. Kuhn。用于指派问题的匈牙利算法。《海军研究后勤季刊》，2(1–2):83–97，1955年。[34] Mohammad Jalali、Haoyu LEI、Amin Gohari 和 Farzan Farnia。SPARKE：通过RKE分数在扩散模型中实现可扩展的提示感知多样性与新颖性引导。载于第三十九届神经信息处理系统年会，2025年。[35] Azim Ospanov、Mohammad Jalali 和 Farzan Farnia。Scendi分数：通过CLIP嵌入的舒尔补进行提示感知多样性评估。载于*IEEE/CVF*国际计算机视觉会议，第16927–16937页，2025年。[36] Mohammad Jalali、Azim Ospanov、Amin Gohari 和 Farzan Farnia。条件Vendi分数：针对文本引导生成式人工智能模型的提示感知多样性评估。载于第*29*届国际人工智能与统计会议论文集，《机器学习研究论文集》。PMLR，2026年5月2日至5日。[37] Pin-Yu Chen、Lingfei Wu、Sijia Liu 和 Indika Rajapakse。快速增量式冯·诺依曼图熵计算：理论、算法与应用。2018年。[38] Parham Rezaei、Farzan Farnia 和 Cheuk Ting Li。比最多样化的模型更加多样化：通过混合UCB多臂老虎机算法实现生成模型的最优混合。载于第十三届国际学习表征会议，2025年。

A附录 A 证明与其他理论结果

A.1 定理 1 的证明

我们在给定假设下证明定理1。全文中， Γ 和 $\mathcal{Q}$ 是有限维实向量空间中的非空、凸且紧的子集，并且 $L:\ \Gamma\times\mathcal{Q}\to\mathbb{R}$ 是连续的，对 ρ而言是仿射的，对 q而言是凸的。

对于每个 $q\in\mathcal{Q}$，定义自然的最佳响应集

$$\mathrm{BR}_\Gamma(q)\ :=\ \arg\max_{\rho\in\Gamma} L(\rho,q),$$

对于每个 $\rho\in\Gamma$，定义决策者的最佳响应集

$$\mathrm{BR}_\mathcal{Q}(\rho)\ :=\ \arg\min_{q\in\mathcal{Q}} L(\rho,q).$$

根据 L 的连续性以及 Γ 和 $\mathcal{Q}$ 的紧性，这两个集合均非空且紧。 $L(\cdot,\ q)$ 的仿射性意味着 $\mathrm{BR}_\Gamma(q)$ 是凸集，而 $L(\rho,\ \cdot)$ 的凸性意味着 $\mathrm{BR}_\mathcal{Q}(\rho)$ 是凸集。

考虑乘积集 $K:=\Gamma\times\mathcal{Q}$ 以及对应 $T\colon K\rightrightarrows K$，其定义如下

$$T(\rho,q)\ :=\ \mathrm{BR}_\Gamma(q)\times\mathrm{BR}_\mathcal{Q}(\rho).$$

对应 T 具有非空、凸且紧的值。此外，由于 L 连续，标准论证表明， T的图是闭的：如果 $(\rho_k,q_k)\ \to(\rho,q)$ 且 $(\tilde\rho_k,\tilde q_k)\ \in T(\rho_k,q_k)$，并且 $(\tilde\rho_k,\tilde q_k)\ \to(\tilde\rho,\tilde q)$，那么 L 的连续性意味着 $(\tilde\rho,\tilde q)\ \in T(\rho,q)$。在有限维空间中，闭图和紧值蕴含上半连续性。

引理1 （存在鞍点）。 存在 $(\rho^\star,q^\star)\ \in\Gamma\times\mathcal{Q}$ 使得

$$\rho^\star\in\mathrm{BR}_\Gamma(q^\star)\qquad and\qquad q^\star\in\mathrm{BR}_\mathcal{Q}(\rho^\star).$$

证明。 由于 K 非空、紧且凸，而 T 具有非空、紧且凸的值并且是上半连续的，因此卡库塔尼不动点定理适用，并给出一个满足 $(\rho^\star,q^\star)\ \in T(\rho^\star,q^\star)$ 的不动点 $(\rho^\star,q^\star)\ \in K$ 。

□

现在我们完成定理1的证明。根据不动点性质， $\rho^\star$ 在 Γ上最大化 $L(\rho,\ q^\star)$，而 $q^\star$ 在 $\mathcal{Q}$上最小化 $L(\rho^\star,\ q)$。等价地，

$$L(\rho,q^\star)\ \le\ L(\rho^\star,q^\star)\ \le\ L(\rho^\star,q)\qquad\forall\,\rho\in\Gamma,\ \forall\,q\in\mathcal{Q},$$

这确立了鞍点性质，并证明了定理的第2项。

极小极大等式立即成立。弱对偶性总是给出

$$\sup_{\rho\in\Gamma}\inf_{q\in\mathcal{Q}} L(\rho,q)\ \le\ \inf_{q\in\mathcal{Q}}\sup_{\rho\in\Gamma} L(\rho,q).$$

利用鞍点不等式，

$$\sup_{\rho\in\Gamma} L(\rho,q^\star)=L(\rho^\star,q^\star)=\inf_{q\in\mathcal{Q}} L(\rho^\star,q),$$

这意味着等号成立，并证明了第1项。

最后，定义广义熵 $H(\rho)\colon=\inf_{q\in\mathcal{Q}} L(\rho,\ \ q)$。由于 $q^\star$ 最小化 $L(\rho^\star,\ \ \cdot)$，

$$H(\rho^\star)=L(\rho^\star,q^\star).$$

对于任意 $\rho\in\Gamma$，

$$H(\rho)=\inf_q L(\rho,q)\le L(\rho,q^\star)\le L(\rho^\star,q^\star)=H(\rho^\star),$$

因此， $\rho^\star$ 在 Γ 上使 $H(\rho)$ 达到最大值。此外，

$$V_\Gamma=\inf_q\sup_\rho L(\rho,q)=\sup_\rho L(\rho,q^\star)=L(\rho^\star,q^\star)=H(\rho^\star),$$

从而证明第3项并完成证明。

A.2 冯·诺依曼熵特化的详细说明

本附录为推论1提供细节，验证量子对数损失满足定理1的假设，且其诱导的广义熵与冯·诺依曼熵一致。

令 $\mathcal{D}_n := \{\rho \succeq 0 : \mathrm{Tr}(\rho) = 1\}$ 表示 $n \times n$ 密度矩阵的集合，并令 $\Gamma \subseteq \mathcal{D}_n$ 为一个凸紧可行集。我们考虑量子对数损失

$$L_{\log}(\rho, \sigma) := -\mathrm{Tr}\big(\rho \log \sigma\big),$$

定义在 $\sigma \in \mathcal{D}_n$ 上，其中 $\sigma \succ 0$。为方便技术处理，我们将行动集限制为 $\Sigma_\varepsilon := \{\sigma \in \mathcal{D}_n : \sigma \succeq \varepsilon I\}$，其中某个 $\varepsilon > 0$，从而确保 $\log \sigma$ 有限且连续。

对于每个固定的 σ，映射 $\rho \mapsto L_{\log}(\rho, \sigma)$ 是仿射的，因为迹是线性的。对于每个固定的 ρ，映射$\sigma \mapsto L_{\log}(\rho, \sigma)$ 在 Σ_ε 上是凸的，因为标量函数 $x \mapsto -\log x$ 在 $(0, \infty)$ 上是算子凸的，并且当左乘 $\rho \succeq 0$ 时，迹保持勒夫纳序。由矩阵对数在 Σ_ε 上的连续性可知，$L_{\log}$ 在 $\Gamma \times \Sigma_\varepsilon$ 上连续。因此，定理1的所有假设均得到满足。

现在我们确定由对数损失诱导的广义熵。定义冯·诺依曼熵$S(\rho) := -\mathrm{Tr}\big(\rho \log \rho\big)$ 以及量子相对熵

$$D(\rho\|\sigma) := \mathrm{Tr}\big(\rho(\log \rho - \log \sigma)\big),$$

约定 为 $D(\rho\|\sigma) = +\infty$，若 $\mathrm{supp}(\rho) \not\subseteq \mathrm{supp}(\sigma)$。直接计算可得分解式

$$L_{\log}(\rho, \sigma) = S(\rho) + D(\rho\|\sigma),$$

只要 $D(\rho\|\sigma)$ 有限，该式便成立。

根据 Klein 不等式，量子相对熵非负，且当且仅当 $\rho = \sigma$ 时取等号。因此，对于任意 $\rho \in \mathcal{D}_n$，

$$\inf_{\sigma \in \mathcal{D}_n} L_{\log}(\rho, \sigma) = S(\rho),$$

下确界在 $\sigma = \rho$ 处取得。

*推论1的证明。*将定理1应用于自然选择 $\rho \in \Gamma$、决策者选择 $\sigma \in \Sigma_\varepsilon$，且损失为 $L_{\log}(\rho, \sigma)$ 的博弈，可得一个鞍点 $(\rho^\star, \sigma^\star)$ 以及极小极大等式。由于由 $L_{\log}$ 诱导的广义熵与冯·诺依曼熵一致，最大广义熵原理便化为最大冯·诺依曼熵原理。相应的行动 $\sigma^\star$ 是极小极大稳健的。

□

A.3 迹熵与熵–散度分解

本附录记录了一类广泛的迹熵所满足的一般熵加散度恒等式，并简要说明其与 Rényi-α 目标的关系。

在讨论中，我们固定 $\varepsilon \in (0, 1/n)$，并回顾 $\mathcal{D}_{n,\varepsilon} := \{\rho \in \mathcal{D}_n : \rho \succeq \varepsilon I\}$。令 $f : (0, \infty) \to \mathbb{R}$ 为凸且连续可微的函数。对于 $\rho \in \mathcal{D}_{n,\varepsilon}$，定义迹熵泛函

$$H_f(\rho) := -\mathrm{Tr}\big(f(\rho)\big). \tag{37}$$

对于 $\rho,\sigma \in \mathcal{D}_{n,\varepsilon}$，定义相应的矩阵布雷格曼散度

$$D_f(\rho\|\sigma) := \mathrm{Tr}\Big(f(\rho) - f(\sigma) - f'(\sigma)\big(\rho - \sigma\big)\Big). \tag{38}$$

非负性源于谱函数 $F(\rho) := \mathrm{Tr}(f(\rho))$ 在厄米矩阵上的凸性：对于凸可微的 F，有

$$F(\rho) \geq F(\sigma) + \mathrm{Tr}\big(f'(\sigma)(\rho - \sigma)\big),$$

这等价于 $D_f(\rho\|\sigma) \geq 0$。

接下来，定义相应的损失

$$L_f(\rho, \sigma) := -\mathrm{Tr}\Big(f(\sigma) + f'(\sigma)\big(\rho - \sigma\big)\Big), \qquad \rho, \sigma \in \mathcal{D}_{n,\varepsilon}. \tag{39}$$

对于每个固定的 σ，映射 $\rho \mapsto L_f(\rho, \sigma)$ 是仿射的。

命题 1. 对于所有 $\rho,\ \sigma \in \mathcal{D}_{n,\ \varepsilon}$，以下结论成立

$$L_f(\rho,\sigma) = H_f(\rho) + D_f\big(\rho\|\sigma\big). \tag{40}$$

因此，下述结论成立，其中下确界在 $\sigma = \rho$ 处取得：

$$\inf_{\sigma\in\mathcal{D}_{n,\varepsilon}} L_f(\rho,\sigma) = H_f(\rho) \tag{41}$$

证明。 由方程 37、方程 38 和方程 39，

$$H_f(\rho) + D_f(\rho\|\sigma) = -\mathrm{Tr}\big(f(\rho)\big) + \mathrm{Tr}\Big(f(\rho) - f(\sigma) - f'(\sigma)\big(\rho-\sigma\big)\Big) = -\mathrm{Tr}\Big(f(\sigma) + f'(\sigma)\big(\rho-\sigma\big)\Big) = L_f(\rho,\sigma).$$

贝叶斯风险恒等式方程 41 由 $D_f(\rho\|\sigma) \geq 0$ 得出，且在 $\sigma = \rho$ 处取等。 □

两个特殊情形。 对于 $f(t) = t\log\ t$，我们有 $H_f(\rho) = -\mathrm{Tr}\big(\rho\log\ \rho\big) = S(\rho)$ 以及 $D_f(\rho\|\sigma) = D(\rho\|\sigma)$（量子相对熵），且方程 40 恢复了标准恒等式 $-\mathrm{Tr}(\rho\log\ \sigma) = S(\rho) + D(\rho\|\sigma)$。对于 $f(t) = t^2$，我们得到 $H_f(\rho) = -\mathrm{Tr}(\rho^2)$ 以及 $D_f(\rho\|\sigma) = \mathrm{Tr}\big((\rho-\sigma)^2\big)$，从而得到正文中使用的二次（2 阶）代理。

与 Rényi 的关系α。对于 $\alpha > 0$、 $\alpha \neq 1$，(Petz) Rényi-α 熵为

$$S_\alpha(\rho) \ := \ \frac{1}{1-\alpha}\log \mathrm{Tr}\big(\rho^\alpha\big). \tag{42}$$

由于 $\log(\cdot)$ 是递增函数，且 $\frac{1}{1-\alpha}$ 在 $\alpha = 1$ 处变号，因此，在可行集 Γ 上最大化 $S_\alpha(\rho)$，等价于对 $0 < \alpha < 1$ 最大化 $\mathrm{Tr}(\rho^\alpha)$，也等价于对 $\alpha > 1$ 最小化 $\mathrm{Tr}(\rho^\alpha)$。因此，可以将最大 Rényi-α 选择视为通过迹幂代理 $H_f(\rho) = -\mathrm{Tr}(f(\rho))$ 进行，其中 $f(t) = t^\alpha$（符号取决于 α），命题 1 适用于该情形。

A.4 线性约束与等化器解

本附录给出一个简单的充分条件，在该条件下，最大冯·诺依曼熵（Max-VNE）解满足等化器性质，即针对所选行动的对数损失在整个可行集上为常数。

设 A_1、……、$A_m \in \mathbb{S}^n$ 和 $\tau \in \mathbb{R}^m$，并定义仿射约束集

$$\Gamma_\tau := \Big\{\rho \in \mathcal{D}_n :\ \mathrm{Tr}\big(\rho A_j\big) = \tau_j,\ j = 1,\ldots,m\Big\}.$$

定理 2。 *假设 $\Gamma_\tau \neq \emptyset$，且存在满秩 $\rho_\tau \in \Gamma_\tau$ 和标量 c、β_1、…、β_m，使得*

$$\log \rho_\tau = cI + \sum_{j=1}^{m} \beta_j A_j. \tag{43}$$

那么，对于量子对数损失 $L_{\log}(\rho,\sigma) = -\mathrm{Tr}\big(\rho\log\ \sigma\big)$：

1. *值 $L_{\log}(\rho,\rho_\tau)$ 对所有 $\rho \in \Gamma_\tau$ 都是常数。*
2. *$\rho_\tau \in \arg\max_{\rho\in\Gamma_\tau} S(\rho)$.*
3. *(ρ_τ,ρ_τ) 是极小极大博弈 $\inf_{\sigma\in\mathcal{D}_n}\sup_{\rho\in\Gamma_\tau} L_{\log}(\rho,\sigma)$的鞍点，因此 ρ_τ是极小极大最优的。*

证明。 对于任意 $\rho \in \Gamma_\tau$，利用公式 43 和约束 $\mathrm{Tr}(\rho A_j) = \tau_j$，

$$L_{\log}(\rho,\rho_\tau) = -\mathrm{Tr}\big(\rho\log\rho_\tau\big) = -\mathrm{Tr}\Big(\rho\Big(cI + \sum_{j=1}^{m}\beta_j A_j\Big)\Big) = -\Big(c + \sum_{j=1}^{m}\beta_j\tau_j\Big),$$

它与 ρ 无关。这就证明了第 1 项。

接下来，我们回顾 $\rho \in \mathcal{D}_n$ 和满秩 $\sigma \in \mathcal{D}_n$的标准分解：

$$L_{\log}(\rho, \sigma) = S(\rho) + D(\rho\|\sigma),$$

其中，$S(\rho) = -\mathrm{Tr}(\rho \log \rho)$ 是冯·诺依曼熵，而 $D(\rho\|\sigma) = \mathrm{Tr}\big(\rho(\log \rho - \log \sigma)\big) \geq 0$ 是量子相对熵。将其应用于 $\sigma = \rho_\tau$，可得对所有 $\rho \in \Gamma_\tau$,

$$L_{\log}(\rho, \rho_\tau) = S(\rho) + D(\rho\|\rho_\tau) \geq S(\rho).$$

由条目 1，$L_{\log}(\rho, \rho_\tau) = L_{\log}(\rho_\tau, \rho_\tau) = S(\rho_\tau)$，因此对所有 $\rho \in \Gamma_\tau$，$S(\rho) \leq S(\rho_\tau)$，从而证明条目 2。最后，条目 1 得出

$$\sup_{\rho \in \Gamma_\tau} L_{\log}(\rho, \rho_\tau) = L_{\log}(\rho_\tau, \rho_\tau) = S(\rho_\tau).$$

另一方面，对于任意 $\sigma \in \mathcal{D}_n$,

$$\sup_{\rho \in \Gamma_\tau} L_{\log}(\rho, \sigma) \geq L_{\log}(\rho_\tau, \sigma) = S(\rho_\tau) + D(\rho_\tau\|\sigma) \geq S(\rho_\tau),$$

因此 ρ_τ 达到极小极大值，且 (ρ_τ, ρ_τ) 满足鞍点不等式。这证明了条目 3. □

A.5 最大 条件冯·诺依曼熵与极小极大解释

本附录在经典-量子情形下探讨最大熵原理的一种条件形式，其中自然选择矩阵值条件态。本文的论述与 [13],的极小极大观点相呼应，关键区别在于以条件密度矩阵取代条件分布。

设 $\mathcal{X}$ 为有限字母表，且 $p \in \Delta(\mathcal{X})$ 是决策者（DM）已知的固定分布。固定 $\varepsilon \in (0, 1/n)$，并定义 $\mathcal{D}_{n,\,\varepsilon} := \{\rho \succeq 0 : \mathrm{Tr}(\rho) = 1, \quad \rho \succeq \varepsilon I\}$。经典-量子态由 $\rho = \{\rho_x\}_{x \in \mathcal{X}}$ 指定，其中$\rho_x \in \mathcal{D}_{n,\,\varepsilon}$，对应的分块对角矩阵为

$$\rho_{XQ} = \sum_{x \in \mathcal{X}} p(x)\, |x\rangle\langle x| \otimes \rho_x.$$

设 Γ 为这类经典-量子态的一个非空、凸且紧的集合。决策者选择行动 $\sigma = \{\sigma_x\}_{x \in \mathcal{X}}$，其中 $\sigma_x \in \mathcal{D}_{n,\,\varepsilon}$，即行动集为 $\mathcal{Q} = (\mathcal{D}_{n,\,\varepsilon})^{|\mathcal{X}|}$。我们考虑如下定义的条件对数损失：

$$L_{\log}(\rho, \sigma) := \sum_{x \in \mathcal{X}} p(x) \left[-\mathrm{Tr}\big(\rho_x \log \sigma_x\big)\right], \tag{44}$$

根据限制条件 $\sigma_x \succeq \varepsilon I$，其在 $\Gamma \times \mathcal{Q}$ 上有限且连续。

损失诱导的条件熵。 对于 $\Gamma \times \mathcal{Q}$ 上的一般损失 $L(\rho, \sigma)$，定义广义（条件）熵和极小极大值如下

$$H(\rho) := \inf_{\sigma \in \mathcal{Q}} L(\rho, \sigma), \qquad V_\Gamma := \inf_{\sigma \in \mathcal{Q}} \sup_{\rho \in \Gamma} L(\rho, \sigma).$$

对于条件对数损失，最小化可分解为关于 x 的逐项最小化：

$$\inf_{\sigma \in \mathcal{Q}} L_{\log}(\rho, \sigma) = \sum_{x \in \mathcal{X}} p(x) \inf_{\sigma_x \in \mathcal{D}_{n,\varepsilon}} \left[-\mathrm{Tr}(\rho_x \log \sigma_x)\right].$$

利用标准恒等式

$$-\mathrm{Tr}(\rho_x \log \sigma_x) = S(\rho_x) + D(\rho_x\|\sigma_x),$$

其中 $S(\rho_x) := -\mathrm{Tr}(\rho_x \log \rho_x)$，而 $D(\cdot\|\cdot) \geq 0$ 是量子相对熵，我们得到

$$H(\rho) = \sum_{x \in \mathcal{X}} p(x)\, S(\rho_x) =: S(Q|X)_\rho, \tag{45}$$

并且对于所有 x，下确界在 $\sigma_x = \rho_x$ 处取得。因此，由条件对数损失诱导的广义熵与经典-量子态的条件冯·诺依曼熵一致。

最大条件熵与极小极大最优性。考虑极小极大博弈 下确界$_{\sigma\in\mathcal{Q}}$ 上确界$_{\rho\in\Gamma}$ $L_{\log}(\rho,\sigma)$。
我们旨在将 $\rho=\{\rho_x\}$ 与有限维乘积空间 $(\mathbb{S}^n)^{|\mathcal{X}|}$中的一个元素对应起来
，并为其赋予加权配对

$$\langle \rho,\rho'\rangle := \sum_{x\in\mathcal{X}} p(x)\,\mathrm{Tr}\big(\rho_x\rho'_x\big),$$

对 σ也作类似处理。在此对应关系下， Γ 和 $\mathcal{Q}$ 仍然是凸且紧的，而 $L_{\log}$ 关于 ρ是仿射的，关于 σ是凸的。将定理1应用于这个乘积空间博弈，可得极小极大等式、鞍点（$\rho^\star$， $\sigma^\star$）的存在性，以及熵刻画

$$\rho^\star \in \arg\max_{\rho\in\Gamma} H(\rho) = \arg\max_{\rho\in\Gamma} S(Q|X)_\rho.$$

具体而言，最大化条件冯·诺依曼熵会在极小极大博弈中选出自然的均衡态，而相应的行动 $\sigma^\star$ 是极小极大稳健的。

注 1. 对于二次型（二阶）这一特殊情形，我们定义二次条件损失

$$L_2(\rho,\sigma) := \sum_{x\in\mathcal{X}} p(x)\,\big[\mathrm{Tr}(\sigma_x^2) - 2\,\mathrm{Tr}(\rho_x\sigma_x)\big].$$

配方可得，对每个 x,

$$\mathrm{Tr}(\sigma_x^2) - 2\,\mathrm{Tr}(\rho_x\sigma_x) = \mathrm{Tr}\big((\sigma_x-\rho_x)^2\big) - \mathrm{Tr}(\rho_x^2).$$

因此，贝叶斯风险为

$$\inf_{\sigma\in\mathcal{Q}} L_2(\rho,\sigma) = -\sum_{x\in\mathcal{X}} p(x)\,\mathrm{Tr}(\rho_x^2),$$

其中最优值在 $\sigma_x=\rho_x$ 处取得。由于 $S_2(\rho_x) = -\log \mathrm{Tr}(\rho_x^2)$，最大化 $\sum_x p(x)S_2(\rho_x)$ 等价于最大化上述广义熵，并且同样的极小极大解释适用。

附录 B 其他相关工作

除正文所讨论的工作之外，更广泛的文献研究了现代机器学习中用于学习、推断与评估的熵与信息论原理。具体而言，基于熵的目标已被用于通过量化学习得到的分布的覆盖度与多样性来评估和比较生成模型。利用熵和受互信息启发的准则对多模态及生成模型进行的信息论评估在 [6], 中提出，而对基于熵的新颖性与多样性度量的可解释分析则在 [18] 中给出。相关的无参考且可扩展的生成模型评估框架进一步强调了谱与基于核的信息度量 [7]。

越来越多的工作将冯·诺依曼熵与基于核的信息度量扩展到提示感知与条件生成建模。利用核熵与 Rényi 型目标进行的提示感知多样性与新颖性引导在 [34], 中得到研究，而通过嵌入协方差结构的舒尔补分解进行的提示感知多样性评估则在 [35]中提出。对基于提示的生成模型中多样性评估的信息论处理在 [36] 中得到进一步探讨。这些工作凸显了条件熵概念与依赖提示的熵概念的重要性，从而补充了正文所讨论的条件视角与极小极大视角。

谱熵度量也出现在图与网络分析中。冯·诺依曼图熵在归一化图拉普拉斯矩阵上定义了一种熵泛函，并被用于量化图复杂度、结构变化和动力学，同时还提供了适用于大型图的可扩展近似 [37]。这条研究路线进一步强化了这样一种观点：将归一化半正定矩阵视为类密度矩阵对象，并直接在谱层面研究熵。

最后，熵与多样性原理自然关联到混合建模以及对多个模型或数据源的聚合。多样性感知混合选择与聚合方法见于学习与决策场景，包括旨在构建比任何单一分量更具多样性的混合的基于老虎机的方法 [38]。这些方法与本文所研究的最大冯·诺依曼熵原理相辅相成；该原理为在部分信息条件下选择混合与核表示提供了静态的信息论依据。

附录 C：MAX-VNE 选定的嵌入混合的其他数值结果

C.1 实现细节

为选择混合权重，我们在 ImageNet-1K 训练集的一个固定子集上构建四个 RBF 核矩阵，分别对应 OpenCLIP、DINOv2、SigLIP 和 UNICOM 嵌入。具体而言，我们随机抽取 200 个类别，每个类别选取 50 张图像，共计 10,000 张图像。径向基函数核的带宽参数针对每个模型单独选取。特别地，我们选择带宽，使得所得核矩阵在单独考虑时具有可比的冯·诺依曼熵值。这一设计选择避免了因核的尖锐程度或尺度效应而使混合优化偏向某一特定模型。混合权重 α 通过投影梯度上升（PGA）最大化冯·诺依曼熵而获得。在每次迭代中，先解析计算关于 α 的梯度，随后将其欧几里得投影到单纯形上。优化以均匀权重初始化，并使用固定学习率 0.1，最多迭代 100 次。

C.1.1 基于 Max-VN 的线性核混合的其他线性探测结果 E

除正文报告的基于 RBF 核的混合结果外，我们还评估了使用线性核得到的 Max-VNE 选定混合。具体而言，对于每个预训练嵌入，我们构建一个线性核，并应用相同的 Max-VNE 优化方案来选择混合权重$\alpha^\star$。所得混合表示通过将每个嵌入按相应权重缩放并拼接而成。随后，使用与正文完全相同的评估方案进行线性探测，包括相同的优化器、学习率、批量大小和训练轮数。其他结果见表 2，结果显示，基于线性核的 Max-VNE 选定混合取得了具有竞争力的性能，表明所提出的混合选择策略对具体的核选择并不敏感。

表 2：嵌入与 Max-VNE 选定的嵌入混合的线性探测准确率（%，5 次运行的均值 ±标准差）。

模型	ImageNet	CIFAR100	DTD	Aircraft
OpenCLIP	78.9 ± 0.14	86.3 ± 0.02	77.7 ± 0.23	49.2 ± 0.39
SigLIP	81.8 ± 0.16	83.3 ± 0.04	80.3 ± 0.05	60.7 ± 0.41
DINOv2	80.0 ± 0.19	89.9 ± 0.13	79.9 ± 0.22	58.9 ± 0.63
UNICOM	77.4 ± 0.08	85.0 ± 0.04	71.6 ± 0.53	55.4 ± 0.85
Max-VNE	83.1 ± 0.09	91.2 ± 0.09	82.3 ± 0.13	63.2 ± 0.37

C.2 Max-VNE 核补全的其他数值结果

C.2.1 实现细节

在主要实验中，我们使用二阶 Rényi 熵目标函数实现 Max-VNE 原理，该目标函数对应于最小化恢复得到的密度矩阵的纯度。为确保能够扩展到大型数据集，我们使用可学习矩阵$U \in \mathbb{R}^{n\times r}$的低秩分解对补全后的核进行参数化，从而通过构造保证其半正定性。这种分解避免了对完整的 $n \times n$ 矩阵进行显式优化，并能实现高效计算。我们使用基于梯度的方法求解优化问题。具体而言，我们使用 Adam 优化秩为 $r = 50$ 的低秩因子，采用固定学习率 1×10^{-3}，直至收敛。在消融研究中，我们还报告了最大化冯 · 诺依曼熵，以及改变分解秩和观测比例时得到的结果。

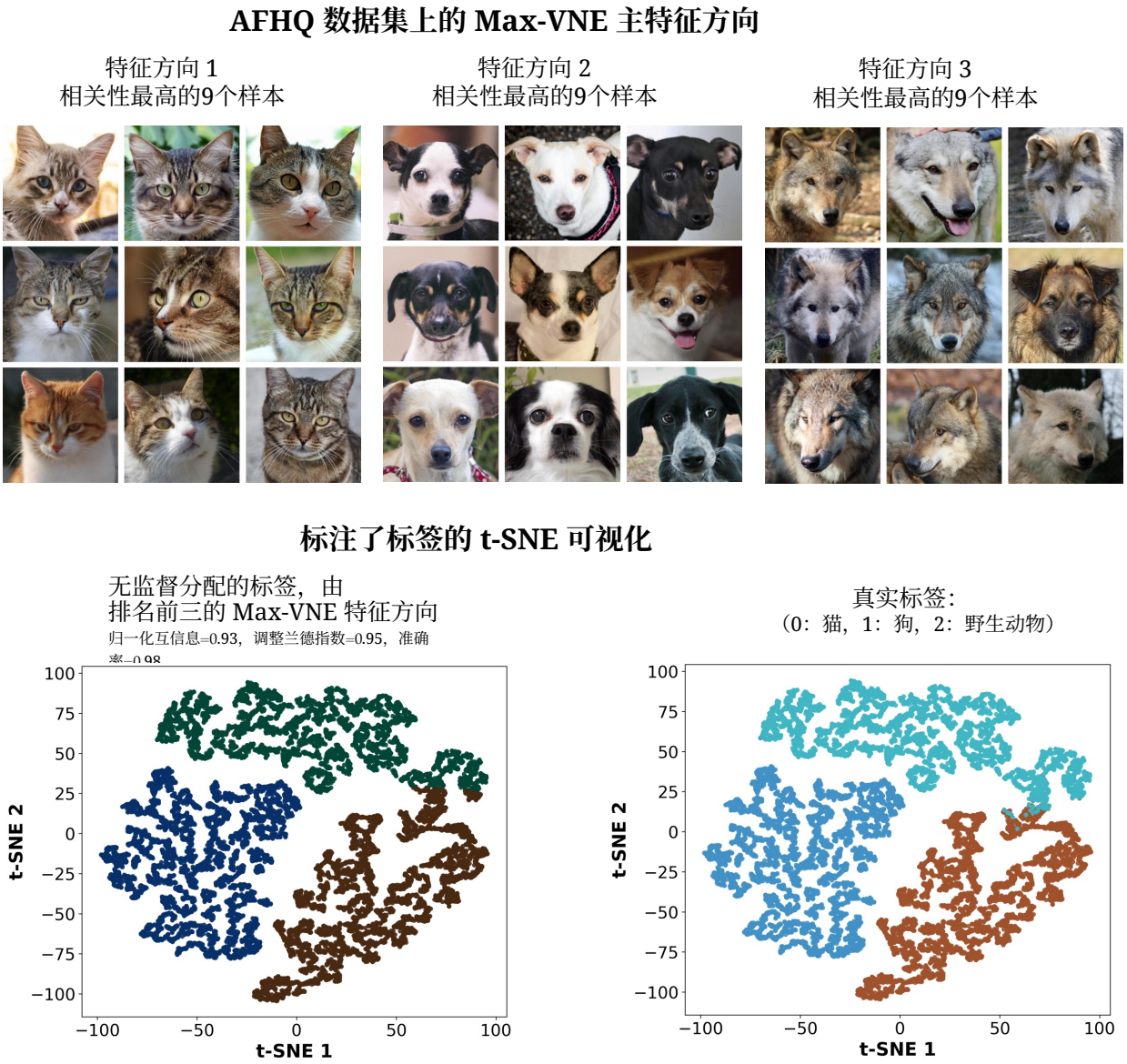


图 3：在 AFHQ 数据集上，使用 10% 的观测条目进行 Max-VNE 核补全的实验结果。使用归一化互信息（NMI）、调整兰德指数（ARI）和准确率（ACC）评估恢复的核矩阵的聚类性能。

C.2.2 最大化冯·诺依曼熵的其他结果

除了正文中使用的二阶Rényi熵目标外，我们还报告了直接最大化冯·诺依曼熵进行核矩阵补全所获得的补充结果。沿用相同的实验设置，将核矩阵参数化为低秩形式$\mathbf{K} = \mathbf{U}\mathbf{U}^\top$，并将其归一化为单位迹。我们使用定义在所得低维协方差上的对数行列式替代函数来优化冯·诺依曼熵，从而避免对完整核矩阵进行显式特征分解。如图 3 所示，恢复出的核在 AFHQ 上得到的聚类结果，无论从定性还是定量角度来看，都与使用二阶目标所获得的结果相当，这表明观察到的行为对熵的形式选择具有稳健性。

C.2.3 消融研究

更具挑战性数据集上的其他结果 我们进一步在 ImageNet-Dogs 数据集上评估所提出的核矩阵补全方法。[30], 该数据集因犬种之间存在细粒度的视觉差异而被视为尤其具有挑战性的基准测试。与 AFHQ 或 MNIST 等数据集相比，ImageNet-Dogs 的类内差异显著更大，且类间差异细微，因此从部分观测的核中进行谱恢复要困难得多。按照正文中的相同实验方案，我们构建了一个密度核，并随机屏蔽了 90% 的非对角元素。图 4 的结果表明，即使在极具挑战性的细粒度场景中，Max-VNE 原理在严重部分观测的情况下仍能够恢复一些有意义的谱结构。

分解秩消融实验 r 我们研究了核矩阵补全中所用分解秩 r 的影响。我们在 AFHQ 数据集上进行实验，固定掩码比例为 90%，同时改变秩 $r \in \{10, 20, 50\}$。其他优化设置均保持不变。从图 5 中可以看到，即使秩很小，恢复性能和下游谱聚类结果也已趋于稳定。

观测比例消融实验。我们在 AFHQ 数据集上研究观测比例的影响，分别随机观测 1%、5% 和 10% 的非对角元素，同时保持其他所有设置不变。从图 6 中可以看到，观测比例从 1% 增至 5% 时，恢复性能有所提升，此后趋于稳定。

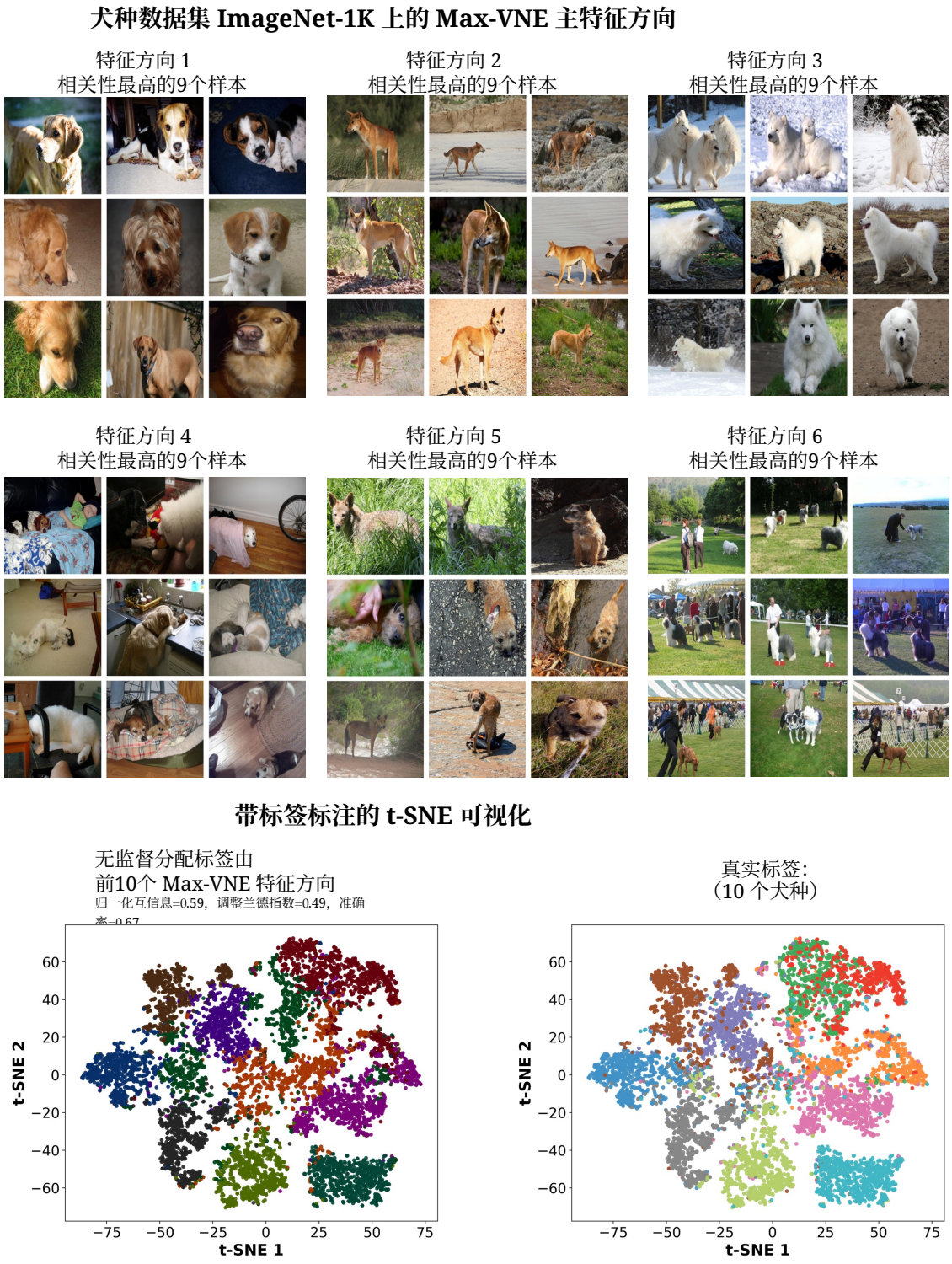


图 4：在 ImageNet-1K 犬种数据集上，基于 10% 已观测条目进行 Max-VNE 核补全的实验结果。使用归一化互信息（归一化互信息）、调整兰德指数（调整兰德指数）和准确率（准确率）评估恢复的核矩阵的聚类性能。

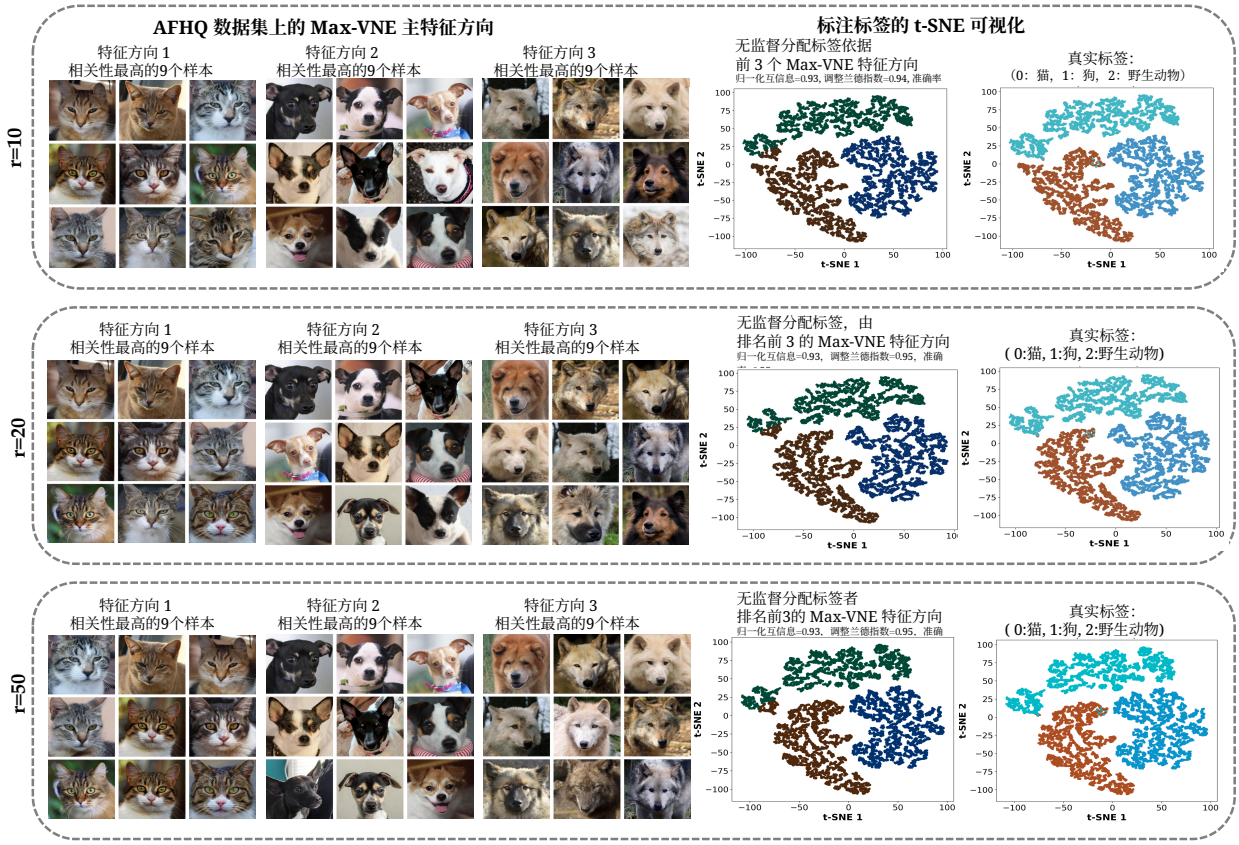


图 5：Max-VNE 核补全的分解秩消融实验 r 在 AFHQ 数据集上进行，观测条目占 10%。使用归一化互信息（NMI）、调整兰德指数（ARI）和聚类准确率（ACC）评估恢复的核矩阵的聚类性能。

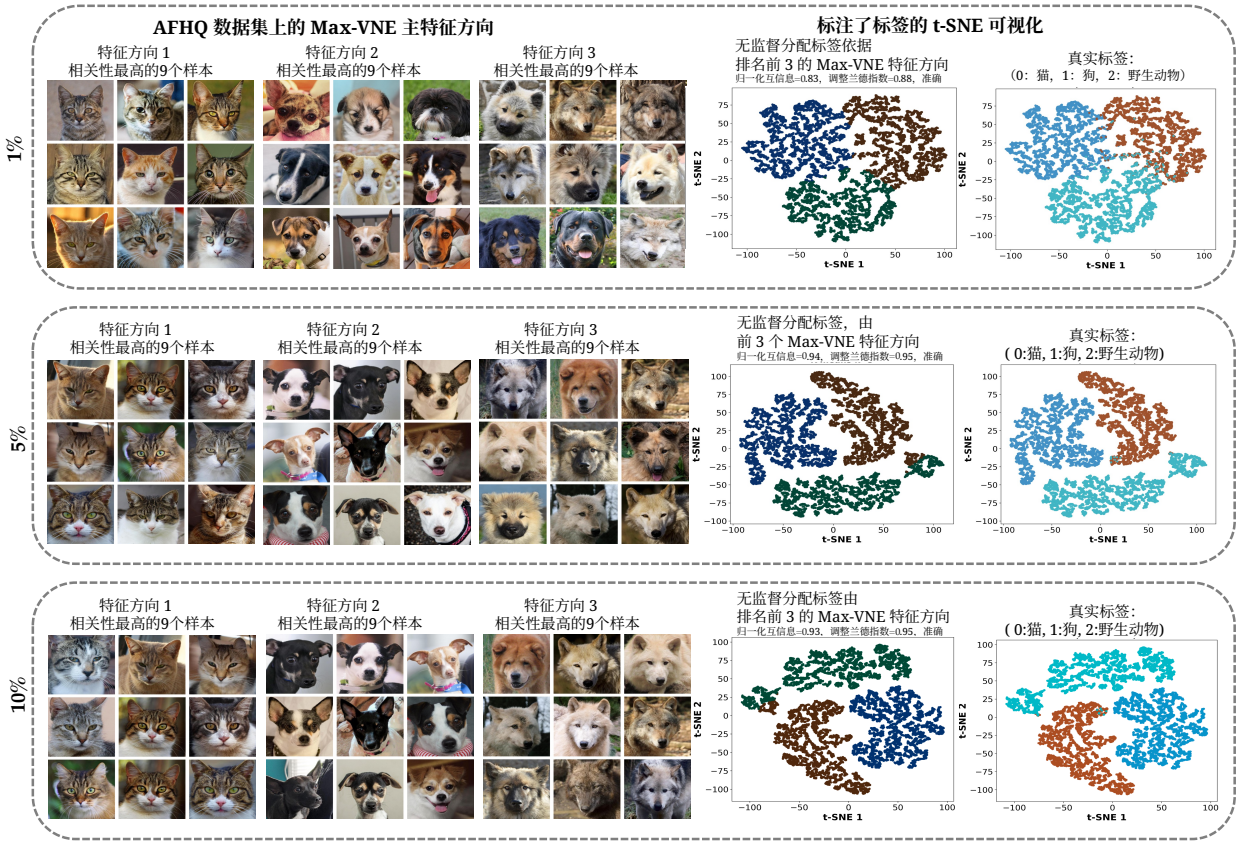


图 6：在分解秩固定为 50 的情况下，对 AFHQ 数据集上核矩阵的不同观测比例进行消融实验。使用归一化互信息（NMI）、调整兰德指数（ARI）和聚类准确率（ACC）评估恢复的核矩阵的聚类性能。