面向速率自适应数字语义通信的零样本知识库调整

ShuminYao[1],Hui Du[1], Lifeng Xie[1], *, Yaping Sun[1], Hao Chen[1],Nan Ma[2,1], IEEE会员，及 Xiaodong Xu[2,1], IEEE高级会员[1]鹏城实验室宽带通信系，中国深圳518066[2]北京邮电大学网络与交换技术国家重点实验室，中国北京100876 {yaoshm, duh, xielf, sunyp,chenh03}@pcl.ac.cn, {manan, xuxiaodong}@bupt.edu.cn

arXiv:2602.01829v1 [cs.IT] 2 Feb 2026

摘要——数字语义通信系统通常采用矢量量化变分自编码器（VQ-VAE）框架，是未来无线网络的关键技术。在基于VQ-VAE的语义通信系统中，传输速率由一种称为知识库（KB）的离散码本大小直接决定。然而，KB大小是固定的超参数，这意味着调整速率需要针对每种所需大小分别训练并存储一个模型——这种做法在计算和存储方面成本过高，无法实现真正精细的速率控制。为解决这一问题，我们提出一种有理论依据的零样本知识库调整方法，无需重新训练即可实现即时速率自适应。我们通过揭示单个大型父知识库中固有的语义层次结构，为其中所有向量建立全局重要性排序。具体而言，该方法采用三步框架：1）将KB向量嵌入双曲空间，以揭示其层次关系；2）利用最小生成树算法构建主语义树；3）通过迭代剪除重要性最低的叶节点，实现即时调整大小。大量仿真表明，与从头训练的专用KB相比，我们的方法能够达到几乎相同的重建质量，而计算预算仅为其一小部分。此外，在传统KB会发生灾难性失效的极低速率下，我们的方法也表现出更强的鲁棒性。本研究解决了基于VQ-VAE的语义通信系统的一项根本性局限，为灵活且速率自适应的语义通信提供了一条实用、高效的途径。

索引词—语义通信、知识库、零样本调整大小

I. 引言

语义通信正逐渐成为一种新范式，有望超越传统的比特级保真度，转而通过优化意义与相关性的传输来实现这一目标，这是应对未来无线网络频谱稀缺问题的关键一步 [1]。为确保与现有数字基础设施兼容，近期研究主要聚焦于数字语义通信系统，这类系统通常基于矢量量化变分自编码器（VQ-VAE）框架构建 [2]。该框架建立了一个共享的离散语义知识库（KB），也称为码本，其中包含一组有限的代表性向量。编码器从源消息中提取连续的语义特征张量后，会执行矢量量化：将张量中的每个特征向量单独映射到一个离散索引该索引对应于知识库中的一个特定向量，该向量是与之最匹配的向量。接收端的解码器随后利用这些传输的索引检索相应的向量，并重建原始消息 [3]。

这种基于矢量量化的方法依赖于传输索引，因此存在一个关键权衡，而这一权衡直接受知识库大小（即向量总数）的影响 K。一方面，具有较高 K 值的大型知识库能够实现更精细的语义表示，从而最大限度地降低量化误差，并显著提升重建质量。另一方面，这种保真度的提升会直接增加传输速率，因为表示每个索引都需要 $\lceil \log_2 K \rceil$ 位。因此，较大的 K 会增加数据载荷总量，需要更多信道带宽。这种固有的矛盾意味着，最优知识库大小并非固定不变；它应动态调整，以平衡重建质量与可用信道资源。然而，这种适应性从根本上与VQ-VAE框架的设计不相容。知识库大小 K是一个在训练前确定的固定超参数，而这种架构上的僵化是所有基于VQ-VAE构建的数字语义通信系统（例如 [3], [4]）所固有的局限。因此，通过改变 K 来调整传输速率，就需要针对每种可能的大小分别训练并存储一个完整模型——这种做法在计算和存储方面的开销都难以承受。

为解决这一局限性，一些研究提出通过从一组预定义的离散知识库配置中进行选择，来近似实现灵活的速率自适应。例如，一些方法通过改变单个大型知识库中处于激活状态的子知识库数量[5] ，或使用其层级数量可变的层次化结构 [6]，来实现不同的速率。另一种方法是训练多个彼此不同的知识库，每个知识库针对率失真曲线上的某个不同点进行定制，然后在推理过程中在它们之间切换 [7]。然而，这种对预定义离散知识库配置的依赖意味着，系统从根本上受限于少量固定的速率等级。尽管近期已有一种无需重新训练的调整大小技术被提出 [8], ，但其核心机制只是一种计算捷径，无法降低实际的

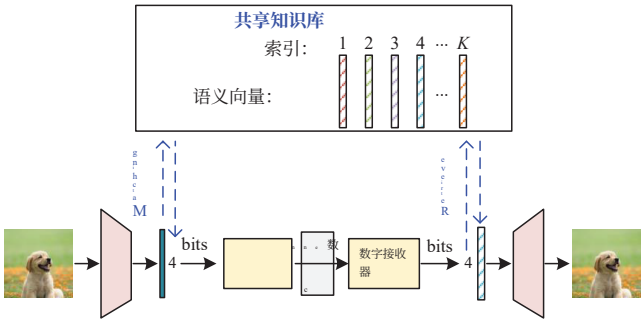


图1. 所考虑的基于 VQ-VAE 的数字语义通信框架。

传输速率。因此，如何将知识库重构为任意大小，以实现真正细粒度的即时自适应，仍是一个亟待解决的重要问题。

本文提出了一种新颖的零样本知识库调整方法，可实现真正的速率自适应语义通信。我们的核心思路是为大型知识库中的所有向量建立全局重要性排序，从而只需保留指定数量的最重要向量，即可构建一个新的、更小的知识库。这里，全局重要性排序的建立利用了这样一个原理：经过良好训练的知识库中的语义关系本质上具有层次结构。我们的贡献总结如下。

1）我们提出了一个完整的三步零样本知识库调整框架。该方法首先利用双曲嵌入揭示知识库向量的内在层次结构，随后通过最小生成树（MST）构建主语义树，以明确呈现这种层次结构，最后通过迭代剪除树的叶节点，实现对任意目标大小的即时调整。

2）我们进行了大量仿真，以证明所提方法的有效性。结果表明，在广泛的知识库规模范围内，我们的零样本调整方法所达到的重建质量几乎与从头训练的专用模型相同，而所需训练预算仅为后者的一小部分。此外，我们还发现，在极低速率下，我们的方法具有更强的鲁棒性，且质量退化更加平缓；而基线方法则会发生灾难性失效。

本文其余部分组织如下。第二节详细介绍基于VQ-VAE的数字语义通信系统模型。第三节提出我们所提出的零样本知识库调整方法，包括其三步框架。在第四节中，我们通过大量仿真评估该方法的性能。最后，第五节对全文进行总结。

II. 系统模型

本研究考虑部署于数字硬件上的语义通信系统，其中信息最终以二进制比特进行处理和传输。系统架构建立在VQ-VAE框架[2], 之上，该框架是端到端数字语义信息传输的代表性模型。如图1所示，系统由发送端和接收端组成，二者共享一个共同的预训练语义知识库（SKB）。

具体而言，发送端采用基于深度神经网络（DNN）的语义编码器，$\mathcal{E}_{\theta_{\text{enc}}}(\cdot)$，其可训练参数为 θ_{enc}。接收端配备相应的基于 DNN 的语义解码器，$\mathcal{D}_{\theta_{\text{dec}}}(\cdot)$，其参数为 θ_{dec}。两端均存储一个共享 SKB，表示为由 K个嵌入向量 $\mathcal{S} = \{\mathbf{s}_k\}_{k=1}^{K}$组成的集合，其中每个 $\mathbf{s}_k \in \mathbb{R}^{C'}$。

为清晰起见，我们聚焦于图像传输场景。设 $\mathbf{I} \in \mathbb{R}^{H\times W\times C}$ 为一幅源图像。发送端首先使用语义编码器提取张量 $\mathbf{X} \in \mathbb{R}^{H'\times W'\times C'}$:

$$\mathbf{X} = \mathcal{E}_{\theta_{\text{enc}}}(\mathbf{I}). \tag{1}$$

其中，$H' < H$和 $W' < W$表示空间维度的语义压缩。

接下来，对于张量 $\mathbf{X}$中的每个向量 $\mathbf{x}_{h',w'} \in \mathbb{R}^{C'}$，发送端执行向量量化。它使用 ℓ_2-范数距离，在 $\mathcal{S}$ 中找到最近的向量，并记录其索引:

$$z_{h',w'} = \underset{k\in\{1,\dots,K\}}{\operatorname{argmin}} ||\mathbf{x}_{h',w'} - \mathbf{s}_k||_2. \tag{2}$$

此过程将连续特征张量 $\mathbf{X}$ 映射为由索引组成的离散整数张量 $\mathbf{Z}_{\text{idx}} \in \mathbb{Z}^{H'\times W'}$，其中每个元素都是 1 到 K 之间的整数。这一量化过程使系统天然具备数字化特性。

随后将索引张量 $\mathbf{Z}_{\text{idx}}$传输至接收端。传输时，每个索引 $z_{h',w'}$ 都用长度为 B 位的二进制码字表示，其中

$$B = \lceil \log_2 K \rceil. \tag{3}$$

因此，该图像的总传输开销为$H' \times W' \times B$位，不考虑任何额外的信道编码。

接收端收到 $\mathbf{Z}_{\text{idx}}$ 后（为定义该模型，我们假设其无差错），会执行查找操作。它通过将每个索引$z_{h',w'}$ 替换为共享 SKB 中对应的向量，重建量化张量 $\mathbf{X}_q \in \mathbb{R}^{H'\times W'\times C'}$: $(\mathbf{X}q)_{h',w'} = \mathbf{s}^{z_{h',w'}}$。最后，语义解码器使用该量化张量重建源图像:

$$\hat{\mathbf{I}} = \mathcal{D}_{\theta_{\text{dec}}}(\mathbf{X}_q). \tag{4}$$

知识库的大小 K 决定了系统性能中的一项根本性权衡。较大的 K 能够实现更精细的语义表示，减少量化ˆ误差，并提升$\mathbf{I}$的重建质量。然而，这种保真度的提升也会带来巨大的代价。首先，传输载荷会随 $B = \lceil \log_2 K \rceil$ 的增加而增大

每索引比特数，从而需要更多信道资源，例如更大的带宽或更长的传输时间。其次，发送端的计算负担会加重，因为式 (2) 中的最近邻搜索需要对每个特征向量进行 $O(K)$ 次距离计算，导致处理时延和能耗增加。相反，较小的 K 能够同时降低传输和计算开销，但代价是语义量化更粗糙，重建质量也会下降。因此， K 的理想取值并非固定不变，而是取决于信道条件和硬件约束等动态因素。遗憾的是，现有的 VQ-VAE 框架不具备这一能力：传统方法需要针对每种可能的知识库大小分别训练并存储模型，因存储和计算成本过高而不切实际。为解决这一根本局限，本文提出了零样本调整方法。

III. 零样本知识库调整

本节详细介绍我们提出的零样本知识库调整方法。我们的核心思路是从一个规模大、训练充分的知识库中，选取一个有原则的向量子集，使其能够根据通信系统的需求，动态适配任意更小的规模——整个过程无需耗费成本进行重新训练或重新优化。因此，关键挑战在于如何选择向量，以确保得到的较小知识库尽可能全面地涵盖各类知识，从而尽可能有效地支持语义通信。

为应对这一挑战，我们利用了一个原理：训练充分的知识库中的语义关系本质上具有层次结构。这一原理体现了语义信息的一项基本属性：宽泛类别自然包含更具体的概念，这种结构在语言和视觉中均显而易见 [9]。这种层次结构使我们能够以有原则的方式调整大小：优先保留界定整体知识结构的粗粒度基础概念，然后在目标大小允许的范围内尽可能纳入细粒度的具体概念。为此，如图 2 所示，我们的方法分为三个关键步骤。首先，我们将大型预训练知识库（大小为 K）中的向量从其原生欧几里得空间嵌入双曲空间；双曲空间的几何特性尤其适合以最小失真表示层次结构。其次，我们基于双曲距离构建最小生成树，以此构造主语义树，从而通过对该树进行确定性剪枝，实现对任意目标大小 $K < K$ 的零样本调整。最后，将剪枝后树中的向量从双曲空间映射回原始欧几里得空间，生成最终可直接使用的知识库。本节余下部分将详细介绍这三个步骤。

A. 用于揭示语义粒度的双曲嵌入

令 $\mathcal{S} = \{\mathbf{s}_{\hat{k}}\}_{\hat{k}=1}^{K}$ 表示大型预训练知识库中的向量集合。我们的第一步是揭示 $\mathcal{S}$ 内部的语义层次结构：将其中的向量嵌入能够自然表征此类结构的空间。我们认为，语义概念之间的关系本质上具有层次性；与欧几里得空间 [9], [10]相比，双曲几何能够以更低的失真更忠实地表示这一特性。将知识库映射到庞加莱球模型后，我们构建出一种几何布局，其中每个向量的位置都反映其对应的语义粒度：语义较粗、具有基础性的向量位于球体中心附近，而语义较细、更加具体的向量则被推向球体边界。

通过指数映射实现嵌入，该映射将每个欧几里得向量 $\mathbf{s}_{\hat{k}}$ 从原点 $\mathbf{0}$ 投影到庞加莱球上。对于曲率为 $c = 1$ 的球，映射定义如下：

$$\hat{\mathbf{p}}_{\hat{k}} = \tanh\left(\left\|\hat{\mathbf{s}}_{\hat{k}}\right\|\right) \frac{\hat{\mathbf{s}}_{\hat{k}}}{\left\|\hat{\mathbf{s}}_{\hat{k}}\right\|}, \quad \text{for } \hat{k} = 1, \ldots, \hat{K}. \tag{5}$$

该变换生成双曲向量集合 $\mathcal{P} = \{\mathbf{p}_{\hat{k}}\}_{\hat{k}=1}^{K}$，其中每个向量到原点的距离现可作为其语义重要性的替代指标。这为下一步构建显式语义树提供了所需的几何基础。

B. 用于零样本调整的分层剪枝

上一步生成的双曲向量集合 $\mathcal{P}$ 是一个按语义粒度在几何上组织的点云。此步骤的目标是将该点云转换为显式的层次结构——主语义树——随后即可对其进行确定性剪枝，使其达到任意所需大小 K。

我们的调整策略基于一个核心原则：必须保留构成知识库骨架的基础性粗粒度向量，同时可控地剪除更具体的细粒度向量。这需要一种能够明确界定向量间父子关系的结构。最小生成树是完成这一任务的理想数学工具；它是一种有理论依据的无参数方法，可用于发现点集最可能的底层结构。通过以尽可能短的总路径长度连接所有向量，最小生成树揭示了点云几何中潜藏的最可能语义层次结构。

使用双曲距离 $d_{\mathcal{H}}$ 作为任意两个向量 $\mathbf{p}_i$ 和 $\mathbf{p}_j$ 之间连边的边权重，构建最小生成树：

$$d_{\mathcal{H}}(\mathbf{p}_i, \mathbf{p}_j) = \operatorname{arccosh}\left(1 + 2\frac{\|\mathbf{p}_i - \mathbf{p}_j\|^2}{(1 - \|\mathbf{p}_i\|^2)(1 - \|\mathbf{p}_j\|^2)}\right). \tag{6}$$

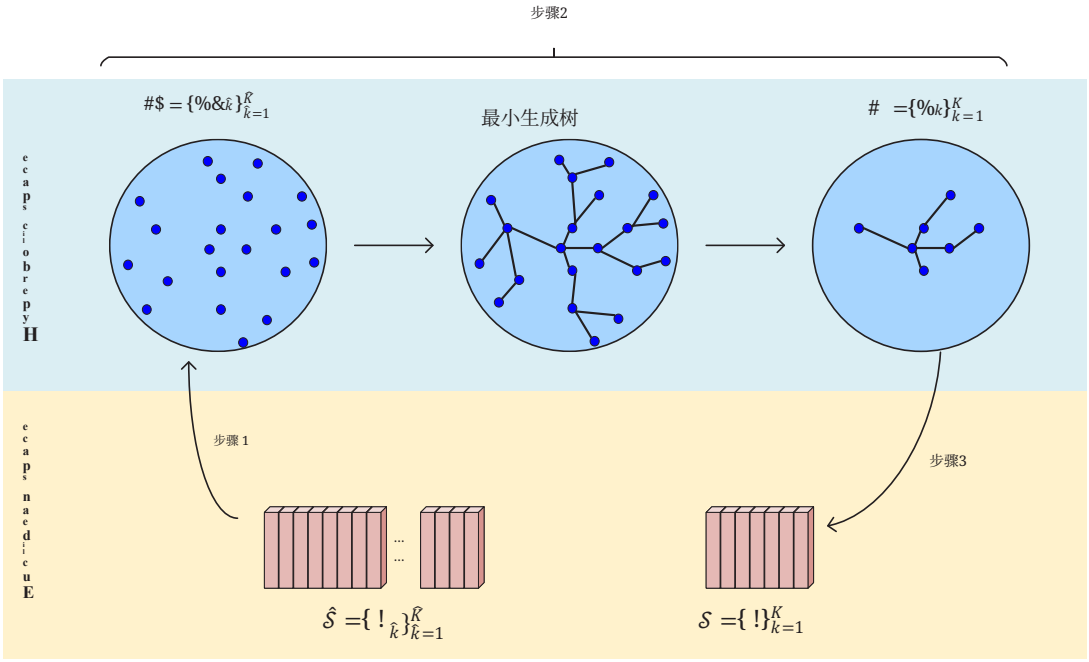


图2. 所提出的零样本知识库流程 调整大小

我们使用Prim算法 [11]构建主语义树。我们将与庞加莱球原点的双曲距离最小的向量指定为树的根节点，因为它代表整个知识库中最一般的概念。随后，算法迭代地将最近的未连接向量添加到不断生长的树ˆ中，直到所有 K 个向量都连接起来。

ˆ 这棵包含所有 K 节点的主树可作为调整大小的蓝图。为了获得特定大小为 ˆ$K < K$ 的知识库，我们执行确定性剪枝过程：迭代移除当前树的叶节点，直至恰好剩下 K 个节点。我们将最终得到的这组K 向量记为 $\mathcal{P}_K$。由于语义树中的叶节点代表最细粒度的概念，这一迭代过程直接体现了我们的核心原则：保留知识库基础性的粗粒度骨架，同时系统地移除最具体的向量。虽然可以设计基于中心性或其他指标的更复杂剪枝策略，但我们选择这种基于叶节点的方法，是因为它计算简单且具有确定性。此外，对于任何目标大小 K，都可以即时执行这一过程，从而兑现零样本调整的承诺。

C. 用于系统兼容性的欧几里得投影

我们框架的最后阶段会为原始通信系统准备调整大小后的知识库。正如第 II 节所述，大多数基于 VQ-VAE 的模型都在欧几里得空间中运行，因此我们必须将剪枝后的双曲向量 $\mathcal{P}_K$ 从庞加莱球映射回欧几里得空间。这通过对数映射实现，它是原点处指数映射的形式逆变换（即 (5)）。每个向量的变换$\mathbf{p}_k \in \mathcal{P}_K$ 定义如下：

$$\mathbf{s}_k = \operatorname{arctanh}(\|\mathbf{p}_k\|)\frac{\mathbf{p}_k}{\|\mathbf{p}_k\|}, \text{for } k = 1, \ldots, K. \quad (7)$$

这一过程得到最终调整大小后的欧几里得知识库， $\mathcal{S} = \{\mathbf{s}_k\}_{k=1}^{K}$，如今它已完全兼容，可集成至数字通信系统。

IV. 评估

本节通过大量仿真评估所提出的零样本知识库调整方法的性能。我们使用 ImageNet 数据集 [12]，在端到端图像语义传输任务上开展实验。语义编码器和解码器均基于 SimVQ [13]，构建；SimVQ 是一种先进的 VQ-VAE框架，可避免表征坍塌。我们将所提方法与一种基线方法进行比较：针对每种大小 K，分别从头训练一个专用知识库。为便于说明，在结果图中，这两种方法分别标记为“零样本调整”和“从头训练”。所有仿真均在一个由 8 个 NVIDIA V100 GPU 组成的集群上完成。

图 3 绘制了不同知识库大小 K 下源图像与重建图像之间的结构相似性指数（SSIM）。从图中可以得出以下两点观察结果。

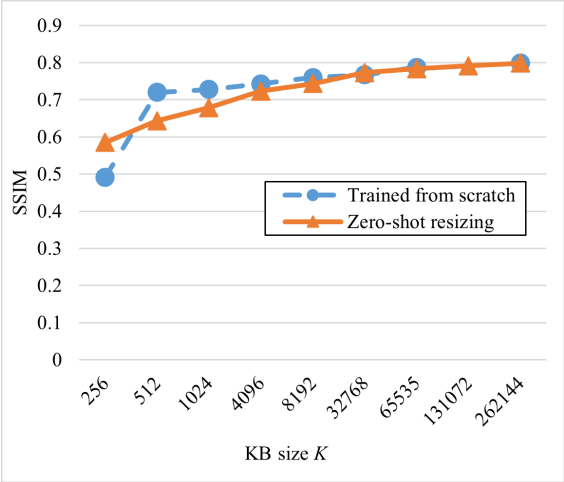


图 3. SSIM 随知识库大小 K 的变化：我们的零样本调整方法取得了与从头训练专用知识库的基线方法几乎相同的性能，且无需针对每种大小重新训练。

1）我们的设计和对比方案中的 SSIM 均随 K 增大而提高，这证实了预期的规律：知识库越大，通常重建质量越高。

2）更重要的是，在所有测试的大小下，我们设计的 SSIM 均与单独训练的基线方法非常接近，平均差异仅为 0.8%。这一结果意义重大，因为它表明，与计算开销巨大的方法（即针对每种特定大小分别训练专用知识库）相比，我们的方法几乎不会造成性能损失，同时还能提供速率自适应的灵活性。

为直观展示结果，图 4 可视化了测试集中的一个样本经重建后的图像。从图中，我们得出以下关键观察：1）对于任意给定的知识库大小 K，我们的零样本调整方法重建的图像在视觉上几乎与专用的从头训练模型重建的图像完全相同。这一结果有力地印证了图 3 中的定量观察，表明我们的调整大小方法无需重新训练，即可达到与计算开销巨大的基线方法相当的性能。

2) 我们方法的一项关键优势在最低码率（$K = 256$）时尤为明显。从头开始训练的基线模型发生了灾难性失效，生成的图像带有严重伪影，颜色也不自然。这表明，使用极小知识库训练基于 VQ-VAE 的端到端语义通信模型本质上并不稳定。与此形成鲜明对比的是，我们的方法保持了语义连贯性，生成的图像虽然模糊，但仍可辨认。这证明了我们分层剪枝方法的鲁棒性：通过保留最大型父知识库中的基础概念，我们的方法确保图像质量逐步下降，而非彻底崩溃。

需要指出的是，我们的方法有一项重大优势：大幅降低了训练开销。“从头训练”基线方法需要针对实验中评估的九种知识库大小，分别训练一个全新且专用的模型。相比之下，我们的“零样本调整”方法只需训练一个父模型（$K = 262,144$）。随后通过分层剪枝进行调整大小，是一种计算开销低廉的后处理步骤，无需进一步训练或微调。通过将所需训练轮次从九次减少到一次，我们的方法将总训练时间和 GPU资源降低了约9× 。因此，我们的方法不仅具有更出色的灵活性和鲁棒性，而且所需的计算预算也仅为传统方法的一小部分。

V. 结论

本文提出了一种新颖的零样本知识库调整方法，解决了数字语义通信中速率自适应这一关键挑战。针对不同传输速率分别训练专用知识库的传统方法，计算成本高昂且缺乏粒度。我们提出的系统性框架利用双曲几何揭示单个大型父知识库中固有的语义层次结构，从而无需重新训练，即可即时创建任意大小的高质量小型知识库。大量评估结果证实，该方法的重建质量几乎与从头训练的专用知识库相当，而所需计算预算仅为其一小部分。此外，我们的方法表现出更强的鲁棒性，在极低速率下性能仍能平稳退化，而传统方法在这种情况下会遭遇灾难性失效。最终，这项工作消除了一个重大的架构障碍，为真正灵活、高效且实用的速率自适应语义通信系统铺平了道路。

致谢

本研究得到国家科技重大专项“移动信息网络”（项目编号：2024ZD1300700）和国家自然科学基金（项目编号：62301471）的资助。

参考文献

[1] J. Ren, Z. Zhang, J. Xu, G. Chen, Y. Sun, P. Zhang, and S. Cui, 《知识库赋能的语义通信：生成式视角》，IEEE Wireless Communications，第 31 卷，第 4 期，第 14–22 页，2024 年。[2] A. Van Den Oord, O. Vinyals 等，《神经离散表示学习》，神经信息处理系统进展，第 30 卷，2017 年。

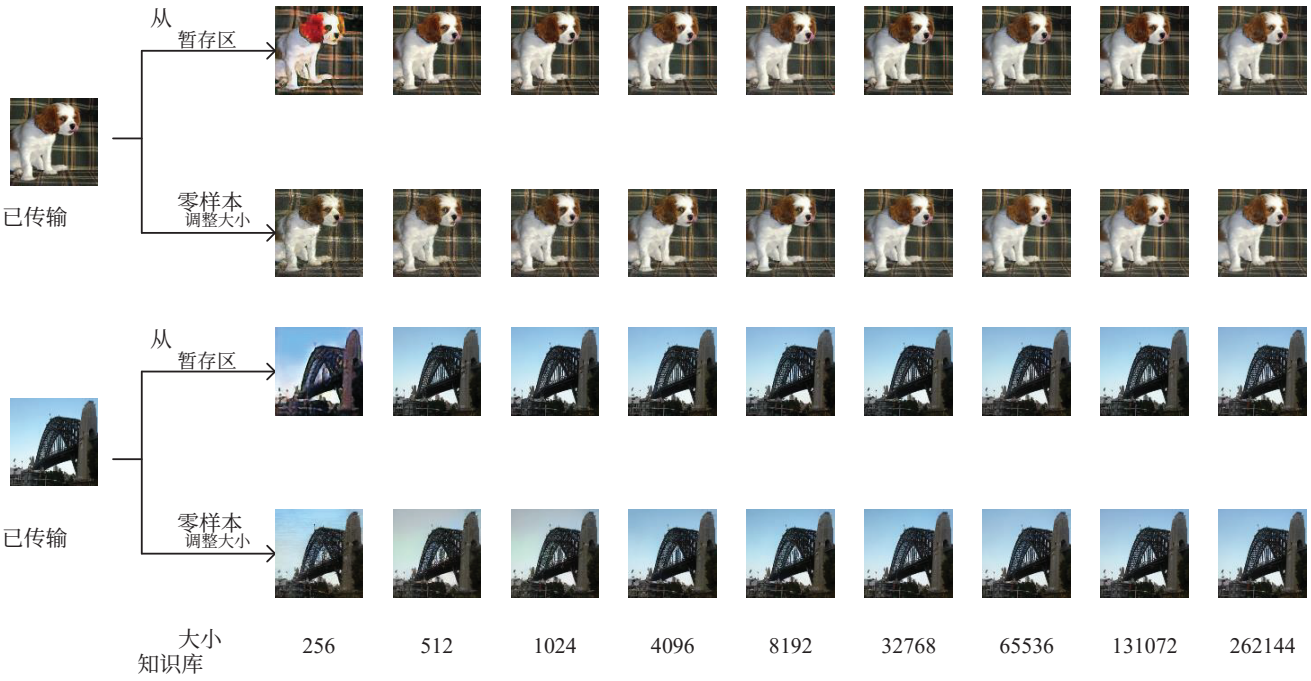


图 4. 不同知识库大小下重建图像的定性比较。

[3] Q. Hu, G. Zhang, Z. Qin, Y. Cai, G. Yu, and G. Y. Li, 《基于掩码向量量化变分自编码器码本的鲁棒语义通信》，IEEE 无线通信汇刊，第 22 卷，第 12 期，第 8707–8722 页，2023 年。[4] Q. Fu, H. Xie, Z. Qin, G. Slabaugh, and X. Tao, 《向量量化语义通信系统》，IEEE 无线通信快报，第 12 卷，第 6 期，第 982–986 页，2023 年。[5] J. Liang, F. Wang, Y. Zheng, W. Xu, X. Xu, and J. Dai, 《采用可变乘积量化进行图像传输的数字语义通信》，载于 2025 IEEE 无线通信与网络会议（WCNC），第 01–06 页，2025 年。[6] Y. Zhou, Y. Sun, G. Chen, X. Xu, H. Chen, B. Huang, S. Cui, and P. Zhang, 《MOC-RVQ：多级码本辅助的数字生成式语义通信》，载于 GLOBECOM 2024 - 2024 IEEE 全球通信大会，第 2078–2083 页，2024 年。[7] J. Shin, Y. Oh, J. Park, J. Park, and Y.-S. Jeon, 《ESC-MVQ：采用多码本矢量量化的端到端语义通信》，2025 年。[8] L. Li, T. Liu, C. Wang, M. Qiu, C. Chen, M. Gao, and A. Zhou, 《无需重新训练即可调整矢量量化码本的大小》，多媒体系统，第 29 卷，第 3 期，第 1499–1512 页，2023 年。[9] V. Khrulkov, L. Mirvakhabova, E. Ustinova, I. Oseledets, and V. Lempitsky, 《双曲图像嵌入》，载于 IEEE/CVF 计算机视觉与模式识别会议论文集，第 6418–6428 页，2020 年。[10] M. Nickel and D. Kiela, 《用于学习层次表示的庞加莱嵌入》，载于神经信息处理系统进展（I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett 编），第 30 卷，Curran Associates, Inc.，2017 年。[11] R. C. Prim, 《最短连接网络及其若干推广》，贝尔系统技术期刊，第 36 卷，第 6 期，第 1389–1401 页，1957 年。[12] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, 《ImageNet：大型层次图像数据库》，载于 2009 年 IEEE 计算机视觉与模式识别会议，第 248–255 页，2009 年。[13] Y. Zhu, B. Li, Y. Xin, Z. Xia, and L. Xu, 《解决单个线性层的向量量化模型中的表征坍塌，”2025。