用于广义零样本多模态信息抽取的双曲多模态生成式表示学习

周宝航 天津工业大学软件学院 中国天津
z
ho
ubaohang@tiangong.edu.cn

Kehui Song*
软件学院
天津工业大学
中国天津
songkehui@tiangong.edu.cn

Rize Jin
软件学院
天津工业大学
中国天津
jinrize@tiangong.edu.cn

Yu Zhao
计算机科学学院
VCIP, DISSec , TMCC, TBI中心
南开大学
中国天津
zhaoyu@dbis.nankai.edu.cn

Xuhui Sui
计算机科学学院
VCIP, DISSec , TMCC, TBI中心
南开大学
中国天津
suixuhui@dbis.nankai.edu.cn

Xinying Qian
计算机科学学院
VCIP, DISSec , TMCC, TBI中心
南开大学
中国天津
qianxinying@dbis.nankai.edu.cn

Xingyue Guo 计算机科学学院
VCIP、DISSec、TMCC、TBI中心
南开大学 中国天津
g
uoxingyue@dbis.nankai.edu.cn

Ying Zhang 计算机科学学院 VCIP、DISSec、TMCC、TBI中心 南开大学 中国天津
yingzhang@nankai.edu.cn

arXiv:2603.16259v1 [cs.MM] 17 Mar 2026

摘要

多模态信息抽取（MIE）是一组重要任务，旨在通过整合图像从网页文本中抽取结构化信息，以促进基于网页的语义知识的结构化构建。为应对网站上不断扩展的类别集合，其中包括新出现的实体类型或关系，先前研究提出了零样本多模态信息抽取（ZS-MIE）任务，旨在利用文本和视觉模态抽取未见结构化知识。然而，零样本多模态信息抽取模型仅限于识别属于未见类别集的样本，难以处理同时包含已见和未见类别的现实场景。现有方法的不足可归因于两个主要方面。一方面，这些方法在欧几里得空间中构建样本和类别的表示，未能捕捉样本内两种模态及其对应类别原型之间的层次语义关系。另一方面，已见类别集合与未见类别集之间的语义相似度分布存在显著间隔，这影响了零样本多模态信息抽取模型的生成能力。为克服上述缺点，我们深入研究广义零样本多模态信息抽取（GZS-MIE）任务，并提出双曲多模态生成式表示学习框架（HMGRL）。变分信息瓶颈和自编码器网络通过双曲空间进行重构，以建模样本与原型之间的多层次层次语义关联。此外，所提出的模型使用解码器生成的未见样本进行训练，并引入语义相似度分布对齐损失，以提升模型的泛化性能。在两个基准数据集上的实验评估表明，与现有基线方法相比，HMGRL具有优越性。

CCS 概念

• **信息系统** → **多媒体与多模态检索**； •**计算方法** → **信息抽**。

关键词

广义零样本学习，多模态信息抽取，多模态双曲表示学习

ACM 参考文献格式：

周宝航、Kehui Song、Rize Jin、Yu Zhao、Xuhui Sui、Xinying Qian、Xingyue Guo 和 Ying Zhang。2026。面向广义零样本多模态信息抽取的双曲多模态生成式表示学习。载于*2026*年 *ACM Web* 会议论文集（*WWW '26*），*2026*年*4*月*13*日至*17*日，阿拉伯联合酋长国迪拜。美国计算机协会，纽约，纽约州，美国，10 页。https://doi.org/10.1145/3774904.3792311

1 引言

多模态信息抽取（MIE） [41] 涵盖一组核心任务，旨在结合文本与图像，从文本中提取结构化信息，以构建多模态知识图谱 [34, 35]。多模态信息抽取任务经过专门设计，以实现不同目标，例如多模态命名实体类型判定（MET） [38]

*通讯作者。

本作品根据知识共享署名国际 4.0 许可协议授权。

WWW '26，*2026*年*4*月*13*日至*17*日，迪拜，阿拉伯联合酋长国。© 2026 版权所有归所有者/作者所有。美国计算机协会
ISBN 979-8-4007-2307-0/2026/04
https://doi.org/10.1145/3774904.3792311

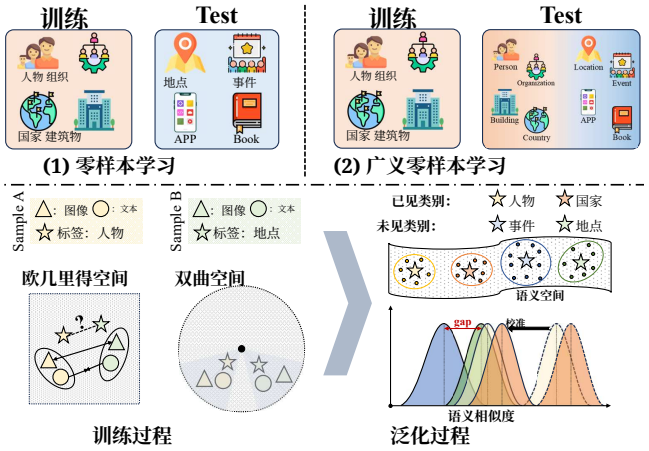


图1：上半部分展示了零样本学习与广义零样本学习的对比。下半部分呈现了广义零样本多模态信息抽取面临的关键挑战。

以及多模态关系抽取（MRE）[36]。与信息抽取（IE）模型相比，多模态方法利用各种融合策略，借助文本信息与视觉信息之间的联系，最终提升不同任务的性能 [39]。为应对实体或关系类型不断涌现所带来的挑战，将零样本学习引入基于文本的信息抽取（ZS-IE）模型后，无需进一步训练即可识别未见实体类型或关系。目前的 ZS-IE 方法聚焦于文本模态，并提取实体特征，以构建类型或关系原型的表示。类型或关系名称被视为原型知识，用于连接样本与类别之间的语义关联 [3, 15]。为识别多模态数据中隐含的新类别，Zhou等人[40] 首次提出了零样本多模态信息抽取（ZS-MIE）任务，并介绍了相应的架构和训练策略。以往的研究基于零样本学习 [19]，其中训练集和测试集的类别集合互不重叠。

当前的零样本信息抽取和零样本多模态信息抽取模型仅限于识别未见类别集中的样本。Zhou等人将零样本学习引入MET和MRE任务，使模型能够在测试阶段仅识别未见实体类型或关系 [38, 40]。然而，在实际场景中，社交媒体上的内容既包含模型已接触过的知识，也包含对模型而言仍然陌生的知识。因此，现有方法难以应对上述情况。一方面，现有模型在整个训练过程中完全依赖欧几里得空间。因此，如图1所示，在对文本相似但图像不同的样本进行建模时，代表不同标签的原型往往会彼此靠近。另一方面，基于零样本学习的方法仅使用来自已见类别的样本进行训练，旨在建立样本与其对应标签之间的语义关联。当推广到同时包含已见和未见类别的测试集时，模型往往会将样本判定为语义相似度较高的已见类别而为未见类别赋予较低分数。即使对已见类别应用校准因子，模型也无法避免所有类别的相似度分布变得分散。由于缺少真实训练样本，未见类别之间存在如图1所示的语义相似度分布差异，这一间隔会对广义零样本学习方法的性能产生不利影响。

为克服上述缺点，我们提出**双**曲**多**模态**生**成式**表**示**学**习（HM-GRL）[1] 框架，以解决广义零样本多模态信息抽取（GZS-MIE）任务。首先，我们采用双曲变分信息瓶颈（HVIB）对齐文本和视觉模态，以弥合二者之间的语义鸿沟。此外，我们设计了双曲多模态条件变分自编码器（HMC-VAE），用于对样本的多模态表示及其与标签原型之间的相关性进行建模。基于上述两个组件，该框架能够借助双曲空间的优势，捕捉样本与原型之间多层次的层次语义关联。在训练过程中，未见类别的原型知识被视为条件，由HMCVAE的解码器生成样本，并使用真实样本和合成样本训练模型。为缓解未见类别样本之间相似度分数的数值差异，我们引入语义相似度分布对齐损失，从而增强GZS-MIE模型的泛化能力。本文的贡献总结如下：

- 据我们所知，我们是首个研究广义零样本多模态信息抽取（GZS-MIE）任务，并分析现有零样本信息抽取和零样本多模态信息抽取模型在这一新任务中的不足之处的研究团队。
- 我们提出了双曲多模态生成式表征学习（HMGRL）框架，以解决GZS-MIE任务。该框架不仅包含HVIB和HMCVAE网络，用于捕捉样本与原型的层次语义关联，还利用合成样本训练模型，并采用语义相似度分布对齐损失，从而获得更强的泛化能力。
- 我们在两个标准多模态信息抽取基准数据集上进行了全面实验，结果表明，我们提出的框架优于基线方法。

2 相关工作

2.1 多模态信息抽取

随着多模态数据量不断增长，研究人员逐渐认识到，从这类数据中提取其中蕴含的复杂语义信息十分必要。因此，他们将传统的信息抽取（IE）任务扩展为多模态信息抽取，旨在构建多模态知识图谱 [41]。Moon等人 [17] 首先将传统的文本命名实体识别扩展为多模态命名实体

[1] https://github.com/ZovanZhou/HMGRL

识别（MER）。为解决这一任务，研究人员提出了多种融合策略，包括模态注意力机制 [17], 、统一多模态变换器 [27]和基于跨度的变分自编码器 [37]，以融合文本和视觉数据，最终提升序列标签预测的性能。Zheng 等人 [36] 提出了多模态关系抽取（MRE）任务，利用视觉信息增强对文本内容的语义理解，并准确识别实体之间的关系。此外，为联合建模上述两项任务，Yuan 等人[28] 提出了多模态实体关系抽取（MERE）任务，并采用词对关系标注策略。为确保提取出的多模态实体与人类知识框架保持一致，Zhang 等人 [32]提出了多模态命名实体类型判定（MET）任务，并使用跨模态Transformer分析视觉对象。

上述多模态信息抽取任务基于监督学习范式，旨在将多模态数据直接映射到预定义标签。然而，我们的研究聚焦于广义零样本多模态信息抽取（GZS-MIE），它无需额外训练，便能识别已见和未见类别。广义零样本学习强调为样本和语义标签构建适应性表示，从而能够将样本分类到此前未训练过的类别中。

2.2 零样本信息抽取

信息抽取涵盖一系列任务，旨在从非结构化文本中提取结构化信息，以构建知识图谱。为在无需进一步训练的情况下提取新颖知识（例如识别未见类别，包括实体类型或关系），零样本学习被引入信息抽取和多模态信息抽取任务。它们面临的关键挑战是学习具有泛化能力的实体表示和类别的原型知识。对于零样本命名实体类型判定（ZS-ET），类型的原型知识通过标签嵌入方法进行编码，Ma等人 [15] 采用该方法来强化实体提及的表示与其对应类型之间的语义联系。在数据方面，Chen 等人 [4], Obeidat等人 [18] 将从网站收集的实体提及及其类型描述等补充信息引入ZS-ET模型，以增强实体提及及其对应类型的表示。为提取上下文中与类型相关的细粒度特征，注意力机制 [20] 或知识记忆能力 [30] 被引入ZS-ET模型。为应对零样本关系抽取（ZS-RE）的挑战，研究者将实体和关系输入预训练语言模型，并通过同时优化它们之间的语义相似度，将其映射到嵌入空间中。Chen和Li [3] 采用了这种方法。Zhao等人 [33] 和Gong和Eldardiry [8]分别提出了语义匹配方法和基于提示的方法，以提升ZS-RE任务的性能。除上述基于文本的信息抽取任务外，Zhou等人 [40]研究了零样本多模态信息抽取（ZS-MIE）任务，并提出了细粒度表征学习框架，其中包括变分混合专家网络和基于多模态图的虚拟对抗学习。

总而言之，现有的零样本信息抽取和零样本多模态信息抽取方法均基于零样本学习。这些方法仅限于从未见类别中识别样本，难以有效地同时识别已见和未见类别。与这些既有研究不同，我们的研究聚焦于广义零样本多模态信息抽取任务，旨在从社交媒体平台产生的多模态数据中提取其中蕴含的基础结构知识和新颖结构知识。

3 洛伦兹线性变换

我们引入双曲空间来建模样本与原型表示之间的层次语义相关性。描述双曲空间的两种有效双曲几何包括庞加莱球 [7] 和洛伦兹模型 [12]。考虑到洛伦兹模型 [11],的优势，我们采用它来定义双曲空间中的线性变换操作。 n维洛伦兹模型 $\mathbb{L}_c^n$ 是满足以下条件的黎曼流形:

$$\mathbb{L}_c^n = \quad \in \mathbb{R}^{n+1} | < \quad , \quad >_{\mathbb{L}} = \{\mathbf{x} \quad \frac{1}{c} \; x_0 > 0 \qquad c \qquad < \; , \; >_{\mathbb{L}}$$

$\bar{\mathbf{xx}}$,}，其中表示负曲率，$\cdot\cdot$表示洛伦兹标量积。黎曼流形$\mathbb{L}_c^n$中的每个点都具有如下形式：$\mathbf{x} = [x_0, \mathbf{x}_s]$，其中 $x_0 \in \mathbb{R}$ 且$\mathbf{x}_s \in \mathbb{R}^n$。点 $\mathbf{x}$ 与$\mathbf{y}$ 的洛伦兹标量积定义如下:

$$< \mathbf{x}, \mathbf{y} >_{\mathbb{L}} = -x_0 y_0 + \sum_{i=1}^{n} x_i y_i.$$

因此，黎曼流形 $\mathbb{L}_c^n$ 在点 $\mathbf{x}$ 处的正交空间可定义为：T

$$_{\mathbf{x}}\mathbb{L}_c^n = \quad \in \mathbb{R}^{n+1} | < \quad , \quad >_{\mathbb{L}} = \{\mathbf{y} \; \mathbf{xy} \},$$

0其中 $\mathbb{T}_{\mathbf{x}}\mathbb{L}_c^n$ 是一个欧几里得子空间。而 $\mathbb{T}_0\mathbb{L}_c^n$ 表示原点处的切空间。

利用指数映射将特征向量从欧几里得空间映射到双曲空间。为了沿着测地线，将点 $\mathbf{x} \in \mathbb{T}_{\mathbf{x}}\mathbb{L}_c^n$ 映射到 $\mathbb{L}_c^n$，指数映射可表示为:

$$\exp_{\mathbf{x}}^c(\mathbf{z}) = \cosh(\alpha)\mathbf{x} + \sinh(\alpha)\frac{\mathbf{z}}{\alpha}$$

其中 $\alpha = \sqrt{-c}\|\mathbf{z}\|_{\mathbb{L}}$，且 $\|\mathbf{z}\|_{\mathbb{L}} = \sqrt{< \mathbf{z}, \; \mathbf{z} >_{\mathbb{L}}}$。作为逆向映射，对数映射用于将向量映射回欧几里得空间。给定双曲空间中的点 $\mathbf{y} \in \mathbb{L}_c^n$，对数映射可定义为:

$$\log_{\mathbf{x}}^c(\mathbf{y}) = \frac{\cosh^{-1}(\beta)}{\sqrt{\beta^2 - 1}}(\mathbf{y} - \beta\mathbf{x})$$

其中 $\beta = c < \mathbf{x}, \; \mathbf{y} >_{\mathbb{L}}$。基于上述映射函数，洛伦兹模型为两个点 $\mathbf{x} \in \mathbb{L}_c^n$ 和 $\mathbf{v} \in \mathbb{T}_0\mathbb{L}_c^n$ 提供了从 $\mathbb{L}_c^n$ 到 $\mathbb{L}_c^m$ 的线性变换。给定一个线性映射 $\mathbf{M}: \mathbb{R}^n \rightarrow \mathbb{R}^m$ 及其权重矩阵，洛伦兹线性变换定义为:

$$\mathcal{M}_c(\mathbf{x}) = \exp_0^c(\hat{\mathcal{M}}(\log_0^c(\mathbf{x}))), \hat{\mathcal{M}} = [0, \mathcal{M}(\mathbf{v}_s)]$$

其中 $\mathbf{v}_s = [v_1, v_2, \ldots, v_n]$，$\exp_0^c: \mathbb{T}_0\mathbb{L}_c^n \rightarrow \mathbb{L}_c^m$和 $\log_0^c: \mathbb{L}_c^n \rightarrow$ T$_0\mathbb{L}_c^m$。为了对样本和原型的层次语义关联进行建模，我们定义洛伦兹线性层，将学习到的特征映射到双曲空间，然后在双曲空间中使用洛伦兹线性变换。最后，我们利用对数映射将变换后的特征映射回欧几里得空间，以完成多模态信息抽取任务。洛伦兹

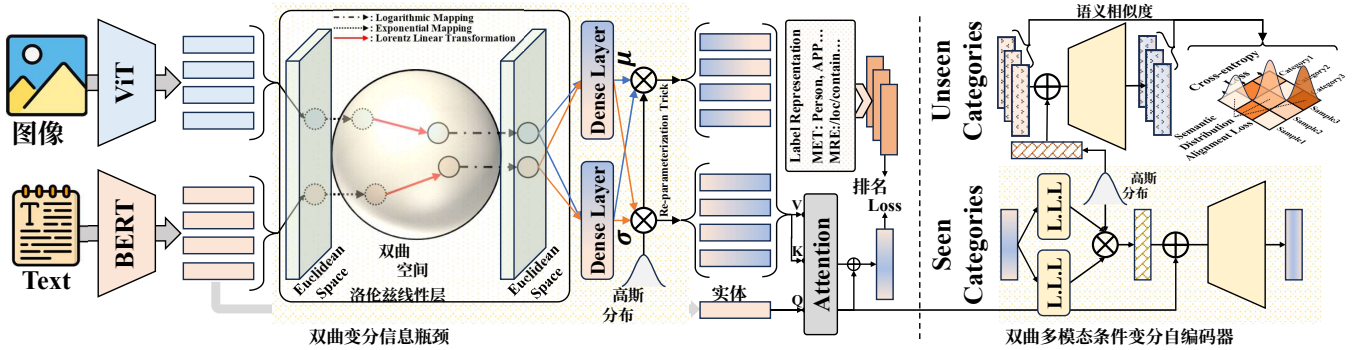


图 2：用于广义零样本多模态信息抽取的双曲多模态生成式表征学习（HMGRL）框架。该框架主要包含双曲变分信息瓶颈和双曲多模态条件变分自编码器（VAE）。"L.L.L"是洛伦兹线性层的缩写。 ⊕表示拼接操作， ⊗ 是重参数化技巧。

线性层（LLL）定义如下：

$$\mathrm{LLL}(\cdot;\mathcal{M}) = \log_0^c(\mathcal{M}_c(\exp_0^c(\cdot))) \in \mathbb{R}^m.$$

4 方法论

在介绍所提出模型的细节之前，我们先形式化广义零样本多模态信息抽取（GZS-MIE）问题。我们将已见类别集合形式化为 $\mathrm{Y}_s = \{y_1^s, y_2^s, \ldots, y_{|\mathrm{Y}_s|}^s\}$，其中包含 $|\mathrm{Y}_s|$ 个已见类型；将未见类别集形式化为 Y_u，$\mathrm{Y}_s \cap \mathrm{Y}_u = \emptyset$，$\mathrm{Y}_u = \{y_1^u, y_2^u, \ldots, y_{|\mathrm{Y}_u|}^u\}$，其中包含 $|\mathrm{Y}_u|$ 个未见类型，且 $\mathrm{Y}_s \cap \mathrm{Y}_u = \emptyset$。为了训练模型，我们获取训练数据集 $\mathrm{D}_{train} = \{(T_i, V_i, E_i, Y_i) | Y_i \in \mathrm{Y}_s, i = 1, 2, \ldots, |\mathrm{D}_{train}|\}$，其中 T 表示自然语言句子，V 是与该句子关联的图像，E 表示句子中的实体，Y 是多模态信息抽取任务的标签。与零样本学习不同，广义零样本学习的测试数据集定义为 $\mathrm{D}_{test} = \{(T_i, V_i, E_i, Y_i) | Y_i \in \mathrm{Y}_s \bigcup \mathrm{Y}_u, i = 1, 2, \ldots, |\mathrm{D}_{test}|\}$。

本节介绍用于广义零样本多模态信息抽取的双曲多模态生成式表示学习（HMGRL）框架，如图 2 所示。该框架主要由以下几个部分构成：（1）首先，我们利用预训练的语言模型和视觉模型从样本中提取多模态输入表征，并采用双曲变分信息瓶颈作为统一的主干网络，旨在对齐这些多模态输入表征。（2）接着，为了捕捉多模态表示中固有的局部语义特征，我们采用注意力机制并引入排序损失，以完成 GZS-MIE 任务。（3）此外，我们提出一种双曲多模态条件变分自编码器，用于对样本的多模态表示建模，并利用样本与原型之间的语义关系，生成属于未见类别的合成样本。（4）最后，为了将样本划分到已见和未见类别中，我们计算标签语义特征与样本多模态特征之间的距离。

4.1 多模态表征对齐

考虑一条由文本和图像对的多模态样本，我们将多模态数据映射为适合深度神经网络处理的稠密表征。对于文本模态，我们将包含 $|S|$ 个词的原始句子表示为 $S = \{w_1, w_2, \ldots, w_{|S|}\}$。为将这些离散词转换为分布式表征，我们使用预训练语言模型（如 BERT）作为文本编码器。句子特征的提取过程可表示为 $\mathbf{T} = \mathrm{BERT}(\tilde{S}) = \{\mathbf{t}_{[\mathrm{CLS}]}, \mathbf{t}_1, \ldots, \mathbf{t}_{[\mathrm{SEP}]}\} \in \mathbb{R}^{|T| \times d}$ 其中，$\tilde{S}$ 是添加了特殊标记的扩展句子，$|T|$ 是其标记数量。对于视觉模态，我们采用预训练视觉Transformer（ViT）[6] 模型来提取图像的特征，可表示为 $\mathbf{V} = \{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_{|V|}\} \in \mathbb{R}^{|V| \times d}$，其中 $|V|$ 表示从中获得的特征向量数量。

为弥合文本与视觉模态之间的语义鸿沟 [1, 26]，我们需要最大化互信息 $\mathrm{MI}(\mathbf{T}, \mathbf{V})$，同时减少它们的噪声信息。因此，我们提出利用双曲变分信息瓶颈（HVIB）作为统一的主干网络，从中提取有效特征。给定文本表征 $\mathbf{T}$ 和视觉表征 $\mathbf{V}$，训练 HVIB 以学习潜变量 $\mathbf{Z}$，从两种模态中保留充分特征以实现对齐。考虑到多模态表征对齐的对称性，两种模态的信息瓶颈 [21] 定义如下：

$$\begin{aligned} \mathcal{L}_{IB}^{T \to V} &= \beta \cdot \mathrm{MI}(\mathbf{T}, \mathbf{Z}^{T \to V}) - \mathrm{MI}(\mathbf{Z}^{T \to V}, \mathbf{V}), \\ \mathcal{L}_{IB}^{V \to T} &= \beta \cdot \mathrm{MI}(\mathbf{V}, \mathbf{Z}^{V \to T}) - \mathrm{MI}(\mathbf{Z}^{V \to T}, \mathbf{T}). \end{aligned} \tag{1}$$

我们最小化 $\mathrm{MI}(\mathbf{T}, \mathbf{Z}^{T \to V})$ 和 $\mathrm{MI}(\mathbf{V}, \mathbf{Z}^{V \to T})$，以减少两种模态中的噪声信息。为确保样本内部两种模态的语义对齐，我们通过预测其对齐的模态特征，最大化 $\mathrm{MI}(\mathbf{Z}^{T \to V}, \mathbf{V})$ 和 $\mathrm{MI}(\mathbf{Z}^{V \to T}, \mathbf{T})$。接着，我们以变分方式对潜在表示进行编码，并通过重参数化技巧 [10]计算高斯变量 $\mathbf{Z}^{T \to V}$ 和 $\mathbf{Z}^{V \to T}$，如下所示：

$$\mathbf{Z}^{T \to V} = \boldsymbol{\mu}_{T \to V} + \boldsymbol{\sigma}_{T \to V} \odot \boldsymbol{\epsilon}, \quad \mathbf{Z}^{V \to T} = \boldsymbol{\mu}_{V \to T} + \boldsymbol{\sigma}_{V \to T} \odot \boldsymbol{\epsilon} \tag{2}$$

其中 $\boldsymbol{\epsilon} \sim \mathrm{N}(\mathbf{0}, \mathbf{I})$，$\odot$ 表示逐元素乘法运算。为对两种模态之间的层次语义相关性进行建模，我们定义均值的计算过程

变量 $\boldsymbol{\mu}$ 和标准差为一的 $\boldsymbol{\sigma}$ 如下：

$$\begin{aligned}\boldsymbol{\mu}_{T\to V}&=\psi(\mathrm{LLL}(\mathbf{T};\mathcal{M}_\mu);\boldsymbol{\theta}_\mu),\quad \boldsymbol{\sigma}_{T\to V}=\psi(\mathrm{LLL}(\mathbf{T};\mathcal{M}_\sigma);\boldsymbol{\theta}_\sigma)\\ \boldsymbol{\mu}_{V\to T}&=\psi(\mathrm{LLL}(\mathbf{V};\mathcal{M}_\mu);\boldsymbol{\theta}_\mu),\quad \boldsymbol{\sigma}_{V\to T}=\psi(\mathrm{LLL}(\mathbf{V};\mathcal{M}_\sigma);\boldsymbol{\theta}_\sigma)\end{aligned}\tag{3}$$

其中，$\psi(\cdot;\theta)$ 表示参数矩阵为 θ的前馈神经网络，$\{\mathrm{M}_\mu,\mathrm{M}_\sigma\}$ 是洛伦兹线性层中的权重矩阵。为了依据信息瓶颈原理优化模型，我们可以使用变分后验分布 $q(\mathbf{Z}^{T\to V}|\mathbf{T})$ 和$q(\mathbf{Z}^{V\to T}|\mathbf{V})$ 来近似后验分布 $p(\mathbf{Z}^{T\to V}|\mathbf{T})$ 和 $p(\mathbf{Z}^{V\to T}|\mathbf{V})$ [5]。假设先验分布 $p(\mathbf{Z}^{T\to V})$ 和 $p(\mathbf{Z}^{V\to T})$ 为标准正态分布，则 HVIB 的正则化损失定义如下：

$$\begin{aligned}\mathcal{L}_{reg}=&\frac{1}{2}(\mathrm{KL}(\mathcal{N}(\boldsymbol{\mu}_{T\to V},\boldsymbol{\sigma}^2_{T\to V})||\mathcal{N}(0,\mathbf{I}))\\&+\mathrm{KL}(\mathcal{N}(\boldsymbol{\mu}_{V\to T},\boldsymbol{\sigma}^2_{V\to T})||\mathcal{N}(0,\mathbf{I}))).\end{aligned}\tag{4}$$

此外，为了最大化项目 $\mathrm{MI}(\mathbf{Z}^{T\to V},\mathbf{V})$ 和 $\mathrm{MI}(\mathbf{Z}^{V\to T},\mathbf{T})$，我们可以通过对比学习 [29] 直接最大化 $\mathrm{MI}(\mathbf{Z}^{T\to V},\mathbf{Z}^{V\to T})$，以对齐这两种模态。给定一个包含$N$样本的批次，我们对其潜在表示进行平均池化，并得到全局表示 $\{(\bar{\mathbf{Z}}_i^{T\to V},\bar{\mathbf{Z}}_i^{V\to T})|i=1,2,\ldots,N\}$，其中 $\bar{\mathbf{Z}}_i^{T\to V}=\frac{1}{|T|}\sum_{j=1}^{|T|}\mathbf{z}_{i,j}^{T\to V}$，且 $\bar{\mathbf{Z}}_i^{V\to T}=\frac{1}{|V|}\sum_{j=1}^{|V|}\mathbf{z}_{i,j}^{V\to T}$。对比学习旨在从图像和文本潜在表示的 $N\times N$种可能组合中区分匹配对，确保配对输入的表示在表示空间中比未配对输入的表示彼此更接近。因此，对比学习损失定义如下：

$$\begin{aligned}\mathcal{L}_{cl}=\sum_{i=1}^{N}-\frac{1}{2}(&\log\frac{\exp(\bar{\mathbf{Z}}_i^{T\to V}\cdot\bar{\mathbf{Z}}_i^{V\to T})}{\sum_{j=1}^{N}\exp(\bar{\mathbf{Z}}_i^{T\to V}\cdot\bar{\mathbf{Z}}_j^{V\to T})}\\&+\log\frac{\exp(\bar{\mathbf{Z}}_i^{V\to T}\cdot\bar{\mathbf{Z}}_i^{T\to V})}{\sum_{j=1}^{N}\exp(\bar{\mathbf{Z}}_i^{V\to T}\cdot\bar{\mathbf{Z}}_j^{T\to V})}).\end{aligned}\tag{5}$$

4.2 广义零样本多模态信息抽取

为完成多模态信息抽取任务，我们根据实体集 E [1]，标记实体的起始和结束位置，并将保留标记 E/E (和 E/E) 插入原始句子中。随后，我们根据具体任务提取实体表示。对于不同任务，实体表示记为：

$$\begin{aligned}&\text{MET:}\quad \mathbf{e}=[\mathbf{t}_{[\mathrm{CLS}]}\oplus\mathbf{t}_{[\mathrm{E1}]}]\in\mathbb{R}^{2d}\\&\text{MRE:}\quad \mathbf{e}=[\mathbf{t}_{[\mathrm{CLS}]}\oplus\mathbf{t}_{[\mathrm{E1}]}\oplus\mathbf{t}_{[\mathrm{E2}]}]\in\mathbb{R}^{3d}\end{aligned}\tag{6}$$

其中，$\mathbf{t}_{[\mathrm{CLS}]}$、$\mathbf{t}_{[\mathrm{E1}]}$ 和 $\mathbf{t}_{[\mathrm{E2}]}$ 分别表示特殊标记 [CLS]、[E1] 和 [E2] 的文本表示，$\oplus$ 表示向量拼接操作。为获得融合多模态表示，我们将对齐后的潜在表示拼接为融合表示 $\mathbf{U}=[\mathbf{Z}^{T\to V},\mathbf{Z}^{V\to T}]\in\mathbb{R}^{(|T|+|V|)\times d}$。为提取与实体相关的融合多模态表示的局部特征，我们对不同模态的特征应用跨模态注意力机制。以文本实体特征 $\mathbf{e}\in\mathbb{R}^{|\mathbf{e}|}$ 作为查询，以融合多模态表示 $\mathbf{U}=\{\mathbf{u}_i|i=1,2,\ldots,|T|+|V|\}$ 作为键和值，注意力分数计算如下：$\beta_i=\frac{\phi([\mathbf{u}_i\oplus\mathbf{e}])}{\sum_{j=1}^{|T|+|V|}\phi([\mathbf{u}_j\oplus\mathbf{e}])}$，其中

$\phi(\mathbf{x})=\exp(\mathbf{W}_A\cdot\mathbf{x}+\mathbf{b}_A)$，且$\mathbf{W}_A\in\mathbb{R}^{(|\mathbf{e}|+d)}$和$\mathbf{b}_A\in\mathbb{R}$是跨模态注意力层中的可训练参数。因此，

实体感知多模态融合表示定义为 $\bar{\mathbf{U}}=\sum_{i=1}^{|T|+|V|}\beta_i\mathbf{u}_i$。然后，我们将上述特征与文本实体特征拼接，得到整体表示 $\tilde{\mathbf{e}}=[\bar{\mathbf{U}}\oplus\mathbf{e}]\in\mathbb{R}^{(|\mathbf{e}|+d)}$。

为确定广义零样本学习下的类别，我们采用标签嵌入方法，将标签名称编码为原型表示。给定已见类别集合 $\mathrm{Y}_s=\{y_1^s,\ y_2^s,\ \ldots\ldots,\ y^s_{|\mathcal{Y}_s|}\}$，我们将每个标签名称视为一个短句，并使用 BERT 对其进行编码，以获得其文本表示 $\mathbf{p}_i^s$。随后，我们对其应用平均池化，该类别的语义表示记为 $\bar{\mathbf{p}}_i^s$。因此，已见类别集合的原型表示可表示为 $\mathbf{P}^s=[\bar{\mathbf{p}}_1^s,\ \bar{\mathbf{p}}_2^s,\ \ldots\ldots,\ \bar{\mathbf{p}}^s_{|\mathcal{Y}_s|}]\in\mathbb{R}^{|\mathcal{Y}_s|\times d}$。为评估样本与类别之间的相关性，我们采用语义相似度 $\mathbf{O}^s=\psi(\tilde{\mathbf{e}};\boldsymbol{\theta}_e)\psi(\mathbf{P}^s;\boldsymbol{\theta}_p)^T\in\mathbb{R}^{|\mathcal{Y}_s|}$。为确保真实标签的相似度高于其他标签，我们通过利用排序损失来优化模型：

$$\mathcal{L}_{rank}=\sum_{i=1}^{|\mathcal{Y}_s|}\max(1-o_+^s+o_i^s,0)\tag{7}$$

其中，o_+^s 表示样本与真实类别之间的相似度。

4.3 双曲多模态条件变分自编码器

考虑到未见类别集中没有样本，我们需要弥合已见类别与未见类别之间的领域差距。为了利用生成模型，我们用它生成合成样本，使模型适应未见类别领域。因此，我们提出双曲多模态条件变分自编码器（HMCVAE），用于建模文本原型与样本多模态表示之间的语义关联。与传统的 VAE 模型相比，HMCVAE 模型包含一个由洛伦兹线性层构成的编码器，用于建模样本与原型之间的层次语义相关性。给定来自已见类别集的样本的多模态表示 $\tilde{\mathbf{e}}^s$，我们可以通过重参数化技巧为 HMCVAE 模型编码潜在表示，如下所示：

$$\begin{aligned}\mathbf{Z}^{VAE}&=\boldsymbol{\mu}_{VAE}+\boldsymbol{\sigma}_{VAE}\odot\boldsymbol{\epsilon};\\ \boldsymbol{\mu}_{VAE}=\mathrm{LLL}(\tilde{\mathbf{e}}^s;\mathcal{M}'_\mu);&\quad \boldsymbol{\sigma}_{VAE}=\mathrm{LLL}(\tilde{\mathbf{e}}^s;\mathcal{M}'_\sigma)\end{aligned}\tag{8}$$

其中，$\boldsymbol{\epsilon}\in\mathbb{R}^h$ 和 h 是潜在表示的维度数。为对来自已见类别集合的样本进行多模态表示建模，我们将实体表示用作不同多模态信息抽取任务的条件原型知识，如下所示：

$$\begin{aligned}&\text{MET:}\quad \mathbf{p}^e=\mathbf{t}_{[\mathrm{E1}]}\in\mathbb{R}^d\\&\text{MRE:}\quad \mathbf{p}^e=\frac{1}{2}(\mathbf{t}_{[\mathrm{E1}]}+\mathbf{t}_{[\mathrm{E2}]})\in\mathbb{R}^d.\end{aligned}\tag{9}$$

然后，我们将潜在表示和条件原型表示输入解码器，以重构输入样本的多模态表示，其形式为 $\hat{\mathbf{e}}^s=\psi([\mathbf{p}^e\oplus\mathbf{Z}^{VAE}];\boldsymbol{\theta}_d)\in\mathbb{R}^{(|\mathbf{e}|+d)}$，其中 $\boldsymbol{\theta}_d$表示解码器中的权重参数集合。HMCVAE 的变分证据下界（证据下界） [10] 可定义如下：

$$\mathrm{ELBO}(\tilde{\mathbf{e}}^s)=\mathrm{KL}(q(\mathbf{Z}^{VAE}|\tilde{\mathbf{e}}^s)||p(\mathbf{Z}^{VAE}))+\log p(\tilde{\mathbf{e}}^s|\mathbf{Z}^{VAE})\tag{10}$$

表1：广义零样本学习设置下MET和MRE数据集上的性能比较。粗体数字表示最佳结果，下划线数字表示次佳结果。

多模态命名实体类型判定							多模态关系抽取						
模型	Seen		未见		总体		模型	Seen		未见		总体	
	准确率	F1	准确率	F1	准确率	F1		准确率	F1	准确率	F1	准确率	F1
Proto	67.9±27.2	73.3±23.5	15.3±10.0	14.2±6.4	21.9±9.0	22.5±8.9	Proto	43.6±3.3	51.0±5.3	31.6±18.5	29.6±15.5	34.6±15.2	36.2±14.9
DBZFET	**76.7±7.1**	**82.4±7.2**	19.8±14.2	20.8±14.1	29.4±15.7	31.2±16.5	ZS-BERT	53.1±15.3	59.7±15.3	17.3±2.8	20.7±3.0	25.7±3.4	30.6±4.9
NZFET	63.3±32.8	66.2±33.1	7.4±5.6	7.5±5.2	11.6±7.6	12.9±8.8	RE-match	49.9±6.8	56.0±4.8	26.9±16.3	28.7±14.1	31.9±14.7	36.3±13.4
MZET	19.8±18.9	20.4±19.4	29.1±20.0	31.2±20.6	14.5±11.4	16.1±13.0	ZS-SKA	53.7±8.8	60.3±7.2	24.6±12.9	23.6±14.6	32.6±14.3	32.3±16.9
MMProto	62.4±30.9	68.9±29.2	18.2±11.3	21.6±11.9	24.0±12.9	30.3±15.3	MMProto	**54.8±0.8**	**62.5±1.1**	29.1±14.8	28.8±10.8	36.5±13.0	38.6±10.8
MOVCNet	65.5±32.8	68.6±34.4	14.0±3.2	15.5±8.5	21.5±2.9	25.2±13.6	MOVCNet	35.1±8.6	42.6±3.4	31.2±9.1	27.4±6.1	31.5±2.7	32.9±3.9
MG-VMoE	55.7±25.8	61.6±28.1	25.3±21.4	29.4±25.1	30.2±20.8	36.0±24.4	MG-VMoE	32.4±6.8	37.5±11.1	26.4±18.7	26.5±16.2	26.1±14.6	30.1±14.7
Ours	62.3±19.6	68.8±24.7	**33.4±3.9**	**38.8±2.6**	**42.8±7.3**	**48.2±7.3**	Ours	49.5±6.7	55.4±5.0	**37.9±12.6**	**35.5±14.7**	**42.6±10.8**	**42.4±12.7**
LLaVA	54.6±12.7	61.2±17.0	37.8±31.1	41.6±23.8	57.7±15.7	49.5±16.2	LLaVA	20.6±3.6	18.4±8.2	12.2±3.7	15.2±1.7	14.9±2.1	15.7±3.1

其中，变分后验分布 $q(\mathbf{Z}^{VAE}|\tilde{\mathbf{e}}^s)$ 是参数为 $\{\boldsymbol{\mu}_{VAE}, \boldsymbol{\sigma}_{VAE}\}$ 的高斯分布，而先验分布 $p(\mathbf{Z}^{VAE})$ 是标准正态分布。因此，HMCVAE 的训练损失定义如下：

$$\mathcal{L}_{vae} = \mathrm{KL}(\mathcal{N}(\boldsymbol{\mu}_{VAE}, \boldsymbol{\sigma}^2_{VAE})||\mathcal{N}(0, \mathbf{I})) + ||\tilde{\mathbf{e}}^s - \hat{\mathbf{e}}^s||^2. \quad (11)$$

借助 HMCVAE 的生成能力，我们可以利用其解码器生成合成样本。给定未见类别集 Y_u，我们同样采用标签嵌入方法对其原型表示进行编码，记为 $\mathrm{P}^u = [\tilde{\mathrm{p}}^u_1, \tilde{\mathrm{p}}^u_2, \ldots, \tilde{\mathrm{p}}^u_{|\mathrm{Y}_u|}] \in \mathbb{R}^{|\mathcal{Y}_u| \times d}$。通过从标准正态分布中采样噪声向量 $\boldsymbol{\epsilon}$，将其与第 i 个未见类别的原型表示拼接起来，生成合成样本 $\hat{\mathrm{e}}^u_i = \psi([\tilde{\mathrm{p}}^u_i \oplus \boldsymbol{\epsilon}]; \theta_d) \in \mathbb{R}^{(|\mathbf{e}|+d)}$。因此，未见类别集的合成多模态表示可定义为 $\hat{\mathrm{E}}^u = [\hat{\mathrm{e}}^u_1, \hat{\mathrm{e}}^u_2, \ldots, \hat{\mathrm{e}}^u_{|\mathrm{Y}_u|}] \in \mathbb{R}^{|\mathcal{Y}_u| \times (|\mathbf{e}|+d)}$。接下来，我们可以计算合成样本与未见类别之间的语义相似度 $\mathrm{O}^u = \psi_{\theta_e}(\psi_{\theta_p}(\hat{\mathrm{E}}^u;); \mathrm{P};); \mathrm{Y}$。为提升模型识别未见类别的性能，我们使用以下交叉熵损失对模型进行优化：

$$\mathcal{L}_{ce} = -\sum_{i=1}^{|\mathcal{Y}_u|} y^u_i \log(\mathrm{softmax}(\mathbf{o}^u_i)) \quad (12)$$

其中，$\mathbf{o}^u_i \in \mathbb{R}^{|\mathcal{Y}_u|}$ 是第 i 个合成样本的语义相似度向量。考虑到交叉熵损失无法确保不同类别之间的相似度分数分布足够接近，平均相似度分数较低的类别将无法被有效识别。为缩小未见类别分数之间的间隔，我们提出如下形式的语义分布对齐损失：

$$\mathcal{L}_{align} = \mathrm{KL}(\mathrm{softmax}(\mathrm{Diag}(\mathbf{O}^u))||\mathrm{softmax}(\mathrm{Diag}(\mathbf{A}))) \quad (13)$$

其中，$\mathbf{A} \in \mathbb{R}^{|\mathcal{Y}_u| \times |\mathcal{Y}^u|}$ 是单位矩阵，$\mathrm{Diag}(\cdot)$表示提取矩阵对角线上的元素并将其作为向量。

4.4 训练和推理流程

为了端到端地训练包含所有损失函数的 HMGRL 框架，我们引入超参数来汇总不同的损失。因此，总体损失函数定义如下：

$$\mathcal{L}_{overall} = \mathcal{L}_{reg} + \mathcal{L}_{cl} + \mathcal{L}_{rank} + \mathcal{L}_{vae} + \eta \cdot \mathcal{L}_{ce} + \zeta \cdot \mathcal{L}_{align} \quad (14)$$

其中，η和 ζ是用于平衡与未见类别集相关损失的超参数。随后，我们采用随机梯度下降（SGD）[9] 算法，根据公式 14 计算出的损失值优化模型权重。

广义零样本多模态信息抽取方法通常对已见类别存在偏向，因为它们训练时只能使用来自已见类别集合的样本，随后却需要在已见和未见类别上进行评估。因此，这种偏向会导致将未见类别的数据误分类为已见类别。在推理过程中，我们引入校准因子 [14]以减轻上述偏向。给定来自测试集 D_{test}的一个样本，我们计算其与所有类别之间的语义相似度，记为 $\hat{\mathbf{O}} = \psi(\tilde{\mathbf{e}}; \theta_e)\psi([\mathbf{P}^s \oplus \mathbf{P}^u]; \theta_p)^T \in \mathbb{R}^{(|\mathcal{Y}_s|+|\mathcal{Y}_u|)}$。利用校准堆叠方法 [2], , 我们按如下函数预测该样本所属的类别：

$$\hat{y} = \underset{y_i \in \mathcal{Y}_s \cup \mathcal{Y}_u}{\arg\max}\ \hat{\mathbf{o}}_i - \gamma \cdot \mathbb{I}(y_i \in \mathcal{Y}_s) \quad (15)$$

其中 γ 是校准因子；若 y_i 属于已见类别集合，则 $\mathbb{I}(\cdot) = 1$，否则为 0。

5 实验

5.1 数据集与实验设置

我们研究广义零样本多模态信息抽取，并针对两项任务开展实验：多模态命名实体类型判定（MET）和多模态关系抽取（MRE）。在 MET 任务中，我们采用 WikiDiverse [24] 作为基准数据集，其中每个样本都包含一个来自 Wikinews 的文本-图像对，并将实体提及标注为 13 种细粒度类型。在 MRE 任务中，我们使用 Zheng 等人 [36], 介绍的数据集，该数据集由 Twitter 帖子构建而成。标注人员选取了涵盖多个主题的样本，构建出一个包含 23 种关系类型的数据集。由于两个数据集都包含非语义类别（例如“其他”或“无”），我们将其排除，仅保留具有语义意义的类别 [40]。

为模拟广义零样本学习条件 [31], , 我们将原始类别集随机划分为三个互不相交的子集。对于 MET 任务，训练集、验证集和测试集各分配 4 个类别。对于 MRE 任务，训练集、验证集和测试集分别分配 8、7 和 7 个类别。随后，我们从每个已见类别中随机抽取 70% 的实例。其中，90% 划入训练集，

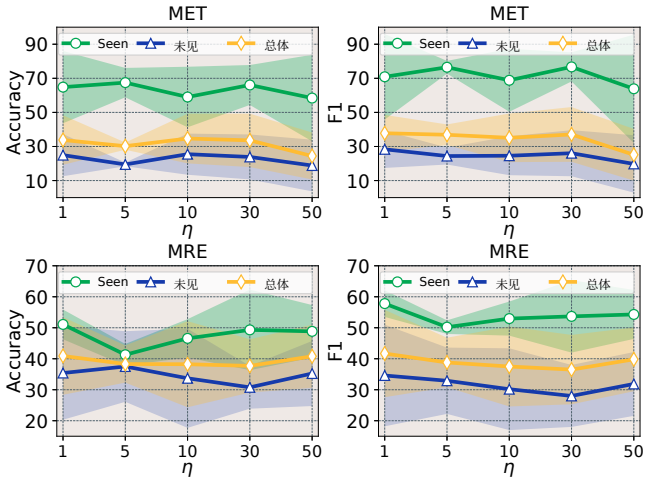


图 3：HM-GRL 在不同超参数 η 值下的性能对比分析。

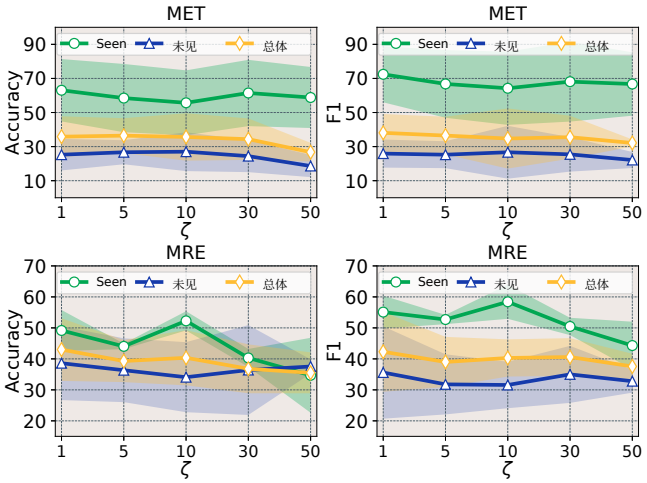


图 4：HM-GRL 在不同超参数 ζ 值下的性能比较分析。

剩余的 10% 划入验证集。测试集包含已见类别中剩余的 30% 实例，以及未见类别中的所有实例。隐藏表示的维度 d 和学习率分别设为 768 和 1e-5。MET 任务设置为 20 个训练轮次、批次大小为 14；MRE 任务设置为 25 个训练轮次、批次大小为 8。潜在表示的维度 h 在 {256,512, 768, 1024} 中搜索，而超参数 η 和 ζ 在 {1, 5, 10, 30, 50} 中搜索。对于 γ 的选择，我们在每个训练轮次结束后使用验证集评估模型。每项实验均使用不同的随机种子运行三次，并报告性能指标的均值和标准差。所有实验均使用 NVIDIA GTX A6000 设备加速。对比方法和评估指标详见附录A.1。

5.2 实验结果

在 MET 和 MRE 基准数据集上，将 HMGRL 框架与基线模型进行评估，详细实验结果如表1所示。在两个任务的大多数评估指标上，基于多模态的方法均优于基于文本的方法，这证实了融入视觉模态的有效性。在 MET数据集的“未见”和“总体”项目上，我们的模型在准确率和 F1分数方面均取得最高性能；与排名第二的结果相比，总体 F1 和准确率分别提升了 12.6% 和 12.2%。上述发现还凸显了所提出框架在识别新型实体类型方面的有效性，尽管其在已见类别上的性能略有下降。在 MRE数据集上，我们的模型在“未见”和“总体”项目的准确率和 F1 分数方面均超过基线方法；与第二优的结果相比，准确率提升了 6.1%，F1分数提升了 3.8%。使用大语言模型进行零样本推理且不经任何训练时，LLaVA 在 MET数据集上的表现优于我们的模型，但在 MRE数据集上则不及我们的模型。由于骨干模型 LLaMA [

22] LLaVA 使用维基百科页面和 MET

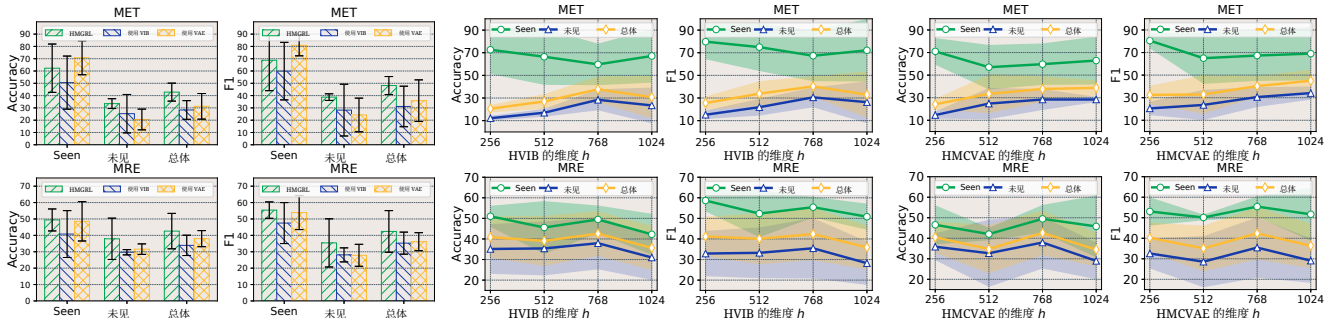
表2：HMGRL 在 MET 和 MRE 基准数据集上的消融研究结果。

模型	MET		MRE	
	准确率	F1	准确率	F1
HMGRL	37.7	40.3	42.6	42.4
不使用 HVIB	23.5	32.5	35.7	37.2
不使用 L_{cl}	34.5	33.9	33.4	37.8
不使用 L_{vae}	22.1	27.5	27.1	30.1
不使用 L_{ce}	26.0	30.4	35.3	36.0
不使用 L^{align}	36.2	35.6	39.8	39.8

数据集中的实体进行训练，并将这些信息存储在其中。总而言之，得益于 HVIB 网络促进的多模态表征对齐以及 HMCVAE 的开发，我们的模型优于纯文本和多模态基线方法。该方法能够有效地对样本的多模态表示进行建模并生成合成数据，最终提升性能。

5.3 消融研究

我们进行了消融研究，以全面评估 HMGRL 中各个模块的影响，结果见表2。值得注意的是，去除 HM-CVAE 的损失函数会导致性能显著下降，这凸显了对样本的多模态表示进行建模并为未见类别生成合成样本的重要作用。为评估多模态表征对齐的影响，我们分别移除了 HVIB 和损失项 L_{cl}，观察到缺少这些组件的模型性能均出现了不同程度的下降。在 MET数据集上，HVIB 的影响大于损失 L_{cl}，而在 MRE数据集上则相反。此外，我们还评估了通过损失 L_{ce} 和 L^{align}利用未见类别的影响。结果表明，利用未见类别对于提升模型的最终性能至关重要。相应结果表明，对未见类别的



（a）洛伦兹线性层的消融实验结果在 HVIB 和 HMCVAE 中。

(b) HMGRL 在不同维度下的性能比较 HVIB 潜在表示的 (h)

(c) HMGRL 在不同维度下的性能比较 HMCVAE 潜在表示的 (h)

Fig图 5：洛伦兹线性层对 HMGRL 框架的影响实验结果

合成样本进行分类对于提升模型的最终性能十分重要。与交叉熵损失相比，语义分布对齐损失在一定程度上提升了我们所提出框架的泛化能力。总之，消融研究的结果验证了将 HVIB 和 HMCVAE 与用于对样本的多模态表示进行建模的辅助损失相结合，能够提升模型在 GZS-MIE任务上的性能。

5.4 超参数 η 和 ζ的影响

在所提出的框架中，未见类别的利用涵盖分类和语义分布对齐。为了平衡真实样本和合成样本的利用，我们引入超参数 η 和 ζ。为评估超参数 η的影响，我们进行了参数敏感性实验，如图 3 所示。超参数 η 控制 HMGRL 对未见类别分类的重视程度。随着模型将更多注意力放在未见类别分类上，由于合成样本质量各异，已见类别上的性能可能会下降。过大的η 值会显著损害 HMGRL 在 MET数据集上的总体性能，因为模型对未见类别的关注会对其处理已见类别的能力产生不利影响。此外，超参数 ζ 会影响语义分布对齐的程度。我们也进行了参数敏感性实验，以评估其影响，如图 4 所示。随着 ζ 值上升，模型性能会下降；过度强调语义分布对齐会导致一定程度的语义匹配失败，这归因于语义表示空间的连续性。

5.5 洛伦兹线性层的影响

为了评估洛伦兹线性层对所提出框架的影响，我们开展了全面实验，对 HVIB 和 HMCVAE 进行分析。首先，我们分别移除了 HVIB 和 HMCVAE 中的洛伦兹线性层，消融实验结果如图 5a 所示。与采用变分信息瓶颈或变分自编码器的模型相比，所提出的框架能够取得更优的结果，尤其是在“未见”和“总体”项目上。上述发现凸显了洛伦兹线性层对最终结果所起的重要作用。此外，鉴于洛伦兹线性层用于编码潜在表示，我们开展了实验，以探究这些潜在表示的维度对 HMGRL 的影响。如图 5b 所示，随着 HVIB 中潜在表示的维度增加，HMGRL 的性能呈现先提升后下降的趋势。此外，如图 5c 所示，HMCVAE 中潜在表示维度的增加会显著影响 HMGRL 的性能，因为这会直接影响模型构建多模态表示并生成合成训练样本。上述实验结果进一步证明，在我们提出的框架中，洛伦兹线性层会影响生成式表示学习。

6 结论

本研究深入探讨了广义零样本多模态信息抽取（GZS-MIE）任务，旨在应对层次语义表示和语义相似度分布间隔所带来的挑战。为克服这些问题，我们提出了双曲多模态生成式表示学习（HMGRL）框架。HMGRL 框架由双曲变分信息瓶颈（HVIB）和带有辅助损失的双曲多模态条件变分自编码器（HMCVAE）组成。HVIB网络利用洛伦兹线性层（L.L.L）捕捉不同模态之间的层次语义联系，从而提取对齐的多模态表示。HMCVAE网络通过L.L.L捕捉样本与原型之间的层次语义相关性，从而能够对样本表示进行建模，并生成未见类别的合成实例。这使得HMGRL能够通过语义分布对齐损失进行训练，从而提升性能。实验结果表明，HMGRL在GZS-MIE 任务上表现优异。

致谢

本研究得到中国国家自然科学基金（编号：72342017）和南开大学中央高校基本科研业务费（63253232）的资助。

参考文献

[1] Hangbo Bao、Wenhui Wang、Li Dong、Qiang Liu、Owais Khan Mohammed、Kriti Aggarwal、Subhojit Som、Songhao Piao 和 Furu Wei。2022。VLMo：基于混合模态专家的统一视觉-语言预训练。载于神经信息处理系统进展第*35*卷：*2022*年神经信息处理系统年会，*NeurIPS 2022*，美国路易斯安那州新奥尔良，*2022*年*11*月*28*日至*12*月*9*日。[2] Wei-Lun Chao、Soravit Changpinyo、Boqing Gong 和 Fei Sha。2016。野外目标识别广义零样本学习的实证研究与分析。载于计算机视觉——*ECCV 2016*——第*14*届欧洲会议（《计算机科学讲义》，第*9906*卷）。Springer，52–68。[3] Chih-Yao Chen 和 Cheng-Te Li。2021。ZS-BERT：通过属性表示学习实现零样本关系抽取。载于*2021*年北美计算语言学协会年会论文集：人类语言技术。计算语言学协会，线上。[4] Yi Chen、Haiyun Jiang、Lemao Liu、Shuming Shi、Chuang Fan、Min Yang 和 Ruifeng Xu。2021。多种信息来源对零样本细粒度实体类型判定的实证研究。载于*2021*年自然语言处理经验方法会议论文集，*EMNLP 2021*，线上会议／多米尼加共和国蓬塔卡纳，*2021*年*11*月*7*日至*11*日。计算语言学协会，2668–2678。[5] Shiyao Cui、Jiangxia Cao、Xin Cong、Jiawei Sheng、Quangang Li、Tingwen Liu 和 Jinqiao Shi。2024。利用变分信息瓶颈增强多模态实体与关系抽取。美国电气电子工程师学会／美国计算机协会音频、语音与语言处理汇刊32（2024），1274–1285。[6] Alexey Dosovitskiy、Lucas Beyer、Alexander Kolesnikov、Dirk Weissenborn、Xiaohua Zhai、Thomas Unterthiner、Mostafa Dehghani、Matthias Minderer、Georg Heigold、Sylvain Gelly、Jakob Uszkoreit 和 Neil Houlsby。2021。一张图像胜过16×16个词：用于大规模图像识别的Transformer。载于第*9*届国际学习表征会议，*ICLR 2021*，线上会议，奥地利，*2021*年*5*月*3*日至*7*日。[7] Octavian-Eugen Ganea、Gary Bécigneul 和 Thomas Hofmann。2018。双曲神经网络。载于神经信息处理系统进展第*31*卷：*2018*年神经信息处理系统年会，*NeurIPS 2018*，加拿大蒙特利尔，*2018*年*12*月*3*日至*8*日。5350–5360。[8] Jiaying Gong 和 Hoda Eldardiry。2024。通过语义知识增强实现基于提示的零样本关系抽取。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*，意大利都灵，*2024*年*5*月*20*日至*25*日。欧洲语言资源协会和国际计算语言学委员会，13143–13156。[9] Diederik P. Kingma 和 Jimmy Ba。2015。Adam：一种随机优化方法。载于第*3*届国际学习表征会议。[10] Diederik P. Kingma 和 Max Welling。2014。自动编码变分贝叶斯。载于第*2*届国际学习表征会议，*ICLR 2014*，加拿大艾伯塔省班夫，*2014*年*4*月*14*日至*16*日，会议论文集。[11] Qiuyu Liang、Weihua Wang、Feilong Bao 和 Guanglai Gao。2024。L\$^2$GC：用于节点分类的洛伦兹线性图卷积网络。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。欧洲语言资源协会和国际计算语言学委员会，9988–9998。[12] Qiuyu Liang、Weihua Wang、Cunda Wang、Feilong Bao 和 Jie Yu。2025。双曲多模态知识图谱嵌入。载于*2025*年*IEEE*声学、语音与信号处理国际会议。电气电子工程师学会，1–5。[13] Haotian Liu、Chunyuan Li、Qingyang Wu 和 Yong Jae Lee。2023。视觉指令微调。载于神经信息处理系统进展第*36*卷：*2023*年神经信息处理系统年会，*NeurIPS 2023*，美国路易斯安那州新奥尔良，*2023*年*12*月*10*日至*16*日。[14] Yang Liu、Lei Zhou、Xiao Bai、Yifei Huang、Lin Gu、Jun Zhou 和 Tatsuya Harada。2021。面向零样本学习的目标导向注视估计。载于*IEEE*计算机视觉与模式识别会议。计算机视觉基金会／电气电子工程师学会，3794–3803。[15] Yukun Ma、Erik Cambria 和 Sa Gao。2016。用于零样本细粒度命名实体类型判定的标签嵌入。载于*COLING 2016*，第*26*届国际计算语言学会议论文集：技术论文，*2016*年*12*月*11*日至*16*日，日本大阪。ACL，171–180。[16] Laurens Van Der Maaten 和 Hinton Geoffrey。2008。使用t-SNE可视化数据。机器学习研究期刊9，2605（2008），2579–2605。[17] Seungwhan Moon、Leonardo Neves 和 Vitor Carvalho。2018。面向短社交媒体帖子的多模态命名实体识别。载于*2018*年北美计算语言学协会年会论文集：人类语言技术，*NAACL-HLT 2018*，美国路易斯安那州新奥尔良，*2018*年*6*月*1*日至*6*日，第*1*卷（长篇论文）。计算语言学协会，852–860。[18] Rasha Obeidat、Xiaoli Z. Fern、Hamed Shahbazi 和 Prasad Tadepalli。2019。基于描述的零样本细粒度实体类型判定。载于2019年北美计算语言学协会年会论文集：人类语言技术，NAACL-HLT 2019，明尼阿波利斯，美国明尼苏达州，*2019*年*6*月*2*日至*7*日，第*1*卷（长篇与短篇论文）. 计算语言学协会，807–814。[19] Farhad Pourpanah、Moloud Abdar、Yuxuan Luo、Xinlei Zhou、Ran Wang、Chee Peng Lim、Xi-Zhao Wang 和 Q. M. Jonathan Wu。2023年。广义零样本学习方法综述。*IEEE*模式分析与机器智能汇刊 45, 4 (2023), 4051–4070。[20] Yankun Ren、Jianbin Lin 和 Jun Zhou。2020年。神经零样本细粒度实体类型判定。载于*2020*年万维网大会伴随会议，*2020*年，中国台湾台北，*2020*年*4*月*20*日至*24*日。美国计算机协会 / 万维网会议国际指导委员会，846–847。[21] Naftali Tishby 和 Noga Zaslavsky。2015年。深度学习与信息瓶颈原理。载于*2015*年*IEEE*信息论研讨会，*ITW 2015*，以色列耶路撒冷，*2015*年*4*月*26*日至*5*月*1*日。电气电子工程师学会，1–5。[22] Hugo Touvron、Thibaut Lavril、Gautier Izacard、Xavier Martinet、Marie-Anne Lachaux、Timothée Lacroix、Baptiste Rozière、Naman Goyal、Eric Hambro、Faisal Azhar、Aurélien Rodriguez、Armand Joulin、Edouard Grave 和 Guillaume Lam- ple。2023年。LLaMA：开放且高效的基础语言模型。计算研究文库 abs/2302.13971 (2023)。[23] Hai Wan、Manrong Zhang、Jianfeng Du、Ziling Huang、Yufei Yang 和 Jeff Z. Pan。2021年。FL-MSRE：一种基于小样本学习的多模态社交关系抽取方法。载于第三十五届*AAAI*人工智能大会，*AAAI 2021*，第三十三届人工智能创新应用大会，*IAAI 2021*，第十一届人工智能教育进展研讨会，*EAAI 2021*，线上会议，*2021*年*2*月*2*日至*9*日。AAAI出版社，13916–13923。[24] Xuwu Wang、Junfeng Tian、Min Gui、Zhixu Li、Rui Wang、Ming Yan、Lihan Chen 和 Yanghua Xiao。2022年。WikiDiverse：具有多样化上下文主题和实体类型的多模态实体链接数据集。载于计算语言学协会第*60*届年会论文集（第*1*卷：长篇论文）。计算语言学协会，爱尔兰都柏林，4785–4797。[25] Yongqin Xian、Christoph H. Lampert、Bernt Schiele 和 Zeynep Akata。2019年。零样本学习——对好、坏与丑的全面评估。*IEEE*模式分析与机器智能汇刊 41, 9 (2019), 2251–2265。[26] Bo Xu 、 Shizhou Huang 、 Chaofeng Sha 和 Hongya Wang。2022年。MAF：面向多模态命名实体识别的通用匹配与对齐框架。载于*WSDM '22*：第十五届*ACM*国际网络搜索与数据挖掘会议，线上会议 / 美国亚利桑那州坦佩，*2022*年*2*月*21*日至*25*日。美国计算机协会，1215–1223。[27] Jianfei Yu、Jing Jiang、Li Yang 和 Rui Xia。2020年。通过统一多模态Transformer的实体跨度检测改进多模态命名实体识别。载于计算语言学协会第*58*届年会论文集。计算语言学协会，线上，3342–3352。[28] Li Yuan、Yi Cai、Jin Wang 和 Qing Li。2023年。基于边增强图对齐网络和词对关系标注的联合多模态实体关系抽取。载于第三十七届*AAAI*人工智能大会。AAAI出版社，11051–11059。[29] Masoumeh Zareapoor、Pourya Shamsolmoali 和 Yue Lu。2025年。BiMAC：对比学习中的双向多模态对齐。载于*AAAI-25*。AAAI出版社，22290–22298。[30] Tao Zhang、Congying Xia、Chun-Ta Lu 和 Philip S. Yu。2020年。MZET：记忆增强型零样本细粒度命名实体类型判定。载于第*28*届国际计算语言学会议论文集，*COLING 2020*，西班牙巴塞罗那（线上），*2020*年*12*月*8*日至*13*日。国际计算语言学委员会，77–87。[31] Xiaotong Zhang、Qianru Zhou、Han Liu 和 Hong Yu。2025年。通过参数高效快速适应进行基于成对提示的调优，以实现广义零样本意图检测。载于计算语言学协会论文集：*NAACL 2025*。计算语言学协会，917–929。[32] Ying Zhang、Wenbo Fan、Kehui Song、Yu Zhao、Xuhui Sui 和 Xiaojie Yuan。2023 年。将对象级视觉上下文融入多模态细粒度实体类型判定。载于计算语言学协会论文集：*EMNLP 2023*。计算语言学协会，新加坡，15380–15390。[33] Jun Zhao、WenYu Zhan、Xin Zhao、Qi Zhang、Tao Gui、Zhongyu Wei、Junzhe Wang、Minlong Peng 和 Mingming Sun。2023年。RE-Matching：一种用于零样本关系抽取的细粒度语义匹配方法。载于计算语言学协会第*61*届年会论文集（第*1*卷：长篇论文），*ACL 2023*，加拿大多伦多，*2023*年*7*月*9*日至*14*日。计算语言学协会，6680–6691。[34] Yu Zhao、Ying Zhang、Xuhui Sui、Baohang Zhou、Haoze Zhu、Jeff Z. Pan 和 Xiaojie Yuan。2025年。模态的暗面：用于多模态知识图谱推理的强化多模态蒸馏。计算研究文库 abs/2507.20738 (2025)。[35] Yu Zhao、Ying Zhang、Baohang Zhou、Xinying Qian、Kehui Song 和 Xiangrui Cai。2024年。先对比后记忆：语义邻居检索增强的归纳式多模态知识图谱补全。载于第*47*届*ACM SIGIR*信息检索研究与发展国际会议，*SIGIR 2024*，美国华盛顿特区，*2024*年*7*月*14*日至*18*日。美国计算机协会，102–111。[36] Changmeng Zheng、Junhao Feng、Ze Fu、Yi Cai、Qing Li 和 Tao Wang。2021年。采用高效图对齐的多模态关系抽取。载于*MM '21*：*ACM*多媒体会议，线上会议，中国，*2021*年*10*月*20*日至*24*日。美国计算机协会，5298–5306。

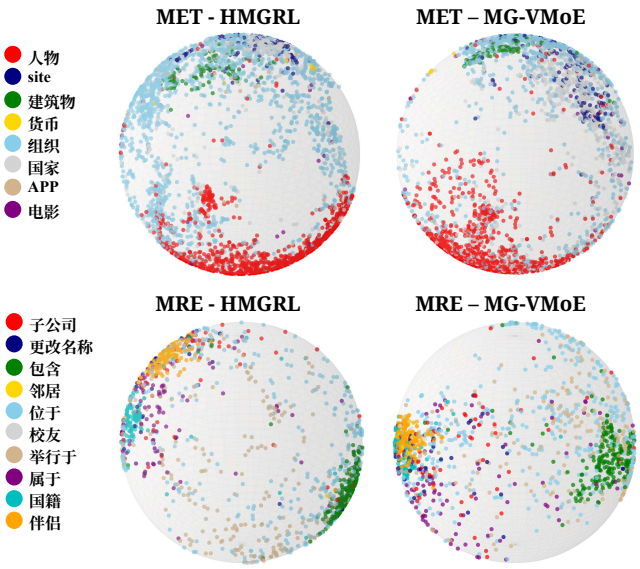


图6：通过 t-SNE 将 GZS-MIE 模型的多模态特征 $\tilde{e}$ 投影到球面流形上进行可视化。

[37] Baohang Zhou、Ying Zhang、Kehui Song、Wenya Guo、赵国庆、王宏斌和 Xiaojie Yuan。2022年。一种用于半监督多模态命名实体识别的基于跨度的多模态变分自编码器。载于*2022*年自然语言处理实证方法会议论文集。计算语言学协会，阿布扎比，阿拉伯联合酋长国，6293–6302。[38] Baohang Zhou、Ying Zhang、Kehui Song、Xuhui Sui、Yu Zhao 和 Xiaojie Yuan。2025年。ZS-MNET：一种基于零样本学习的多模态命名实体类型判定方法。神经网络 186 (2025), 107264。[39] Baohang Zhou、Ying Zhang、Kehui Song、Hongru Wang、Yu Zhao、Xuhui Sui 和 Xiaojie Yuan。2024年。MCIL：面向低资源实体级多模态信息抽取的多模态反事实实例学习。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*）。欧洲语言资源协会和国际计算语言学委员会，都灵，意大利，11101–11110。[40] Baohang Zhou、Ying Zhang、Yu Zhao、Xuhui Sui 和 Xiaojie Yuan。2025年。用于零样本多模态信息抽取的基于多模态图的变分混合专家网络。载于*ACM*网络会议论文集*2025*，*WWW 2025*，澳大利亚新南威尔士州悉尼，*2025*年*4*月*28*日至*5*月*2*日。美国计算机协会，4823–4831。[41] 朱祥儒、Zhixu Li、王晓丹、蒋雪瑶、孙鹏磊、Xuwu Wang、Yanghua Xiao 和袁景。2024年。多模态知识图谱构建与应用：一项综述。*IEEE*知识与数据工程汇刊 36，2（2024年），715–735。

附录A

A.1 对比方法和评估指标

对于零样本信息抽取（ZS-IE），原型（Proto）[15] 网络是一种基于标签嵌入的方法，用于解决这两个任务。此外，我们选用基于注意力机制的DBZFET [18] 和NZFET [20] 模型，以及基于记忆增强的MZET [30] 作为零样本实体类型（ZS-ET）基线。除此之外，我们将基于表示的ZS-BERT [3]模型、细粒度语义匹配模型RE-match [33],和基于提示的ZS-SKA [8] 模型作为零样本关系抽取（ZS-RE）基线。对于零样本多模态信息抽取（ZS-MIE），我们选用多模态原型网络（MMProto） [23], 、基于目标视觉上下文的模型MOVCNet [32] 以及基于细粒度多模态表示学习的模型MG-VMoE [40] 作为多模态基线。我们还选用7B版本的LLaVA（v1.5） [13]作为执行广义零样本多模态信息抽取（GZS-MIE）任务的基线。在方法评估中，我们使用准确率和加权F1分数作为指标，并分别报告测试集中已见和未见类别的结果。总体指标通过计算已见和未见类别上的准确率与F1分数的调和平均数得到 [25]。

A.2 可视化分析

为评估表示学习在GZS-MIE任务中的有效性，我们对HMGRL和MG-VMoE模型提取的表示 $\tilde{\mathbf{e}}$ 进行了可视化，如图6所示。将测试集中来自已见和未见类别的样本输入两个模型，以获取各自的多模态表示。随后，使用t-SNE将这些特征投影到球面流形上进行降维 [16] 。结果表明，与MG-VMoE相比，HMGRL在特定类别内生成的表示聚类更加紧密。例如，在MET数据集上，“人物”和“国家”等类别的多模态表示呈现出更紧密的类别内聚类。此外，与MRE数据集上的MG-VMoE相比，HMGRL的表示展现出更强的类别级可区分性。因此，可视化分析证实了HMGRL在为GZS-MIE任务建模多模态表示方面的有效性。