arXiv:2601.14732v1 [cs.CV] 21 Jan 2026

DEEPMOLM：利用视觉与几何结构信息进行分子-文本建模

Jing Lan[1*], **Hexiao Ding**[1*], **Hongzhao Chen**[1*], **Yufeng Jiang**[1], **Nga-Chun Ng**[1,2], **Gwing Kei Yip**[1,3], **Gerald W.Y. Cheng**[1], **YunlinMao**[1], **Jing Cai**[1], **Liang-ting Lin**[1], **Jung Sun Yoo**[1†]健康技术与信息学系，香港理工大学[2]核医学与正电子发射断层显像系，香港养和医院[3]诊断及介入放射学系，伊丽莎白女王医院，中国香港特别行政区 {jing-hti.lan, hexiao.ding, hongzhao.chen, yufeng.jiang, yunlin.mao}@connect.polyu.hk {wai-yeung.cheng, jing.cai, ltlin, jungsun.yoo}@polyu.edu.hk sam.nc.ng@hksh.com, gwinky.yip@ha.org.hk

摘要

用于药物发现和化学文献挖掘的人工智能模型必须能够解读分子图像，并生成符合三维几何和立体化学的输出。大多数分子语言模型依赖字符串或图，而视觉语言模型往往会遗漏立体化学细节，也难以将连续的3D结构映射为离散标记。我们提出**DeepMoLM**：深度分子语言建模，这是一种双视图框架，可利用从分子构象中提取的几何不变量，为高分辨率分子图像提供依据。DeepMoLM 保留来自 1024×1024 输入的高频信息，将构象邻域编码为离散的扩展三维指纹，并通过交叉注意力融合视觉与几何流，从而无需原子坐标即可生成符合物理规律的结果。与表现最强的通用基线相比，DeepMoLM 在 PubChem 分子描述生成任务上的 METEOR 相对提升了 12.3%，同时保持了与专科模型相当的竞争力。对于所有属性查询，它都能生成有效的数值输出；在专科模型设置下，其分子量预测的 MAE 为 13.64 g/mol，复杂度预测的 MAE 为 37.89。在基于图像生成 ChEBI-20 描述的任务中，它优于通用基线，并达到与最先进视觉语言模型相当的水平。代码可在 https://github.com/1anj/DeepMoLM 获取。

***K*关键词** 药物发现、光学内容识别、分子性质预测、分子描述生成、视觉语言模型

1 引言

分子的物理本质遵循三维（3D）几何规律和量子约束 [1], ，但大量化学知识以分子图像的形式存储在科学文章中，包括二维绘图和渲染结构图 [2]。然而，许多人工智能模型无法将文本与其底层三维几何相对齐。因此，这些分子图像通常被视为无用的插图，而不是结构化输入。在这种情况下，键布局和立体化学标记等视觉线索无法与一致的物理构象关联起来，分子图像中的信息也未得到充分利用。

分子图像与语言之间的双视图一致性取决于三个方面的差距。首先，普遍采用的多模态设计会冻结单模态编码器，并依赖轻量级适配器进行后期融合 [3]。这种弱耦合会助长捷径对齐，因此模型关注粗粒度语义，却忽视几何结构和立体化学约束。其次，将连续的三维结构投影到离散的语言空间中，可能会破坏几何结构 [4]。物理无效性

*共同第一作者
†通讯作者

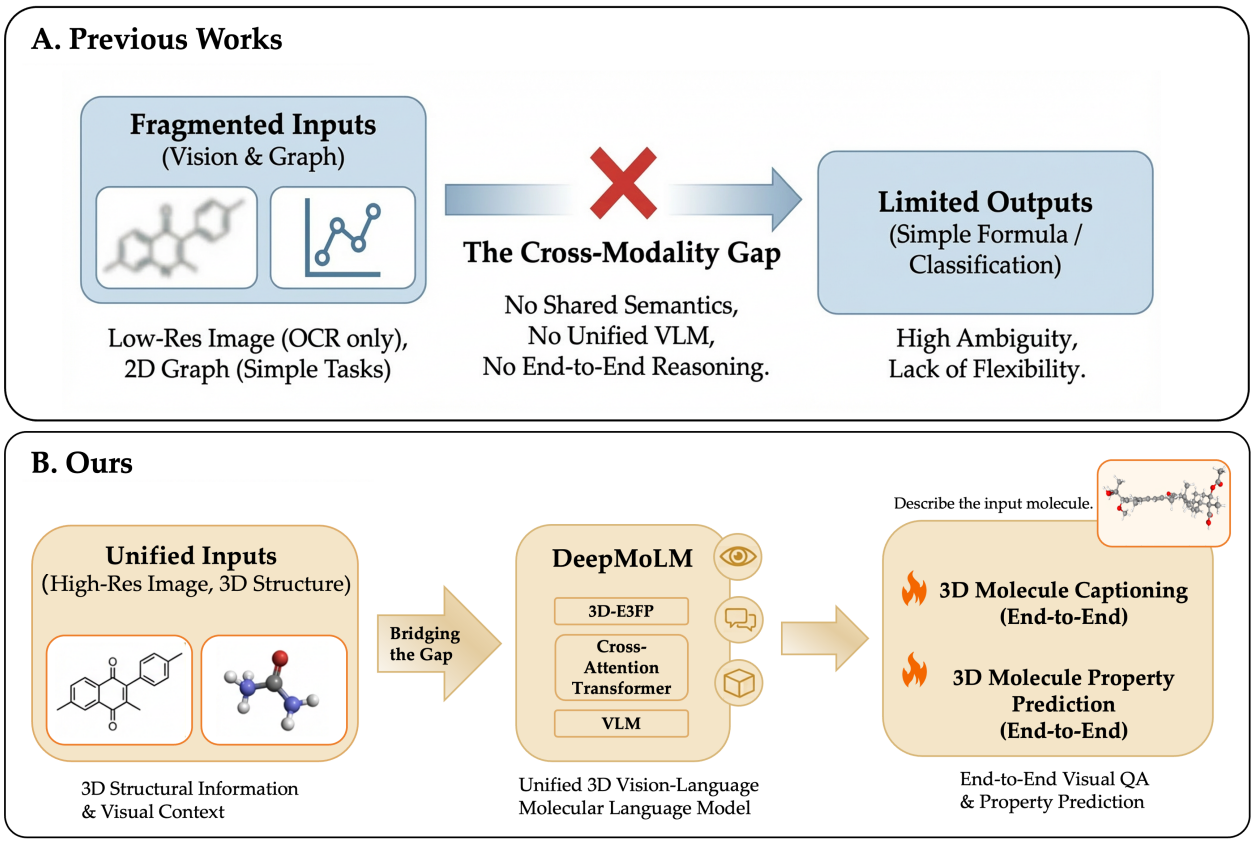


图1：**以往方法与 DeepMoLM 的比较。(A)** 此前的研究使用经光学化学结构识别处理的低分辨率图像或二维分子图来表示分子。**(B)** DeepMoLM 将高分辨率分子图像与显式的基于三维构象的结构描述符相结合。通过在视觉语言模型中建立视觉模态与几何模态之间联系的交叉注意力机制，DeepMoLM 支持端到端执行三维分子描述生成和性质预测等复杂任务。

会在未保留刚体运动等变性和立体化学不变量时发生。同一分子的旋转或平移可能改变其嵌入。这样一来，对映异构体或密切相关的立体异构体可能无法区分。

在存在构象歧义时，多种构象会坍缩为相似的表示，因此生成的文本虽然流畅，却可能错误描述手性或空间关系 [5, 6]。这是几何表示坍缩，而非能量相关问题。因此，第三个缺口显而易见。分子示意图通过局部高频线索（如立体键和环闭合）来编码语义。在标准分辨率下，视觉编码器往往会丢失这些细节；而提高保真度的代价很高，因为自注意力的计算量会随序列长度呈二次方增长 [7]。DeepMoLM 通过融合模块将图像特征与三维感知指纹分支相结合，从而在构象不确定时提升立体化学稳定性，并在推理时无需原子坐标即可进行跨模态推理。

我们提出 DeepMoLM，在端到端视觉语言框架中统一分子图像理解与三维结构。在 DeepMoLM 中，DeepEncoder 是一个双路径模块，可保留分子图像中的高频线索，并通过卷积式词元压缩实现高效的 1024×1024 处理。该架构最初是为光学字符识别 [8]开发的。局部路径捕捉精细的立体化学细节，全局路径则通过聚合长程上下文来确保结构连贯性。为连接连续的三维几何与语言模型，我们使用扩展三维指纹（E3FP） [9]引入显式结构不变量。不同于拼接，交叉注意力融合投影器使视觉令牌能够查询几何描述符，从而在解码前对图像特征进行锚定。

贡献如下。**(1)** 我们集成了一种双路径 DeepEncoder，用于处理 1024×1024 分子图像，并通过卷积式词元压缩保留细粒度的立体化学标记。**(2)** 我们提出了一种交叉注意力融合投影器，将视觉表征与离散的 3D E3FP 指纹对齐，在统一的多模态嵌入空间中提供明确的几何基础。**(3)** 实验结果表明，

DeepMoLM 中的几何基础提升了分子描述生成和性质预测的表现，而这两者都与立体化学有关。

2 相关工作

2.1 分子表征

在药物发现中，分子表示学习通常以符号形式为基础。一种常见方法是将分子表示为线性字符串，例如 SMILES 或 SELFIES。随后，在这些字符串上预训练 Transformer，以学习语法、子结构和生成模式。MolT5 将分子到文本和文本到分子任务联系起来，从而支持在多种下游场景中的迁移 [10]。面向化学的指令微调和领域适配也遵循相同思路，包括 ChemLLM 和 Galactica 等大型科学语言模型 [11, 12]。这些方法仍依赖一维字符串或二维描述符，因为语言模型中的分词和词汇表无法自然地编码完整的分子结构。图神经网络通过将原子和化学键建模为二维图来应对这一挑战，其中节点状态通过消息传递进行迭代更新 [13]。自监督图预训练（例如 GROVER）通过从大量无标签图中学习化学规律，提升了低标签场景下的泛化能力 [14]。多视图预训练将几何信息融入表示学习。GraphMVP 将二维拓扑与三维构象监督相结合，以提升迁移能力 [15]。

2.2 分子图像的视觉语言模型

基于图像的方法加入视觉信息，以增强图表示 [16], ；对分子图像进行自监督训练表明，二维渲染图能够承载丰富的化学内容 [17]。光学化学结构识别（OCSR）旨在将文档中的化学结构图转换为机器可理解的图表示，包括 PDF 和扫描件中的化学结构图 [18]。MolScribe 通过同步进行符号解析和分子图构建来识别结构，确保在解码过程中明确建模连接关系 [19]。SwinOCSR 和 ChemVLM 则增强视觉主干网络，以处理更复杂的输入 [20, 21]。大多数 OCSR 方法仍会生成二维图或字符串，而不应用三维几何约束。此外，当图像模糊、质量较差或绘制风格罕见时，立体化学仍难以确定 [18, 19]。OCSR 将显式分子结构数字化，而视觉语言模型则捕捉隐式语义，以支持更高层次的推理。分子图像通过原子标记、键线和立体化学标记编码信息。这些信号稀疏且局部化，并非分布于纹理之中。要保留这些信息，需要高分辨率，但视觉 Transformer 的计算成本较高，因为自注意力的复杂度会随词元数量呈平方增长 [7]。DeepEncoder 采用带有卷积式词元压缩的双路径编码器，即使在处理 1024×1024 分子图像 [8]时也能节省词元预算。

2.3 几何学习与多模态融合

Uni-Mol 等方法以及 SchNet、DimeNet 和 GemNet 等三维模型，将物理几何中的归纳偏置融入模型。研究表明，三维不变量对于量子级性质预测至关重要 [1, 22, 23, 24]。SE(3) Transformer 和 EGNN 强制满足对称性，使模型能够从构象和分子点云中学习，同时保持对空间变换的不变性 [25]。近期研究还将三维编码器与语言模型相结合，以实现分子—文本交互。3D-MoLM 通过学习得到的投影器，将三维分子编码器与语言模型对齐，从而提升模型在三维分子—文本数据上的指令遵循能力 [4]。该方法假设能够获取结构化三维输入，并不旨在实现基于分子图像的强大视觉感知。3D-MolT5 将细粒度三维子结构标记化，并在预训练过程中将其与序列词元对齐。这种设计加强了几何与语言之间的耦合 [11]。

3 方法论

DeepMoLM 的架构如图2所示。它包括用于高分辨率视觉特征提取的 **分子深度编码器** 、用于对齐视觉与几何模态的 **多模态融合投影器** ，以及**基于 Qwen2-VL 构建的视觉语言解码器** [26]。

文本生成被构建为一项多模态自回归任务。每个分子 m 由一张高分辨率图像 $\mathbf{I}_m \in \mathbb{R}^{H\times W\times C}$ 表示，包含 $H = W = 1024$ 和 $C = 3$；同时还由一个结构标记序列 $\mathbf{S}_m$ 表示，该序列由规范 SELFIES 标记和 E3FP 标识符构成。给定提示词$\mathbf{X}^{\text{prompt}}$，模型通过最大化

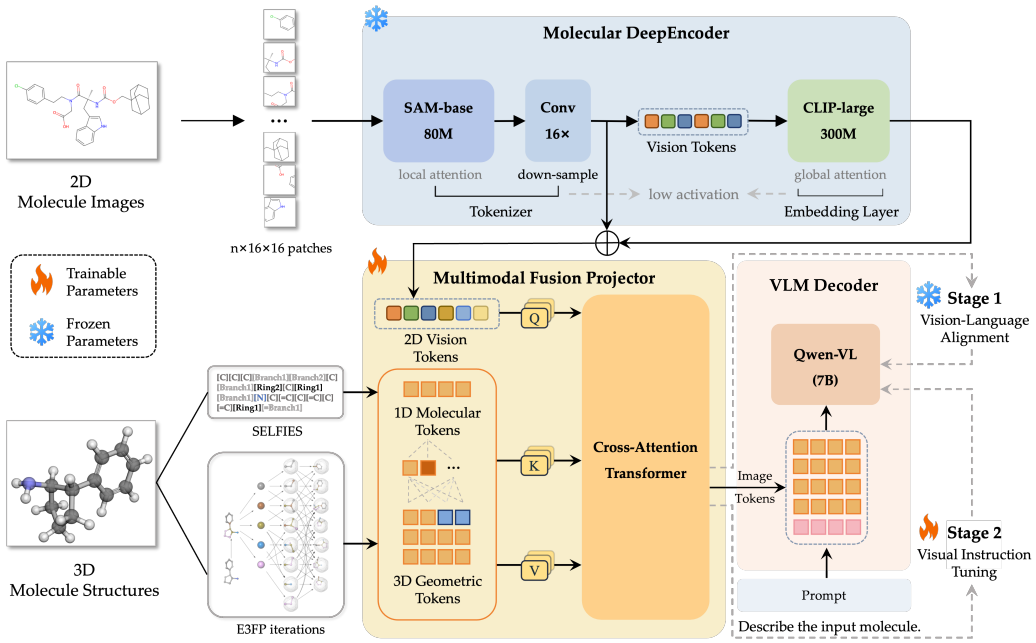


图2：**DeepMoLM 的架构**。该框架分三个阶段实现双视图分子理解。(1) **DeepEn-coder** 使用 SAM-Base 局部编码器、卷积式词元压缩和 CLIP-Large 全局编码器，从高分辨率分子图像中提取多尺度特征。(2) **融合投影器**通过交叉注意力，将二维视觉标记与离散的三维几何指纹融合。(3) **VLM解码器** 根据融合表示生成描述文本和性质文本。

$\mathbf{y} = [y_1, \ldots, y_T]$ 来生成目标序列

$$p(\mathbf{y} \mid \mathbf{X}_{\text{prompt}}, \mathbf{I}_m, \mathbf{S}_m) = \prod_{t=1}^{T} p(y_t \mid \mathbf{y}_{<t}, \mathbf{X}_{\text{prompt}}, \mathbf{I}_m, \mathbf{S}_m)\,. \tag{1}$$

3.1 分子结构表示

3.1.1 结构感知一维分子标记

我们使用自引用嵌入字符串（SELFIES） [27]表示分子拓扑。对于分子 m，规范SELFIES生成一个词元序列

$$\mathcal{T}_m^{(1\text{D})} = \{s_1, s_2, \ldots, s_L\}. \tag{2}$$

由于存在非原子控制词元，其长度 L 可能超过重原子的数量。我们定义一个原子位置索引集 $\mathcal{A} \subseteq \{1, \ldots, L\}$ 以及一个双射

$$\phi : \{1, \ldots, N_{\text{ha}}\} \to \mathcal{A}, \tag{3}$$

其中， N_{ha} 表示重原子的数量。映射 ϕ 通过将规范SELFIES解析为分子图，并记录引入重原子的SELFIES位置来获得。

3.1.2 三维几何标记

我们使用扩展三维指纹（E3FP） [9]来表示构象几何。构象坐标表示为$\mathbf{X}_m \in \mathbb{R}^{N_{\text{ha}} \times 3}$。对于每个重原子 a_i，E3FP会生成哈希标识符$\hat{d}_{i,\,j}$，覆盖 $K+1$ 个半径。$\hat{}$初始标识符$d_{i,\,0}$由原子级不变量导出。在迭代 $j \in \{1, \ldots, K\}$ 中，对于半径为 $R_j = r \cdot j$的范围，E3FP根据$\mathbf{X}_m$中的连接关系和立体化学构型，收集距离在 R_j 以内的邻居，然后对$\hat{}$聚合结果进行哈希，得到$d_{i,\,j}$。我们将每个标识符映射到大小为 $|F|$ 的离散词汇表中，方法是

$$d_{i,j} \triangleq \hat{d}_{i,j} \bmod |F|, \qquad j \in \{0, \ldots, K\}. \tag{4}$$

由此得到一个原子级三维标记元组

$$\mathbf{d}_i = [d_{i,0}, d_{i,1}, \ldots, d_{i,K}], \qquad i \in \{1, \ldots, N_{\text{ha}}\}, \tag{5}$$

以及一个带索引的三维标记表

$$\mathbf{D}_m \in \{0, \ldots, |F|-1\}^{N_{\text{ha}} \times (K+1)}, \quad \mathbf{D}_m[i,j] = d_{i,j}. \tag{6}$$

SELFIES 提供离散的拓扑骨架 $\mathcal{T}_{(1\text{D})\,m}$，E3FP 则提供按重原子索引、反映分子构象的 3D 描述符。我们通过将 $\mathbf{d}_i$ 分配给 SELFIES 位置 $t = \phi(i)$，使二者对齐。位置 $t \notin \mathcal{A}$没有对应的重原子，也没有关联的 3D 元组。

3.1.3 词元嵌入融合

我们使用可训练嵌入$\mathbf{E}_{1\text{D}}$： $\mathcal{V}_{1\text{D}} \to \mathbb{R}^{d_s}$和$\mathbf{E}_{3\text{D}}$： $\{0, \ldots, |F|-1\} \to \mathbb{R}^{d_s}$。对于每个 SELFIES 位置 t，我们计算

$$\mathbf{e}_t^{(1\text{D})} = \mathbf{E}_{1\text{D}}(s_t) \in \mathbb{R}^{d_s}. \tag{7}$$

我们定义 $m_t \triangleq \mathbb{I}[t \in \mathcal{A}]$。如果 $m_t = 1$，令 $i = \phi^{-1}(t)$，并计算

$$\mathbf{e}_t^{(3\text{D})} = \frac{1}{K+1} \sum_{j=0}^{K} \mathbf{E}_{3\text{D}}(d_{i,j}) \in \mathbb{R}^{d_s}. \tag{8}$$

如果 $m_t = 0$，我们设$\mathbf{e}_t^{(3\text{D})} = \mathbf{0} \in \mathbb{R}^{d_s}$。我们通过以下方式融合 1D 和 3D 嵌入：

$$\mathbf{s}_t = \frac{\mathbf{e}_t^{(1\text{D})} + m_t\, \mathbf{e}_t^{(3\text{D})}}{1 + m_t} \in \mathbb{R}^{d_s}. \tag{9}$$

结构序列为 $\mathbf{S}_m = [\mathbf{s}1, \ldots, \mathbf{s}_L]$，它作为多模态融合投影器的输入。

3.2 分子深度编码器

分子图像包含决定化学含义的高频线索，包括原子字形、键级、环闭合和立体化学楔形键。我们使用由 SAM-Base 局部视觉Transformer [28], 、卷积标记压缩器和 CLIP-Large 全局视觉Transformer [29]组成的分层编码器对 $\mathbf{I}_m \in \mathbb{R}^{1024 \times 1024 \times 3}$ 进行编码。

3.2.1 局部窗口视觉Transformer

我们将 $\mathbf{I}_m$ 划分为大小为 $p = 16$的非重叠图像块，得到长度为 $N = (1024/p)^2 = 4096$的标记序列。一个 12 层的 SAM-Base 视觉变换器使用大小为 $w = 14$的窗口注意力对这些标记进行编码。输出为

$$\mathbf{H}_{\text{local}} \in \mathbb{R}^{N \times d}, \quad N = 4096, \quad d = 1024. \tag{10}$$

3.2.2 卷积标记压缩器

我们将$\mathbf{H}_{\text{local}}$ 重塑为网格 $\tilde{\mathbf{H}}_{\text{local}} \in \mathbb{R}^{64 \times 64 \times d}$。我们使用两个 3×3 步幅为 2、填充为 1 的卷积，将网格下采样至 16×16，同时保持通道维度为 d。压缩序列为

$$\mathbf{H}_{\text{cmp}} \triangleq \text{Compress}(\mathbf{H}_{\text{local}}) \in \mathbb{R}^{M \times d},\ M = 256. \tag{11}$$

3.2.3 全局注意力变换器

一个 24 层的 CLIP-Large 视觉变换器对$\mathbf{H}^{\text{cmp}}$ 执行密集的全局自注意力，并输出

$$\mathbf{H}_{\text{global}} \in \mathbb{R}^{M \times d}, \quad M = 256, \quad d = 1024. \tag{12}$$

在此阶段，长 距离依赖关系以及环和官能团等结构基元均可被识别。

3.2.4 局部与全局特征融合

我们沿通道维度将全局词元与局部词元拼接起来

$$\mathbf{H}_{\text{vis}} = \text{Concat}(\mathbf{H}_{\text{global}}, \mathbf{H}_{\text{local}}) \in \mathbb{R}^{M \times 2d} = \mathbb{R}^{256 \times 2048}. \tag{13}$$

融合表示$\mathbf{H}_{\text{vis}}$ 由多模态融合投影器处理。

3.3 多模态融合投影器

在输入视觉语言解码器之前，多模态融合投影器会将三维结构信息添加到二维视觉令牌流中。它是一个带交叉注意力的后归一化Transformer块，以视觉令牌为查询，以结构词元为键和值。DeepEncoder输出$\mathbf{H}_{\text{vis}} \in \mathbb{R}^{N_v \times d_v}$，其中包含 $N_v = 256$ 和$d_v = 2048$。第3.1节中的分子结构序列记为$\mathbf{S}_m \in \mathbb{R}^{L \times d_s}$，其中 L 是SELFIES长度， d_s 是结构嵌入维度。为支持批处理，我们将结构序列填充至长度$L_{\max}$，并记为$\mathbf{S}_m^{\text{pad}} \in \mathbb{R}^{L_{\max} \times d_s}$。

融合过程并非自回归过程，因此不应用因果掩码。我们将两种模态投影到共享的隐藏维度 d_h，使其与解码器隐藏层大小相匹配。投影后的特征为

$$\mathbf{H}_V = \mathbf{H}_{\text{vis}}\mathbf{W}_V \in \mathbb{R}^{N_v \times d_h}, \quad \mathbf{H}_S = \mathbf{S}_m^{\text{pad}}\mathbf{W}_S \in \mathbb{R}^{L_{\max} \times d_h}, \tag{14}$$

其参数为可学习的$\mathbf{W}_V \in \mathbb{R}^{d_v \times d_h}$ 和$\mathbf{W}_S \in \mathbb{R}^{d_s \times d_h}$。

按照算法1中的步骤，构造加性掩码$\mathbf{M}$，以区分有效数据和填充。具体而言，对于实际结构位置，我们设置$\mathbf{M}[:\, t] = 0$；对于填充（$t > L$），设置$\mathbf{M}[:\, t] = -\infty$，确保softmax运算忽略填充元素。

令 N_h 表示注意力头的数量， $d_k = d_h/N_h$ 表示每头维度。对于每个注意力头 $h \in \{1, \ldots, N_h\}$，我们计算

$$\mathbf{Q}^{(h)} = \mathbf{H}_V \mathbf{W}_Q^{(h)} \in \mathbb{R}^{N_v \times d_k}, \tag{15}$$

$$\mathbf{K}^{(h)} = \mathbf{H}_S \mathbf{W}_K^{(h)} \in \mathbb{R}^{L_{\max} \times d_k}, \tag{16}$$

$$\mathbf{V}^{(h)} = \mathbf{H}_S \mathbf{W}_V^{(h)} \in \mathbb{R}^{L_{\max} \times d_k}. \tag{17}$$

其中$\mathbf{W}_Q^{(h)}$、$\mathbf{W}_K^{(h)}$、$\mathbf{W}_V^{(h)} \in \mathbb{R}^{d_h \times d_k}$ 是可学习参数。注意力权重和注意力头输出为

$$\mathbf{A}^{(h)} = \text{Softmax}_{\text{last}}\left(\frac{\mathbf{Q}^{(h)}\mathbf{K}^{(h)\top}}{\sqrt{d_k}} + \mathbf{M}\right) \in \mathbb{R}^{N_v \times L_{\max}}, \tag{18}$$

$$\mathbf{O}^{(h)} = \mathbf{A}^{(h)}\mathbf{V}^{(h)} \in \mathbb{R}^{N_v \times d_k}. \tag{19}$$

多头交叉注意力的输出为

$$\text{MHA}_{\text{cross}}(\mathbf{H}_V, \mathbf{H}_S) = \text{Concat}\left(\mathbf{O}^{(1)}, \ldots, \mathbf{O}^{(N_h)}\right)\mathbf{W}_O, \tag{20}$$

其中$\mathbf{W}_O \in \mathbb{R}^{d_h \times d_h}$。

我们采用残差连接和后归一化。交叉注意力更新为

$$\mathbf{H}_{\text{cross}} = \text{LayerNorm}(\mathbf{H}_V + \text{MHA}_{\text{cross}}(\mathbf{H}_V, \mathbf{H}_S)). \tag{21}$$

随后，逐位置前馈网络会进一步优化该表示

$$\text{FFN}(\mathbf{H}_{\text{cross}}) = \text{GELU}(\mathbf{H}_{\text{cross}}\mathbf{W}_1 + \mathbf{b}_1)\,\mathbf{W}_2 + \mathbf{b}_2, \tag{22}$$

其中$\mathbf{W}_1 \in \mathbb{R}^{d_h \times d_{\text{ff}}}$ 和$\mathbf{W}_2 \in \mathbb{R}^{d_{\text{ff}} \times d_h}$。融合输出为

$$\mathbf{H}_{\text{fused}} = \text{LayerNorm}(\mathbf{H}_{\text{cross}} + \text{FFN}(\mathbf{H}_{\text{cross}})) \in \mathbb{R}^{N_v \times d_h}. \tag{23}$$

对于70亿参数解码器， $d_h = 4096$ 和 $N_v = 256$。输出保留了视觉 token 预算，同时通过交叉注意力，让每个视觉 token 都以源自分子构象的结构描述符为条件。

3.4 视觉语言模型（VLM）解码器

我们使用 Qwen2-VL 作为视觉语言解码器。解码器接收融合视觉令牌 $\mathbf{H}_{\text{fused}} \in \mathbb{R}^{N_v \times d_h}$以及 $N_v = 256$ 和提示令牌序列 $\mathbf{X}_{\text{prompt}} = \{x_1, \ldots, x_P\}$。令 $\mathbf{E}_{\text{txt}}(\cdot)$ 表示 Qwen2-VL 的令牌嵌入查找，在 $\mathbb{R}^{d_h}$ 中生成嵌入。嵌入提示为 $\mathbf{H}_{\text{txt}} = [\mathbf{E}_{\text{txt}}(x_1) \ldots, \mathbf{E}_{\text{txt}}(x_P)] \in \mathbb{R}^{P \times d_h}$。拼接上下文为

$$\mathbf{H}_{\text{in}} = [\mathbf{H}_{\text{fused}}; \mathbf{H}_{\text{txt}}] \in \mathbb{R}^{(N_v+P) \times d_h}. \tag{24}$$

算法1 三维词元化与序列填充

输入： 分子 m（原子 $\{a_i\}_{i=1}^{N_{\text{ha}}}$），SELFIES 长度 L
超参数： E3FP 迭代次数 K，半径 r，哈希空间 $|F|$，最大长度 $L_{\max}$
输出： 结构词元$\mathbf{S}_m$，填充掩码 $\mathbf{M}$

对于 $i = 1$**到** N_{ha}，**执行**
  $A_i \leftarrow$原子不变量(a_i)
  $\hat{d}_{i,\,0} \leftarrow$MurmurHash3$(A_i)$
结束循环
对于 $j = 1$**到** K，**执行**
  $R_j \leftarrow r \cdot j$
  对于 $i = 1$**到** N_{ha}，**执行**
    $\mathcal{N}_{i,\,j} \leftarrow \emptyset$
    对于 $k = 1$**到** N_{ha} **执行**
      如果 $\|\mathbf{x}_k - \mathbf{x}_i\|_2 \le R_j$ **则**
        $c_k^i \leftarrow$ 连接关系$(a_k,\ a_i)$
        $\sigma_k^i \leftarrow$ 立体化学$(a_k,\ a_i)$
        $\mathcal{N}_{i,j} \leftarrow \mathcal{N}_{i,j} \cup \{(c_k^i,\ \hat{d}_{k,j-1}, \sigma_k^i)\}$
      结束 if
    结束 for
    $L_{i,\,j} \leftarrow [\,j, \hat{d}_{i,\,j-1}\,] \oplus$ 排序$(\mathcal{N}_{i,\,j})$
    $\hat{d}_{i,\,j} \leftarrow$MurmurHash3$(L_{i,\,j})$
  结束 for
结束循环
对于 $i = 1$ **到** N_{ha}，**执行**
  $\mathbf{d}_i \leftarrow [\hat{d}_{i,0},\ \hat{d}_{i,\,1},\ \ldots,\ \hat{d}_{i,\,K}]$ 模块 $|F|$
结束循环
$\mathbf{S}_m \leftarrow \mathbf{0} \in \mathbb{R}^{L_{\max} \times d_s}$
对于 $t = 1$**到** $L_{\max}$ **do**
  根据公式（9）生成$\mathbf{s}_t$，并设置$\mathbf{S}_m[t] \leftarrow \mathbf{s}_t$
  $\mathbf{M}[$: 如果 $t \le L$，否则 $-\infty$
结束循环

解码器自回归地生成输出词元序列 $\mathbf{y} = [y_1, \ldots, y_T]$。对于步骤 t，令 $\mathbf{y}_{<t}$ 为前缀。模型计算

$$\mathbf{h}_t = f_{\text{vlm}}(\mathbf{H}_{\text{in}}, \mathbf{y}_{<t}) \in \mathbb{R}^{d_h}, \tag{25}$$

$$\boldsymbol{\pi}_t = \text{Softmax}(f_{\text{vocab}}(\mathbf{h}_t)) \in \mathbb{R}^{|\mathcal{V}|}, \tag{26}$$

其中 f_{vocab}：$\mathbb{R}^{d_h} \to \mathbb{R}^{|\mathcal{V}|}$ 是词汇投影。条件概率为 $p(y_t = w \mid \mathbf{H}_{\text{in}},\ \mathbf{y}_{<t}) = \boldsymbol{\pi}_t[w]$，适用于 $w \in \mathcal{V}$。在评估时，我们通过解码

$$\tilde{y}_t = \arg\max_{w \in \mathcal{V}} p(y_t = w \mid \mathbf{H}_{\text{in}}, \mathbf{y}_{<t}), \tag{27}$$

或从 $\boldsymbol{\pi}_t$ 中采样。

3.5 训练策略

我们采用两阶段训练流程。第一阶段使用冻结的解码器对齐多模态融合投影器，使融合表示$\mathbf{H}_{\text{fused}}$ 成为兼容的前缀。第二阶段通过多模态指令微调，使解码器适应以$\mathbf{H}_{\text{fused}}$为条件的分子提示。两个阶段均优化公式 1 中的同一自回归负对数似然目标。

3.5.1 第一阶段预训练

视觉-语言对齐 我们冻结 DeepEncoder 和所有解码器参数，仅在分子图像–文本对上训练多模态融合投影器。

3.5.2 第二阶段微调

视觉指令微调 从第一阶段检查点开始，我们保持 DeepEncoder 冻结，并在分子描述生成和性质预测等指令风格数据上联合优化多模态融合投影器和解码器，使模型能够在 $\mathbf{H}_{\text{fused}}$ 的条件下生成任务特定文本输出。

4 实验

4.1 实现细节

我们采用了训练策略中描述的两阶段训练流程，并报告实现细节和超参数，以确保结果可复现。所有模型均使用 PyTorch 实现，并在配备 8 个 NVIDIA H800 GPU（80GB PCIe）的单节点上训练。除非另有说明，解码器主干网络为 *Qwen2-VL-7B-Instruct*，分子图像编码器为经过预训练的 *SAM-CLIP*分子深度编码器。

第一阶段预训练 我们训练了 5 个 epoch，全局批次大小为 128（每个 GPU 为 16），使用 AdamW [30]，学习率为 10^{-4}，采用余弦衰减，预热比例为 0.03。训练使用 BF16，最大上下文长度为 4096。

第二阶段微调 除非另有说明，第二阶段微调均从第一阶段检查点初始化，并更新解码器主干网络和多模态融合投影器，同时冻结分子图像编码器。我们使用 DeepSpeed ZeRO-3 [31] 在八张 H800 GPU 上进行全参数微调，训练了10 个周期，最大学习率为 5×10^{-5}，全局批次大小为 64。我们报告了两个变体。通用模型采用统一多任务目标，在所有下游数据集的混合数据上训练；专用模型则使用任务特定监督，针对每个下游任务分别进行微调，预热比例为 0.01。

4.2 分子描述生成

我们在分子描述生成任务上评估了 DeepMoLM，以衡量其对三维分子结构的理解能力。我们使用了 PubChem 数据集 [32], ，其中包含约 15,000 对三维分子结构和 PubChem 文本描述。每条描述都包含分子名称和三维构象相关陈述。此设置综合评估了遵循 Favre 和 Powell 方法的名称预测，以及遵循 Ed wards 方法的描述预测 [10],，比单独评估性质描述更具挑战性。
我们将模型与两组基线模型进行了比较。专用模型针对分子任务进行了优化，包括 MolT5-Large [10], MoMu-Large [33], UniMoT [34], 和 3D-MoLM [4], ；此外，我们还报告了一个未经过 GPT-3.5 增强预训练的变体。通用模型则面向范围更广的多模态生成任务，或依赖大语言模型，包括 2D-MoLM [4], Llama2-7B [35], 和 Qwen2-VL-7B [26]。我们报告了 BLEU-2、BLEU-4、ROUGE-1、ROUGE-2、ROUGE-L 和 METEOR 指标。

表 1 显示，DeepMoLM 与专用基线和通用基线相比均具有竞争力。在专用模型中，它的表现与 UniMoT 相当出色；UniMoT 使用了预训练的分子编码器。尽管 DeepMoLM 仅使用分子图像，但它捕捉到了细粒度的结构线索，其 ROUGE-1 和 METEOR 得分均高于专用基线。在通用模型中，DeepMoLM 明显优于标准大语言模型和视觉语言模型，其 METEOR 得分比排名第二的基线模型相对高出 10% 以上。总的来说，DeepMoLM 将视觉分子表征与自然语言对齐，无需显式提供原子坐标即可生成信息丰富的描述文本。

4.3 分子性质预测

我们使用 PubChem 评估了 DeepMoLM 在分子性质预测方面的表现，涵盖分子量、脂水分配系数、拓扑极性表面积和复杂度。我们将其与 Uni-Mol、2D-MoLM、3D-MoLM 以及包括 Llama2 [35]和 Qwen2-VL [26]在内的通用模型进行了比较。DeepMoLM 对所有查询均返回了有效输出，有效率达到 100%，避免了通用模型中频繁出现的格式错误。

DeepMoLM 仅使用二维分子图像，表现便与领先的三维方法相当或更优。在专科模型设置下，它在分子量和复杂度上的 MAE 最低，优于包括 3D-MoLM 在内的几何感知基线模型，在 LogP 和 TPSA 上也保持了竞争力。这表明，视觉编码器无需显式的三维坐标，也能捕捉关键的结构线索。在通才设置下，DeepMoLM 大幅超越了视觉语言模型和纯文本模型。

表 1：分子描述生成效果。† 指一种 3D-MoLM 变体，该变体最初使用未经 GPT-3.5 增强的原始 PubChem 文本进行预训练。最佳结果以**粗体**显示，次佳结果带下划线。

TYPE	模型	BLEU-2	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L	METEOR
专科模型	MolT5-Large	25.87	17.28	34.07	16.42	23.41	28.04
	MoMu-Large	26.34	18.01	34.75	16.86	24.76	28.73
	3D-MoLM†	29.82	22.39	37.23	22.49	31.07	32.69
	3D-MoLM	30.32	22.52	36.84	22.32	31.23	33.06
	UniMoT	31.30	23.80	37.50	**23.70**	**33.60**	34.80
	DeepMoLM	**33.78**	**25.53**	**44.49**	22.57	29.04	**35.87**
通才	2D-MoLM	27.15	21.19	36.02	20.76	29.12	32.28
	3D-MoLM†	29.25	22.07	36.48	21.80	**30.95**	33.12
	3D-MoLM	28.95	21.63	36.51	21.26	30.02	33.55
	Llama2-7B	27.01	20.94	35.76	20.68	28.88	32.11
	Qwen2-VL-7B	18.91	8.33	36.38	10.90	19.55	23.44
	DeepMoLM	**35.87**	**25.53**	**47.07**	**23.07**	30.76	**37.66**

表 2：计算属性预测的性能。报告平均绝对误差 (MAE)，括号内为有效答案率。† 表示一种在原始 PubChem 文本上进行预训练、未使用 GPT-3.5 增强的 3D-MoLM 变体。* 表示模型未经微调进行评估。最佳结果以**粗体**显示，次佳结果带下划线。

TYPE	模型	分子量（G/MOL）↓	脂水分配系数 ↓	拓扑极性表面积（Å²）↓	复杂度 ↓
非语言模型	Uni-Mol	20.35	**0.59**	13.48	57.24
专科模型	2D-MoLM	21.48（94%）	0.88（96%）	13.52（92%）	55.74（94%）
	3D-MoLM†	16.18（96%）	0.95（96%）	10.26（94%）	49.15（95%）
	3D-MoLM	14.79（95%）	**0.66**（97%）	**9.71**（93%）	44.85 (94%)
	Llama2-7B	22.10 (96%)	1.45 (95%)	15.87 (92%)	69.74 (93%)
	DeepMoLM	**13.64** (100%)	0.73 (100%)	10.04 (100%)	**37.89** (100%)
通才	2D-MoLM	20.80 (92%)	1.36 (94%)	12.47 (89%)	52.70 (91%)
	3D-MoLM†	19.54 (93%)	0.92 (92%)	11.14 (92%)	54.68 (90%)
	3D-MoLM	16.58 (92%)	**0.78** (95%)	**10.90** (90%)	45.49 (89%)
	Llama2-7B*	42.18 (82%)	2.10 (85%)	27.11 (84%)	121.87 (76%)
	Llama2-7B	27.42 (92%)	1.78 (93%)	17.07 (90%)	78.16 (92%)
	Qwen2-VL-7B	42.13 (86%)	3.75 (99%)	50.89(32%)	103.07 (100%)
	DeepMoLM	**14.63** (100%)	1.36 (100%)	11.14 (100%)	**41.73** (100%)

4.4 分子描述

遵循 OCSU [36], ，我们将 CheBI-20 [10] 转换为指令微调格式。为防止数据泄漏，我们移除了 CheBI-20 测试分子中也出现在用于预训练的 PubChem 三维分子-文本对里的分子。此外，我们从目标文本中移除了分子名称，以避免模型学习到从名称到序列的简单映射。基线模型取自 MolT5 [10], 、MolReGPT [37],、MolFM [38], 、GIT-Mol [39], 、MolXPT [40], 和 BioT5 [41]。

表 3 报告了分子描述生成的结果。我们将 DeepMoLM 与以 SMILES 为输入的专用模型和通用模型，以及以图像为输入的视觉语言模型进行了比较。DeepMoLM 在大多数指标上取得了最佳结果，并始终优于所有通用基线。BioT5+ 等专用模型得分更高，这是因为它们基于大规模一维化学序列进行了预训练；然而，DeepMoLM 缩小了这一差距，同时在推理时无需使用 SMILES。这表明 DeepMoLM 能够直接从二维图像中学习有用的分子表征，这在缺少文本表征或文本表征不可靠时尤为有用。

4.5 消融研究

为验证预训练、三维分子编码器和多模态融合的效果，我们在 PubChem [32] 和 ChEBI-20 [10] 上进行了消融实验。我们将 DeepMoLM 与三个未经过预训练的变体进行了比较。其中一个从头开始训练。另一个将融合投影器替换为特征拼接。第三个变体移除了

表 3：分子描述任务的表现。最佳结果以**粗体**显示，次佳结果带下划线 _____.

方法	输入	BLEU-2	BLEU-4	ROUGE-2	ROUGE-L	METEOR
分子输入（专科模型）						
MoMu	SMILES	54.9	46.2	47.9	57.5	57.6
MolFM	SMILES	58.5	49.8	50.8	59.4	60.7
BioT5	SMILES	63.5	55.6	55.9	63.3	65.6
MolCA	SMILES	62.0	53.1	53.7	61.8	65.1
BioT5+	SMILES	**66.6**	**59.1**	**58.4**	**65.0**	**68.1**
分子输入（通用型）						
3D-MoLM	SMILES	6.7	3.0	4.2	8.6	18.3
Mol-Instructions	SMILES	24.9	17.1	20.3	28.9	27.1
BioMedGPT	SMILES	30.6	19.8	25.7	38.3	35.0
GIT-Mol	SMILES	**35.2**	**26.3**	**48.5**	**56.0**	**43.0**
视觉输入						
MolScribe 与 BioT5+	图像	53.43	44.64	46.48	55.30	57.09
DoubleCheck 与 BioT5+	图像	54.40	45.56	47.21	55.94	58.07
Mol-VL-2B	图像	50.57	40.26	41.99	52.25	53.34
Mol-VL-7B	图像	55.73	46.14	47.26	**56.61**	58.14
Qwen2-VL-7B	图像	6.34	1.13	4.53	14.15	19.95
DeepMoLM	图像	**55.76**	**46.29**	**48.07**	55.96	**58.23**

3D-E3FP分支，并采用线性投影。表5显示，移除预训练会导致所有指标明显下降，这表明预训练对于使分子表征与语言语义对齐至关重要。架构选择也会影响生成质量。将融合投影器替换为拼接后，表现甚至不如未经过预训练的普通模型，其中 BLEU-2 和 METEOR 的降幅最大，这表明需要专门的融合模块。移除 3D 指纹后，ROUGE-L 得分最低，说明语义保真度有所下降。这些结果表明，3D立体化学线索和有效融合对于生成准确且化学上合理的描述都不可或缺。

5 结论

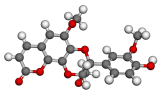
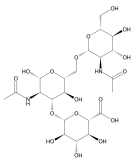
我们研究了双视图分子理解，使语言与分子图像和3D结构相对齐。DeepMoLM 增强了跨模态交互，编码了立体化学不变量，并通过双路径编码器提升了感知能力。在分子描述生成、描述和性质预测任务中，DeepMoLM取得了具有竞争力的结果，以有效输出超越了通用基线。消融实验验证了预训练、融合和3D指纹的必要性。因此，DeepMoLM 为分子理解提供了一个稳健的框架。

附录

可复现性声明

我们已尽一切努力确保本文所呈现的结果可复现。所有代码和数据集均已在匿名仓库中公开，以便他人复现和验证。本文详细介绍了实验设置，包括训练步骤、模型配置和硬件详情。我们还完整介绍了DeepMoLM，以帮助他人复现我们的实验。此外，补充数据（如二维分子图像和 E3FP分子指纹）也已公开，确保评估结果一致且可复现。我们相信，这些措施将使其他研究人员能够复现我们的工作，并进一步推动该领域的发展。

表 4：分子描述生成与描述的案例研究。

分子	指令	DEEPMOLM	真实标注
	描述输入分子。	具有以下 SMILES 编码的分子 COC1=C(C=CC(=C1)[C@@H]2C@@HCO)O 是一种复杂的天然产物-，名为8'-epi-cleomiscosinA，它分离自地上部分的*Rhododendroncollettianum*。它具有一种高度复杂的结构，由多种环构成，包含多个环和官能团。该分子含有若干个羟基和甲氧基，这使其具有相对较高的极性以及形成氢键的能力。这一性质使其能够通-过分子间作用力与其他分子相互作用，从而可能增-影响其溶解度和反应性。此外，含氧杂环和内酯基团的存在表明，该分子可能具有生物活性，因为这类结构特征很常见在生物活性天然产物中。 The 分子的立体化学由.....决定“@”符号的存在，这些符号出现在 SMILES 编码中，对应于明确的三维排列原子。立体化学..... 学可以影响其生物活性以及与酶的相互作用或再-受体。值得注意的是，8'-epi-cleomiscosin A 据报道显示出抑-制酪氨酸酶的活性，表明其在以下领域具有潜在应用药物化学和化妆品科学—— ence.	8'-epi-cleomiscosinA是一种复杂的有机分子，具有一个 SMILES 代码为 COC1=C(C=CC(=C1)[C@@H]2C@@HCO)O。它属于酚类，并归类为一种芳香醚和一种有机杂三环化合物。在结构上，8'-epi-cleomiscosin A 是一种取代的 2,3-二氢-9H-[1,4]二氧杂环[2,3-h]色烯-9-酮。 它含有一个 4-羟基-3-甲氧基苯基在3位有一个羟甲基在2位有一个甲氧基在5位。 此外，它是 2S,3R 立体异构体。该分子是从.....的地上部分分离得到的 *Rhododendroncollettianum*。它具有表现出 抑制 活性针对 the 酶 酪氨酸酶，使其成为一种 EC 1.14.18.1 抑制剂。酪氨酸酶参与黑色素合成，因此抑制活性关于8'-epi-cleomiscosin A表明其在皮肤美白或抗色素沉着应用中的潜力。 In 总之， 8'-epi-cleomiscosin A 是一种独特的有机化合物，具有有趣的化学和 物理性质。它的结构以及其对酪氨酸酶的抑制活性，使其成为一种有价值的分子，应用于领域的药物化学以及化妆品科学。
	请描述这一内容药物。	该分子是一种支链氨基四糖，包含N-乙酰-β-D-氨基葡萄糖 位于还原端端 带有 一个 β-D-半乳糖基残基连接于 6位以及一个连接于（1→3）的 β-连接于该处的 D-半乳糖基部分 3位。	该分子是一种支链氨基三糖，由以下部分组成：N-乙酰基-β-D-氨基葡萄糖，其具有 β-D-葡萄糖醛酸基和 N-乙酰基-β-D-氨基葡萄糖基残基分别连接在 3 位和 6 位。 It 是一种氨基三糖和一种葡萄糖胺寡糖。

B 大语言模型的使用

大语言模型（LLM）仅用于句子改写、语法检查和提升文本流畅度。它们未参与构思、方法论或实验；所有科学内容均由作者完成。作者对稿件承担全部责任，所有经大语言模型辅助生成的文本均符合伦理准则，且未导致抄袭或不端行为。

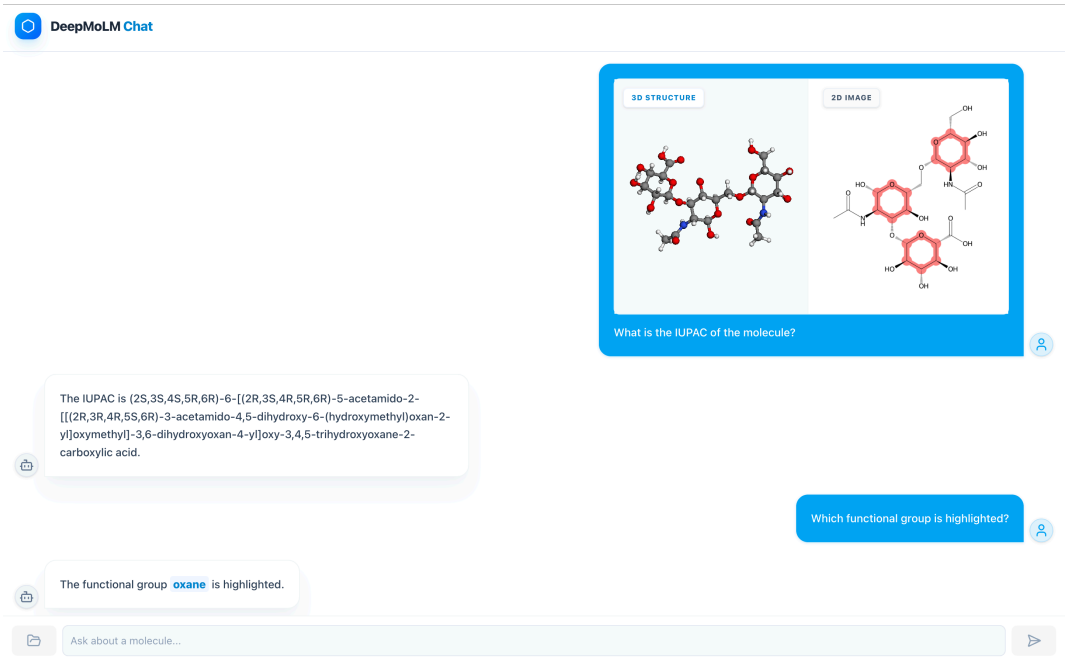


图3：开放文本分子问答的视觉语言模型接口

表5：消融研究。移除预训练、3D-E3FP分支或融合投影器后，在分子描述生成（PubChem）和分子描述（ChEBI-20）任务上的结果。**粗体** 表示最佳分数。↓表示各项指标中的最大降幅。

模型	BLEU-2	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L	METEOR
分子描述生成						
无预训练、3D-E3FP 和融合（线性）	18.91	8.33	36.38 ↓	10.90 ↓	19.55 ↓	23.44
DeepMoLM	**33.78**	**25.53**	**44.49**	**22.57**	**29.04**	**35.87**
分子描述						
不使用预训练、3D-E3FP 和融合（线性）	24.56	15.00	40.03 ↓	20.22 ↓	33.25 ↓	29.58
不使用预训练和融合（拼接）	19.94 ↓	12.96 ↓	42.33	22.18	36.67	27.29 ↓
无预训练	26.28	15.99	43.29	24.09	38.19	31.60
DeepMoLM	**55.76**	**46.29**	**61.64**	**48.07**	**55.96**	**58.23**

伦理声明

本研究遵循 IJCAI-ECAI 伦理准则。研究未涉及人体或动物实验。所有数据集，包括 PubChem 和 CCheBI-20，均在遵守相关指南且未侵犯隐私的前提下使用。研究未包含任何个人身份信息，所采用的程序也未造成隐私或安全风险。本研究以透明、诚信的方式开展。

致谢

本研究获得香港理工大学内部拨款（项目编号：P0051278，Jung Sun Yoo）以及中国香港特别行政区研究资助局优配研究金（项目编号：PolyU 15101422，Jung SunYoo）的资助。

参考文献

[1] G Zhou、Z Gao、Q Ding、H Zheng、H Xu、Z Wei、L Zhang 和 G Ke。《通用三维分子表示学习框架》。发表于 ICLR 2023 大会，2023年。[2] Kohulan Rajan、Henning Otto Brinkhaus、M I sabel Agea、Achim Zielesny 和 Christoph Steinbeck。Decimer.ai：一个用于科学出版物中化学结构自动光学识别、分割和辨认的开放平台。《自然通讯》，14(1):5045，2023年。[3] Zhiyuan Liu、Sihang Li、Yanchen Luo、Hao Fei、Yixin Cao、Kenji Kawaguchi、Xiang Wang 和 Tat-Seng Chua。《Molca：结合跨模态投影器与单模态适配器的分子图-语言建模》。arXiv 预印本 arXiv:2310.12798，2023年。[4] Sihang Li、Zhiyuan Liu、Yanchen Luo、Xiang Wang、Xiangnan He、Kenji Kawaguchi、Tat-Seng Chua 和 Qi Tian。《迈向语言模型中的三维分子-文本解释》。arXiv 预印本 arXiv:2401.13923，2024年。[5] Fabian Fuchs、Daniel Worrall、Volker Fischer 和 Max Welling。《Se (3)-变换器：三维旋转平移等变注意力网络》。神经信息处理系统进展，33:1970–1981，2020年。[6] Kumail Alhamoud、Yasir Ghunaim、Abdulelah S Alshehri、Guohao Li、Bernard Ghanem 和 Fengqi You。《利用二维分子图预训练，通过图神经网络改进三维构象生成》。计算机与化学工程，183:108622，2024年。[7] Alexey Dosovitskiy。《一幅图像胜过 16×16 个词：用于大规模图像识别的变换器》。arXiv 预印本 arXiv:2010.11929，2020年。[8] Haoran Wei、Yaofeng Sun 和 Yukun Li。《Deepseek-ocr：上下文光学压缩》。arXiv 预印本 arXiv:2510.18234，2025年。[9] Seth D Axen、Xi-Ping Huang、Elena L Cáceres、Leo Gendelev、Bryan L Roth 和 Michael J Keiser。《三维分子结构的一种简明表示》。药物化学杂志，60(17):7393–7409，2017年。[10] Carl Edwards、Tuan Lai、Kevin Ros、Garrett Honke、Kyunghyun Cho 和 Heng Ji。《分子与自然语言之间的转换》。arXiv 预印本 arXiv:2204.11817，2022年。[11] Qizhi Pei、Rui Yan、Kaiyuan Gao、Jinhua Zhu 和 Lijun Wu。《3D-MolT5：利用离散结构信息进行分子-文本建模》。arXiv 预印本 arXiv:2406.05797，2024年。[12] Di Zhang、Wei Liu、Qian Tan、Jingdan Chen、Hang Yan、Yuliang Yan、Jiatong Li、Weiran Huang、Xiangyu Yue、Wanli Ouyang 等。《ChemLLM：化学大语言模型》。arXiv 预印本 arXiv:2402.06852，2024年。[13] Justin Gilmer、Samuel S Schoenholz、Patrick F Riley、Oriol Vinyals 和 George E Dahl。《用于量子化学的神经消息传递》。发表于国际机器学习大会，第1263–1272页。PMLR，2017年。[14] Yu Rong、Yatao Bian、Tingyang Xu、Weiyang Xie、Ying Wei、Wenbing Huang 和 Junzhou Huang。《在大规模分子数据上进行自监督图变换器预训练》。神经信息处理系统进展，33:12559–12571，2020年。[15] Shengchao Liu、Hanchen Wang、Weiyang Liu、Joan Lasenby、Hongyu Guo 和 Jian Tang。《利用三维几何预训练分子图表示》。arXiv 预印本 arXiv:2110.07728，2021年。[16] Hongxin Xiang、Shuting Jin、Jun Xia、Man Zhou、Jianmin Wang、Li Zeng 和 Xiangxiang Zeng。《一种融合图像增强的分子图表示学习框架》。发表于第三十三届国际人工智能联合会议论文集，第6107–6115页，2024年。[17] Xiangxiang Zeng、Hongxin Xiang、Linhui Yu、Jianmin Wang、Kenli Li、Ruth Nussinov 和 Feixiong Cheng。《利用自监督图像表示学习框架准确预测分子性质和药物靶点》。自然·机器智能，4(11):1004–1016，2022年。[18] Aleksei Krasnov、Shadrack J Barnabas、Timo Boehme、Stephen K Boyer 和 Lutz Weber。《比较光学化学结构识别软件工具》。数字发现，3(4):681–693，2024年。[19] Yujie Qian、Jiang Guo、Zhengkai Tu、Zhening Li、Connor W Coley 和 Regina Barzilay。《MolScribe：利用图像到图生成实现稳健的分子结构识别》。化学信息与建模杂志，63(7):1925–1934，2023年。[20] Zhanpeng Xu、Jianhua Li、Zhaopeng Yang、Shiliang Li 和 Honglin Li。《SwinOCSR：利用Swin变换器进行端到端光学化学结构识别》。化学信息学杂志，14(1):41，2022年。[21] Junxian Li、Di Zhang、Xunzhi Wang、Zeying Hao、Jingdi Lei、Qian Tan、Cai Zhou、Wei Liu、Yaotian Yang、Xinrui Xiong 等。《ChemVLM：探索多模态大语言模型在化学领域的能力》。发表于 AAAI 人工智能会议论文集，第39卷，第415–423页，2025年。

[22] Kristof Schütt、Pieter-Jan Kindermans、Huziel Enoc Sauceda Felix、Stefan Chmiela、Alexandre Tkatchenko 和 Klaus-Robert Müller。Schnet：一种用于建模量子相互作用的连续滤波卷积神经网络。神经信息处理系统进展，30，2017年。[23] Johannes Gasteiger、Janek Groß 和 Stephan Günnemann。用于分子图的定向消息传递。*arXiv* 预印本 *arXiv:2003.03123*，2020年。[24] Johannes Gasteiger、Florian Becker 和 Stephan Günnemann。Gemnet：适用于分子的通用定向图神经网络。神经信息处理系统进展，34：6790–6802，2021年。[25] Vıctor Garcia Satorras、Emiel Hoogeboom 和 Max Welling。E (n) 等变图神经网络。载于国际机器学习大会，第9323–9332页。PMLR，2021年。[26] Peng Wang、Shuai Bai、Sinan Tan、Shijie Wang、Zhihao Fan、Jinze Bai、Keqin Chen、Xuejing Liu、Jialin Wang、Wenbin Ge 等。Qwen2-vl：提升视觉语言模型在任意分辨率下对世界的感知能力。*arXiv* 预印本 *arXiv:2409.12191*，2024年。[27] Mario Krenn、Florian Häse、AkshatKumar Nigam、Pascal Friederich 和 Alan Aspuru-Guzik。自引用嵌入字符串（SELFIES）：一种100%鲁棒的分子字符串表示。机器学习：科学与技术，1(4)：045024，2020年。[28] Alexander Kirillov、Eric Mintun、Nikhila Ravi、Hanzi Mao、Chloe Rolland、Laura Gustafson、Tete Xiao、Spencer Whitehead、Alexander C Berg、Wan-Yen Lo 等。任意分割。载于*IEEE/CVF*国际计算机视觉大会论文集，第4015–4026页，2023年。[29] 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。载于国际机器学习大会，第8748–8763页。PMLR，2021年。[30] Ilya Loshchilov 和 Frank Hutter。解耦权重衰减正则化。*arXiv* 预印本 *arXiv:1711.05101*，2017年。[31] Samyam Rajbhandari、Jeff Rasley、Olatunji Ruwase 和 Yuxiong He。ZeRO：面向万亿参数模型训练的内存优化。载于*SC20*：高性能计算、网络、存储与分析国际会议，第1–16页。IEEE，2020年。[32] Sunghwan Kim、Jie Chen、Tiejun Cheng、Asta Gindulyte、Jia He、Siqian He、Qingliang Li、Benjamin A Shoemaker、Paul A Thiessen、Bo Yu 等。PubChem 2021年：新增数据内容与改进的网页界面。《核酸研究》，49(D1)：D1388–D1395，2021年。[33] Bing Su、Dazhao Du、Zhao Yang、Yujie Zhou、Jiangmeng Li、Anyi Rao、Hao Sun、Zhiwu Lu 和 Ji-Rong Wen。一种将分子图与自然语言关联起来的分子多模态基础模型。*arXiv* 预印本 *arXiv:2209.05481*，2022年。[34] Juzheng Zhang、Yatao Bian、Yongqiang Chen 和 Quanming Yao。UniMoT：具有离散词元表示的统一分子文本语言模型。*arXiv* 预印本 *arXiv:2408.00863*，2024年。[35] Hugo Touvron、Louis Martin、Kevin Stone、Peter Albert、Amjad Almahairi、Yasmine Babaei、Nikolay Bashlykov、Soumya Batra、Prajjwal Bhargava、Shruti Bhosale 等。Llama 2：开放式基础模型与微调聊天模型。*arXiv* 预印本 *arXiv:2307.09288*，2023年。[36] Siqi Fan、Yuguang Xie、Bowen Cai、Ailin Xie、Gaochao Liu、Mu Qiao、Jie Xing 和 Zaiqing Nie。OCSU：面向以分子为中心的科学发现的光学化学结构理解。*arXiv* 预印本 *arXiv:2501.15415*，2025年。[37] Jiatong Li、Yunqing Liu、Wenqi Fan、Xiao-Yong Wei、Hui Liu、Jiliang Tang 和 Qing Li。借助大语言模型，通过分子描述文本翻译赋能分子发现：ChatGPT视角。《*IEEE*知识与数据工程汇刊》，36(11)：6071–6083，2024年。[38] Yizhen Luo、Kai Yang、Massimo Hong、Xing Yi Liu 和 Zaiqing Nie。MolFM：多模态分子基础模型。*arXiv* 预印本 *arXiv:2307.09484*，2023年。[39] Pengfei Liu、Yiming Ren、Jun Tao 和 Zhixiang Ren。GIT-Mol：一种融合图像、图和文本的分子科学多模态大语言模型。《生物学与医学中的计算机》，171：108073，2024年。[40] Zequn Liu、Wei Zhang、Yingce Xia、Lijun Wu、Shufang Xie、Tao Qin、Ming Zhang 和 Tie-Yan Liu。MolXPT：以文本包装分子以进行生成式预训练。*arXiv* 预印本 *arXiv:2305.10688*，2023年。[41] Qizhi Pei、Wei Zhang、Jinhua Zhu、Kehan Wu、Kaiyuan Gao、Lijun Wu、Yingce Xia 和 Rui Yan。BioT5：借助化学知识和自然语言关联，增强生物学中的跨模态整合。*arXiv* 预印本 *arXiv:2310.07276*，2023年。