MPM-LLM4DSE：通过高层次综合实现帕累托前沿与多模态学习及大语言模型驱动的探索

Lei Xu、Shanshan Wang、Chenglong Xiao*汕头大学计算机科学系，中国
{24lxu,sswang,chlxiao}@stu.edu.cn

arXiv:2601.04801v1 [cs.AR] 8 Jan 2026

摘要——**高层次综合（HLS）设计空间探索（DSE）旨在庞大的编译指示配置空间中寻找帕累托最优设计。为加速HLS设计空间探索，通常采用图神经网络（GNN）作为高层次综合工具的代理模型，以预测结果质量（QoR）指标，同时利用多目标优化算法加速探索。然而，基于图神经网络的预测方法可能无法充分捕捉行为描述中蕴含的丰富语义特征，而传统多目标优化算法通常不会明确考虑有关pragma指令如何影响结果质量的领域知识。为解决这些局限，本文提出MPM-LLM4DSE框架，该框架采用多模态预测模型（MPM），同时融合行为描述以及控制流图和数据流图中的特征。此外，该框架使用大语言模型（LLM）作为优化器，并配套定制的提示工程方法。该方法将pragma影响分析纳入结果质量评估，以引导大语言模型生成高质量配置（LLM4DSE）。实验结果表明，我们的多模态预测模型显著优于当前最先进的方法ProgSG，最高提升10.25×。此外，在DSE任务中，所提出的LLM4DSE较以往方法平均提升39.90%，验证了所用提示方法的有效性。代码和模型见https://github.com/wslcccc/MPM-LLM4DSE。**

索引词—**图神经网络、高层次综合、设计空间探索、多模态特征、语言模型**

I. 引言

高层次综合（HLS）已成为现代超大规模集成电路设计中的一种范式转变方法。该方法使设计人员能够利用高级语言（例如 C/C++）中的算法规格进行硬件生成，同时提供可调的综合参数，以优化实现质量。然而，高层次综合工具生成结果质量（QoR）指标所需的大量时间，仍然是一个关键瓶颈，尤其是在硬件优化过程中，因为设计空间通常具有组合爆炸性。在设计空间探索（DSE）中，大多数算法由于无法解读综合编译指示对最终实现的影响，在这两个方面都面临困难。因此，HLSDSE 的关键目标是实现快速、准确的 QoR预测，并建立智能的设计空间探索范式。

如图 1 所示，HLS设计空间探索流程从添加指令注释的 C/C++行为描述开始，经高层次综合生成 RTL实现并提取相应的 QoR指标。随后，多目标

*通讯作者：Chenglong Xiao（chlxiao@stu.edu.cn）。本工作部分受广东省基础与应用基础研究基金（2022A1515110712、2025A1515010272 和 2023A1515010077）以及中国广东省高等学校科学研究项目（项目编号：2021ZDZX1027）资助。

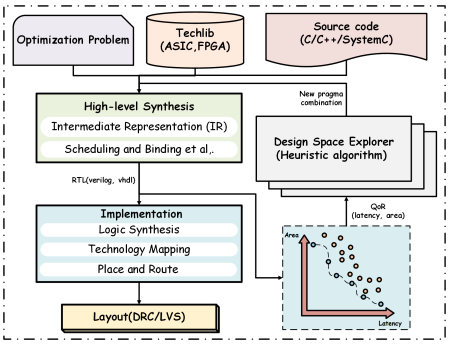

图 1：HLS设计空间探索概览。

优化引擎会探索新的指令组合，逐步细化设计空间，最终收敛到能够平衡 QoR指标的帕累托最优配置。值得注意的是，该流程采用深度学习技术快速预测 QoR指标，从而加快优化过程，减少每次迭代都进行完整综合的需求。

结果质量预测。结果质量预测方法的发展呈现出清晰的演进轨迹：从基于控制数据流图（CDFG）的图神经网络（GNN）特征提取这一基础方法起步，逐渐迈向架构和范式日益复杂的方向。早期研究确立了图神经网络作为高层次综合结果质量估计代理模型的可行性（Wu 等人 [1]）；后续创新则通过以边为中心的聚合（Lin 等人 [2]）和层次化池化（Kuang 等人 [3]）增强了对拓扑结构的感知。随后出现的突破性方法绕开了传统的EDA依赖：Sohrabizadeh 等人通过代理建模实现了毫秒级评估 [4]，；Ferretti 等人在无需编译器内部机制的情况下，达到了与工业级模拟器相当的准确度 [5]，；Gao 等人则率先直接处理 C/C++源代码，并采用多粒度嵌入 [6]。最近，Qin 等人提出了一种结合源代码序列与控制数据流图的表示学习方法（ProgSG [7]）。

DSE方法。以往的DSE方法主要基于进化算法和机器学习构建。Wang 等人利用机器学习优化元启发式参数，证明其性能显著优于手动调优方法 [8]。Wu 等人建立了一个由强化学习驱动的框架，用于在相互竞争的目标之间实现帕累托最优资源分配 [9]，；而 Yao 等人则通过 MOEA/D 和 EDA概率建模分解 DSE 问题，以减少综合运行次数 [10]。Xu 等人提出了 LLMMH框架，将大语言模型（LLM）作为解算子融入元启发式算法，以提高 DSE 的精度 [11]。

尽管基于图神经网络的结果质量预测已展现出潜力，但仍存在若干局限。控制数据流图只能向图神经网络提供控制流和数据流信息，无法明确捕捉对预测准确率至关重要的结构语义或程序员注释。例如，对于 #pragma HLS UNROLL factor=4 这类指令，控制数据流图中包含与相应代码块相连的循环展开节点。然而，图特征提取器难以同时捕捉循环展开指令的精确语义及其作用范围。插入循环展开节点虽能补充抽象控制数据流图的信息，却无法准确表示源代码中使用 pragma 的隐含意图。尽管 ProgSG[7] 引入了词元-节点对齐来缓解这一问题，但其单体式 Transformer架构可能无法充分捕捉复杂的源代码特征，也缺乏融合图与文本表示的稳健机制。例如，ProgSG 的方法只是将循环展开指令转换为节点，并使其与相应词元对齐，却未能捕捉该指令如何影响代码语句块。此外，虽然设计空间探索算法旨在迭代改进解，但许多现有方法往往未能系统地考虑 pragma 指令对最终结果质量指标（如延迟和资源利用率）的重大影响。为应对上述挑战，我们提出了 MPM-LLMDSE，这是一个具备以下特点的自动化框架：

1) **多模态数据集**：构建图-文本数据集，将控制数据流图的结构特征与使用预训练语言模型（LM）从添加 pragma 的源代码中提取的语义嵌入相结合。

2) **多模态预测模型（MPM）**：提出一种混合架构，分别利用语言模型和图神经网络从源代码和控制数据流图中提取互补特征，并通过多头注意力机制进行动态融合。

3) **LLM驱动的设计空间探索引擎（LLM4DSE）**：引入大语言模型作为设计空间探索的优化器，并提出一种先进的提示工程方法（PEODSE），将 pragma 影响分析和高质量示例纳入其中，以引导大语言模型生成高质量的设计配置。

II. 预备知识

A. 高层次综合*QoR*预测中的多模态图学习

现有基于图神经网络的高层次综合 QoR预测方法依赖编译工具生成的控制数据流图 [12], [13],，主要利用结构流信息，却未能充分捕捉源代码中丰富的语义特征。这一问题促使我们探索如何更有效地整合源代码级语义，以提高预测准确率。Tang等人 [14] 证明了大语言模型嵌入与多层感知机头结合所具备的内在回归能力，而Joshi等人 [15] 则将Transformer视为消息传递图神经网络。基于Transformer的语言模型（如BERT [16]）能够利用其固有的多头注意力机制，从文本中学习语义和句法特征 [17]。基于这些理论基础，我们认为，融合图表示与文本表示，能够生成同时包含语义信息和图结构信息的嵌入

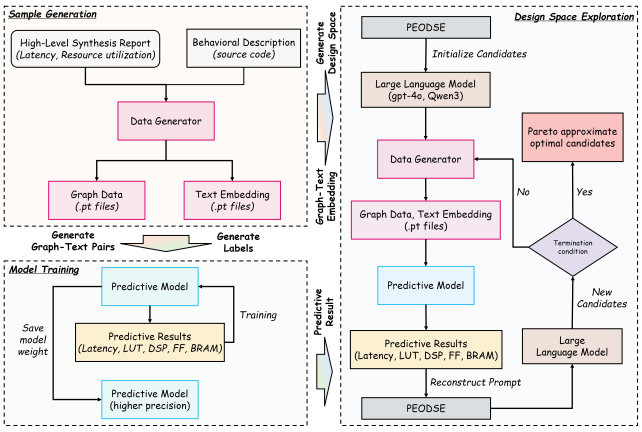

图2：MPM-LLM4DSE的框架。

，从而实现更准确的结果质量预测。然而，图表示和文本表示处于不同的表示空间，因此如何有效整合二者成为关键挑战。基于上述分析，我们将面向结果质量预测的多模态图学习概念化如下：

$$Target_i = MLP_i\left(Fuse(h_{\mathcal{G}}, h_{\mathcal{S}})\right) \tag{1}$$

其中，i 表示QoR预测目标的索引，$h_{\mathcal{G}}$和 $h_{\mathcal{S}}$ 分别表示由图神经网络提取的全局结构表示和由语言模型得到的全局语义表示，$Fuse(\cdot)$表示参数化融合机制，用于将 $h_{\mathcal{G}}$ 和 $h_{\mathcal{S}}$ 整合为联合表示

.

B. 用于*HLS QoR*预测的多模态数据集

传统的单一CDFG数据集与我们提出的预测模型不兼容，因此我们引入了一种专为QoR预测设计的新型多模态数据集。给定一个CDFG $\mathcal{G}$、其设计配置 d_c和行为描述b_d，我们首先将配置 d_c 和描述 b_d 合并为复合文本输入。经过分词后，这些输入将被送入语言模型。 $h_{\mathcal{S}}$ 通过对语言模型最后一层中CLS标记的隐藏状态求和后取平均得到（CLS标记本身表示输入文本的全局语义表示）。该过程表示为：

$$h_{\mathcal{S}} = \frac{1}{l}\sum_{k=0}^{l} CLS(LM_{hidden_k}(tokenizer\,(merge\,(d_c, b_d)))) \tag{2}$$

其中，$hidden_k$ 是语言模型第 k 层最后一个隐藏状态的输出，$merge(\cdot)$用于合并行为描述和设计配置，$tokenizer(\cdot)$用于将文本数据转换为标记，$CLS(\cdot)$则从语言模型的隐藏状态中提取 CLS标记特征。将获得的特征表示 $h_{\mathcal{S}}$ 与相应的控制数据流图 $\mathcal{G}$ 结合，即可生成各个可训练数据实例。

III. 方法论 Y

图 2 展示了 MPM-LLM4DSE 框架，该框架集成了三个紧密协作的模块：样本生成模块使用编译工具和图神经网络从源代码中提取图表示，同时利用语言模型提取

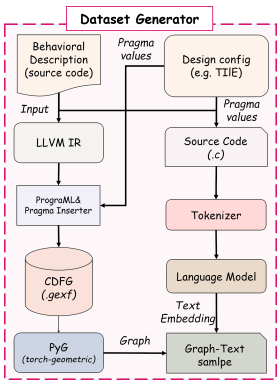

图 3：数据集生成器的工作流程。

表 I：图数据的节点特征。

特征	描述	值类型
节点类型	节点类型索引	long
指令类型	指令索引	long
函数类型	函数索引	long
块类型	块索引	long
延迟	节点中的时钟周期数	int
LUT	节点中使用的查找表数量	int
DSP	节点中的数字信号处理器数量	int
FF	节点中 FF 的数量	int

文本表示，从而形成多模态嵌入表示。随后，将这些表示与从 HLS 报告中提取的 QoR指标相结合，生成训练样本。模型训练模块初始化模型权重，并通过在生成的多模态数据集上进行训练来优化权重，持续保存能够使预测误差最小化的权重。设计空间探索模块负责组织迭代循环，在该循环中，特定任务的提示词引导大语言模型生成设计配置，直至满足给定的终止条件（例如收敛阈值或迭代上限）。每次迭代都包括数据生成器将这些内容与源代码合成为图文本嵌入表示，以及预测模型评估相应的 QoR指标。

A. 数据生成器

图 3 概述了数据生成器的处理流程。源代码首先由 LLVM 处理，以生成中间表示（IR），随后通过 ProGraML 将其转换为控制数据流图（CDFG）。与此同时，根据 GNN-DSE [4], pragmainserter 构造 icmp 节点，以使用 pragma 相关信息丰富 CDFG。另一方面，如公式 2 所述，我们通过整合 pragma 获得完整的源代码。

尽管大语言模型的参数量迅速增长至数千亿，但大多数语言模型仅包含数百万个参数。然而，对于专门的下游任务，在领域专用语料库上训练的语言模型已被证明优于大语言模型。此外，由于部署速度更快且计算需求更低，这类语言模型非常适合用于结果质量预测应用。在本研究中，我们选择预训练的 CodeBERT-c [18] 作为语言模型，因为行为描述通常以 C 或 C++ 编写。CodeBERT-c 专门使用 C 代码进行训练，

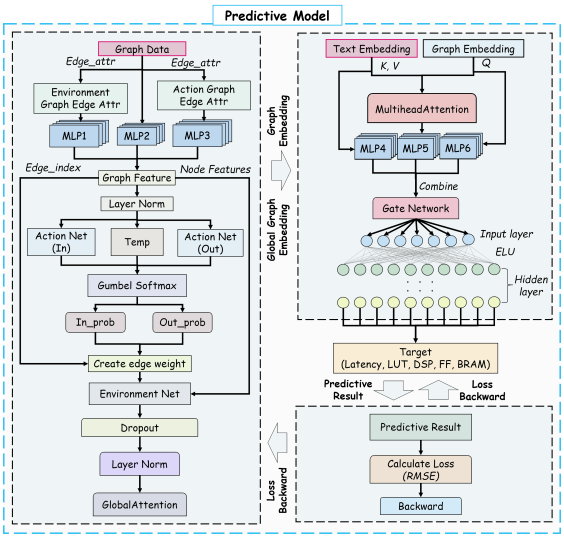

图 4：所提出的预测模型的架构。

因此对 C 语言的理解能力更强。由此，选择 CodeBERT-c 能够更好地从源代码中提取特征。我们使用分词器将源代码转换为词元表示，并利用 CodeBERT 从这些词元表示中提取并保存特征。通过整合 CDFG 和文本嵌入，我们获得了可训练的图文本表示。图文本表示与 HLS 报告中的 QoR 指标或预测结果共同构成训练样本。此外，CDFG 的节点特征列于表 I。类别特征（如节点类型和指令类型）采用独热编码表示，而预处理后的数值特征则直接使用。

B. 多模态预测模型

图 4 展示了所提出的多模态预测模型的架构，该模型由三个核心部分组成：使用图神经网络进行图特征提取、融合图特征与文本嵌入，以及预测模型的端到端训练。首先，在进行图神经网络处理之前，图数据会经过初步转换，将原始边特征矩阵映射为两个不同的边特征矩阵。这些矩阵与原始图结构共同构成增强的图特征表示，并用于后续阶段。

其次，我们受 CoGNN [19]框架启发，设计了一种用于 QoR预测的增强型 CoGNN（ECoGNN）变体。由于基线 CoGNN 仅处理无向图信息流，因此我们扩展了其功能。我们将节点状态集合 Φ 重新定义如下：

$$\Phi = \{S, L_{in}, L_{out}, B, I\} \tag{3}$$

其中， S、 B和 I 表示信息聚合时的节点状态： S 接收邻居信息，同时广播自身状态； B 仅广播自身状态而不接收信息； I 既不广播也不接收信息； L_{in}仅从通过入

边连接的邻居接收信息，而 L_{out} 仅从通过出边连接的邻居接收信息。

在第 k 个 ECoGNN 层中，我们将节点特征 h_{k-1} 输入层归一化 [20]，以获得正则化后的特征 h'_{k-1}。由于节点动作遵循分类分布，训练过程不可微。因此，我们采用 Gumbel-Softmax估计器 [21] 实现可微训练。根据 [11], 中的对比实验，环境网络和动作网络分别使用 MEANGUNs [22]和 SUMGNNs[22] 实现，其中 Temp（可学习的 Gumbel-Softmax 温度）主要通过多层感知机实现。边$edge_weights$的权重根据边的索引 $edge_indexs$ 以及信息流入和流出的概率 $prob_{in}$, $prob_{out}$，分别为出边和入边构建；将其与 $\mathcal{G}'_f$ 整合后输入环境网络，得到 h_i，随后通过层归一化进行正则化。重复上述过程，直至从最后一个 ECoGNN 层输出节点特征，即可获得最终的节点级表示 h_f。为了得到图级嵌入 $h_\mathcal{G}$，可将其直接输入多层感知机进行 QoR预测。然而，使用平均池化会造成显著的信息损失。为此，我们利用全局节点注意力 [23]生成最终的图级嵌入 $h_\mathcal{G}$。在获得源代码的文本嵌入和图级嵌入后，我们优先考虑多模态特征融合。直接求和会引入过多特征噪声。在图像描述任务中，Sun等人采用多头注意力将图像区域与词元对齐 [24]。对于面向 QoR 的多模态图学习，我们的融合旨在为图数据中的节点补充语义表示，并增强子区域的结构信息。多头注意力 [25]通过融合图级嵌入和文本嵌入，有效实现了这一目标。

然后，我们提出使用图嵌入 $h_\mathcal{G}$ 作为查询矩阵 Q，将文本嵌入作为键矩阵 K 和值矩阵 V。随后执行多头注意力计算，并聚合结果以获得融合特征 h_{fuse}。该过程表示为：

$$head_w = Attention(QW_w^Q, KW_w^K, VW_w^V) \tag{4}$$

$$h_{fuse} = Concat(head_1, ..., head_w)W^O \tag{5}$$

其中，W_w^Q、W_w^K 和 W_w^V 分别是 Q、K 和 V 的可学习权重矩阵，W^O 用于融合多个头的输出。多层感知机首先对齐图嵌入 $h_\mathcal{G}$ 和融合特征 h_{fuse} 的维度。随后，将这些对齐后的特征输入门控网络 [23]，以动态控制它们之间的信息流。

最后，将动态融合特征 h'_{fuse} 输入多层感知机预测头，以生成预测结果。计算均方根误差（RMSE）损失，以更新模型参数。虽然平均绝对百分比误差（MAPE）也能反映预测准确率，但对于延迟这类数值超过数十万的目标，1%的误差就意味着数千个时钟周期。因此，RMSE 能更准确地量化预测误差。

图 5 展示了用于 QoR 预测的多模态特征融合过程。图 5 左侧展示了 ECoGNN 从

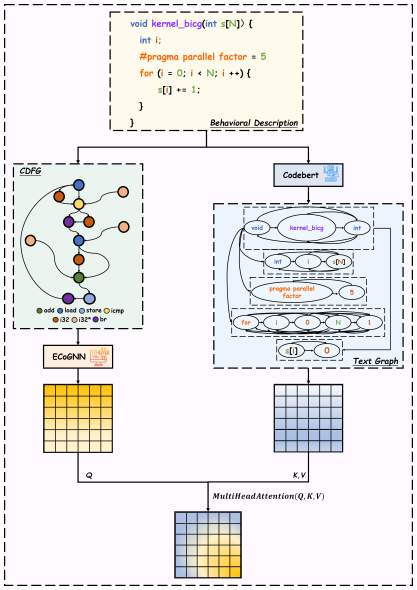

图 5：用于 QoR预测的多模态特征融合。

CDFG 中提取特征的过程，提取出的特征构成查询矩阵。右侧展示了 LM 从源代码中提取特征的过程，其中文本图 [15] 说明，该过程本质上是一种类似于图神经网络的消息传递过程。获得的文本嵌入作为键矩阵和值矩阵，通过多头注意力机制与查询矩阵融合，从而生成兼具语义信息和图结构信息的特征表示。

C. 设计空间探索

为充分发挥大语言模型在智能设计空间探索中的语义能力，我们提出了 LLM4DSE 方法。该方法包含三个核心步骤：利用大语言模型作为优化器来初始化或生成新的解，评估这些解，并相应地更新最优解和提示词。如图 2 所示，迭代过程首先采用面向任务的提示方法 PEODSE，引导大语言模型初始化解。随后，数据生成器对这些解进行处理，以生成图嵌入和文本嵌入，并将其输入预测模型进行评估。评估结果会更新最优解集，并触发 PEODSE 的重构，进而引导大语言模型生成新解。该循环持续进行，直至满足迭代终止条件（探索设计配置的最大数量）。

为什么选择 PEODSE? 大语言模型具有强大的语义理解能力，但其在目标任务上的表现高度依赖精心设计的提示词。尽管 LLMMH [11]创新性地引导大语言模型为元启发式算法生成解，但大语言模型参数的指数级增长，意味着需要更复杂的提示工程，才能充分利用领域知识。LLMMH 的提示设计依赖简短的任务描述，这限制了模型对任务的深入理解。尽管零样本 [26]、少样本 [27], 、通过提示进行优化

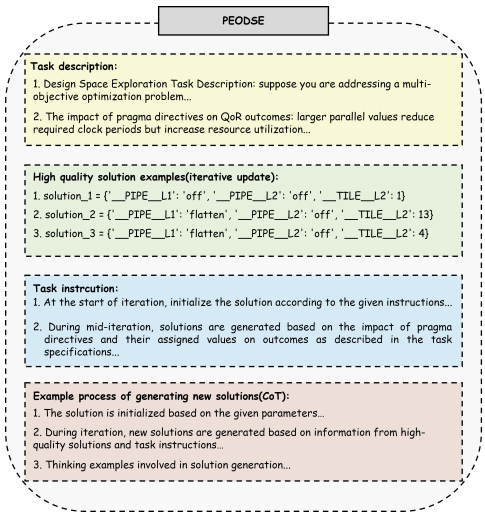

图 6：所提出的设计空间探索提示工程策略示意图。

（OPRO） [28], 和思维链（CoT） [29]等成熟技术在提示工程方面已展现出良好前景，我们认为，为提示词补充高质量的上下文信息，能够显著提升大语言模型生成解的质量。为此，我们提出 PEODSE，这是一种专门设计的提示构建方法，可提供全面的设计空间探索（DSE）任务信息、动态更新的高质量解以及用于生成解的推理轨迹，从而降低大语言模型生成错误解的概率。

如图 6 所示，PEODSE 由四个部分组成：任务描述、高质量解决方案示例、任务指令和解决方案生成示例。任务描述介绍任务背景，以及 pragma 指令及其取值对结果质量的影响。受 OPRO 方法启发，高质量解决方案示例提供典型配置，并在迭代过程中动态更新。任务指令为大语言模型生成解提供明确指引。在任务指令中，我们纳入 pragma 指令对结果质量的影响，以实现更智能的设计空间探索（例如，将流水线指令设为“off”会降低查找表等资源的利用率，同时相对增加延迟；而设为“flatten”则会产生相反的效果）。解决方案生成示例纳入了 CoT 推理，以展示逐步推导解的过程。

IV. 实验

A. 实验设置

基于 GNN-DSE [4], 在 III-A 节中详述的数据集，我们构建了一个用于 QoR预测的多模态图学习数据集，并采用了比 GNN-DSE 更具区分度的划分方案。如表 II 所示，我们从 MachSuite [30] 中选取五个目标内核，并从多面体基准测试 [31] 中选取十个目标内核，作为训练集。训练集的 70% 用于训练，15% 用于测试，15% 用于验证。所有图神经网络模型的隐藏层维度均为 128，并使用 Adam 进行优化，批次大小为 64，学习率为 0.001，迭代 500+ 次。实验在 AMD Ultrascale+ MPSoC ZCU104 平台上进行。

表 II：训练集包含共 15 个基准测试，总计 4,353 个图文样本。数据集大小表示每个基准测试用于训练或推理的图文样本数量。

内核	描述	数据集大小
adi	交替方向隐式求解器	322
aes	一种常见的分组密码	45
atax	矩阵转置和向量乘法	227
bicg	双共轭梯度稳定法线性求解器的双共轭梯度子内核	347
doitgen	多分辨率分析内核	74
fdtd-2d	二维有限差分时域内核	289
gemm-blocked	分块版本的矩阵乘法	243
gemm-ncubed	稠密矩阵乘法	362
gemver	向量乘法和矩阵加法	464
gemm-p	矩阵乘法	303
gesummv	标量、向量和矩阵乘法	159
mvt	矩阵向量乘积和转置	476
spmv-crs	稀疏矩阵-向量乘法（可变长度）	73
spmv-ellpack	稀疏矩阵-向量乘法（固定大小）	114
2mm	2 次矩阵乘法	388
3mm	3 个矩阵乘法	467

表 III：用于推理和设计空间探索的内核。

内核	描述	数据集大小 # 设计配置	
heat-3d	三维数据域上的热方程	225	71511
jacobi-1d	一维 Jacobi stencil 计算	222	2871
jacobi-2d	二维 Jacobi 模板计算	524	7609187
nw	动态规划算法	386	6615
seidel-2d	二维 Seidel 模板计算	56	10919
stencil	二维模板计算	524	7591

使用 Vitis-HLS 2022.1 和 Vivado 2022.1 对基准测试进行综合，以收集真实值延迟和资源利用率（查找表、数字信号处理器、触发器、块随机存取存储器）指标，并将其用作训练标签。

B. HLS QoR 预测准确率评估

我们首先评估所提出的 MPM 的 HLS QoR预测性能。模型训练采用表 II 中详细说明的数据集。为展示泛化能力，我们在完全未见过的内核上进行推理，数据集详情见表 III。我们的比较分析涵盖了最先进（SOTA）方法：GNN-DSE [4],HGBO [3], IronMan-Pro [9], PROGSG [7]。实验结果表明，MPM 在所有目标上的预测误差均最低：0.3870（延迟）、0.0004（查找表）、0.0004（数字信号处理器）、0.0015（触发器）、0.0005（块随机存取存储器）。与 ProgSG 相比，我们的方法显著优于它，最高可达 10.25×。MPM 的卓越性能源于其增强的多模态特征融合能力，以及采用了语义理解能力更强的文本特征提取器，从而有效建立文本特征与源代码特征之间的对应关系。

我们还通过将仅 GNN 和仅 LM 与 MPM 进行比较，开展了消融研究。结果表明，单独使用 LM 的表现优于 ECoGNN，这凸显了代码语义在 QoR预测中的潜在价值。在多模态图学习框架下，MPM 将来自 CDFG（通过 ECoGNN）的图特征与文本特征（通过 LM）融合，显著优于仅 GNN 的方法。

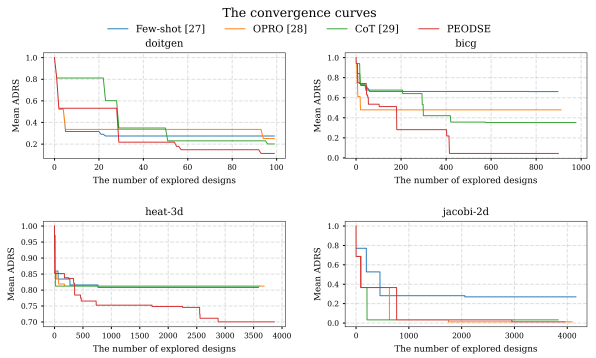

图7：不同提示策略实现的收敛曲线。

表IV：MPM 和最先进模型在未见过的应用上的均方根误差损失。

模型	延迟	LUT	DSP	FF	BRAM	All
GNN-DSE [4]	0.7759	0.0025	0.0023	0.0078	0.0023	0.7909
HGBO [3]	0.5754	0.0127	0.0041	0.0064	0.0069	0.6082
IronMan-Pro [9]	0.4201	0.0012	0.0015	0.0016	0.0007	0.4251
ProgSG [7]	0.4061	0.0041	0.0018	0.0081	0.0025	0.4226
仅 ECoGNN	0.4019	0.0016	0.0015	0.0016	0.0015	0.4084
仅 LM	0.3920	0.0011	0.0011	0.0015	0.0009	0.3965
MPM	**0.3870**	**0.0004**	**0.0004**	**0.0015**	**0.0005**	**0.3898**

C. 设计空间探索的评估

为证明我们提出的 PEODSE 方法在 DSE 任务中优于其他提示方法，我们使用 LLM4DSE 框架评估了各种提示策略。我们使用与参考集的平均距离（ADRS）指标 [32] 对近似帕累托最优集的质量进行量化：

$$ADRS(\Gamma,\Omega)=\frac{1}{|\Gamma|}\sum_{\lambda\in\Gamma}\min_{\mu\in\Omega}f(\lambda,\mu) \tag{6}$$

其中，Γ 表示参考帕累托最优集，Ω 表示近似帕累托最优集，函数 f 用于计算 λ 与 μ之间的距离。ADRS 值越低，表示近似帕累托最优集相对于参考集的准确性越高。我们使用 GPT-3.5-turbo [33]作为大语言模型，通过 OpenAI 的 API [1]对提示策略进行对比实验。不同提示方法得到的收敛曲线如图7所示。结果表明，PEODSE 优于其他提示方法。这一优势源于 PEODSE 能够结合现有方法的长处，并利用 pragma 指令如何影响结果质量的领域知识。因此，大语言模型能够更深入地理解设计空间探索任务，并持续找到高质量的设计配置。

我们采用提出的 MPM-LLM4DSE 框架，并使用 PEODSE 提示进行对比实验，将其与多种成熟方法进行比较：多目标遗传算法（NSGA-II） [34], 模拟退火（SA） [35], 蚁群优化（ACO） [36], 以及 LLMMH [11]。这些实验针对的是在表 III 所示训练中未见过的内核，

[1]https://openai.com/index/openai-api/

表V：未见过的应用上的 ADRS 结果和运行时间。

基准测试	NSGA-II [34]	模拟退火 [35]	ACO [36]	LLMMH [11]	LLM4DSE（GPT-4o）	LLM4DSE（Qwen3）
heat-3d	0.0756	0.0672	0.0784	0.0645	0.0602	**0.0486**
jacobi-1d	0.0320	0.0253	0.0255	0.0298	**0.0212**	0.0228
jacobi-2d	0.0111	**0.0014**	0.0201	0.0029	0.0086	0.0035
seidel-2d	0.0517	0.0551	0.0606	0.0492	**0.0391**	0.0406
stencil	0.0946	0.0742	0.0893	0.0742	**0.0636**	0.0666
nw	0.0028	0.1197	0.1812	0.0121	0.0018	**0.0010**
平均 ADRS	0.0447	0.0572	0.0758	0.0388	0.0324	**0.0305**
总运行时间（秒）	1081	664	1216	10941	7222	9847
LLM4DSE（Qwen3）相较于 NSGA-II 的改进					**31.77%**	
LLM4DSE（Qwen3）相较于 SA 的改进					**46.68%**	
LLM4DSE（Qwen3）相较于 ACO 的提升					**59.76%**	
LLM4DSE（Qwen3）相较于 LLMMH 的提升					**21.39%**	

表VI：ADRS 值比较分析
在相同条件下，GNN-DSE 和 LLM4DSE 在大型基准测试上的表现
时间约束 .

基准测试 #	设计配置	时间（秒）	LLM4DSE（Qwen3）	GNN-DSE [4]	改进（%）
heat-3d	71511	3068	**0.0486**	0.0689	30.78
jacobi-2d	7609187	4462	**0.0035**	0.0074	52.70
seidel-2d	10919	581	**0.0666**	0.0779	14.51

并采用高性能大语言模型（GPT-4o [37], Qwen3-235B-A22B-Thinking-2507 [38],，结果汇总于表 V。为确保比较公平，所有算法探索的设计数量均相同。实验结果表明，两种 LLM4DSE 组合的探索准确性均优于传统元启发式算法和 LLMMH。LLM4DSE（Qwen3）取得了最低的 ADRS 值，平均表现比传统元启发式算法提升 46.07%，比 LLMMH 提升 21.39%。此外，由于优化后的提示设计降低了大语言模型推理的延迟，LLM4DSE 的运行时间也比 LLMMH 更短。当前 LLM4DSE 的实现基于 API调用。通过在本地部署大语言模型，可以大幅降低通信延迟，进而显著缩短 LLM4DSE 的运行时间。

为进一步评估 LLM4DSE 的有效性，我们将其与 GNN-DSE 的精确设计空间探索算法进行比较 [4]，测试对象为设计空间庞大、包含多达 700 万种设计配置的基准测试。对于每个基准测试，我们记录 LLM4DSE 的运行时间，然后在相同时间内运行 GNN-DSE，并报告其 ADRS。表 VI 显示，在相同的时间约束下，LLM4DSE 找到的设计配置质量更高，其 ADRS 比 GNN-DSE 低 32.66%，证明了它在大型设计空间中的优越性能。

五、结论

本文提出了一种用于HLS QoR预测的新型语言模型-图神经网络框架，为传统高层次综合工具提供了一种高效、准确的替代方案。我们的分析表明，在结果质量预测方面，语言模型的表现优于图神经网络架构。此外，所提出的PEODSE提示方法显著增强了LLM4DSE的任务理解能力，从而改善了设计空间探索结果。本研究为HLS设计空间探索确立了一种新范式，凸显了语言模型和大语言模型在硬件设计领域尚未充分挖掘的潜力。未来的工作将着重采用更小型、经过微调且面向特定任务的模型进行本地执行，以减轻计算负担，并探索其在跨平台综合中的应用。

参考文献

[1] N. Wu 等人，“Ironman：通过强化学习实现高层次综合中的图神经网络辅助设计空间探索”，载于 GLSVLSI。美国计算机协会，2021，第39–44页。[2] Z. Lin 等人，“Powergear：通过异构边中心图神经网络进行FPGA高层次综合早期功耗估计”，载于 DATE，2022，第1341–1346页。[3] H. Kuang 等人，“HGBO-DSE：基于分层图神经网络和贝叶斯优化的高层次综合设计空间探索”，载于 ICFPT，2023，第106–114页。[4] A. Sohrabizadeh 等人，“图神经网络辅助的自动化加速器优化”，载于 DAC。美国计算机协会，2022，第55–60页。[5] L. Ferretti 等人，“用于高级综合设计空间探索的图神经网络”，《美国计算机协会设计自动化与电子系统汇刊》，第28卷，第2期，2022年。[6] M. Gao 等人，“利用图神经网络进行高层次综合中的分层源代码到布线后结果质量预测”，载于 DATE，2024，第1–6页。[7] Z. Qin 等人，“面向高层次综合电子设计自动化的跨模态程序表示学习”，载于 MLCAD。美国计算机协会，2024。[8] Z. Wang 等人，“利用机器学习为高层次综合设计空间探索设置特定于元启发式算法的参数”，载于 DAC。电气电子工程师学会，2020。[9] N. Wu 等人，“Ironman-Pro：通过强化学习和基于图神经网络的建模实现HLS中的多目标设计空间探索”，《IEEE集成电路与系统计算机辅助设计汇刊》，第42卷，第3期，第900–913页，2023。[10] Y. Yao 等人，“用于高层次综合设计空间探索的基于分解的分布估计算法”，《I ntegration》，第100卷，第102292页，2025。[11] L. Xu 等人，“Intelligent4DSE：利用图神经网络和大语言模型优化高层次综合设计空间探索”，arXiv预印本 arXiv:2504.19649，2025。[12] C. Lattner 等人，“LLVM：用于持续程序分析与转换的编译框架”，载于《国际代码生成与优化研讨会》。电气电子工程师学会，2004，第75–86页。[13] C. Cummins 等人，“ProGraML：用于数据流分析和编译器优化的基于图的程序表示”，载于 ICML，2021，第2244–2253页。[14] E. Tang 等人，“理解用于回归的LLM嵌入”，《机器学习研究汇刊》，2025。[15] C. K. Joshi，“Transformer就是图神经网络”，arXiv预印本 arXiv:2506.22084，2025。[16] J. Devlin 等人，“BERT：用于语言理解的深度双向Transformer预训练”，载于《2019年计算语言学协会北美分会会议论文集》，2019，第4171–4186页。[17] K. Clark 等人，“BERT关注什么？对BERT注意力机制的分析”，载于 ACL Workshop BlackboxNLP。ACL，2019，第276–286页。[18] S. Zhou 等人，“CodeBERTScore：使用代码预训练模型评估代码生成”，载于 EMNLP。ACL，2023，第13 921–13 937页。[19] B. Finkelshtein 等人，“协作式图神经网络”，载于 ICML，2024。[20] J. L. Ba 等人，“层归一化”，arXiv预印本 arXiv:1607 06450，2016。[21] E. Jang 等人，“使用Gumbel-Softmax进行类别重参数化”，载于 ICLR，2017。[22] W. L. Hamilton，《图表示学习》。Morgan & Claypool出版社，2020。[23] Y. Li 等人，“门控图序列神经网络”，载于 ICIR，2016。[24] H. Bao 等人，“VLMo：采用模态专家混合的统一视觉语言预训练”，载于 NeurIPS，第35卷，2022，第32 897–32 912页。[25] A. Vaswani 等人，“注意力就是一切”，载于 NeurIPS，2017，第6000–6010页。[26] A. Radford 等人，“语言模型是无监督多任务学习器”，2019。 [在线]。可访问：https://api.semanticscholar.org/CorpusID: 160025533[27] T. B. Brown 等人，“语言模型是少样本学习器”，载于 NeurIPS，2020。[28] C. Yang 等人，“大语言模型作为优化器”，载于第十二届国际学习表征会议，2024。[29] J. Wei 等人，“思维链提示促使大语言模型进行推理”，载于 NeurIPS，2022。[30] B. Reagen 等人，“MachSuite：加速器设计和定制架构的基准测试”，载于 IISWC。电气电子工程师学会，2014，第110–119页。[31] T. Yuki 等人，“PolyBenchC，4.2.1”，2010，https://github.com/ MatthiasJReisinger/PolyBenchC-4.2.1于2025-3-5访问。

[32] B. C. Schafer 等，“高层次综合设计空间探索：过去、现在与未来”，*IEEE*集成电路与系统计算机辅助设计汇刊，第39卷，第10期，第2628–2639页，2019年。[33] L. Ouyang 等，“通过人类反馈训练语言模型遵循指令”，载于*NeurIPS*，2022年。[34] B. C. Schafer，“面向精确延迟行为级IP的并行高级综合设计空间探索”，*ACM*设计、自动化与电子系统汇刊，第22卷，第4期，2017年。[35] B. C. Schafer 等，“用于高层次综合设计空间探索的自适应模拟退火算法”，载于国际超大规模集成电路设计、自动化与测试研讨会，2009年，第106–109页。[36] B. Carrion Schafer，“概率多旋钮高级综合设计空间探索加速”，*IEEE*集成电路与系统计算机辅助设计汇刊，第35卷，406，2016年3月，第394页。[37] OpenAI等，“Gpt-4*21276*，2024*2410*o系统卡”，*arXiv*预印本，*arXiv*：。[38] A.Yang 等，“Qwen3*09388*，2025*2505*技术报告”，*arXiv*预印本，*arXiv*：。