硕士学位论文

面向大模型的推理任务批式调度和弹性伸缩研究

RESEARCH ON BATCH SCHEDULING AND ELASTIC SCALING FOR LARGE-MODEL INFERENCE TASKS

研　　究　　生：何忆源

指　导　教　师：徐敏贤副研究员

南方科技大学

二〇二六年一月

国内图书分类号：XXxxx.x　　学校代码：14325
国际图书分类号：xx-x　　密级：公开

工学硕士专业学位论文

面向大模型的推理任务批式调度和弹性伸缩研究

学 位 申 请 人：何忆源
指 导 教 师：徐敏贤副研究员
专 业 类 别：计算机技术
答 辩 日 期：2026 年 3 月
培 养 单 位：深圳理工大学
学位授予单位：南方科技大学

RESEARCH ON BATCH SCHEDULING AND ELASTIC SCALING FOR LARGE-MODEL INFERENCE TASKS

A dissertation submitted to
Southern University of Science and Technology
in partial fulfillment of the requirement
for the professional degree of
Master of Engineering

by
He Yiyuan

Supervisor: Associate Researcher Minxian Xu

March, 2026

学位论文公开评阅人和答辩委员会名单

公开评阅人名单

刘 XX	教授	南方科技大学
陈 XX	副教授	XXXX 大学
杨 XX	研究员	中国 XXXX 科学院 XXXXXXX 研究所

答辩委员会名单

主席	赵 XX	教授	南方科技大学
委员	刘 XX	教授	南方科技大学
	杨 XX	研究员	中国 XXXX 科学院 XXXXXXX 研究所
	黄 XX	教授	XXXX 大学
	周 XX	副教授	XXXX 大学
秘书	吴 XX	助理研究员	南方科技大学

南方科技大学学位论文原创性声明和使用授权说明

南方科技大学学位论文原创性声明

本人郑重声明：所提交的学位论文是本人在导师指导下独立进行研究工作所取得的成果。除了特别加以标注和致谢的内容外，论文中不包含他人已发表或撰写过的研究成果。对本人的研究做出重要贡献的个人和集体，均已在文中作了明确的说明。本声明的法律结果由本人承担。

作者签名：　　日期：

南方科技大学学位论文使用授权书

本人完全了解南方科技大学有关收集、保留、使用学位论文的规定，即：

1. 按学校规定提交学位论文的电子版本。

2. 学校有权保留并向国家有关部门或机构送交学位论文的电子版，允许论文被查阅。

3. 在以教学与科研服务为目的前提下，学校可以将学位论文的全部或部分内容存储在有关数据库提供检索，并可采用数字化、云存储或其他存储手段保存本学位论文。

（1）在本论文提交当年，同意在校园网内提供查询及前十六页浏览服务。

（2）在本论文提交 □ 当年/□ __ 年以后，同意向全社会公开论文全文的在线浏览和下载。

4. 保密的学位论文在解密后适用本授权书。

作者签名：　　日期：

指导教师签名：　　日期：

摘 要

随着大语言模型（LLMs）在云计算环境中的广泛部署，其推理服务面临着资源消耗巨大、服务等级目标（SLO）要求严格、以及预填充与解码阶段资源需求异构等多重挑战。本文围绕“面向大模型的推理任务批式调度和弹性伸缩研究”这一主题，提出了系统化的资源管理与调度优化方案，显著提升了推理服务的吞吐量、资源利用率与服务质量。

首先，针对传统推理系统中批处理策略粗放、请求组合不合理、以及部署配置静态化等问题，本文提出了 UELLM 统一高效推理框架。该框架通过基于微调大模型的资源画像方法，精确预测请求输出长度；设计了 SLO 与输出长度驱动的动态批处理算法（SLO-ODBS），在满足服务等级目标的前提下优化请求组合，减少 KV Cache 冗余；同时提出了基于动态规划的高效低延迟资源分配算法（HELR），根据集群硬件拓扑自动优化模型层到 GPU 的映射策略。实验结果表明，UELLM 相比现有方法可降低推理延迟 72.3% 至 90.3%，提升 GPU 利用率 1.2 倍至 4.1 倍，吞吐量提高 1.92 倍至 4.98 倍。

其次，面向 PD 分离架构中固有的计算–内存负载不均衡、静态资源配置适应性差、以及前缀缓存感知路由导致的热点倾斜等问题，本文提出了 BanaServe 动态编排框架。该框架创新性地引入了模块级细粒度迁移机制，包括层级权重迁移与注意力级 KV Cache 迁移，实现了计算与内存资源的在线动态重平衡；设计了全局 KV Cache 存储与层间流水线重叠传输机制，消除了前缀缓存对调度策略的约束；并提出了基于实时负载感知的请求调度算法。实验表明，BanaServe 相比 vLLM 吞吐量提升 1.2 倍至 3.9 倍、延迟降低 3.9% 至 78.4%；相比 DistServe 吞吐量提升 1.1 倍至 2.8 倍、延迟降低 1.4% 至 70.1%，并在 Azure 生产环境 traces 下展现出优异的鲁棒性。

本文所提出的方法在真实 GPU 集群和公开基准数据集上得到了充分验证，为大模型推理服务的高效部署与资源管理提供了新的理论依据和技术路径。

关键词：大语言模型推理；批式调度；弹性伸缩；资源管理；动态迁移

Abstract

With the widespread deployment of Large Language Models (LLMs) in cloud computing environments, inference services face critical challenges including massive resource consumption, strict Service Level Objective (SLO) requirements, and heterogeneous resource demands between the prefill and decode phases. This thesis focuses on "Batch Scheduling and Elastic Scaling for Large Language Model Inference," proposing systematic resource management and scheduling optimization schemes that significantly improve inference throughput, resource utilization, and service quality.

Firstly, to address the issues of coarse-grained batching strategies, inefficient request combinations, and static deployment configurations in existing systems, this thesis proposes UELLM (Unified and Efficient LLM Inference Serving), a comprehensive inference framework. UELLM employs a fine-tuned LLM-based resource profiler to accurately predict request output lengths. It introduces the SLO and Output-Driven Dynamic Batch Scheduler (SLO-ODBS), which optimizes request combinations while meeting SLO constraints and reducing KV Cache redundancy. Additionally, the High-Efficiency Low-Latency Resource Allocation (HELR) algorithm based on dynamic programming is proposed to automatically optimize layer-to-GPU mapping strategies according to cluster hardware topology. Experimental results demonstrate that UELLM reduces inference latency by 72.3% to 90.3%, improves GPU utilization by 1.2× to 4.1×, and increases throughput by 1.92× to 4.98× compared to state-of-the-art methods.

Secondly, targeting the inherent compute-memory load imbalance, poor adaptability of static resource configurations, and hotspot skews caused by prefix cache-aware routing in PD (Prefill-Decode) disaggregated architectures, this thesis presents BanaServe, a dynamic orchestration framework. BanaServe introduces novel fine-grained module-level migration mechanisms, including layer-level weight migration and attention-level KV Cache migration, enabling online dynamic rebalancing of computational and memory resources. It designs a Global KV Cache Store with layer-wise pipelined transmission to eliminate constraints imposed by prefix caching on scheduling decisions, and proposes a real-time load-aware request scheduling algorithm. Experiments show that BanaServe achieves 1.2×–3.9× higher throughput with 3.9%–78.4% lower latency compared to vLLM, and 1.1×–2.8× throughput improvement with 1.4%–70.1% latency re-

duction compared to DistServe, demonstrating superior robustness under Azure production traces.

The proposed methods have been extensively validated on real GPU clusters and public benchmarks, providing novel theoretical foundations and technical pathways for efficient deployment and resource management of LLM inference services.

Keywords: Large Language Model Inference; Batch Scheduling; Elastic Scaling; Resource Management; Dynamic Migration

目 录

符号和缩略语说明

As-PPT	聚苯基不对称三嗪
DFT	密度泛函理论 (Density Functional Theory)
DMAsPPT	聚苯基不对称三嗪双模型化合物（水解实验模型化合物）
E_a	化学反应的活化能 (Activation Energy)
HMAsPPT	聚苯基不对称三嗪模型化合物的质子化产物
HMPBI	聚苯并咪唑模型化合物的质子化产物
HMPI	聚酰亚胺模型化合物的质子化产物
HMPPQ	聚苯基喹噁啉模型化合物的质子化产物
HMPY	聚吡咙模型化合物的质子化产物
HMSPPT	聚苯基对称三嗪模型化合物的质子化产物
HPCE	高效毛细管电泳色谱 (High Performance Capillary lectrophoresis)
HPLC	高效液相色谱 (High Performance Liquid Chromatography)
IRC	内禀反应坐标 (Intrinsic Reaction Coordinates)
LC-MS	液相色谱-质谱联用 (Liquid chromatography-Mass Spectrum)
MAsPPT	聚苯基不对称三嗪单模型化合物，3,5,6-三苯基-1,2,4-三嗪
MPBI	聚苯并咪唑模型化合物，N-苯基苯并咪唑
MPI	聚酰亚胺模型化合物，N-苯基邻苯酰亚胺
MPPQ	聚苯基喹噁啉模型化合物，3,4-二苯基苯并二嗪
MPY	聚吡咙模型化合物
MSPPT	聚苯基对称三嗪模型化合物，2,4,6-三苯基-1,3,5-三嗪
ONIOM	分层算法 (Our own N-layered Integrated molecular Orbital and molecular Mechanics)
PBI	聚苯并咪唑
PDT	热分解温度
PES	势能面 (Potential Energy Surface)
PI	聚酰亚胺
PMDA-BDA	均苯四酸二酐与联苯四胺合成的聚吡咙薄膜
PPQ	聚苯基喹噁啉
PY	聚吡咙
S-PPT	聚苯基对称三嗪
SCF	自洽场 (Self-Consistent Field)

SCRF	自洽反应场 (Self-Consistent Reaction Field)
TIC	总离子浓度 (Total Ion Content)
TS	过渡态 (Transition State)
TST	过渡态理论 (Transition State Theory)
ZPE	零点振动能 (Zero Vibration Energy)
ab initio	基于第一原理的量子化学计算方法，常称从头算法
$\Delta G^{\neq}$	活化自由能 (Activation Free Energy)
κ	传输系数 (Transmission Coefficient)
ν_i	虚频 (Imaginary Frequency)

第 1 章　绪论

1.1　研究背景与意义

1.1.1　研究背景

近年来，以 GPT-4[1]、LLaMA[2]、Claude[3]等为代表的大语言模型（Large Language Models, LLMs）在自然语言处理、代码生成、知识检索和内容创作等领域展现出卓越的性能，推动了人工智能技术的跨越式发展。然而，随着模型参数规模从数十亿增长至数千亿甚至万亿级别，其训练和推理过程对计算资源提出了极高的要求。据统计，云平台中约 90% 的人工智能计算资源被用于模型推理服务而非训练[4]，且推理成本随着模型规模呈指数级增长。例如，为维持 ChatGPT 日均 7000 万次访问的服务规模，需要部署超过 6 万张 NVIDIA A100 GPU，初始投资成本高达 16 亿美元，每日电费支出约 10 万美元[5]。

在典型的机器学习即服务（MLaaS）架构中，开发者首先利用大规模数据集离线训练 LLMs，随后将训练好的模型部署到分布式云环境中提供在线推理服务。由于单张 GPU 的显存容量有限（通常为 40GB 或 80GB），而千亿级模型的参数和激活值往往超出单机显存容量，因此必须采用分布式部署策略。然而，随着部署 GPU 数量的增加，设备间的通信开销和同步延迟显著上升；反之，若 GPU 数量不足，则会导致显存溢出（Out-of-Memory, OOM）错误或极长的推理延迟。此外，大模型推理具有独特的两阶段执行特性：预填充（Prefill）阶段和解码（Decode）阶段。Prefill 阶段计算密集，需并行处理整个输入序列以生成第一令牌（Time to First Token, TTFT）；Decode 阶段则受限于自回归特性，逐令牌生成输出，呈现内存密集型特征，其关键指标为每输出令牌时间（Time Per Output Token, TPOT）。这种计算-内存需求的固有不对称性，使得传统单一架构难以同时优化两个阶段，导致严重的资源利用率低下和负载不均衡问题。

另一方面，生产环境的推理负载具有高度动态性和不可预测性。请求到达率（Requests Per Second, RPS）随时间剧烈波动，输入/输出序列长度分布呈现重尾特性（Heavy-tailed）。传统静态资源配置策略在负载低谷期造成严重的资源浪费（GPU 利用率仅 20%-40%），而在突发流量（Bursty Traffic）下又因扩容滞后导致服务等级目标（Service Level Objective, SLO）违约甚至服务中断。现有系统如 vLLM[6]通过 PagedAttention 优化显存管理，DistServe[7]采用 PD 分离架构消除阶段间干扰，但这些方案仍受限于粗粒度的资源配置（实例级或 GPU 池级）和缓存感知的静态

路由策略，无法适应快速变化的负载模式。因此，如何在保障严格 SLO 的前提下，实现计算与内存资源的细粒度弹性调度，成为大模型推理服务面临的核心挑战。

1.1.2　研究意义

本研究围绕大模型推理服务的批式调度与弹性伸缩展开，具有以下重要的理论价值和实践意义：

从科学价值角度，本研究针对 LLM 推理中计算-内存协同优化这一基础科学问题，揭示了静态资源配置与动态负载需求之间的结构性矛盾，提出了从请求级批处理到模块级弹性伸缩的分层优化理论框架。通过建立考虑 KV Cache 动态增长、网络拓扑异构性和 SLO 约束的数学模型，探索了在离散配置空间中寻找最优部署策略的算法边界，为大规模分布式推理系统的资源管理提供了新的理论依据。

从工程实践角度，研究成果可直接应用于云原生 AI 基础设施，具有显著的经济效益和社会价值。首先，通过精确的输出长度预测和 SLO 感知的批处理算法，可减少无效填充（Padding）和冗余计算，预计可降低推理成本 30% 以上；其次，模块级细粒度迁移技术打破了传统副本级扩缩容的粒度限制，将资源响应时间从分钟级降至秒级，显著提升系统对突发流量的鲁棒性；再者，跨实例的 KV Cache 共享机制消除了缓存局部性对调度策略的约束，解决了前缀缓存感知路由导致的热点倾斜问题，提高了集群整体资源利用率。这些技术对推动大模型在智能客服、自动驾驶、金融风控等延迟敏感场景的普及应用具有重要意义，同时通过提升资源利用效率减少了能源消耗，符合绿色计算和可持续发展的战略目标。

此外，本研究提出的统一优化框架兼顾算法效率与系统可实现性，已在真实 GPU 集群和 Azure 生产环境 traces 中验证了其有效性，为学术界和工业界提供了可复现、可部署的技术路径，有助于填补当前 LLM 推理服务在细粒度资源管理方面的研究空白。

1.2　国内外研究现状及分析

本节从大模型推理优化的技术路径出发，系统梳理了批处理调度、资源配置优化和 PD 分离架构三个维度的研究现状，分析现有方案的局限性，并引出本文的研究切入点。

1.2.1　批处理与请求调度优化

批处理（Batching）是提升 LLM 推理吞吐量的关键技术。近期的 vLLM[6] 提出了 PagedAttention 机制，通过块级 KV Cache 管理减少显存碎片，支持动态批处

理（Continuous Batching）。SGLang[8]通过前缀缓存（Prefix Caching）和缓存感知路由减少预填充阶段的冗余计算。然而，这些系统的批处理策略主要依据请求到达顺序（FIFO）或简单启发式规则，缺乏对请求输出长度的先验知识，导致批内请求长度差异过大时产生严重的填充浪费（Padding Waste）。

针对此问题，S^3[9]将批处理建模为装箱问题，通过预测输出长度优化请求组合，但仅考虑内存优化而未考虑 SLO 约束。BATCH[10]提出了针对无服务器平台的自适应批处理框架，但采用穷举搜索配置，时间复杂度高。Tabi[11]针对判别式模型优化，难以应用于生成式 LLMs。现有方案普遍存在三个缺陷：一是缺乏对请求 SLO 的显式建模，导致高优先级请求可能被延迟；二是未考虑 KV Cache 随生成过程动态增长的特性；三是静态批处理策略无法适应负载变化。

1.2.2　模型部署与资源配置优化

在资源配置方面，MArk[12]提出了面向 SLA 的推理服务系统，集成 IaaS 和无服务器计算以降低成本，但主要针对小型传统模型。Morphling[13]采用元学习（Meta-learning）快速搜索最优配置，但需要对每个候选配置进行压力测试，在 LLM 场景下引入显著延迟。Splitwise[14]和 DistServe[7]探索了 PD 分离架构，将 Prefill 和 Decode 阶段分配到不同 GPU 实例以消除阶段间干扰，但采用静态的资源配比，无法适应动态负载。

现有 PD 分离方案存在固有资源失衡问题：Prefill 实例通常计算利用率超过 95% 但显存利用率仅 35%，而 Decode 实例则相反。Mooncake[15]和 MemServe[16]尝试通过分布式 KV Cache 池解决内存瓶颈，但引入跨节点查找开销和一致性协调复杂性。总体而言，当前系统缺乏细粒度的在线资源重平衡机制，难以在请求级动态调整资源配置。

1.2.3　弹性伸缩与负载均衡

弹性伸缩（Auto-scaling）是应对动态负载的关键技术。传统方法如 Kubernetes 的 Horizontal Pod Autoscaler（HPA）采用副本级（Replica-level）扩缩容，粒度粗糙，启动延迟长达 1-30 分钟，无法应对 LLM 推理的突发流量。Llumnix[17]提出了动态负载均衡策略，但未针对 PD 分离架构中的计算-内存异构性进行优化。SpotServe[18]利用可抢占实例降低成本，但缺乏对 LLM 特定访存模式的优化。

在细粒度资源管理方面，Choi 等[19]提出了 GPU 时空共享的 gpulet 抽象，但仅适用于小型传统模型。现有研究尚未解决 PD 分离架构中模块级（Layer-level）的动态迁移问题，无法在 Prefill 和 Decode 实例之间实时转移计算负载或显存压力，导致资源碎片化严重。

1.2.4　现有研究的不足与分析

综合上述分析，现有 LLM 推理系统在以下三个方面存在显著局限：

（1）调度策略与资源状态紧耦合：现有系统（如 SGLang、Mooncake）的调度决策严重依赖 KV Cache 的物理位置，前缀缓存感知路由导致负载热点倾斜。路由器被迫在计算负载均衡与缓存命中率之间做困难权衡，无法基于实时负载做出最优调度。

（2）资源配置静态化与粗粒度：现有 PD 分离系统（DistServe、Splitwise）在部署时固定 Prefill 与 Decode 实例比例，无法在运行期间根据实际负载动态调整。副本级扩缩容响应滞后，难以处理突发流量，且模型加载开销巨大。

（3）缺乏跨阶段的细粒度资源协同：Prefill 与 Decode 阶段的资源需求（计算 vs 内存）互补，但现有系统缺乏在两个阶段之间动态迁移工作负载或显存数据的机制，导致严重的资源利用率失衡。

针对上述问题，本文拟从请求级批处理优化和模块级弹性伸缩两个层面，研究大模型推理服务的统一资源管理框架。

1.3　论文主要工作

针对大模型推理服务中资源利用率低、SLO 保障困难、动态适应性差等挑战，本文围绕批式调度与弹性伸缩两个核心问题，系统性地提出了资源画像、动态批处理、智能部署与细粒度迁移/复制的优化方法，主要贡献包括以下两个方面：

1.3.1　统一高效的批式调度与部署框架 UELLM

针对传统推理系统批处理策略粗放、部署配置静态化、缺乏 SLO 感知等问题，本文提出了 UELLM（Unified and Efficient LLM Inference Serving）框架，实现了请求级资源画像、动态批处理与高效部署的协同优化。

（1）基于微调的推理任务资源画像方法：针对 LLM 推理中输出长度不确定导致的资源需求难以预测问题，本文采用基于微调大模型（ChatGLM3-6B）的方法对请求输出长度进行预测。通过在 Alpaca、Natural Questions 等指令遵循数据集上微调，模型对长度分桶的预测准确率达到 99.51%，为后续调度提供先验知识。

（2）SLO 与输出长度驱动的动态批处理算法（SLO-ODBS）：突破了传统 FIFO 批处理的局限，建立了综合考虑请求 SLO 约束和输出长度差异的批处理优化模型。算法通过权重化目标函数平衡总延迟与总输出长度，采用贪心策略将请求组合成批，显著减少 KV Cache 冗余和填充浪费。实验表明，SLO-ODBS 在保持低延迟的同时，将 SLO 违约率降低至接近 0%。

（3）高效低延迟资源分配算法（HELR）：针对 LLM 分布式部署中设备映射（Device Map）搜索空间巨大、静态配置性能次优的问题，提出了基于动态规划的高效资源分配算法。算法综合考虑集群网络拓扑（NVLink、PCIe 带宽异构）、GPU 计算能力差异和模型层间依赖关系，自动求解最优的层到设备映射策略。通过调整权重系数，可同时优化 GPU 利用率（HE 模式）或推理延迟（LR 模式）。

UELLM 框架在真实 4-GPU 集群上的验证表明，相比 Morphling 和 S^3 等先进方案，系统降低推理延迟 72.3% 至 90.3%，提升 GPU 利用率 1.2 倍至 4.1 倍，吞吐量提高 1.92 倍至 4.98 倍，且实现了零 SLO 违约的推理服务。

1.3.2　面向 PD 分离架构的细粒度弹性伸缩框架 BanaServe

针对 PD（Prefill-Decode）分离架构中固有的计算-内存负载失衡、静态资源配置适应性差、前缀缓存导致路由倾斜等问题，本文提出了 BanaServe 动态编排框架，实现了模块级细粒度迁移与全局 KV Cache 共享。

（1）模块级细粒度动态迁移机制：突破传统副本级扩缩容的粒度限制，提出了层级（Layer-level）权重迁移与注意力级（Attention-level）KV Cache 迁移两种机制。层级迁移支持将连续的 Transformer 层动态迁移至负载较轻的 GPU，实现粗粒度负载重平衡；注意力级迁移将 KV Cache 按注意力头维度切分，将部分头的计算卸载至辅助 GPU（Cold GPU），主 GPU（Hot GPU）与辅助 GPU 并行计算，在不迁移模型权重的情况下实现细粒度负载分担。数学上证明了注意力分解的正确性，确保分布式计算与单卡计算数值等价。

（2）全局 KV Cache 存储与层间流水线传输：为消除前缀缓存对路由决策的约束，设计了跨 Prefill 实例的全局 KV Cache 存储层，结合 CPU/SSD 作为持久化后端。针对全局存储引入的访问延迟，提出了层间流水线重叠传输机制，利用 Transformer 逐层计算特性，将第 i 层计算与第 $i+1$ 层 KV Cache 预取重叠，隐藏通信开销。理论分析与实验验证表明，当 KV Cache 传输时间（约 0.082ms）远小于层计算时间（约 4.22ms）时，可实现近透明的缓存访问。

（3）负载感知请求调度算法：基于全局 KV Cache 存储，实现了完全基于实时负载的调度策略，无需考虑缓存局部性。算法周期性地测量各 Prefill 实例的综合负载（计算 + 内存），将新请求分发至负载最轻的实例，并配合动态迁移机制实现快速负载均衡。

BanaServe 在 13B 参数模型（LLaMA-13B、OPT-13B）和公开基准（Alpaca、LongBench、Azure 生产环境 traces）上的评估表明，相比 vLLM，系统吞吐量提升 1.2 倍至 3.9 倍，延迟降低 3.9% 至 78.4%；相比 DistServe，吞吐量提升 1.1 倍至 2.8 倍，延迟降低 1.4% 至 70.1%，并在突发流量下展现出卓越的鲁棒性。

1.3.3　主要创新点

本文的主要创新点可总结为：

（1）提出了 SLO 感知的动态批处理理论框架：首次将输出长度预测与 SLO 约束联合建模，突破了传统批处理仅优化吞吐量的局限，实现了延迟违约率与资源利用率的联合优化。

（2）提出了基于网络拓扑感知的 LLM 部署优化方法：将模型层分配问题形式化为带约束的动态规划问题，综合考虑异构互联拓扑和计算能力差异，解决了传统方法部署配置次优或搜索开销过高的问题。

（3）提出了模块级细粒度弹性伸缩新范式：将扩缩容粒度从实例级下沉至 Transformer 层/注意力头级，实现了 PD 分离架构中计算与内存资源的在线重平衡，填补了细粒度在线资源迁移研究空白。

（4）提出了全局 KV Cache 存储与计算-通信重叠机制：通过解耦缓存状态与计算位置，消除了前缀缓存对调度的约束，结合流水线传输解决了全局存储的延迟瓶颈。

1.4　论文组织架构

本文共分为六章，各章内容安排如下：

第一章绪论：介绍大模型推理服务的研究背景与意义，分析批式调度和弹性伸缩领域的国内外研究现状，阐述本文的主要工作与创新点。

第二章相关技术与理论基础：介绍大语言模型 Transformer 架构、KV Cache 机制、PD 分离架构等背景知识；阐述资源管理中的关键 metrics（TTFT、TPOT、SLO 等）；分析现有系统架构（vLLM、DistServe 等）的技术细节与局限性。

第三章 UELLM：统一高效的批式调度与部署：详细介绍 UELLM 系统架构，包括资源画像模块的设计与实现、SLO-ODBS 批处理算法的数学建模与算法流程、HELR 部署优化算法的动态规划求解过程，以及系统集成与实现细节。

第四章 BanaServe：面向 PD 分离架构的细粒度弹性伸缩：阐述 BanaServe 的整体架构设计，详细论述层级迁移与注意力级迁移的数学原理与实现机制，介绍全局 KV Cache 存储的设计与层间流水线传输优化，以及负载感知调度算法与动态决策流程。

第五章实验验证与性能评估：介绍实验环境搭建（硬件配置、测试模型、数据集），设计对比实验验证 UELLM 在批处理、部署优化方面的性能提升，验证 BanaServe 在不同负载模式、上下文长度、生产环境 traces 下的吞吐量和延迟表现，并进行消融实验分析各组件的贡献。

第六章总结与展望：总结本文的主要研究成果，讨论存在的问题与局限性，展望未来研究方向，包括异构硬件感知调度、预测性弹性伸缩、跨地域分布式推理等。

第 2 章　图、表及条目示例

2.1　插图

图片通常在 figure 环境中使用 `\includegraphics` 插入，如图 2-1 的源代码。建议矢量图片使用 PDF 格式，比如数据可视化的绘图；照片应使用 JPG 格式；其他的栅格图应使用无损的 PNG 格式。注意，LaTeX 不支持 TIFF 格式；EPS 格式已经过时。

建议图的大小一般为宽 6.67 cm × 高 5.00 cm。特殊情况下，也可为宽 9.00 cm × 高 6.75 cm，或宽 13.5 cm × 高 9.00 cm。总之，一篇论文中，同类图片的大小应该一致，编排美观、整齐。图应尽可能显示在同一页（屏）。

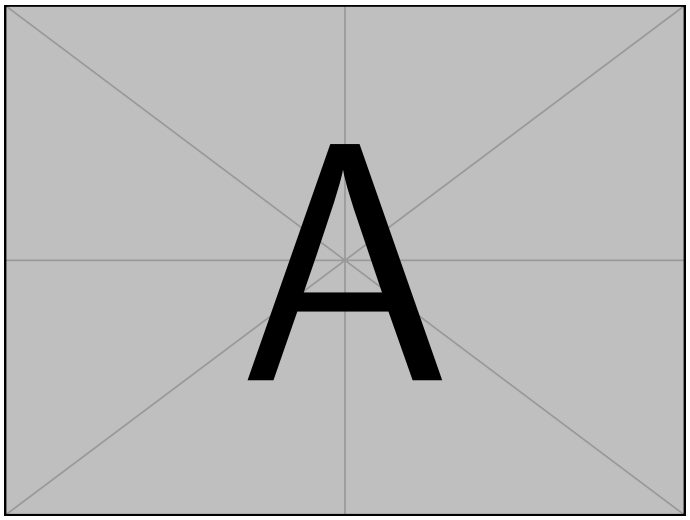


图 2-1　示例图片

若图或表中有附注，采用英文小写字母顺序编号，附注写在图或表的下方。

如果一个图由两个或两个以上分图组成时，各分图分别以 (a)、(b)、(c)..... 作为图序，并须有分图题。推荐使用 subcaption 宏包来处理，比如图 2-3(a) 和图 2-3(b)。

2.2　表格

表应具有自明性。为使表格简洁易读，尽可能采用三线表，如表 2-1。三条线可以使用 booktabs 宏包提供的命令生成。

表格如果有附注，尤其是需要在表格中进行标注时，可以使用 threeparttable 宏包。使用英文小写字母 a、b、c……顺序编号。

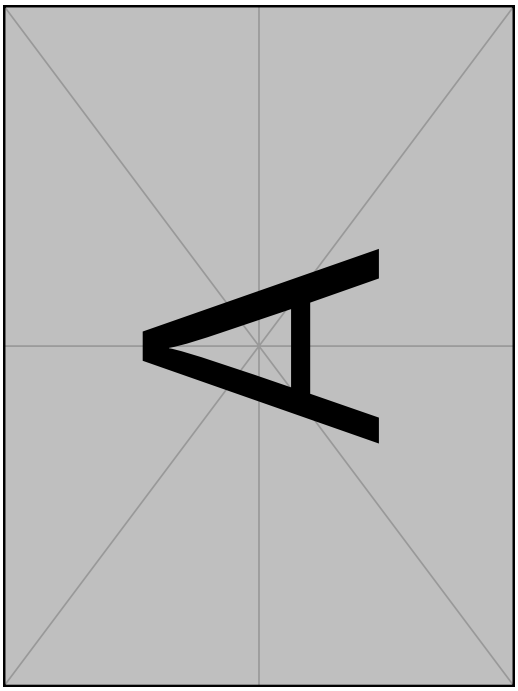
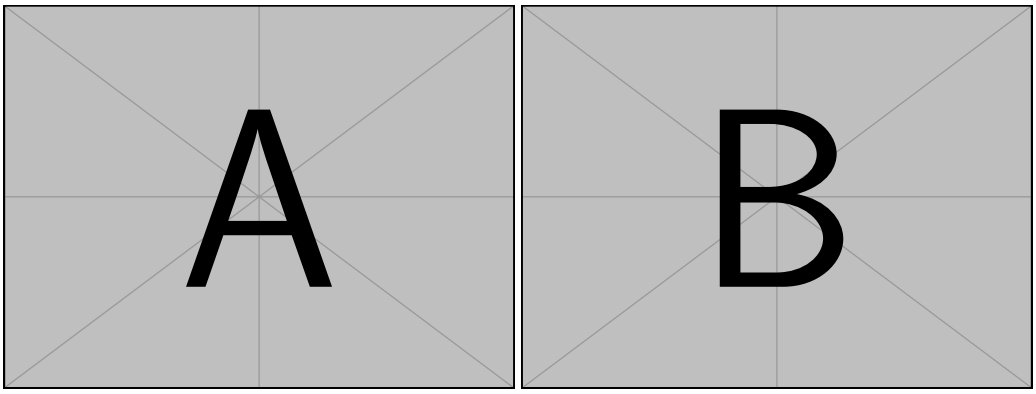


图 2-2　示例图片旋转 90 度



(a) 分图 A　　(b) 分图 B

图 2-3　多个分图的示例

表 2-1　三线表示例

文件名	描述
thuthesis.dtx	模板的源文件，包括文档和注释
thuthesis.cls	模板文件
thuthesis-*.bst	BibTeX 参考文献表样式文件
thuthesis-*.bbx	BibLaTeX 参考文献表样式文件
thuthesis-*.cbx	BibLaTeX 引用样式文件

表 2-2　带附注以及调整列宽的的表格示例

2cm	4cm	6cm
左右居中的 2cm 宽度左右居中的 2cm 宽度[a]	左右居左的 4cm 宽度左右居左的 4cm 宽度	左右居右的 6cm 宽度左右居右的 6cm 宽度
左右居中的 2cm 宽度左右居中的 2cm 宽度[b]	左右居左的 4cm 宽度左右居左的 4cm 宽度	左右居右的 6cm 宽度左右居右的 6cm 宽度

[a] A 的注释
[b] B 的注释

如果需要调整表格列宽度，可以改用命令 `L`, `R`, 或者 `C`, 如 `C{2cm}` 代表居中列宽 2cm。

表 2-3　合并单元格的三线表

Metaclass	A-B		C-D	
Class	A	B	C	D
L1	1	2	3	4
L2	1	2	3	4

如有辅助线要求可以使用 `\cmidrule` 命令。在连续使用时，可以使用一组圆括号括起来的参数 `l`、`r`或`l<距离>`、`r<距离>`表示间距的表格线可以在左右向内缩短一小段，表2-3展示了效果。

表格如果想要与页面等宽，可以使用 tabularx 宏包，如表格2-4所示。模版定义了一些扩展命令，实现一些排版需求。`X` 两端对齐, `Y` 左对齐, `Z` 右对齐，或者 `A` 居中对齐。

如果表格横向宽度不够，可以使用 sidewaystable 将表格旋转 90 度，如表 2-5。

如果您要排版的表格长度超过一页，那么推荐使用 longtable 或者 supertabular 宏包，模板对 longtable 进行了相应的设置，所以用起来可能简单一些。表 2-6 就是 longtable 的简单示例。

表 2-4　同页宽的表格实例

Cell with text aligned to the left	1	2	3
4	Cell with justified text	5	6
7	8	Cell with centered text	9
10	11	12	Cell with text aligned to the right

表 2-6　实验数据（超长表格示例）

测试程序	正常运行时间 (s)	同步时间 (s)	检查点时间 (s)	卷回恢复时间 (s)	进程迁移时间 (s)	检查点文件（KB）
CG.A.2	23.05	0.002	0.116	0.035	0.589	32491
CG.A.4	15.06	0.003	0.067	0.021	0.351	18211
CG.A.8	13.38	0.004	0.072	0.023	0.210	9890
CG.B.2	867.45	0.002	0.864	0.232	3.256	228562
CG.B.4	501.61	0.003	0.438	0.136	2.075	123862
CG.B.8	384.65	0.004	0.457	0.108	1.235	63777
MG.A.2	112.27	0.002	0.846	0.237	3.930	236473
MG.A.4	59.84	0.003	0.442	0.128	2.070	123875
MG.A.8	31.38	0.003	0.476	0.114	1.041	60627
MG.B.2	526.28	0.002	0.821	0.238	4.176	236635
MG.B.4	280.11	0.003	0.432	0.130	1.706	123793
MG.B.8	148.29	0.003	0.442	0.116	0.893	60600
LU.A.2	2116.54	0.002	0.110	0.030	0.532	28754
LU.A.4	1102.50	0.002	0.069	0.017	0.255	14915
LU.A.8	574.47	0.003	0.067	0.016	0.192	8655
LU.B.2	9712.87	0.002	0.357	0.104	1.734	101975
LU.B.4	4757.80	0.003	0.190	0.056	0.808	53522
LU.B.8	2444.05	0.004	0.222	0.057	0.548	30134

续下页

续表 2-6　实验数据（超长表格示例）

测试程序	正常运行时间 (s)	同步时间 (s)	检查点时间 (s)	卷回恢复时间 (s)	进程迁移时间 (s)	检查点文件（KB）
EP.A.2	123.81	0.002	0.010	0.003	0.074	1834
EP.A.4	61.92	0.003	0.011	0.004	0.073	1743
EP.A.8	31.06	0.004	0.017	0.005	0.073	1661
EP.B.2	495.49	0.001	0.009	0.003	0.196	2011
EP.B.4	247.69	0.002	0.012	0.004	0.122	1663
EP.B.8	126.74	0.003	0.017	0.005	0.083	1656
EP.A.2	123.81	0.002	0.010	0.003	0.074	1834

2.3　源代码

使用 listings 环境高亮代码。参数较为复杂，请自行搜索或查阅文档。引用效果如代码 2-1。示例使用 minipage 环境嵌套一层的原因是防止换页中被插入其他浮动体，结合实际情况，按需使用 minipage，例如如需要跨页代码就无需使用 minipage。

```
1 class HelloWorldApp {
2     public static void main(String[] args) {
3         System.out.println("Hello World!"); // Display the
              string.
4         for (int i = 0; i < 100; ++i) {
5             System.out.println(i);
6         }
7     }
8 }
```

代码 2-1　Java 代码示例（使用 listings 高亮）

2.4　伪代码

推荐使用 algorithm2e 宏包中的 algorithm 环境书写伪代码。algorithm2e 可选参数 `linesnumbered` 控制代码行号显示。引用效果如算法 2-1。

表 2-5　旋转 90 度的三线表示例

文件名	描述
thuthesis.dtx	模板的源文件，包括文档和注释
thuthesis.cls	模板文件
thuthesis-*.bst	BibTeX 参考文献表样式文件
thuthesis-*.bbx	BibLaTeX 参考文献表样式文件
thuthesis-*.cbx	BibLaTeX 引用样式文件

算法 2-1　Simulation-optimization heuristic

Data: current period t, initial inventory I_{t-1}, initial capital B_{t-1}, demand samples
Result: Optimal order quantity Q_t^*

$r \leftarrow t$;
$\Delta B^* \leftarrow -\infty$;
while $\Delta B \leqslant \Delta B^*$ and $r \leqslant T$ **do**
    $Q \leftarrow \arg\max_{Q \geqslant 0} \Delta B_{t,r}^{Q}(I_{t-1}, B_{t-1})$;
    $\Delta B \leftarrow \Delta B_{t,r}^{Q}(I_{t-1}, B_{t-1})/(r-t+1)$;
    if $\Delta B \geqslant \Delta B^*$ **then**
        $Q^* \leftarrow Q$;
        $\Delta B^* \leftarrow \Delta B$;
    end
    $r \leftarrow r+1$;
end

算法 2-2　SumExample

Result: s
$s \leftarrow 0$; /* 这是默认多行注释 */
/* 这是默认独占一行的注释 */
/* 这是在取消独占一行后的注释 */
/* 这是恢复独占一行的注释 */
foreach $i \in [1, 100]$ **do**
    if $i\%3 = 0$ **then**
        $s \leftarrow s + i$; // 这是单行注释，一个没有 end 的 if
    else if $i\%3 = 1$ **then**
        break ; ▷ 这是三角形的单行注释，一个没有 end 的 else if
    else
        continue ; /* 这是超长多行注释，关于伪代码的 if-then-else 详细查看 https://texdoc.org/serve/algorithm2e/0 的 10.4 */
    end
end
return s ;

2.5　图表格式测试

图题在图之下，段前空 6 磅，段后空 12 磅。图整体前后距离未定义，目前默认距离：段前空 12 磅，段后空 12 磅。

图前，图前，图前，图前，图前，图前，图前，图前，图前，图前，图前。

图 2-4　图高度为 12bp vs 6bp

图后，图后，图后，图后，图后，图后，图后，图后，图后，图后，图后。

表题在表之上，段前空 12 磅，段后空 6 磅。表整体前后距离未定义，目前默认距离：段前空 12 磅，段后空 12 磅。

表前，表前，表前，表前，表前，表前，表前，表前，表前，表前，表前。

表 2-7　简单表格

column1	column2
column1	column2

表后，表后，表后，表后，表后，表后，表后，表后，表后，表后，表后。图表前后是否有空行不影响图表与正文之间的距离。

2.6　条目编写

2.6.1　支持三级目录显示

支持三级目录显示

2.6.2　条目要求

条目要求首行左缩进 2 个汉字符，避免悬挂缩进。如需使用带括号的条目列表，请自行添加 `label=<style>` 参数。下面是两个例子，还有更多用法，查阅 enumitem 宏包的文档。

默认条目序号：

```
\begin{enumerate} ... \end{enumerate}
```

（1）一级

① 二级

a. 三级

A. 四级，《写作要求》未定义，请自行定义或者选择。

自定义序号样式定义如表 2-8。

表 2-8　条目样式选项

Code	Description
\alph	Lowercase letter (a, b, c, ...)
\Alph	Uppercase letter (A, B, C, ...)
\arabic	Arabic number (1, 2, 3, ...)
\roman	Lowercase Roman numeral (i, ii, iii, ...)
\Roman	Uppercase Roman numeral (I, II, III, ...)

2.6.2.1　条目测试

条目前文字，条目前文字，条目前文字，条目前文字，条目前文字，条目前文字，条目前文字，条目前文字，条目前文字，条目前文字。

（1）一级条目，超长行。南方科技大学，南方科技大学，南方科技大学，南方科技大学，南方科技大学，南方科技大学，南方科技大学，南方科技大学，南方科技大学，南方科技大学，南方科技大学，南方科技大学，南方科技大学。

（2）一级条目，南方科技大学，南方科技大学，南方科技大学。

（3）一级条目，南方科技大学，南方科技大学，南方科技大学。

① 二级条目，超长行。南方科技大学，南方科技大学，南方科技大学，南方科技大学，南方科技大学，南方科技大学，南方科技大学，南方科技大学，南方科技大学，南方科技大学，南方科技大学，南方科技大学，南方科技大学。

② 二级条目，南方科技大学，南方科技大学，南方科技大学。

③ 二级条目，南方科技大学，南方科技大学，南方科技大学。

④ 二级条目，南方科技大学，南方科技大学，南方科技大学。

⑤ 二级条目，南方科技大学，南方科技大学，南方科技大学。

⑥ 二级条目，南方科技大学，南方科技大学，南方科技大学。

a. 三级条目，南方科技大学，南方科技大学，南方科技大学。

b. 三级条目，南方科技大学，南方科技大学，南方科技大学。

c. 三级条目，南方科技大学，南方科技大学，南方科技大学。

A. **四级条目及之后的条目无规定序号格式，请自行设定选择。**

条目后文字，条目后文字，条目后文字，条目后文字，条目后文字，条目后文字，条目后文字，条目后文字，条目后文字，条目后文字。

第 3 章　数学符号和公式

3.1　数学符号

模板中使用 unicode-math 宏包来配置数学符号，

研究生《写作指南》要求量及其单位所使用的符号应符合国家标准《国际单位制及其应用》（GB 3100—1993）、《有关量、单位和符号的一般原则》（GB/T 3101—1993）的规定，但是与 TEX 默认的美国数学学会（AMS）的符号习惯有所区别。

英文论文的数学符号使用 TEX 默认的样式。论文以中文为主要撰写语言按照国标建议的配置数学字体格式：

（1）大写希腊字母默认为斜体，如

$$\Gamma\Delta\Theta\Lambda\Xi\Pi\Sigma\Upsilon\Phi\Psi\Omega.$$

注意有限增量符号 Δ 固定使用正体，模板提供了 `\increment` 命令。

（2）小于等于号和大于等于号使用倾斜的字形 ⩽、⩾。

（3）积分号使用正体，比如 ∫、∮。

（4）行间公式积分号的上下限位于积分号的上下两端，比如

$$\int_a^b f(x)\,\mathrm{d}x.$$

行内公式为了版面的美观，统一居右侧，如 $\int_a^b f(x)\,\mathrm{d}x$ 。

（5）偏微分符号 ∂ 使用正体。

（6）省略号 `\dots` 按照中文的习惯固定居中，比如

$$1,2,\cdots,n \quad 1+2+\cdots+n.$$

（7）实部 Re 和虚部 Im 的字体使用罗马体。

以上数学符号样式的差异可以在模板中统一设置。另外国标还有一些与 AMS 不同的符号使用习惯，需要用户在写作时进行处理：

（1）数学常数和特殊函数名用正体，如

$$\mathrm{\pi} = 3.14\cdots; \quad \mathrm{i}^2 = -1; \quad \mathrm{e} = \lim_{n\to\infty}\left(1+\frac{1}{n}\right)^n.$$

（2）微分号使用正体，比如 $\mathrm{d}y/\mathrm{d}x$。

（3）向量、矩阵和张量用粗斜体（`\mathbfit`），如 $\boldsymbol{x}$、$\boldsymbol{\Sigma}$、$\boldsymbol{T}$。

（4）自然对数用 $\ln x$ 不用 $\log x$。

关于数学符号更多的用法，参考 unicode-math 宏包的使用说明，全部数学符号命的令参考 unimath-symbols，也可以参考 Stack Overflow 上的答案 What are all the font styles I can use in math mode?。

关于量和单位推荐使用 siunitx 宏包，可以方便地处理希腊字母以及数字与单位之间的空白，比如：$6.4 \times 10^6\,\mathrm{m}$，$9\,\mu\mathrm{m}$，$\mathrm{kg \cdot m \cdot s^{-1}}$，10 ℃～20 ℃。

3.2　数学公式

数学公式可以使用 equation 和 equation* 环境。注意数学公式的引用应前后带括号，建议使用 `\eqref` 命令，比如式(3-1)。

$$\frac{1}{2\pi \mathrm{i}} \int_{\gamma} f = \sum_{k=1}^{m} n(\gamma; a_k) \mathcal{R}(f; a_k) \tag{3-1}$$

注意公式编号的引用应含有圆括号，可以使用 `\eqref` 命令。

晶体衍射基础的著名公式 —— 布拉格方程：

$$2d \sin\theta = k\lambda, \quad k = 1, 2, 3 \cdots \tag{3-2}$$

式中 d —— 晶面间距（nm）；

θ —— 入射线与晶面的夹角（rad）；

λ —— X 射线波长（nm）。

k —— 公式中第一次出现的物理量代号应给予注释，注释的转行应与破折号“——”后第一个字对齐。

多行公式尽可能在“=”处对齐，推荐使用 align 环境。

$$a = b + c + d + e \tag{3-3}$$

$$= f + g \tag{3-4}$$

此外需要注意：公式需紧挨段前文字，不可空行，不然会导致公式独立成段，如下**错误**效果。公式前文字公式前文字公式前文字公式前文字公式前文字。

$$\frac{1}{2\pi \mathrm{i}} \int_{\gamma} f = \sum_{k=1}^{m} n(\gamma; a_k) \mathcal{R}(f; a_k) \tag{3-5}$$

公式后文字公式后文字公式后文字公式后文字公式后文字公式后文字公式后

文字公式后文字公式后文字公式后文字公式后文字。**正确**效果，如下：

$$\frac{1}{2\pi \mathrm{i}} \int_{\gamma} f = \sum_{k=1}^{m} n(\gamma; a_k) \mathcal{R}(f; a_k) \tag{3-6}$$

公式后文字公式后文字公式后文字公式后文字公式后文字公式后文字公式后。

3.3　数学定理

定理环境的格式可以使用 amsthm 或者 ntheorem 宏包配置。用户在导言区载入这两者之一后，模板会自动配置 thoerem、proof 等环境。

定理 3.1 (Lindeberg–Lévy 中心极限定理)：设随机变量 $X_1, X_2, \cdots, X_n$ 独立同分布，且具有期望 μ 和有限的方差 $\sigma^2 \neq 0$，记 $\bar{X}_n = \frac{1}{n}\sum_{i+1}^{n} X_i$，则

$$\lim_{n \to \infty} P\left(\frac{\sqrt{n}\left(\bar{X}_n - \mu\right)}{\sigma} \leqslant z \right) = \Phi(z), \tag{3-7}$$

其中 $\Phi(z)$ 是标准正态分布的分布函数。

证明：Trivial. ■

同时模板还提供了 assumption、definition、proposition、lemma、theorem、axiom、corollary、exercise、example、remar、problem、conjecture 这些相关的环境。

3.4　数学字体

按照《撰写规范》表达式字体可以采用 Times New Roman、Xits Math 或 Cambria Math（MS Word 默认字体）。Cambria Math 缺少部分样式，例如：积分符号设定为 upright 也看起来没有变化。

TeX Gyre Termes Math 字体（Times New Roman 的 TeX 克隆版）样例：

$$\frac{1}{2\pi \mathrm{i}} \int_{\gamma} f = \sum_{k=1}^{m} n(\gamma; a_k) \mathcal{R}(f; a_k) \tag{3-8}$$

Cambria Math 字体样例：

$$\frac{1}{2\pi \mathrm{i}} \int_{\gamma} f = \sum_{k=1}^{m} n(\gamma; a_k) \mathcal{R}(f; a_k) \tag{3-9}$$

Xits Math 字体样例：

$$\frac{1}{2\pi \mathrm{i}} \int_{\gamma} f = \sum_{k=1}^{m} n(\gamma; a_k) \mathcal{R}(f; a_k) \tag{3-10}$$

STIX Math 字体样例：

$$\frac{1}{2\pi \mathrm{i}} \int_{\gamma} f = \sum_{k=1}^{m} n(\gamma; a_k) \mathcal{R}(f; a_k) \tag{3-11}$$

第 4 章　引用文献的标注

模板支持 BibTeX 和 BibLaTeX 两种方式处理参考文献。下文主要介绍 BibTeX 配合 natbib 宏包的主要使用方法。

4.1　顺序编码制

在顺序编码制下，默认的 \cite 命令同 \citep 一样，即序号置于方括号中，引文页码会放在括号外。统一处引用的连续序号会自动用短横线连接。如多次引用同一文献，可能需要标注页码，例如：引用第二页[20]2，引用第五页[20]5。

```
\cite{zhangkun1994}                  ⇒ [20] 不带页码的上标引用
\citet{zhangkun1994}                 ⇒ 张昆 等[20]
\citep{zhangkun1994}                 ⇒ [20]
\cite[42]{zhangkun1994}              ⇒ [20]42 手动带页码的上标引用
\cite{zhangkun1994,zhukezhen1973}    ⇒ [20-21] 一次多篇文献的上标引用
```

4.2　著者-出版年制

著者-出版年制下的 \cite 跟 \citet 一样。

```
\cite{zhangkun1994}                  ⇒ 张昆 等 (1994)
\citet{zhangkun1994}                 ⇒ 张昆 等 (1994)
\citep{zhangkun1994}                 ⇒ (张昆 等, 1994)
\cite[42]{zhangkun1994}              ⇒ (张昆 等, 1994)42
\citep{zhangkun1994,zhukezhen1973}   ⇒ (张昆 等, 1994; 竺可桢, 1973)
```

4.2.1　其他引用注意事项

注意，引文参考文献的每条都要在正文中标注[20-53]。

引用测试：2 个连续引用[20-21]，2 个间隔[20,22]，3 个连续[20-22]。

如参考文献中需要使用上标或者下标，使用数学环境书写 `$\mathrm{Ba}_{3}\mathrm{CoSb}_{2}\mathrm{O}_{9}$`，例如该文献[54]。根据 gbt7714 规定著者姓名自动转为大写。西文的题名、期刊名的大小写不自动处理，需要自行处理以符合信息资源本身文种的习惯用法。

第 5 章　English and lower-case Example

If your supervisor is a foreign resident, or if your supervisor or defense committee specifically allows writing in English, the thesis may be written in English as the primary language. Please check with your supervisor or department secretary to confirm if you can write in English.

5.1　Reference guide

Writing in English still requires the Chinese reference standard GB/T 7714-2015.

结　论

学位论文的结论作为论文正文的最后一章单独排写，但不加章标题序号。

结论应是作者在学位论文研究过程中所取得的创新性成果的概要总结，不能与摘要混为一谈。博士学位论文结论应包括论文的主要结果、创新点、展望三部分，在结论中应概括论文的核心观点，明确、客观地指出本研究内容的创新性成果（含新见解、新观点、方法创新、技术创新、理论创新），并指出今后进一步在本研究方向进行研究工作的展望与设想。对所取得的创新性成果应注意从定性和定量两方面给出科学、准确的评价，分（1）、（2）、（3）…条列出，宜用“提出了”、“建立了”等词叙述。

在评价自己的研究工作成果时，要实事求是，除非有足够的证据表明自己的研究是“首次”、“领先”、“填补空白”的，否则应避免使用这些或类似词语

参考文献

[1] ACHIAM J, ADLER S, AGARWAL S, et al. GPT-4 Technical Report[A]. 2023.

[2] TOUVRON H, LAVRIL T, IZACARD G, et al. LLaMA: Open and Efficient Foundation Language Models[A]. 2023.

[3] Anthropic. Claude[EB/OL]. 2025. https://claude.ai.

[4] WU C J, RAGHAVENDRA R, GUPTA U, et al. Sustainable AI: Environmental Implications, Challenges and Opportunities[J]. Proceedings of Machine Learning and Systems (MLSys), 2022, 4: 795-813.

[5] HE Y, XU M, WU J, et al. UELLM: A Unified and Efficient Approach for Large Language Model Inference Serving[C/OL]//Proceedings of the 22nd International Conference on Service-Oriented Computing (ICSOC 2024). Berlin, Heidelberg: Springer-Verlag, 2024: 218-235. DOI: 10.1007/978-3-031-72312-0_14.

[6] KWON W, LI Z, ZHUANG S, et al. Efficient Memory Management for Large Language Model Serving with PagedAttention[C]//Proceedings of the 29th Symposium on Operating Systems Principles (SOSP). ACM, 2023: 611-626.

[7] ZHONG Y, LIU S, CHEN J, et al. DistServe: Disaggregating Prefill and Decoding for Goodput-optimized Large Language Model Serving[C]//Proceedings of the 18th USENIX Symposium on Operating Systems Design and Implementation (OSDI). 2024: 193-210.

[8] ZHENG L, YIN L, XIE Z, et al. SGLang: Efficient Execution of Structured Language Model Programs[C]//Advances in Neural Information Processing Systems (NeurIPS): Vol. 37. 2024: 62557-62583.

[9] JIN Y, WU C F, BROOKS D, et al. S^3: Increasing GPU Utilization during Generative Inference for Higher Throughput[C]//Advances in Neural Information Processing Systems (NeurIPS): Vol. 36. 2023: 18015-18027.

[10] ALI A, PINCIROLI R, YAN F, et al. BATCH: Machine Learning Inference Serving on Serverless Platforms with Adaptive Batching[C]//Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis (SC). IEEE, 2020: 1-15.

[11] WANG Y, CHEN K, TAN H, et al. Tabi: An Efficient Multi-Level Inference System for Large Language Models[C]//Proceedings of the Eighteenth European Conference on Computer Systems (EuroSys). 2023: 233-248.

[12] ZHANG C, YU M, WANG W, et al. MArk: Exploiting Cloud Services for Cost-Effective, SLO-Aware Machine Learning Inference Serving[C]//Proceedings of the USENIX Annual Technical Conference (ATC). 2019: 1049-1062.

[13] WANG L, YANG L, YU Y, et al. Morphling: Fast, Near-Optimal Auto-Configuration for Cloud-Native Model Serving[C]//Proceedings of the ACM Symposium on Cloud Computing (SoCC). 2021: 639-653.

[14] PATEL P, CHOUKSE E, ZHANG C, et al. Splitwise: Efficient Generative LLM Inference Using Phase Splitting[C]//Proceedings of the 51st Annual International Symposium on Computer Architecture (ISCA). ACM, 2024: 118-132.

[15] QIN R, LI Z, HE W, et al. Mooncake: A KVCache-centric Disaggregated Architecture for LLM Serving[C]//2024.

[16] HU C, HUANG H, HU J, et al. MemServe: Context Caching for Disaggregated LLM Serving with Elastic Memory Pool[A]. 2024.

[17] SUN B, HUANG Z, ZHAO H, et al. Llumnix: Dynamic Scheduling for Large Language Model Serving[C]//Proceedings of the 18th USENIX Symposium on Operating Systems Design and Implementation (OSDI). 2024: 173-191.

[18] MIAO X, SHI C, DUAN J, et al. SpotServe: Serving Generative Large Language Models on Preemptible Instances[C]//Proceedings of the 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS). 2024: 1112-1127.

[19] CHOI S, LEE S, KIM Y, et al. Serving Heterogeneous Machine Learning Models on Multi-GPU Servers with Spatio-Temporal Sharing[C]//Proceedings of the USENIX Annual Technical Conference (ATC). 2022: 199-216.

[20] 张昆, 冯立群, 余昌钰, 等. 机器人柔性手腕的球面齿轮设计研究[J]. 清华大学学报: 自然科学版, 1994, 34(2): 1-7.

[21] 竺可桢. 物理学论[M]. 北京: 科学出版社, 1973: 56-60.

[22] DUPONT B. Bone marrow transplantation in severe combined immunodeficiency with an unrelated MLC compatible donor[C]//WHITE H J, SMITH R. Proceedings of the third annual meeting of the International Society for Experimental Hematology. Houston: International Society for Experimental Hematology, 1974: 44-46.

[23] 郑开青. 通讯系统模拟及软件[D]. 北京: 清华大学无线电系, 1987.

[24] 姜锡洲. 一种温热外敷药制备方案: 中国, 88105607.3[P]. 1980-07-26.

[25] 中华人民共和国国家技术监督局. GB3100-3102. 中华人民共和国国家标准-量与单位[S]. 北京: 中国标准出版社, 1994.

[26] MERKT F, MACKENZIE S R, SOFTLEY T P. Rotational Autoionization Dynamics in High Rydberg States of Nitrogen[J]. J Chem Phys, 1995, 103: 4509-4518.

[27] MELLINGER A, VIDAL C R, JUNGEN C. Laser reduced fluorescence study of the carbon monoxide nd triplet Rydberg series - Experimental results and multichannel quantum defect analysis[J]. J Chem Phys, 1996, 104: 8913-8921.

[28] BIXON M, JORTNER J. The dynamics of predissociating high Rydberg states of NO[J]. J Chem Phys, 1996, 105: 1363-1382.

[29] 马辉, 李俭, 刘耀明, 等. 利用 REMPI 方法测量 BaF 高里德堡系列光谱[J]. 化学物理学报, 1995, 8: 308-311.

[30] CARLSON N W, TAYLOR A J, JONES K M, et al. Two-step polarization-labeling spectroscopy of excited states of Na2[J]. Phys Rev A, 1981, 24: 822-834.

[31] TAYLOR A J, JONES K M, SCHAWLOW A L. Scanning pulsed-polarization spectrometer applied to Na2[J]. J Opt Soc Am, 1983, 73: 994-998.

[32] TAYLOR A J, JONES K M, SCHAWLOW A L. A study of the excited 1Σg+ states in Na2[J]. Opt Commun, 1981, 39: 47-50.

[33] SHIMIZU K, SHIMIZU F. Laser induced fluorescence spectra of the a 3Πu–X 1Σg+ band of Na2 by molecular beam[J]. J Chem Phys, 1983, 78: 1126-1131.

[34] ATKINSON J B, BECKER J, DEMTRÖDER W. Experimental observation of the a 3Πu state of Na2[J]. Chem Phys Lett, 1982, 87: 92-97.

[35] KUSCH P, HESSEL M M. Perturbations in the A 1Σu+ state of Na2[J]. J Chem Phys, 1975, 63: 4087-4088.

[36] 广西壮族自治区林业厅. 广西自然保护区[M]. 北京: 中国林业出版社, 1993.

[37] 霍斯尼. 谷物科学与工艺学原理[M]. 李庆龙, 译. 2 版. 北京: 中国食品出版社, 1989: 15-20.

[38] 王夫之. 宋论[M]. 刻本. 金陵: 曾氏, 1865（清同治四年）.

[39] 赵耀东. 新时代的工业工程师[M/OL]. 台北: 天下文化出版社, 1998[1998-09-26]. http://www.ie.nthu.edu.tw/info/ie.newie.htm.

[40] 全国信息与文献工作标准化技术委员会出版物格式分委员会. GB/T 12450-2001 图书书名页[S]. 北京: 中国标准出版社, 2002.

[41] 全国出版专业职业资格考试办公室. 全国出版专业职业资格考试辅导教材: 出版专业理论与实务 • 中级[M]. 2014 版. 上海: 上海辞书出版社, 2004: 299-307.

[42] World Health Organization. Factors Regulating the Immune Response: Report of WHO Scientific Group[R]. Geneva: WHO, 1970.

[43] PEEBLES P Z, Jr. Probability, Random Variables, and Random Signal Principles[M]. 4th ed. New York: McGraw Hill, 2001.

[44] 白书农. 植物开花研究[M]//李承森. 植物科学进展. 北京: 高等教育出版社, 1998: 146-163.

[45] WEINSTEIN L, SWERTZ M N. Pathogenic Properties of Invading Microorganism[M]// SODEMAN W A, Jr, SODEMAN W A. Pathologic physiology: mechanisms of disease. Philadelphia: Saunders, 1974: 745-772.

[46] 韩吉人. 论职工教育的特点[C]//中国职工教育研究会. 职工教育研究论文集. 北京: 人民教育出版社, 1985: 90-99.

[47] 中国地质学会. 地质评论[J]. 1936, 1(1)–. 北京: 地质出版社, 1936–.

[48] 中国图书馆学会. 图书馆学通讯[J]. 1957(1)–1990(4). 北京: 北京图书馆, 1957–1990.

[49] American Association for the Advancement of Science. Science[J]. 1883, 1(1)–. Washington, D.C.: American Association for the Advancement of Science, 1883–.

[50] 傅刚, 赵承, 李佳路. 大风沙过后的思考[N/OL]. 北京青年报, 2000-04-12(14)[2002-03-06]. http://www.bjyouth.com.cn/Bqb/20000412/B/4216%5ED0412B1401.htm.

[51] 萧钰. 出版业信息化迈入快车道[EB/OL]. (2001-12-19)[2002-04-15]. http://www.creader.com/news/20011219/200112190019.htm.

[52] Online Computer Library Center, Inc. About OCLC: History of Cooperation[EB/OL]. 2000 [2000-01-08]. http://www.oclc.org/about/cooperation.en.htm.

[53] Scitor Corporation. Project scheduler[CP/DK]. Sunnyvale, Calif.: Scitor Corporation, 1983.

[54] KAMIYA Y, GE L, HONG T, et al. The nature of spin excitations in the one-third magnetization plateau phase of $Ba_3CoSb_2O_9$[J]. Nature communications, 2018, 9(1): 1-11.

附录 A　补充内容

附录是与论文内容密切相关、但编入正文又影响整篇论文编排的条理和逻辑性的资料，例如某些重要的数据表格、计算程序、统计表等，是论文主体的补充内容，可根据需要设置。

A.1　图表示例

A.1.1　图

附录中的图片示例（图 A-1）。

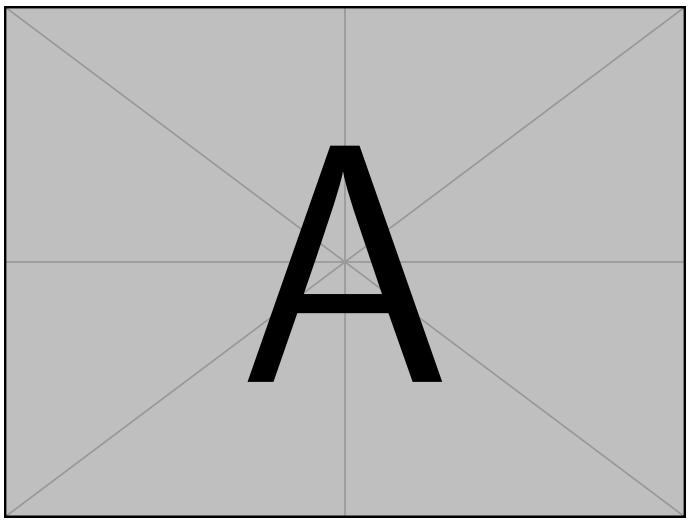


图 A-1　附录中的图片示例

A.1.2　表格

附录中的表格示例（表 A-1）。

A.2　数学公式

附录中的数学公式示例（公式(A-1)）。

$$\frac{1}{2\pi \mathrm{i}} \int_{\gamma} f = \sum_{k=1}^{m} n(\gamma; a_k) \mathcal{R}(f; a_k) \tag{A-1}$$

表 A-1　附录中的表格示例

文件名	描述
sustechthesis.dtx	模板的源文件，包括文档和注释
sustechthesis.cls	模板文件
thuthesis-*.bst	BibTeX 参考文献表样式文件
thuthesis-*.bbx	BibLaTeX 参考文献表样式文件
thuthesis-*.cbx	BibLaTeX 引用样式文件

A.3　源代码

附录中的代码示例：代码A-1。

```
1 class HelloWorldApp {
2     public static void main(String[] args) {
3         System.out.println("Hello World!"); // Display the
              string.
4         for (int i = 0; i < 100; ++i) {
5             System.out.println(i);
6         }
7     }
8 }
```

代码 A-1　Java 代码示例（使用 listings 高亮）

A.4　伪代码

附录中的伪代码示例（算法A-1）。

算法 A-1　Simulation-optimization heuristic

Data: current period t, initial inventory I_{t-1}, initial capital B_{t-1}, demand samples
Result: Optimal order quantity Q_t^*

$r \leftarrow t$;
$\Delta B^* \leftarrow -\infty$;
while $\Delta B \leqslant \Delta B^*$ *and* $r \leqslant T$ **do**
    $Q \leftarrow \arg\max_{Q \geqslant 0} \Delta B_{t,r}^{Q}(I_{t-1}, B_{t-1})$;
    $\Delta B \leftarrow \Delta B_{t,r}^{Q}(I_{t-1}, B_{t-1})/(r - t + 1)$;
    if $\Delta B \geqslant \Delta B^*$ **then**
        $Q^* \leftarrow Q$;
        $\Delta B^* \leftarrow \Delta B$;
    end
    $r \leftarrow r + 1$;
end

致　谢

衷心感谢导师 ××× 教授对本人的精心指导。他的言传身教将使我终生受益。

感谢 ××× 教授，以及实验室全体老师和同窗们的热情帮助和支持！

本课题承蒙 ×××× 基金资助，特此致谢。

以下内容为提示，仔细阅读后删除。

致谢应另起页，放置在参考文献、附录之后，标题和页眉均为“致谢”。语言要诚恳、恰当、简短。

致谢对象可以包括指导教师，在研究工作中提出建议和提供帮助的人，给予转载和引用权的资料、图片、文献、研究和调查的所有者，其他应感谢的组织和个人，资助研究工作的项目基金、奖学金基金、合同单位、资助或支持的企业、组织或个人，协助完成研究工作和提供便利条件的组织或个人。致谢字数以不超过一页纸为宜。

学位论文应由学生在导师（组）的指导下独立完成；**若涉及团队工作，应注明属于团队成果，并明确个人独立完成的内容**，科学严谨，恪守规范。

个人简历、在学期间完成的相关学术成果

个人简历

×××× 年 ×× 月 ×× 日出生于 ××××。

×××× 年 ×× 月考入 ×× 大学 ×× 院（系）×× 专业，×××× 年 ×× 月本科毕业并获得 ×× 学学士学位。

×××× 年 ×× 月——×××× 年 ×× 月，在 ×× 大学 ×× 院（系）×× 学科学习并攻读（获得）×× 学硕士学位。**【注：博士生已获得硕士学位写“获得”，硕士生申请硕士学位应写“攻读”，本括号在使用时请删除】**

获奖情况：如获三好学生、优秀团干部、× 奖学金等（不含科研学术获奖）。

工作经历：……

在学期间完成的相关学术成果

特别注意，下面的引用文献部分需要使用半角括号，例如 [J]，(已被 xxxx 录用)。(本行在使用时请删除)。

学术论文

[1] Pei S, Huang L L, Li G, et al. Magnetic Raman continuum in single-crystalline $H_3LiIr_2O_6$[J]. Physical Review B, 2020, 101(20): 201101. (SCI 收录, IDS 号为 LJ4UN, IF=3. 575, 对应学位论文 2.2 节和第 5 章.)

[2] Pei S, Tang J, Liu C, et al. Orbital-fluctuation freezing and magnetic-nonmagnetic phase transition in $\alpha-TiBr_3$[J]. Applied Physics Letters, 2020, 117(13): 133103. (SCI 收录, IDS 号为 NY3GK, IF=3. 597, 对应学位论文 2.2 节和第 3 章.)

申请及已获得的专利（无专利时此项不必列出）

[3] 任天令, 杨轶, 朱一平, 等. 硅基铁电微声学传感器畴极化区域控制和电极连接的方法: 中国, CN1602118A[P]. 2005-03-30.

[4] Ren T L, Yang Y, Zhu Y P, et al. Piezoelectric micro acoustic sensor based on ferroelectric materials: USA, No.11/215, 102[P]. (美国发明专利申请号.)

参与的科研项目及获奖情况（无获奖时此项不必列出）

[5] 姜锡洲，××××× 研究，×× 省自然科学基金项目。课题编号：××××，长长长。

[6] ×××，××××× 研究，×× 省自然科学基金项目。课题编号：××××。

[7] ×××，××××× 研究，×× 省自然科学基金项目。课题编号：××××。