

跨平台智能算法模型转化工具
采购项目

报 价 书

供应商： 西南交通大学

2025 年 9 月 17 日

目 录

一、 报 价 函	1
二、 授权委托书	2
三、 报价一览表	3
四、 报价方案	4
4.1 技术响应偏离表	4
4.2 技术方案	5
4.2.1 系统设计	5
4.2.2 架构设计	8
4.2.3 基础功能——模型管理	21
4.2.4 功能指标 1: 可支持模型的跨平台、跨底层架构的格式转化	24
4.2.5 功能指标 2: 可根据模型格式进行模型可视化	29
4.2.6 跨平台推理测试模块	33
4.2.7 任务监控模块	36
4.3 运行环境	38
4.4 售后服务	39
4.5 进度安排	41
4.6 交付成果清单	42
4.7 项目团队	42
4.8 配套能力	48
五、 资质文件	49
5.1 事业单位法人证书副本	49
5.2 GJB 质量管理体系证书	50
5.3 2022 年至 2024 年资产负债表	51
5.4 2022 年至 2024 年收入支出表	53
5.5 三年内承担过同类或类似项目	55
5.6 近一年中任意一个月依法纳税的凭证	59
5.7 近一年中任意一个月缴纳社会保障资金的凭证	60
5.8 “信用中国”信息查询截图	61
5.9 不存在控股、管理等关联关系的承诺书	62

5.10 不出具第三方审计报告情况说明	63
---------------------------	----

一、 报 价 函

中国船舶集团有限公司第七一六研究所:

1. 我方已仔细研究了 跨平台智能算法模型转化工具 询价文件的全部内容, 愿意以人民币 (大写) 伍拾陆万元 (¥ 560000) 的报价 (其中不含税价为: ¥543689; 增值税税额为: ¥16311) 完成服务, 并按照合同预定履行义务。

2. 我方的报价书包括下列内容:

- (1) 报价函;
- (2) 授权委托书;
- (3) 报价一览表;
- (4) 报价方案;
- (5) 资质文件。

报价书的上述组成部分如存在内容不一致的, 以报价函为准。

3. 我方承诺除商务和技术偏差表列出的偏差外, 我方响应询价文件的全部要求。

4. 我方承诺在询价文件规定的报价书有效期内不撤销报价书。

5. 如我方成交, 我方承诺:

- (1) 响应询价文件的全部要求
- (2) 在收到成交通知书后, 在成交通知书规定的期限内与你方签订合同;
- (3) 在签订合同时不向你方提出附加条件;
- (4) 按照询价文件要求提交履约保证金;
- (5) 在合同约定的期限内完成合同规定的全部义务。

6. 我方在此声明, 所递交的报价书及有关资料内容完整、真实和准确, 且不存在第一章“询价采购公告/询价采购邀请书”中规定的供应商不得存在的情形。

7. 无 (其他补充说明)。

供 应 商: 西南交通大学 (盖单位章)
法定代表人 (单位负责人) 或其委托代理人: (签字)
地 址: 四川省成都市郫都区犀安路 999 号
网 址: https://www.swjtu.edu.cn/
电 话: 028-66366272
传 真: 028-66366272
邮政编码: 611756

_____年_____月_____日

二、 授权委托书

(适用于有委托代理人的情况)

本人闫学东系西南交通大学的法定代表人(单位负责人),现委托李敏为我方代理人。代理人根据授权,以我方名义签署、澄清确认、递交、撤回、修改跨平台智能算法模型转化工具询价采购项目报价书、签订合同和处理有关事宜,其法律后果由我方承担。

委托期限:自本委托书签署之日起至跨平台智能算法模型转化工具询价采购项目签订采购合同之日止。

代理人无转委托权。

附:法定代表人(单位负责人)身份证复印件及委托代理人身份证复印件



供 应 商: 西南交通大学
法定代表人(单位负责人): _____ (签字)
身份证号码: 110105197509062533
委托代理人: _____ (签字)
身份证号码: 320322199009016544

_____年_____月_____日

三、 报价一览表

经测算明细费用如下表:

经费项目		合计	比例	备注
计 价 成 本	材料费	2.8000	5.00%	材料采购和设备租赁费等
	专用费	0.0000	0.00%	
	外协费	2.8000	5.00%	含测试费
	燃料动力费	0.0000	0.00%	
	事务费	2.8000	5.00%	含出差和评审费等
	固定资产折旧费	0.0000	0.00%	
	工资及劳务费	40.3200	72.00%	
管理费/税费		7.2800	13.00%	
收益		0.0000	0.00%	
其他支出		0.0000	0.00%	
合计		56.0000	100.00%	

四、 报价方案

4.1 技术响应偏离表

序号	条目	招标文件要求	投标文件响应	偏离情况
1	功能指标	可支持模型的跨平台、跨底层架构的格式转化	支持模型的跨平台、跨底层架构的格式转化	无偏离
2		可根据模型格式进行模型可视化	具备模型可视化功能，可根据模型格式选择	无偏离
3	性能指标	可支持不少于 20 类主流目标检测、目标分类、目标跟踪、自然语言大模型等智能体模型的网络结构可视化	支持目标检测、目标分类、目标跟踪、自然语言大模型等 20 类主流智能体模型的网络结构可视化	无偏离
4		可支持不少于 20 类主流目标检测、目标分类、目标跟踪、自然语言大模型等智能体模型的格式转化	支持目标检测、目标分类、目标跟踪、自然语言大模型等 20 类主流智能体模型的格式转化	无偏离
5		支持不少于 4 种包括华为、瑞星微、英伟达、寒武纪平台模型的转化	支持华为、瑞星微、英伟达、寒武纪平台模型的转化	无偏离
6	硬件运行环境	可适配英特尔、鲲鹏 CPU，英伟达、华为、寒武纪 GPU 进行部署	适配英特尔、鲲鹏 CPU，英伟达、华为、寒武纪 GPU 进行部署	无偏离
7	软件运行环境	操作系统支持 Linux（Ubuntu20.04 及以上）、麒麟 v10 或 Windows Server 2019 及以上	支持 Linux（Ubuntu20.04 及以上）、麒麟 v10 或 Windows Server 2019 及以上操作系统	无偏离

4.2 技术方案

4.2.1 系统设计

4.2.1.1 B/S 架构基础支撑系统

B/S 架构基础支撑系统是整个工具的运行骨架，从用户交互到数据存储再到硬件调度，实现全流程协同。

用户层：支持 Chrome、Edge、Firefox 等主流浏览器访问，无需额外安装客户端插件。界面采用响应式设计，适配 1920×1080 等常见分辨率，满足试验场调试机不同显示设备需求。核心交互功能包括模型上传（大文件尽量本地工具上传）、转换配置（可视化选择目标平台、精度模式（FP32/FP16/INT8））、任务进度查看（实时进度条+百分比显示，支持任务新增查询和删除）、结果下载（转换后模型文件下载），确保用户操作便捷高效。

应用服务层：采用微服务架构拆分功能模块，各模块通过 RESTful API 实现通信，支持独立部署与弹性扩展。前端应用基于 Vue3+TypeScript 开发，搭配 Element Plus 组件库构建轻量化交互界面，集成 ECharts 实现任务状态、性能指标等数据可视化展示；后端 RESTful API 服务基于常见开源框架开发，采用 Swagger 生成接口文档，方便后期维护与二次开发。模块拆分遵循“高内聚、低耦合”原则，将模型解析、格式转换、可视化处理等核心功能拆分为独立服务。

数据存储层：采用混合存储架构满足不同数据类型需求。MySQL 8.0 数据库用于存储结构化数据，包括用户信息（账号、权限、操作日志）、任务数据（任务 ID、模型名称、目标平台、转换状态、创建时间）、系统配置（支持的模型类型、平台参数、精度阈值），保障数据安全性；MinIO 分布式对象存储用于存储非结构化数据，包括原始模型文件、转换后模型文件、测试数据集、日志文件，支持断点续传与版本控制，可根据需求扩展存储节点，满足试验场日益增长的模型数据存储需求；Redis 用于缓存高频访问数据，如用户登录令牌、任务实时状态、模型元数据（输入输出维度、算子数量），缓存有效期可配置（默认 24 小时），大幅提升系统响应速度，减少数据库访问压力。

硬件资源层：部署英伟达（RTX 3060 或 Jetson AGX Orin）、华为昇腾（Atlas 某型号）、瑞芯微（RK3588 或 RK3599）、寒武纪（思元 370 或思元 590）等 4 类平台硬件测试节点，形成分布式测试集群。通过 Docker 容器化技术或者 Conda 虚拟化环境等方式统一管理各节点运行环境，为每类平台预置专用镜像或者执行环境（含驱动、SDK、依赖库），如华为昇腾节点镜像包含 CANN 7.0 SDK、MindSpore 2.0 框架，英伟达节点镜像包含 CUDA、TensorRT，确保环境一致性。

4.2.1.2 模型跨平台转换系统

模型跨平台转换系统是工具的核心功能模块，解决不同硬件平台间模型格式不兼容问题，实现“一次开发，多端部署”。

转换工具集：集成 ONNX Runtime 作为基础转换框架，搭配各平台专用转换工具形成完整工具链。华为昇腾平台采用 ATC (Ascend Tensor Compiler) 工具将 ONNX 模型转换为 OM 格式，支持算子融合、常量折叠等优化；瑞芯微平台采用 RKNN Toolkit 2 工具实现 ONNX 到 RKNN 格式转换，提供量化工具（支持 KL 散度、Min-Max 等量化算法）；英伟达平台采用 TensorRT 工具将 ONNX 模型转换为 TensorRT Engine 格式，支持动态形状推理与混合精度计算；寒武纪平台采用 MagicMind 工具完成 ONNX 到 CNModel 格式转换，优化算子调度策略提升推理性能。同时，集成 ONNX Simplifier 工具对输入模型进行简化（移除冗余算子、合并常量节点），减少转换复杂度，提升转换效率。

多模型适配模块：针对 20 类主流智能模型（例如 YOLO、Faster R-CNN、SSD、ResNet50/101、EfficientNet-B0/B4、MobileNetV3 等等）开发专用适配模板，每个模板包含模型结构解析规则、算子映射表、精度优化策略。例如，针对 YOLO 系列模型，模板中预设检测头结构解析逻辑，确保转换时保留 Anchor 信息；针对 BERT 类自然语言模型，优化 Transformer 层算子调用顺序，减少跨平台精度损失。

跨平台转换逻辑模块：采用“原生格式→ONNX→目标平台格式”的两步转换流程，ONNX 作为中间格式实现“解耦”，降低不同原生格式与目标平台间的适配复杂度。第一步，通过模型解析引擎读取 PyTorch (.pth/.pt)、TensorFlow (.pb/.saved_model)、TensorFlow Lite (.tflite) 等原生格式模型，提取模型拓扑结构、算子类型、输入输出维度、权重数据，转换为标准 ONNX 格式，转换过程中自动处理格式差异（如 TensorFlow 的 NHWC 格式转 ONNX 的 NCHW 格式）；第二步，根据目标平台选择对应转换工具，加载该平台专用适配模板，执行格式转换与优化，如华为昇腾平台转换时启用算子融合优化（将 Conv+BN+ReLU 合并为一个算子），瑞芯微平台转换时根据硬件算力选择合适量化精度（RK3588 支持 INT8 量化，RK3599 支持 FP16/INT8 混合量化）。转换过程中实时记录日志（包含每一步操作、耗时、错误信息），便于后期问题排查。

转换流程管控模块：构建任务队列机制实现批量任务有序处理，采用 RabbitMQ 作为消息队列，用户提交的转换任务先进入队列等待，调度器根据节点负载（CPU/内存/GPU 利用率）分配任务至空闲测试节点。实时监控转换进度，通过 Redis 记录任务状态（等待中、处理中、成功、失败），前端通过 WebSocket 或者轮询等方式实时获取状态更新，无需手动刷新页面。转换完成后自动进行基础校验，对比转换前后模型的输入输出维度、权重数据哈希值，若校验不通过则标记任务失败并生成错误报告（含失败原因、解决方案建议）；校验通过则将转换后模型文件存储至 MinIO，任务状态更新为“成功”并通知用户。

4.2.1.3 模型可视化展示系统

模型可视化展示系统帮助用户直观理解模型结构与参数特征,降低智能模型“黑箱”属性,辅助模型调试与优化。

网络结构可视化模块: 基于 Netron 开源框架进行二次开发, 增强对复杂模型(尤其是自然语言大模型)的解析与展示能力。支持自动提取模型网络层信息(输入层、卷积层、池化层、全连接层、Transformer 层等)、节点关系(层间数据流向)、算子参数(卷积核大小、步长、填充方式、激活函数类型), 以拓扑图形式展示, 不同类型层采用不同颜色标识(如卷积层、全连接层、Transformer 层颜色设定不同), 便于用户快速区分。支持交互操作: 鼠标滚轮缩放拓扑图、拖拽移动整个拓扑图或单个节点、双击节点展开查看层内详细参数(如权重矩阵维度、偏置值范围)、框选多个节点查看层间数据依赖关系。针对大型模型, 提供分层展示功能, 可按功能模块(如编码器层、解码器层)折叠/展开, 避免界面混乱。

可视化配置适配模块: 针对不同类型模型(计算机视觉模型、自然语言模型)开发专用可视化模板, 满足差异化展示需求。计算机视觉模型模板重点展示特征提取流程, 如 YOLO 模型展示从输入层到 backbone(如 CSPDarknet)、neck(如 FPN)、head(如检测头)的完整结构, 标注各层输出特征图尺寸; 自然语言模型模板突出 Transformer 层细节, 展示多头注意力机制(每个注意力头的权重分布、Q/K/V 矩阵维度)、Feed Forward 网络结构(隐藏层神经元数量、激活函数)。

渲染优化模块: 针对大规模模型可视化时易出现的页面卡顿、加载缓慢问题, 采用多项优化技术。采用分层加载策略, 优先加载模型整体拓扑结构, 用户点击展开节点时再加载该节点的详细参数, 减少初始加载数据量; 利用 WebWorker 在后台线程处理模型解析与数据计算, 避免阻塞主线程导致界面无响应; 采用 Canvas 绘图替代 DOM 元素渲染, 提升图形绘制效率, 支持每秒 60 帧稳定渲染; 对权重数据等大规模数值信息采用抽样展示(默认抽样 10%数据), 用户可点击“查看全部”获取完整数据, 平衡展示效果与性能。

4.2.1.4 硬件集成验证系统

硬件集成验证系统用于验证转换后模型在目标硬件平台的实际运行效果, 确保模型功能正常、性能达标、精度满足需求。

硬件测试集群: 搭建包含 4 类平台的硬件测试集群, 确保测试连续性。具体设备配置如下: 华为昇腾平台采用 Atlas 800 (Ascend 910 NPU, 32GB 显存)、Atlas 200I (Ascend 310 NPU, 4GB 显存); 瑞芯微平台采用 RK3588 开发板 (NPU 算力 6TOPS, 8GB 内存)、RK3599 开发板 (NPU 算力 10TOPS, 16GB 内存); 英伟达平台采用 RTX 4090 (24GB 显存)、Jetson AGX Orin (32GB 显存); 寒武纪平台采用思元 370 服务器 (2 颗思元 370NPU, 每颗 16GB 显存)、思元 590 服务

器（4 颗思元 590 NPU，每颗 32GB 显存）。所有设备接入试验场内网，通过远程控制工具（如 SSH、VNC）实现统一管理，支持远程部署模型、执行测试任务、采集测试数据。实际硬件环境根据建设方选型进行调整。

验证环境管理模块：开发统一的验证环境管理界面，支持查看各硬件节点的实时状态（在线/离线、CPU/内存/GPU/NPU 利用率、磁盘空间）、已部署模型列表、历史测试任务记录。提供环境配置手册，可依据参考远程为各节点安装/更新驱动、SDK、依赖库，如为华为昇腾节点更新 CANN SDK 版本，为英伟达节点安装最新 TensorRT 插件，配置过程自动生成日志，便于追溯。支持环境快照功能，可保存当前节点的环境状态（驱动版本、软件依赖），当后续环境出现问题时，可快速恢复至快照状态，减少环境配置时间。

多维度验证模块：从功能、性能、精度三个维度开展验证，全面评估模型适配效果。

功能验证：上传测试数据集（目标检测采用 COCO 2017 验证集、目标分类采用 ImageNet 验证集、自然语言模型采用 GLUE 数据集），执行模型推理，检查是否能正常输出结果（如目标检测输出 bounding box、类别、置信度；分类模型输出类别概率；NLP 模型输出文本预测结果），验证结果格式是否符合目标平台要求，无推理崩溃、输出为空等异常情况。

性能测试：针对不同输入数据量（如目标检测测试不同分辨率图像、NLP 模型测试不同长度文本）执行多轮推理，采集关键性能指标：推理延迟（单条数据从输入到输出的耗时，取 100 轮平均值）。例如，测试 YOLOv8 模型在华为 Atlas 800 上的推理性能，记录输入图像分辨率 640×640 时的延迟与吞吐量。

精度对比：以英伟达平台（例如 RTX3060，具体以实际研发环境为准）（FP32 精度）的推理结果为基准，计算转换后模型在其他平台的精度损失。目标检测模型采用 mAP（mean Average Precision）、Precision（精确率）、Recall（召回率）作为评价指标；目标分类模型采用 Top-1 Accuracy、Top-5 Accuracy；自然语言模型采用 GLUE Score（含 CoLA、SST-2 等子任务得分）。精度损失阈值可配置（默认 $\leq 5\%$ ），若超过阈值则分析原因（如量化精度过高、算子替换导致误差累积），并提供优化建议（如调整量化算法、更换适配模板）。

报告生成模块：验证完成后自动生成可视化验证报告，采用 HTML/PDF 两种格式存储，报告包含测试概况（测试模型、目标平台、测试时间、测试人员）、功能验证结果（是否通过、异常信息）、性能数据图表（推理延迟对比图、吞吐量趋势图、资源占用热力图）、精度对比表格（各指标基准值、测试值、损失率）、优化建议（针对性能不达标或精度损失超标的问题）。

4.2.2 架构设计

本系统采用分层架构设计，从用户交互到硬件执行，依次分为用户层、应用服务层、数据存储层、硬件资源层，各层职责清晰、协同运作，实现智能模型跨

平台适配与可视化全流程管理。



4.2.2.1 用户层：便捷交互入口

作为系统与用户的交互界面，聚焦模型管理全流程操作，提供直观、简洁的功能入口：

核心功能覆盖

模型上传下载：支持本地文件上传（支持拖拽上传、点击选择文件）、内网文件导入（从试验场文件服务器选择模型文件，通过路径映射直接读取）；支持转换后模型文件下载（可选择单个模型下载）、原始模型文件重新下载（保留上传记录，便于后续重新转换），下载速度取决于内网带宽。

可视化查看：提供模型网络结构拓扑图展示（支持缩放、拖拽、节点展开/折叠）、层参数详情查看（点击节点显示算子类型、输入输出维度、权重数据范围、计算量）、参数分布可视化（以直方图展示权重值分布，以热力图展示注意力权重（NLP 模型）），支持将可视化结果导出为 PNG、SVG 格式，用于文档编写或汇报展示。

模型转换：支持选择目标平台（华为昇腾、瑞芯微、英伟达、寒武纪）、精度模式（FP32（高精度）、FP16（平衡精度与性能）、INT8（高性能，适合边缘设备））、输入形状适配（固定形状、动态形状（设置最大/最小/步长）），提供转换参数预设功能（可保存常用配置，如“瑞芯微 RK3588-INT8”配置，后续直接调用），提交转换任务后实时显示进度（百分比+进度条），支持任务开始（任务配置和运行状态分开）、取消（取消后释放资源）。

推理测试：支持上传测试数据（图像文件（JPG/PNG/BMP）、文本文件（TXT）、批量测试数据集（ZIP 压缩包），使用 MinIO 存储系统），选择测试硬件节点，执行推理测试，实时展示推理结果（目标检测显示带 bounding box 的图像、分类显示类别与置信度、NLP 显示生成文本或分类结果），支持推理结果截图保存、推理日志查看（包含每一步推理操作、耗时）。

任务监控：展示所有任务列表（任务 ID、模型名称、目标平台、提交时间、状态（等待中/处理中/成功/失败）），支持按状态（全部/等待中/处理中/成功/失败）、时间（自定义时间段）、目标平台筛选任务，点击任务查看详情（转换参数、日志、结果链接），失败任务显示失败原因（如模型解析错误、算子不兼容）及解决方案建议。

4.2.2.2 应用服务层：业务逻辑核心

作为系统业务逻辑的核心承载层，应用服务层围绕模型处理全流程需求，按“高内聚、低耦合”原则拆分为前端服务、转换系统、可视化系统、支撑服务四大模块，各模块通过标准化接口协同工作，同时具备独立扩展与维护能力，可应对试验场调试机场景下模型处理的多样性与高可靠性需求。

前端服务：轻量化交互与数据可视化载体。前端服务聚焦用户操作体验与数据直观呈现，基于 Web 技术栈构建适配试验场调试机终端的交互界面，实现“操作-反馈-分析”的闭环体验。

交互式界面开发：采用 Vue3 框架结合 TypeScript 语言开发，确保代码可维护性与类型安全性；集成 Element Plus 组件库构建标准化界面元素（如按钮、表单、弹窗、表格），统一视觉风格；通过响应式布局适配调试机常见的 1920 × 1080 等分辨率，同时支持窗口缩放时界面元素自适应调整，避免出现内容截断或空白问题。针对模型上传、转换配置等核心操作，设计联动式配置（配置项按“目标平台→精度模式→输入形状”逻辑层级约束），降低操作复杂度；支持操作历史记录（保留近 30 天的模型上传、转换任务记录），用户可快速回溯历

史操作或重复发起相似任务。

可视化引擎集成：以 Netron 开源框架为核心二次开发，增强对 20 类目标模型（含自然语言大模型）的解析适配能力。针对计算机视觉模型（如 YOLO、ResNet），优化卷积层、池化层、检测头等结构的渲染逻辑，确保拓扑图中层与层的连接关系清晰；针对自然语言大模型（如 BERT、GPT），新增 Transformer 层专用渲染模板，单独展示多头注意力模块、Feed Forward 网络的内部结构，并用不同颜色标识 Q/K/V 矩阵流向。同时，开发引擎适配接口，支持将解析后的模型结构数据转换为前端可渲染的 JSON 格式，接口响应时间短（针对 1000 层以内模型），避免用户等待过长。

EChart 图表可视化：集成 ECharts 实现多维度数据展示，覆盖任务进度、性能指标、精度对比等核心场景。任务进度展示采用“环形进度条+实时日志流”组合形式，进度条实时更新转换/测试任务完成百分比，下方日志流按时间顺序显示任务关键节点（如“模型解析完成”“开始 ONNX 格式转换”“目标平台插件加载成功”），日志支持关键词搜索（如搜索“错误”快速定位问题）；性能指标展示采用多折线图对比不同平台的推理延迟（如华为昇腾、英伟达平台处理同一份数据的延迟曲线），图表支持鼠标悬浮查看具体数值、缩放时间范围、切换显示指标（如从“延迟”切换为“吞吐量”）；精度对比采用柱状图展示转换前后模型的 mAP、Top-1 Accuracy 等指标差异，超标数据（如精度损失 > 5%）自动标红提示，辅助用户快速识别异常。

转换系统：跨平台格式适配核心引擎。转换系统是实现“跨平台、跨底层架构格式转化”功能指标的核心，通过整合多平台工具链、开发适配模块，确保 20 类模型在 4 类目标平台上的高效、精准转换。

工具集工厂构建：整合 ONNX Runtime 作为基础转换框架，统一处理不同原生格式（PyTorch、TensorFlow、TensorFlow Lite）到 ONNX 中间格式的转换，支持 ONNX 多个版本兼容；针对 4 类目标平台分别集成专用转换工具，形成完整工具链：华为昇腾平台集成 ATC（Ascend Tensor Compiler）工具，支持算子融合、常量折叠、数据排布优化（NHWC 转 NCHW）；瑞芯微平台集成 RKNN Toolkit 2 工具，提供 KL 散度、Min-Max、ADMM 等多种量化算法，适配 RK3588、RK3599 等硬件的 NPU 算力特性；英伟达平台集成 TensorRT 工具，支持动态形状推理、混合精度计算（FP32/FP16/INT8），并优化 CUDA 内核调用逻辑；寒武纪平台集成 MagicMind 1.8 工具，通过算子自动映射与调度优化，提升 CNModel 格式模型的推理性能。同时，开发工具集管理模块，实现工具版本监控（如检测到 ATC 工具版本过低时自动提示更新）、工具异常重试（单次转换失败时自动重新调用工具，最多重试 3 次），保障转换工具链稳定性。

多模型适配模块开发：针对 20 类主流智能模型（例如目标检测：YOLO、Faster R-CNN、SSD 目标分类：ResNet、EfficientNet-B0/B4、MobileNetV3；目标跟踪：SORT、DeepSORT、ByteTrack；自然语言大模型：BERT-base、RoBERTa、GPT-2、T5-small，具体模型框架在项目实施过程中按照实际需求调整），开发专用适配

模板，每个模板包含模型结构特征库、算子映射规则、精度优化策略三部分。模型结构特征库存储该类模型的典型层结构（例如如 YOLO 的 CSPDarknet backbone、FPN neck、Detect head），用于快速识别模型类型并定位关键层；算子映射规则定义原生框架算子（如 PyTorch 的 nn.Conv2d）到目标平台算子（如华为昇腾的 Conv2D 算子）的对应关系，解决算子不兼容问题（如将 TensorFlow 的 Unique 算子替换为瑞芯微平台支持的 Sort+Deduplicate 组合算子）；精度优化策略针对不同模型类型配置参数，如自然语言大模型的 Transformer 层采用 FP16 精度以平衡精度与性能，目标检测模型的检测头采用 FP32 精度确保定位准确性。模块支持模板动态更新，用户可通过界面上传新模型的适配模板（JSON 格式），无需修改系统代码即可扩展模型支持范围。

跨平台转换逻辑实现：采用“原生格式→ONNX 中间格式→目标平台格式”的两步转换流程，降低不同原生格式与目标平台的适配复杂度。第一步，原生格式转 ONNX：通过模型解析引擎读取 PyTorch（.pth/.pt）、TensorFlow（.pb/.saved_model）等原生格式文件，提取模型拓扑结构、输入输出维度、权重数据，自动处理格式差异（如 TensorFlow 的 NHWC 数据排布转为 ONNX 标准的 NCHW 排布，PyTorch 的动态 BatchSize 转为固定 BatchSize 或标记为动态维度），生成标准 ONNX 模型；第二步，ONNX 转目标平台格式：根据用户选择的目标平台，加载对应平台的转换工具与适配模板，执行格式转换与优化。例如，转换为华为 OM 格式时，调用 ATC 工具并传入适配模板中的算子映射规则，完成 ONNX 模型到 OM 模型的编译，同时启用算子融合优化（将 Conv+BN+ReLU 合并为一个算子，减少计算量）；转换为瑞芯微 RKNN 格式时，调用 RKNN Toolkit 2 工具，根据模板中的量化策略执行 INT8 量化，并生成量化校准报告。转换过程中实时记录关键数据（如每一步耗时、算子转换数量、优化项），形成转换日志，便于后续问题排查。

可视化系统：模型结构与参数直观呈现载体。可视化系统聚焦解决模型“黑箱”问题，通过结构解析、模板管理、渲染优化，实现 20 类模型的清晰展示与交互，辅助试验场技术人员理解模型结构、调试转换问题。

结构解析模块开发：基于 Netron 核心解析能力扩展，开发模型结构深度解析功能，支持提取模型的层信息、节点关系、参数细节三类数据。层信息包含层名称、层类型（如 Conv2d、Linear、MultiHeadAttention）、输入输出张量形状（如 [1, 3, 640, 640]）、计算量（FLOPs）；节点关系包含层间数据流向（如输入层→卷积层 1→卷积层 2→池化层）、依赖关系（如某卷积层依赖前一层的输出作为输入）；参数细节包含卷积核大小、步长、填充方式、激活函数类型、权重数据范围（如权重值分布在 [-1.2, 1.5]）。针对自然语言大模型，额外解析 Transformer 层的注意力头数量、隐藏层维度、序列长度等关键参数；针对目标检测模型，解析 Anchor 尺寸、检测层数量、类别数等信息。解析模块支持批量处理，可同时解析多个模型文件并对比结构差异（如展示两个 YOLO 模型的层数量、算子类型差异），解析成功率≥90%（针对 20 类目标模型）。

模板管理模块设计: 为不同类型模型设计专用可视化模板, 实现差异化展示。计算机视觉模型模板采用“分层展开式”布局, 将模型按功能模块 (backbone、neck、head) 分组, 用户可单独展开某一模块查看内部层结构, 同时标注各模块的功能描述 (如“backbone: 提取图像特征”); 自然语言大模型模板采用“横向链式”布局, 按 Transformer 层顺序横向排列, 突出层间的序列依赖关系, 同时用悬浮窗展示注意力头的权重热力图; 目标跟踪模型模板新增“跟踪流程可视化”功能, 展示检测器输出结果如何通过卡尔曼滤波、匈牙利算法进行轨迹关联。模板支持用户自定义配置, 如调整层的显示颜色 (默认卷积层蓝色、全连接层绿色)、设置是否显示计算量 (FLOPs)、隐藏无关层 (如忽略 BN 层仅显示卷积层), 配置结果可保存为个性化模板, 供后续使用。

渲染优化模块实现: 针对大规模模型 (如包含上千层的 GPT-2、ResNet101) 可视化时易出现的页面卡顿、加载缓慢问题, 采用三项核心优化技术。一是分层加载: 优先加载模型整体拓扑结构 (仅显示层名称与连接关系), 用户点击某一层时再动态加载该层的详细参数 (如权重数据), 初始加载时间控制在 10s 以内; 二是异步渲染: 利用 WebWorker 在后台线程处理模型解析与图形绘制, 避免阻塞主线程导致界面无响应, 支持同时渲染多个模型窗口; 三是图形压缩: 对模型拓扑图采用矢量图形 (SVG) 格式存储, 减少文件大小, 同时支持图形缩放时不失真; 对大规模权重数据采用抽样展示 (默认显示 10% 的权重值, 用户可点击“查看全部”加载完整数据), 降低数据传输与渲染压力。优化后, 支持 1000 层以上模型的流畅可视化, 界面操作响应时间 $\leq 1000\text{ms}$ 。

支撑服务: 系统稳定运行基础保障。

支撑服务为前端服务、转换系统、可视化系统提供基础能力支撑, 涵盖用户管理、任务调度、接口管理、日志监控, 确保系统在试验场多用户、多任务场景下稳定运行。

注册登录与权限管理: 实现用户身份认证, 用户通过账号密码登录后, 系统生成有效期为 24 小时的 Token, 后续请求携带 Token 验证身份; 一般系统划分“管理员”“操作员”“查看员”三类角色 (但本系统主要是工具应用, 暂时不区分角色, 登录用户均具备所有权限), 对应三类角色权限如下: 管理员拥有系统配置 (如添加目标平台、更新转换工具)、用户管理 (新增/删除用户、分配角色) 权限; 操作员拥有模型上传、转换任务提交、推理测试权限; 查看员仅拥有任务结果查看、可视化展示权限。权限管理模块支持细粒度控制, 如限制某操作员仅能使用英伟达平台进行转换, 无法访问华为昇腾平台, 确保试验场数据安全与操作合规。

任务调度模块开发: 采用“消息队列+分布式调度”架构, 处理模型转换、推理测试等耗时任务。基于 RabbitMQ 构建任务队列, 用户提交的任务先进入队列等待, 调度中心 (基于 XXL-Job 开发) 根据任务优先级 (高/中/低)、硬件节点负载 (CPU/内存/GPU 利用率) 分配任务至空闲节点。任务优先级支持动态调整, 试验场紧急调试任务可设为高优先级, 优先占用资源; 硬件节点负载监控采

用 Prometheus 采集数据，当某节点 CPU 利用率 > 80% 或内存利用率 > 90% 时，自动将后续任务调度至其他节点，避免节点过载。同时，开发任务管理功能，支持任务暂停（暂停后保留当前进度，恢复时继续执行）、取消（取消后释放占用的硬件资源）、重试（失败任务可一键重试，无需重新配置参数），任务成功率 ≥ 98%（排除硬件故障、模型文件损坏等不可抗因素）。

网关代理与接口管理：实现 API 网关，统一管理所有后端接口，实现“入口统一、路由转发、权限校验”三大功能。入口统一：前端所有请求均通过网关转发至对应服务（如转换任务请求转发至转换系统，可视化请求转发至可视化系统），避免直接访问后端服务；路由转发：通过配置路由规则（如“/api/convert/**”转发至转换系统），支持动态调整路由（如服务扩容后新增节点时，无需修改前端请求地址）；权限校验：网关拦截所有请求，验证用户 Token 有效性与接口访问权限（如普通操作员无法访问管理员接口），拒绝非法请求。网关支持接口调用日志记录（包含请求参数、响应结果、耗时），便于后期接口性能分析与问题排查。

日志监控模块设计：构建全链路日志监控体系，覆盖用户操作、系统运行、任务执行三大场景。用户操作日志记录用户的登录、模型上传、转换配置等操作（包含操作人、操作时间、操作内容、IP 地址），用于安全审计与操作追溯；系统运行日志记录服务启动/停止、接口调用异常、硬件节点状态变化等信息（如“华为昇腾节点离线”“转换工具调用失败”），支持日志级别划分（DEBUG/INFO/WARN/ERROR），ERROR 级日志实时推送至管理员告警列表；任务执行日志记录转换、测试任务的每一步细节（如“模型解析开始时间：2025-09-10 10:00:00”“ONNX 转换完成，耗时 120s”“目标平台 OM 格式生成成功”），日志支持按任务 ID 查询、关键词搜索（如搜索“算子不兼容”定位转换失败原因）和导出。

4.2.2.3 数据存储层：数据流转中枢

作为系统数据全生命周期管理的核心载体，数据存储层采用“混合存储架构”，针对结构化数据、非结构化数据、高频访问数据的不同特性，分别选用 MySQL、MinIO、Redis 三类存储组件，实现数据分类存储、安全管控与高效调用，同时满足试验场调试机场景下模型文件大、任务数据多、访问频率高的存储需求。

关系数据存储：MySQL 结构化数据管理。MySQL 数据库主要用于存储系统运行过程中产生的结构化数据，通过“表分区+索引优化”保障数据查询效率，支撑用户管理、任务调度、系统配置等核心业务逻辑的有序运行。

核心数据表设计

用户信息表（sys_user）：存储用户账号、密码（采用加密算法并加盐存储，防止明文泄露）、角色类型（管理员/操作员/查看员，预留字段，暂不生效）、所属部门、联系方式、账号状态（启用/禁用）、创建时间、最后登录时间等字

段。新增“登录 IP 白名单”字段,仅允许试验场内网指定 IP 段(如 192.168.1.0/24)的设备登录,提升账号安全性。

任务信息表(task-info):记录模型转换、推理测试任务的核心数据,包括任务 ID(采用 UUID 生成唯一标识)、模型名称、源模型格式(PyTorch/TensorFlow/ONNX)、目标平台(华为昇腾/瑞芯微/英伟达/寒武纪)、精度模式(FP32/FP16/INT8)、任务状态(等待中/处理中/成功/失败)、提交人、提交时间、开始时间、完成时间、失败原因(任务失败时填写,如“算子不兼容”“硬件节点离线”)。针对任务状态查询场景,建立“状态+提交时间”联合索引,将任务列表查询耗时控制在 50ms 以内;支持按“目标平台”“任务状态”进行表分区(如按平台分为 4 个分区,按状态分为 4 个分区),提升大批量任务数据的查询与统计效率。

模型元数据表(model-meta):存储上传模型的基础信息,包括模型 ID、模型名称、文件大小(单位:MB)、源格式、输入维度(如[1,3,640,640])、输出维度(如[1,80,8400])、算子数量、模型类型(目标检测/分类/跟踪/NLP 大模型)、上传人、上传时间、存储路径(关联 MinIO 中的模型文件路径)。新增“模型 MD5”字段,用于检测重复上传(若 MD5 值已存在,提示用户“该模型已上传,是否直接使用”),避免存储空间浪费;针对模型类型筛选场景,建立“模型类型”索引,支持快速筛选出某类模型(如所有目标检测模型)。

系统配置表(sys-config):存储系统运行所需的配置参数,包括支持的模型格式列表、目标平台参数(如华为昇腾 ATC 工具路径、瑞芯微 RKNN Toolkit 版本)、任务优先级规则(如“紧急任务优先级高于普通任务”)、精度损失阈值(默认 $\leq 5\%$)、日志保留天数(默认 30 天)。采用“键-值”存储结构(key 字段为配置项名称,value 字段为配置值),支持通过接口动态修改配置(如管理员调整精度损失阈值),无需重启系统即可生效,提升系统灵活性。

数据安全与备份策略

定时备份:采用 xtrabackup 工具(或者类似工具)实现 MySQL 全量备份与增量备份,全量备份每日凌晨 2 点执行(此时系统负载最低),增量备份每 6 小时执行 1 次;备份文件存储至独立的存储服务器(与数据库服务器物理隔离),并保留最近 30 天的备份文件;定期(每周日)进行备份恢复测试,验证备份文件有效性,确保数据可恢复率 100%。

对象存储:MinIO 非结构化数据管理。MinIO 作为分布式对象存储组件,主要用于存储模型文件(原始模型、转换后模型)、测试数据集(图像/文本文件)、日志文件(任务执行日志、系统运行日志)等非结构化数据,通过“分布式部署+版本控制+访问控制”实现数据的安全存储与高效访问,满足试验场大文件存储与共享需求。

存储结构与桶规划

按数据类型创建 4 个核心存储桶(Bucket),每个桶配置独立的访问策略与生命周期规则:

raw-model 桶：存储用户上传的原始模型文件，桶内按“上传人/上传时间（年-月-日）/模型名称”路径存储（如“user1/2025-09-10/yolov8.pt”）。针对重复上传场景，通过“模型 MD5 值”校验（与 MySQL 模型元数据表关联），若 MD5 值已存在，直接复用已有文件，仅更新元数据，节省存储空间；设置生命周期规则：原始模型文件长期保留（考虑到工具系统特性），避免无效数据占用空间。

converted-model 桶：存储转换后的目标平台模型文件，桶内按“任务 ID/目标平台/精度模式”路径存储（如“task123/华为昇腾/FP16/model.om”）。每个转换后模型文件关联对应的任务 ID（通过文件名前缀标识），支持通过任务 ID 快速查询到对应的模型文件；设置访问权限：允许提交该任务的用户及管理员（当前所有用户都具备管理员权限）下载。

test-data 桶：存储用户上传的测试数据集（如 COCO 验证集、GLUE 数据集），桶内按“数据集类型/上传人”路径存储（如“目标检测数据集/user2/coco_val2017.zip”）。支持数据集共享功能（用户可将数据集设置为“公开”，供其他用户使用），共享数据集需经过管理员审核（防止恶意上传非测试数据）；针对大尺寸数据集（如超过 10GB），通过 MinIO 系统的配置优化，并形成配置手册文档，支持分片上传（将数据集拆分为 5MB/片），完成大文件高效管理。

log 桶：存储任务执行日志与系统运行日志，桶内按“日志类型/日期（年-月-日）”路径存储（如“任务日志/2025-09-10/task123.log”“系统日志/2025-09-10/sys.log”）。任务日志包含转换/测试过程中的每一步操作（如“2025-09-10 10:00:00 开始模型解析”“2025-09-10 10:05:00 目标平台转换完成”），系统日志包含服务启动/停止、接口调用异常、硬件节点状态变化等信息；设置生命周期规则：本系统作为内部工具使用，为了方便问题定位和后期优化，日志永久保存。

性能优化与访问控制

分布式部署：采用 4 节点 MinIO 集群（1 个纠删码组，3 个数据块+1 个校验块），支持数据冗余存储（任意 1 个节点故障，数据可通过其他节点恢复），同时提升并发访问性能（支持每秒 100+模型文件上传/下载）；针对试验场内网环境，配置 10Gbps 以太网连接，将大模型文件（如 20GB NLP 大模型）的上传/下载速度提升至 500MB/s 以上，缩短文件传输时间。特别注意：考虑到工具系统的硬件投入，实际上线采用单节点 MinIO 存储即可。

缓存加速：集成 Redis 作为 MinIO 的缓存层，将高频访问的模型信息（非模型文件本身）缓存至 Redis，用户再次访问时直接从缓存读取，减少 MinIO 集群的访问压力；设置缓存有效期（默认 24 小时），缓存过期后自动从 MinIO 重新加载数据，确保数据一致性。

访问控制：基于 MinIO 的 IAM（Identity and Access Management）功能，为不同角色用户分配不同的桶访问权限：管理员拥有所有桶的读写权限；操作员

拥有 raw-model 桶的上传/读取权限、converted-model 桶的读取权限（仅自己提交任务的模型）、test-data 桶的上传/读取/共享权限；查看员仅拥有 converted-model 桶的读取权限（仅已完成任务的模型）、test-data 桶的读取权限（仅公开数据集）。所有访问请求均通过 HTTPS 协议加密传输，防止数据在传输过程中被窃取或篡改。

缓存服务：Redis 高频数据加速。Redis 作为高性能缓存组件，主要用于存储高频访问数据（如任务实时状态、模型元数据、用户登录 Token），通过“内存存储+数据结构优化”将数据访问延迟控制在 100ms 以内，大幅提升系统响应速度，减轻 MySQL 与 MinIO 的访问压力，支撑试验场多用户并发操作场景。

核心缓存数据结构设计

任务状态缓存（Hash 类型）：采用“task: {taskId}”作为 Key，存储任务的实时状态（如“status”：“处理中”，“currentStep”：“ONNX 格式转换”）。应用服务层（如转换系统、任务调度模块）实时更新缓存中的任务状态（如每完成一步更新 progress 字段），前端通过 WebSocket 订阅任务状态变化（订阅 Key 为“task: {taskId}”），无需频繁查询 MySQL 即可获取最新进度，将任务状态更新延迟控制在 1000ms 以内；设置缓存有效期为 24 小时（任务完成后），超过有效期自动删除，释放内存空间。

模型元数据缓存（String 类型）：采用“model: {modelId}”作为 Key，存储模型元数据表中的核心字段（如模型名称、输入维度、目标类型、存储路径），格式为 JSON 字符串。当用户查看模型详情、发起转换任务时，优先从 Redis 读取元数据，若缓存未命中（如模型刚上传），再从 MySQL 查询并写入 Redis，将模型元数据查询耗时从 50ms 缩短至 5ms 以内；针对高频访问的热门模型（如 YOLO、ResNet），设置缓存有效期为 7 天，避免频繁查询 MySQL。

用户登录 Token 缓存（String 类型）：采用“token: {tokenValue}”作为 Key，存储用户 ID 与角色类型（如“userId:1001,role:operator”），缓存有效期与 Token 有效期一致（24 小时）。当用户发起请求时，网关代理模块从请求头中获取 Token，查询 Redis 验证 Token 有效性及用户角色，若 Token 不存在或已过期，拒绝请求；通过 Redis 缓存 Token，避免每次请求都查询 MySQL，将用户身份验证耗时从 100ms 缩短至数十 ms 以内，支撑每秒 1000+ 的并发请求验证。

硬件节点状态缓存（Set 类型）：采用“hardware:online”作为 Key，存储当前在线的硬件节点 ID（如“nvidia-node1”“huawei-node2”）；采用“hardware: {nodeId}”作为 Key（Hash 类型），存储单个节点的实时负载（如“cpuUsage: 30%”，“gpuUsage: 50%”，“memoryUsage: 40%”）。任务调度模块调度任务时，先从“hardware:online”集合中筛选在线节点，再查询各节点负载缓存，选择负载最低的节点分配任务，将节点筛选与负载查询耗时控制在 10ms 以内；硬件监控模块每 10 秒更新一次节点负载缓存，确保负载数据实时性。

缓存优化与一致性保障

数据结构优化：针对任务状态、模型元数据等结构化缓存数据，采用 Hash 类型存储（而非 String 类型的 JSON 字符串），支持单独更新某一字段（如仅更新任务进度 progress，无需重新写入整个 JSON），减少数据传输量与内存占用；针对硬件节点在线列表，采用 Set 类型存储，支持快速判断节点是否在线（通过 SISMEMBER 命令），查询耗时 $\leq 10\text{ms}$ 。

缓存穿透防护：针对不存在的模型 ID、任务 ID 查询请求（如恶意请求 “model: 9999”），采用“空值缓存”策略（若 MySQL 查询结果为空，在 Redis 中写入空值，有效期为 5 分钟），避免频繁查询 MySQL 导致数据库压力增大；同时，在网关层增加请求参数校验（如校验模型 ID 是否符合 UUID 格式），过滤无效请求。

缓存一致性保障：采用“更新数据库后更新缓存”的策略，当 MySQL 中的数据发生变更（如任务状态从“处理中”改为“成功”、模型元数据更新）时，先更新 MySQL，再同步更新 Redis 缓存（若缓存存在），确保缓存与数据库数据一致；针对并发更新场景（如多个线程同时更新同一任务状态），采用 Redis 分布式锁（基于 SET NX EX 命令），避免缓存数据覆盖错误，确保数据更新的原子性。

内存管理：采用 Redis 的内存淘汰策略（allkeys-lru），当内存使用率超过 90% 时，自动淘汰最近最少使用的缓存数据（如长期未访问的模型元数据、过期任务状态），确保 Redis 始终有足够内存存储高频数据；定期（每日凌晨）执行缓存碎片整理（通过 MEMORY PURGE 命令），减少内存碎片占用，提升 Redis 内存利用率。

4.2.2.4 硬件资源层：算力支撑和验证底座

硬件资源层作为系统算力输出与跨平台验证的核心载体，整合容器化环境与异构硬件集群，覆盖华为、瑞芯微、英伟达、寒武纪 4 类目标平台，通过标准化部署、精细化监控与高效调度，为模型转换、推理测试提供稳定算力支撑，同时确保转换后模型在真实硬件环境下的功能与性能验证效果，匹配试验场内场调试机的实际应用场景。

容器化环境：标准化运行底座搭建。基于 Docker 与 Conda 构建“硬件-环境-工具”三层隔离的标准化运行环境，解决不同平台工具链依赖冲突问题，实现环境快速部署与复用，保障模型转换、推理测试的一致性。

Docker 容器或者 Conda 环境部署：为 4 类目标平台分别构建专用 Docker 容器或者 Conda 环境镜像，内置转换工具链（如 TensorRT 8.6、ATC 7.0、RKNN Toolkit 2.5、MagicMind 1.8）及基础依赖库（Python3.9、PyTorch2.0、TensorFlow2.12）。镜像采用分层构建策略，将基础系统层、驱动层、工具层分开打包，后续更新工具版本时仅需重构工具层。

Conda 虚拟环境隔离：针对不同模型类型（如目标检测、NLP 大模型）创建专属 Anaconda 虚拟环境，实现 Python 依赖库的精细化隔离。例如，为自然语言

大模型创建“nlp-env”环境，安装 transformers、tokenizers 等专用库；为目标检测模型创建“cv-env”环境，安装 mmdetection、opencv-python 等库。通过 conda env export/import 功能实现环境配置快速迁移，当新增硬件节点时，可直接导入已配置好的环境文件，避免重复安装依赖，环境迁移时间 ≤ 30 分钟。同时，虚拟环境支持版本锁定（如指定 torch==2.0.1），防止依赖库版本更新导致工具链兼容性问题。

环境一致性校验：开发环境校验脚本，每次容器启动时自动检查硬件驱动版本（如 CUDA 版本是否匹配 TensorRT 要求）、工具链路径（如 ATC 工具是否在系统环境变量中）、依赖库完整性（如是否缺少 RKNN Toolkit 依赖的 libomp.so 库），若存在异常则生成校验报告并提示修复方案（如“CUDA 版本不匹配，需升级至 12.0”）。校验通过后，生成环境唯一标识（基于驱动+工具版本哈希计算），与模型转换任务关联存储，当后续出现转换失败时，可快速定位是否因环境版本差异导致，环境校验通过率 $\geq 99.5\%$ 。

异构硬件集群：4 类平台算力部署。围绕项目要求的 4 类目标平台，部署专用硬件设备形成分布式算力集群，每类平台配置“主测试节点+备用节点”，确保硬件故障时业务不中断，同时满足 20 类模型的跨平台转换与推理测试需求。

硬件网络部署：所有硬件节点接入试验场内网，采用 10Gbps 以太网交换机（华为 CloudEngine S5735-S24T4X-A）实现节点间高速通信，模型文件传输速度 $\geq 500\text{MB/s}$ （大模型传输时间 ≤ 40 秒/20GB）。为硬件节点分配固定 IP 地址（如华为昇腾节点段：192.168.1.100-120，瑞芯微节点段：192.168.1.121-140），通过 DHCP 静态绑定确保 IP 不漂移，同时配置防火墙规则，仅允许应用服务层服务器（IP：192.168.1.50）访问硬件节点的 SSH（22 端口）、推理服务（5000 端口），禁止外部网络直接访问，保障硬件安全。

硬件监控与调度：算力高效管理。开发硬件监控与调度系统，实时采集硬件状态，确保硬件集群高效、稳定运行，支撑多任务并发处理。

实时监控模块：基于 Prometheus+Grafana 构建硬件监控面板，每 10 秒采集一次硬件指标，覆盖 CPU 利用率、内存/显存占用率、磁盘空间、网络带宽、NPU/GPU 温度等核心参数。针对不同硬件类型定制监控指标：如英伟达 GPU 新增“Tensor Core 利用率”“显存带宽占用”；华为昇腾 NPU 新增“算子执行效率”“AICore 利用率”；瑞芯微 NPU 新增“NPU 负载百分比”“数据吞吐量”。监控面板支持按平台分组查看（如单独展示所有瑞芯微节点状态）、异常指标告警（如 GPU 温度超过 85℃ 时触发邮件告警，内存利用率超过 90% 时触发短信告警），告警响应时间 ≤ 1 分钟，同时支持历史数据查询（保留 30 天监控数据），可生成硬件利用率趋势图（如某节点一周内 CPU 平均利用率），辅助算力资源优化。

硬件验证适配：跨平台兼容性保障。针对 4 类平台的硬件特性，开发专用适配插件与验证脚本，确保 20 类模型在不同硬件上的转换兼容性与推理正确性，同时优化算子调用逻辑提升性能。

平台专用适配插件：为每类平台开发模型转换适配插件，解决硬件算子不兼

容、数据格式不匹配问题。华为昇腾插件：针对 Ascend NPU 不支持的 TensorFlow “Unique” 算子，开发 “Sort+Deduplicate” 算子替换逻辑；优化数据排布（将 ONNX 的 NCHW 格式转为昇腾高效的 NHWC 格式），提升算子执行效率。瑞芯微插件：针对 RK3588 NPU 对大 kernel 卷积（如 11×11 ）支持不足的问题，开发卷积拆分算法（将 11×11 拆分为 $3 \times 3 + 3 \times 3 + 5 \times 5$ ）；支持 INT8 量化校准数据自动生成（基于测试集抽样 100 张图像），减少量化精度损失。英伟达插件：利用 TensorRT 的 Plugin 机制，自定义 YOLO 检测头算子，将检测头的“卷积+激活+解码”流程融合为单个算子，推理速度提升 30%；支持动态形状推理（如输入图像尺寸从 640×640 动态调整为 320×320 ），适配不同场景需求。寒武纪插件：针对思元 NPU 的算子调度特性，优化 Transformer 层的注意力计算顺序（将多头注意力并行计算改为串行调度），减少内存占用；支持模型分片加载（将超过 NPU 显存的大模型拆分为多段，分段加载执行），实现 30GB 大模型在 16GB 显存节点上的推理。

硬件验证脚本库：针对 20 类模型开发标准化验证脚本，覆盖功能验证、性能测试、精度对比三类场景。功能验证脚本：加载转换后模型，输入标准测试数据（如 COCO 验证集图像、GLUE 文本数据），检查是否输出合法结果（如目标检测模型输出 bounding box 坐标在 0-1 范围内，分类模型输出概率和为 1），若输出异常则定位错误算子（如“Conv2d 算子输出 NaN”）并生成错误日志。性能测试脚本：执行 100 轮推理，采集单轮推理延迟（取平均值）、硬件资源占用峰值，生成性能报告（如“YOLO 在 RK3588 上 INT8 推理延迟 20ms”），同时对比不同精度模式（FP32/FP16/INT8）的性能差异。精度对比脚本：以英伟达 RTX 4090（FP32 精度）的推理结果为基准，计算其他平台的精度损失（目标检测用 mAP 差值，分类用 Top-1 Accuracy 差值，NLP 用 GLUE Score 差值），若损失超过预设阈值（默认 10%），则可以调整转换参数（如降低量化精度、更换算子映射规则）并重新验证，精度达标率 $\geq 90\%$ （针对 20 类目标模型）。

硬件特性优化：结合各平台硬件优势，针对性优化模型推理性能。华为昇腾：利用 Ascend NPU 的 AICore 并行计算能力，将模型的卷积、池化等算子拆解为并行任务，提升计算效率；开启昇腾的“异构计算架构”，实现 CPU 与 NPU 的数据流水线传输，减少数据等待时间。瑞芯微：针对 RK3588 的 Arm 架构特性，采用 NEON 指令集优化模型前处理（如图像 Resize、归一化），前处理速度提升 40%；关闭非必要的后台服务（如图形界面），释放 CPU 资源用于模型推理。英伟达：利用 TensorRT 的混合精度计算，在 FP16 精度下保留关键层（如检测头）的 FP32 计算，平衡性能与精度；开启 GPU 的“持久化模式”，避免每次推理重新初始化 GPU 上下文，减少启动延迟。寒武纪：利用思元 NPU 的“张量压缩”技术，对模型权重进行压缩存储（如 FP16 转 FP8），减少显存占用；优化 NPU 与内存的数据交互频率（如批量读取输入数据），降低数据传输开销。优化后，各平台模型推理性能较基础转换提升 20%-50%，满足试验场调试机的实时性需求。

4.2.3 基础功能——模型管理

模型管理作为系统核心基础功能，覆盖智能模型从“接入-维护-版本控制-存储-检索”的全生命周期管理，需兼容主流光模块及交换机（网络传输基础能力不是本项目重点），同时满足 20 类智能模型（目标检测/分类/跟踪/NLP 大模型）的高效管理需求，为后续跨平台转换、可视化展示提供标准化模型数据源。

4.2.3.1 功能描述：全流程模型管控能力

模型管理功能围绕试验场调试机场景下的模型使用需求，构建“接入-维护-版本-存储-检索”五大核心能力，确保模型数据的完整性、可追溯性与易用性。

模型接入：支持多渠道、多格式模型文件接入，适配试验场本地调试机与内网环境。本地接入支持通过浏览器拖拽上传、点击选择文件两种方式，兼容 PyTorch (.pth/.pt)、TensorFlow (.pb/.saved_model/.h5)、ONNX (.onnx)、TensorFlow Lite (.tflite) 等原生格式，单次上传支持多文件批量导入（最多 10 个文件），单个文件最大支持 2GB；内网接入支持从试验场文件服务器（如华为 OceanStor）通过路径映射直接读取模型文件，无需重复上传，适配 25G 组网环境下的高速数据传输（传输速率 $\geq 2\text{GB/s}$ ，1GB 模型文件读取时间 ≤ 0.5 秒）。上传过程中自动触发格式校验（如检测文件后缀、解析文件头部标识），若格式不支持则实时提示“当前文件格式为.xlsx，仅支持.pth/.pb/.onnx 等格式”，并列出支持的格式清单，降低操作失误率。

模型信息维护：提供精细化模型元数据编辑与补充功能，确保模型信息完整可追溯。上传后自动解析模型基础信息（如模型名称、文件大小、源框架、输入输出维度、算子数量），用户可补充业务属性信息：模型类型（目标检测/目标分类/目标跟踪/自然语言大模型）、训练数据集（如 COCO2017/ImageNet/GLUE）、精度指标（如 mAP 值、Top-1 准确率、BLEU 分数）、适用场景（如安防监控/智能质检/文本生成）、备注说明（如“2024 年 3 月训练，适用于小目标检测”）。支持信息批量编辑（选中多个同类型模型统一修改“适用场景”），编辑后自动同步至数据存储层的模型元数据表（model-meta），并生成修改日志（记录修改人、修改时间、修改前后内容），便于后续追溯信息变更历史。

模型版本控制：实现模型迭代过程的版本化管理，避免版本混乱。当同一模型（通过模型名称+MD5 值联合判定）再次上传时，系统自动识别为新版本，生成连续版本号（如 V1.0 \rightarrow V2.0），并关联历史版本（显示“基于 V1.0 更新”）。每个版本保留完整元数据（含该版本的训练数据、精度指标）与对应的模型文件，支持版本回溯（如从 V3.0 回退至 V2.0 进行转换测试）、版本对比（可视化展示不同版本的输入输出维度、算子数量、精度指标差异，如“V2.0 比 V1.0 算子数量减少 20%，mAP 提升 3%”）。版本生命周期支持手动标记（如“当前版本 V2.0 为生产版本，V1.0 为历史版本”），系统默认保留最近 10 个版本，超过数量的

历史版本可手动归档至冷存储（MinIO 归档桶），节省热存储资源。

模型存储管理：基于混合存储架构实现模型文件的安全存储与高效管理，适配多光模块兼容需求。模型物理文件存储于 MinIO 分布式对象存储的“raw-model”桶，按“用户 ID/上传日期（年-月-日）/模型名称-版本”路径结构化存储（如“user001/2024-05-20/yolov8-v1.0.pth”），支持文件权限控制（仅上传者与管理员可修改/删除，其他用户仅可查看）；模型元数据存储于 MySQL 的 model-meta 表，通过“模型 ID”与 MinIO 文件路径关联，确保元数据与物理文件一一对应。针对重复上传的模型（哈希值已存在），系统自动复用已有物理文件，仅新增版本元数据，避免存储空间浪费（经测试，重复上传率达 30% 时，可节省 25% 存储资源）。模型文件备份暂时不考虑，但基于 MinIO 的分布式存储，后续如果使用规模扩大，数据安全要求提高，可以扩展为冗余备份模式，水平扩展为 3 节点模式，保障数据安全性。

模型检索与筛选：提供多维度快速检索功能，提升模型查找效率。支持精准检索（输入模型名称/模型 ID 精确匹配）、模糊检索（输入关键词如“yolo”，匹配所有含“yolo”的模型）、高级筛选（按模型类型、源框架、上传时间、精度指标等组合条件筛选，如“目标检测+PyTorch 框架+2024 年 5 月后上传+mAP ≥ 0.85 ”）。检索结果以列表形式展示，包含模型名称、版本号、类型、精度指标、上传时间等关键信息，支持按“上传时间（降序/升序）”“精度指标（降序）”排序，列表支持分页（每页 10/20/50 条可选）。针对常用模型，支持“收藏”功能（点击星标标记），收藏模型在“我的收藏”列表置顶展示，减少重复检索操作，平均模型查找时间从 30 秒缩短至 5 秒以内。

4.2.3.2 技术实现方案：高兼容、高可靠的技术架构

围绕“多光模块兼容、25G 组网稳定、全生命周期管理”需求，采用分层技术架构实现模型管理功能，覆盖前端交互、后端服务、数据存储、兼容性适配四大层面，确保功能稳定与性能达标。

前端交互层：易用性与兼容性设计

基于 Vue3+Element Plus 构建响应式界面，适配试验场调试机的 1920×1080 分辨率，支持 Chrome/Edge/Firefox 主流浏览器（版本 ≥ 90.0 ），界面操作支持键盘快捷键（如 Ctrl+U 快速上传、Ctrl+F 快速检索），提升操作效率。

模型上传采用分片上传技术（基于 WebUploader 组件），将大文件（如 5GB NLP 模型）拆分为 5MB/片，通过 25G 组网环境实现并行传输，同时支持断点续传（网络中断后重新连接可从上次中断分片继续上传），上传成功率 $\geq 99.9\%$ 。上传进度实时展示（进度条+百分比+剩余时间），上传完成后弹出“上传成功”提示，并自动跳转至模型信息编辑页面。

模型检索功能采用“前端预筛选+后端精准查询”结合方式：前端缓存最近 30 天的检索历史与常用筛选条件，用户输入关键词时先在前端预匹配，若匹配

结果不足再发起后端查询；结果通过 WebSocket 或者前端轮询实时推送到前端，避免页面刷新，提升交互流畅度。

后端服务层：高可用的业务逻辑支撑

基于成熟框架构建微服务接口，采用 RESTful 风格设计模型管理 API（如 POST/api/model/upload 上传模型、GET/api/model/list 查询模型列表、PUT/api/model/info 更新模型信息），接口支持 HTTPS 加密传输（基于 TLS1.2 协议），保障数据传输安全。

模型解析服务集成 Netron 核心解析引擎与 ONNX Runtime 解析工具，实现多格式模型元数据提取：对于 PyTorch/TensorFlow 模型，通过加载模型权重文件解析输入输出维度、算子类型；对于 ONNX 模型，直接解析 ONNX ProtoBuf 结构获取模型信息，解析成功率 $\geq 99.5\%$ （覆盖 20 类目标模型）。解析过程采用异步处理（基于 RabbitMQ 消息队列），避免大模型解析阻塞主线程，解析耗时 ≤ 3 秒（1GB 模型）。

版本控制服务基于“模型唯一标识+版本号”机制实现：通过 MD5 算法计算模型文件哈希值，结合模型名称生成唯一标识（如“yolov8_7a3f2d”），同一唯一标识的模型视为同一基础模型，再次上传时自动递增版本号（如 V1.0 \rightarrow V2.0）。版本关联关系存储于 MySQL 的 model_version 表，通过“基础模型 ID”关联所有版本，支持高效的版本回溯与对比查询。

数据存储层：混合存储与安全保障

模型文件存储采用 MinIO 分布式对象存储，本项目部署单节点即可（可支持 3 节点集群 1 主 2 从扩展，支持数据冗余备份，每份文件保存 3 个副本），适配多光模块兼容需求。MinIO 桶策略配置为“私有访问”，仅通过后端服务的访问密钥（Access Key/Secret Key）授权访问，防止未授权下载。

模型元数据与版本信息存储于 MySQL 数据库，采用“分表分索引”优化性能：model-meta 表按“上传日期”分表（每月 1 张表），减少单表数据量；建立“模型名称+哈希值”联合唯一索引（防止重复基础模型）、“模型类型+上传时间”联合索引（优化筛选查询），查询耗时 $\leq 50\text{ms}$ 。

数据安全保障：模型文件上传后自动生成访问令牌（基于 UUID），令牌有效期 24 小时，用户下载时需验证令牌有效性；数据库中敏感信息（如模型训练数据集路径）采用 AES-256 加密存储，解密密钥通过配置中心动态管理，避免密钥硬编码；所有操作日志（上传/编辑/删除/下载）记录至 MySQL 的 model-operation-log 表，包含操作人、操作时间、IP 地址、操作内容，日志永久保留，便于安全审计。

兼容性适配层：25G 组网与多光模块兼容。应用系统不验证网络传输层适配性，由甲方保证网络传输层性能和稳定性。

性能优化与监控

采用 Redis 缓存高频访问数据：将最近 7 天的热门模型元数据（如访问量 Top20 的模型）、用户检索历史、常用筛选条件缓存至 Redis，缓存有效期 24

小时，缓存命中率 $\geq 90\%$ ，减少 MySQL 查询压力（MySQL 查询量降低 60%）。

部署 Prometheus+Grafana 监控模型管理服务性能：监控指标包括 API 接口响应时间（目标 $\leq 300\text{ms}$ ）、模型解析成功率（目标 $\geq 99.5\%$ ）、文件上传速率（目标 $\geq 2\text{GB/s}$ ）、数据库查询耗时（目标 $\leq 50\text{ms}$ ），设置阈值告警（如接口响应时间超过 500ms 时发送邮件告警），确保服务稳定运行。

4.2.4 功能指标 1：可支持模型的跨平台、跨底层架构的格式转化

作为系统核心功能指标，模型跨平台、跨底层架构格式转化需实现“多源模型兼容、多平台适配、高精度转换、全流程可控”，严格匹配“支持不少于 20 类主流智能体模型、覆盖华为/瑞芯微/英伟达/寒武纪 4 类平台”的性能要求，同时满足试验场内场调试机对转换效率、结果可靠性的实际应用需求，为后续模型部署测试提供标准化格式支撑。

4.2.4.1 功能描述，支持性能指标 2 和 3，覆盖全场景的转换能力

围绕“跨平台、跨架构”核心目标，构建从“转换配置-任务执行-结果校验-日志追溯”的全流程功能体系，确保 20 类模型（目标检测/分类/跟踪/NLP 大模型）在 4 类平台上的高效、精准转换，同时提供灵活的参数配置与可视化的任务管控能力。

多源模型格式兼容：全面支持 20 类主流智能体模型的原生格式导入，覆盖计算机视觉与自然语言处理两大领域。其中目标检测模型支持 YOLOv5/v7/v8 (.pt)、Faster R-CNN (.pth/.pb)、SSD (.h5)；目标分类模型支持 ResNet50/101 (.pth)、EfficientNet-B0/B4 (.h5)、MobileNetV3 (.tflite)；目标跟踪模型支持 SORT (.pt)、DeepSORT (.pb)、ByteTrack (.onnx)；自然语言大模型支持 BERT-base (.bin)、RoBERTa (.pt)、GPT-2 (.tf)、T5-small (.h5)。导入时自动识别模型类型（通过文件后缀+模型结构特征双重判定），如检测到 YOLOv8 的 .pt 文件时，自动标记为“目标检测模型”，并关联对应的转换模板，无需用户手动选择，降低操作复杂度。

4 类平台格式输出适配：针对华为、瑞芯微、英伟达、寒武纪 4 类平台的底层架构差异，输出对应原生格式，确保模型可直接部署。具体适配如下：华为昇腾平台输出 OM 格式（适配 Atlas 800/200I 设备，基于 CANN 架构编译）；瑞芯微平台输出 RKNN 格式（适配 RK3588/RK3599 开发板，支持 NPU 硬件加速）；英伟达平台输出 TensorRT Engine 格式（适配 RTX4090/Jetson AGX Orin，优化 CUDA 内核调用）；寒武纪平台输出 CNModel 格式（适配思元 370/590 服务器，匹配 Neuware 架构）。支持同一模型批量转换至多平台格式（如一次提交将 YOLOv8 模型同时转换为 OM、RKNN、TensorRT Engine 格式），批量任务数量上限为 10 个，满足试验场多平台并行测试需求。

灵活转换参数配置：提供精细化参数配置功能，用户可根据平台特性与应用场景调整转换策略，平衡精度与性能。核心配置项包括：

精度模式：支持 FP32（高精度，适用于医疗影像等对精度敏感场景）、FP16（平衡精度与性能，适用于多数计算机视觉场景）、INT8（高性能低功耗，适用于边缘端瑞芯微/RK3588 平台），选择 INT8 时自动启用量化校准（支持 KL 散度/Min-Max 两种校准算法，默认抽样 100 张测试集图像生成校准数据）。

输入形状适配：支持固定形状（如设置输入图像尺寸为 640×640 ）与动态形状（设置尺寸范围如 $320 \times 320 \sim 1280 \times 1280$ ，步长 16），动态形状适配自然语言大模型的可变序列长度（如 BERT 模型支持序列长度 $16 \sim 512$ ），避免模型因输入尺寸固定导致的场景限制。

算子优化策略：提供“性能优先”“精度优先”“兼容优先”三种策略，性能优先模式启用算子融合（如 Conv+BN+ReLU 合并）、常量折叠，提升推理速度；精度优先模式禁用激进量化、保留关键层 FP32 计算，降低精度损失；兼容优先模式替换稀有算子（如将 TensorFlow 的“Unique”算子替换为平台支持的组合算子），确保转换成功率。

转换任务全流程管控：支持转换任务的提交、监控、暂停、取消、重试全生命周期操作。用户提交任务时生成唯一任务 ID（UUID 格式），关联模型 ID、目标平台、参数配置等信息；任务列表实时展示任务状态（等待中/处理中/成功/失败）、进度百分比（如“60%（ONNX 格式转换中）”）、预计剩余时间；处理中任务支持暂停（暂停后保留当前进度，恢复时从断点继续）与取消（取消后释放硬件资源，删除临时文件）；失败任务显示失败原因摘要（如“算子 XXX 不支持”“硬件节点离线”），并提供修复建议（如“建议切换为兼容优先策略”），支持一键重试（自动沿用原配置参数，无需重新填写）。

转换结果自动校验：转换完成后触发多维度校验，确保输出模型的完整性与可用性，避免无效转换结果流入后续测试环节。校验内容包括：

格式完整性校验：检查输出文件是否符合目标平台格式规范（如 OM 文件是否包含完整的昇腾算子信息、RKNN 文件是否通过 RKNN Toolkit 的格式验证），若格式损坏则标记“校验失败”，并重新执行转换（最多重试 2 次）。

结构一致性校验：对比转换前后模型的输入输出维度（如原模型输入 $[1, 3, 640, 640]$ ，转换后应保持一致）、算子数量（允许 $\pm 5\%$ 偏差，因平台优化导致的算子融合/拆分），若偏差超过阈值则生成“结构异常报告”，定位差异算子。

基础功能校验：在目标平台测试节点加载转换后模型，执行 10 轮基础推理（输入标准测试数据），检查是否输出合法结果（如目标检测模型输出有效 bounding box、NLP 模型输出非空文本），推理失败则标记“功能异常”，并记录错误日志（如“显存溢出”“算子执行超时”）。

转换日志追溯：自动生成详细转换日志，记录任务全流程关键信息，便于问题排查与审计。日志内容包括：任务基础信息（任务 ID、提交人、时间）、参

数配置（精度模式、输入形状、优化策略）、转换步骤详情（每一步操作、耗时、涉及工具版本，如“2024-05-20 10:05:00 开始 ONNX 转换，使用 ONNX Runtime 1.15，耗时 120s”）、校验结果（各校验项是否通过、异常信息）。日志支持按任务 ID 查询、关键词搜索（如搜索“算子不兼容”）、导出为 TXT/PDF 格式，日志保留时间与模型版本一致（默认 30 天，可手动延长至 90 天）。

4.2.4.2 技术实现方案：高兼容、高精度的转换架构

基于“中间格式解耦+平台插件化+自动化校验”的技术思路，构建分层转换架构，覆盖模型解析、中间格式转换、平台适配、结果校验四大环节，解决跨平台算子不兼容、精度损失、转换效率低等核心问题，确保 20 类模型在 4 类平台上的转换需求落地。

分层转换架构设计：采用“原生格式→ONNX 中间格式→目标平台格式”的两步转换流程，以 ONNX 作为“桥梁”解耦原生模型与目标平台，降低适配复杂度。第一步实现“多源归一”：将 PyTorch/TensorFlow 等原生格式统一转换为 ONNX 标准格式（支持 ONNX1.8~1.15 版本），解决不同框架的格式差异；第二步实现“一分多”：基于 ONNX 中间格式，通过平台专用插件转换为 OM/RKNN/TensorRT Engine/CNModel 格式，适配不同硬件架构。

原生格式解析引擎开发：集成 Netron 核心解析能力与框架专用解析工具，实现 20 类模型的精准解析，提取模型拓扑结构、算子类型、输入输出维度、权重数据。具体解析策略如下：

PyTorch 模型（.pth/.pt）：通过 torch.load() 加载模型权重，结合 torch.onnx.export() 将模型导出为 ONNX 格式，导出时设置 do_constant_folding=True 启用常量折叠，优化模型结构；针对 YOLOv8 等自定义模型，加载对应模型定义文件（如 yolov8.py），确保自定义算子（如 Detect 头）可正常解析。

TensorFlow 模型（.pb/.saved_model）：使用 tensorflow.saved_model.load() 加载模型，通过 tf.saved_model.save() 转换为 SavedModel 格式，再调用 tf2onnx 工具导出 ONNX，处理 TensorFlow 特有的 NHWC 数据排布（自动转换为 ONNX 标准的 NCHW 格式），避免后续转换维度错误。

TensorFlow Lite 模型（.tflite）：利用 tflite.Interpreter 读取模型，解析输入输出张量信息，通过 tflite2onnx 工具转换为 ONNX，针对量化模型（如 INT8.tflite），保留量化参数，确保后续平台转换时可复用量化信息。

自然语言大模型（如 BERT/GPT-2）：针对 Transformer 层结构复杂的特点，开发专用解析适配器，提取注意力头数量、隐藏层维度、序列长度等关键参数，导出 ONNX 时启用 dynamic-axes 配置，支持动态形状，避免因序列长度固定导致的场景限制。

ONNX 中间格式优化：在原生格式转 ONNX 后，新增优化环节提升后续平台转

换效率与模型性能，优化操作包括：

模型简化：调用 ONNX Simplifier 工具移除冗余算子（如无用的 Reshape、Identity 算子）、合并常量节点（如将多个 Constant 算子合并为单个）、消除死代码（未被使用的层），简化后模型体积平均减少 20%，转换耗时缩短 15%。

算子兼容性修正：基于 ONNX 算子规范，修正非标准算子（如将 PyTorch 的“torch.nn.functional.interpolate”算子替换为 ONNX 标准的“Resize”算子），避免后续平台插件因算子不识别导致转换失败，算子修正覆盖率 $\geq 98\%$ （覆盖 20 类模型常用算子）。

精度保持优化：针对 FP16/INT8 量化场景，在 ONNX 层面插入量化节点（如 QuantizeLinear/DequantizeLinear），确保量化参数传递至目标平台，减少精度损失；对于自然语言大模型的 Transformer 层，保留 FP32 精度计算，仅对非关键层（如 Embedding 层）进行量化，平衡精度与性能。

平台专用转换插件开发：采用插件化设计，为 4 类平台开发独立转换插件，集成平台专用工具链，实现 ONNX 到目标格式的精准转换，同时解决硬件架构差异导致的兼容性问题。各插件技术实现如下：

华为昇腾 OM 插件：集成华为 ATC（Ascend Tensor Compiler）工具，封装 ATC 编译接口，支持通过配置文件传入算子融合、数据排布优化参数（如 `--input-format=NCHW--fusion-switch=true`）。针对昇腾不支持的算子（如 TensorFlow 的“Unique”），开发算子替换库：将“Unique”算子替换为“Sort+Deduplicate”组合算子，确保功能等价；针对大模型（如 GPT-2），启用昇腾的“模型并行”功能，将模型拆分至多个 AICore 执行，解决显存不足问题。插件支持 ATC 版本自动适配（检测到 ATC 7.0/8.0 版本时自动调整编译参数），转换成功率 $\geq 95\%$ （20 类模型测试）。

瑞芯微 RKNN 插件：基于 RKNN Toolkit2.5 开发，封装 rknn.api 接口，实现 ONNX 模型加载、量化、编译全流程自动化。针对瑞芯微 NPU 对大 kernel 卷积（如 11×11 ）支持不足的问题，开发卷积拆分算法：将 11×11 卷积拆分为 $3 \times 3 + 3 \times 3 + 5 \times 5$ 的组合卷积，确保功能正常且性能损失 $\leq 10\%$ ；支持 RK3588/RK3599 平台差异化配置（如 RK3588 启用 INT8 量化，RK3599 支持 FP16 混合量化），插件自动识别硬件型号并加载对应配置。

英伟达 TensorRT 插件：集成 TensorRT 8.6 C++ API，实现 ONNX 模型解析、构建优化、序列化流程。开发自定义 Plugin 插件，适配 YOLOv8 检测头、BERT 注意力层等专用算子，将检测头的“卷积+激活+解码”流程融合为单个 TensorRT Plugin，推理速度提升 30%；支持动态形状推理（通过 IBuilderConfig 设置 dynamicShapes），适配自然语言大模型的可变序列长度；启用 TensorRT 的“BuilderOptimizationLevel=5”，采用最激进的优化策略（如层融合、内核自动调优），平衡转换时间与推理性能。

寒武纪 CNModel 插件：基于 MagicMind 1.8 SDK 开发，调用 `magicmind::Builder` 接口构建模型，支持算子自动映射（MagicMind 内置

1000+ONNX 算子映射规则），未映射算子通过 MagicMind 的“自定义算子”功能实现（基于 C++ 编写算子执行逻辑，编译为动态链接库供插件调用）。针对寒武纪思元 NPU 的内存特性，优化模型内存分配策略（如按层顺序分配内存、复用中间张量内存），显存占用降低 25%；支持多 NPU 并行转换（如思元 590 的 4 颗 NPU 同时处理 4 个模型转换任务），转换效率提升 3 倍。

转换任务调度与资源管理：基于 Kubernetes+RabbitMQ 构建分布式任务调度系统，实现转换任务的高效分配与硬件资源管控，避免节点过载导致的转换失败。调度逻辑如下：

任务队列管理：用户提交的转换任务先进入 RabbitMQ 队列，按“优先级（高/中/低）+提交时间”排序，试验场紧急调试任务标记为“高优先级”，优先占用资源（优先级权重：高=3，中=2，低=1）。

节点负载感知：通过 Prometheus 采集各硬件节点的实时负载（CPU 利用率、内存/显存占用率），调度器仅将任务分配至负载低于阈值的节点（CPU ≤ 70%、显存 ≤ 60%），如检测到英伟达 RTX 4090 节点显存占用 80% 时，自动将任务调度至备用节点。

资源隔离：采用 Docker 容器或者 Anaconda 环境（具体隔离环境按照工具和平台要求来搭建）隔离不同平台的转换环境（如华为昇腾容器仅运行 OM 插件，瑞芯微容器仅运行 RKNN 插件），避免工具链依赖冲突；为每个容器分配固定资源配额（如 CPU 8 核、内存 16GB、显存 12GB），防止单个任务占用过多资源。

失败重试机制：任务转换失败时（如硬件节点临时离线、算子适配错误），调度系统自动将任务重新加入队列（最多重试 2 次），重试时优先分配至同类型其他节点，若连续失败则标记“任务失败”并通知用户，任务成功率 ≥ 98%（排除模型文件损坏、硬件故障等不可抗因素）。

自动化校验与异常处理：开发校验脚本库与异常处理模块，实现转换结果的自动校验与问题修复，确保输出模型可用。具体实现如下：

格式完整性校验：调用平台专用工具验证输出格式（如华为 `atc--check-model` 验证 OM 文件、瑞芯微 `rknn.load_rknn` 验证 RKNN 文件），校验不通过则触发重新转换（最多 2 次），重新转换时自动调整参数（如更换量化算法、禁用算子融合）。

结构一致性校验：通过 ONNX Runtime 加载转换前后的模型（原生转 ONNX 后的模型、目标平台格式反编译为 ONNX 的模型），对比输入输出维度、算子类型、权重数据哈希值，差异超过阈值时生成“结构差异报告”，标记差异算子（如“原模型 Conv2d → 目标模型 ConvTranspose2d”），辅助开发人员定位适配问题。

基础功能校验：在目标平台测试节点部署校验脚本，加载转换后模型执行推理测试（输入标准测试数据，如 COCO 验证集图像、GLUE 文本），检查推理结果是否合法（如目标检测模型输出 bounding box 坐标在 0-1 范围内、分类模型输出概率和为 1），推理失败则记录错误日志（如“算子 XXX 执行超时”“显存不足”），并提供修复建议（如“建议降低精度模式至 INT8”“增大节点显存配

额”)。

异常处理模块：建立异常案例库，收录常见转换异常（如算子不兼容、显存溢出、量化精度损失超标）及解决方案，当校验发现异常时，自动匹配案例库给出修复建议。

4.2.5 功能指标 2：可根据模型格式进行模型可视化

作为系统核心交互功能指标，模型可视化需实现“多格式模型结构解析、差异化场景展示、交互式参数探索”，严格匹配“支持不少于 20 类主流智能体模型网络结构可视化”的性能指标要求，将复杂的模型“黑箱”转化为直观的图形化表达，辅助试验场技术人员快速理解模型拓扑、定位转换异常、优化模型结构，为跨平台转换效果验证与模型调试提供可视化支撑。

4.2.5.1 功能描述，支持性能指标 1

围绕“结构清晰、参数可查、交互灵活”核心目标，构建从“模型解析-场景化展示-交互探索-结果导出”的全流程可视化功能体系，针对 20 类模型（目标检测/分类/跟踪/NLP 大模型）的结构特性提供差异化展示方案，同时支持多维度参数查询与自定义交互操作，满足试验场调试机场景下的模型分析需求。

多格式模型结构解析与适配：全面支持 20 类模型原生格式的结构解析，覆盖计算机视觉与自然语言处理领域主流模型。其中目标检测模型（YOLOv5/v7/v8、Faster R-CNN、SSD）需解析 backbone（特征提取网络）、neck（特征融合网络）、head（检测头）三层结构及层间连接关系；目标分类模型（ResNet50/101、EfficientNet-B0/B4、MobileNetV3）需解析残差块、注意力模块等核心组件；目标跟踪模型（SORT、DeepSORT、ByteTrack）需解析检测器、卡尔曼滤波、匈牙利匹配等流程模块；自然语言大模型（BERT-base、RoBERTa、GPT-2、T5-small）需解析 Transformer 编码器/解码器层、多头注意力模块、Feed Forward 网络等关键结构。解析过程自动识别模型格式（通过文件后缀+结构特征双重判定，如检测到含“Detect”层的.pt 文件时标记为 YOLO 系列模型），并加载对应的可视化模板，无需用户手动配置，解析成功率 $\geq 99.5\%$ （针对 20 类目标模型）。

差异化场景化展示：针对不同类型模型的结构特性，设计专属可视化视图，突出核心模块与功能逻辑，避免信息冗余。

计算机视觉模型视图：采用“分层折叠式”布局，按功能模块（backbone/neck/head）纵向分组，每个模块用不同颜色标识（backbone 蓝色、neck 绿色、head 橙色），模块内按层顺序横向排列（如 ResNet50 的 backbone 按 conv1→layer1→layer2→layer3→layer4 顺序展示）。支持模块独立折叠/展开（点击模块标题可隐藏/显示内部层结构），层节点标注核心参数（如卷积层标注“Conv2d(3→64, 7×7, stride=2)”，池化层标注“MaxPool(3×

3, stride=2)”),鼠标悬浮时显示该层的输入输出维度(如“输入[1, 3, 224, 224]→输出[1, 64, 112, 112]”)与计算量(FLOPs),帮助快速掌握层功能。

自然语言大模型视图:采用“横向链式”布局,按Transformer层顺序横向排列(如BERT-base的12个编码器层依次横向展开),每个Transformer层内部用“环形图”展示多头注意力模块(不同颜色标识不同注意力头),用“矩形块”展示Feed Forward网络,层间用箭头标注数据流向(如“词嵌入层→编码器层1→编码器层2→...→分类层”)。支持注意力权重热力图展示(点击注意力模块弹出热力图,横轴/纵轴为序列位置,颜色深浅代表权重大小),直观呈现模型对文本序列的关注重点;针对GPT-2等生成式模型,额外展示“解码流程”(如自回归生成时的token预测顺序),辅助理解生成逻辑。

目标跟踪模型视图:新增“流程可视化”功能,除展示检测器的网络结构外,用流程图形式展示跟踪流程(检测器输出bbox→卡尔曼滤波预测→匈牙利算法匹配→轨迹更新),每个流程节点标注核心参数(如卡尔曼滤波标注“状态维度4,观测维度2”),支持模拟输入视频帧,动态演示轨迹生成过程(如不同目标的bbox与轨迹线实时更新),直观呈现跟踪逻辑。

交互式参数探索:提供丰富的交互操作,支持技术人员深入探索模型参数细节,定位潜在问题。

层节点交互:支持鼠标拖拽层节点调整位置(便于查看复杂连接关系)、双击节点展开详细参数面板(显示权重分布直方图、偏置值范围、激活函数类型,如卷积层显示“权重均值0.2、标准差0.1,激活函数ReLU”)、框选多个节点对比参数(如同时选中两个残差块,对比其卷积核大小、输出通道数差异)。

结构导航与搜索:提供“全局缩放”(鼠标滚轮缩放视图,缩放范围10%~500%)、“定位跳转”(输入层名称如“conv2d_1”快速定位至对应节点)、“功能筛选”(勾选“仅显示卷积层”“隐藏BN层”等选项,过滤无关信息),针对1000层以上的大模型(如GPT-2),支持“缩略图导航”(右侧显示模型整体缩略图,红色框标识当前视图位置,拖动红色框可快速切换视图区域),避免操作迷失。

可视化结果导出与分享:支持将可视化结果导出为多种格式,便于文档编写、汇报展示与问题沟通。可导出模型拓扑图(PNG/SVG格式,SVG格式支持无限缩放不失真)、参数详情表(Excel格式,包含所有层的名称、类型、参数、输入输出维度)、注意力热力图(PNG/PDF格式);针对目标跟踪模型,可导出轨迹模拟视频(MP4格式,帧率25fps)。导出时支持自定义配置(如PNG格式可选择分辨率300dpi/600dpi,Excel表可选择是否包含计算量数据),导出文件自动命名(如“YOLOv8-结构拓扑图-20240520.png”),并关联模型ID与版本号,便于追溯。同时支持生成可视化报告(HTML/PDF格式),包含模型基础信息、结构分析、关键参数统计(如总层数、总参数规模、总计算量)与优化建议(如“检测头参数规模较大,建议轻量化以提升推理速度”),报告可直接分享至试验场内部系统(如OA、项目管理平台)。

跨转换环节联动展示:与模型跨平台转换功能联动,支持对比转换前后的模

型结构差异，辅助验证转换效果。转换完成后，可在可视化界面同时加载“转换前 ONNX 模型”与“转换后目标平台模型（反编译为 ONNX 格式）”，系统自动标记结构差异（如算子替换、层顺序调整，差异层用红色边框标识），并生成差异报告（如“转换后 Conv2d 算子替换为昇腾平台的 GeMM 算子，输入维度保持一致”）。针对精度损失超标的模型（如 mAP 损失 > 5%），支持定位差异算子并分析影响（如“量化后的注意力模块权重分布偏移，导致精度下降”），为转换参数优化（如调整量化算法）提供依据。

4.2.5.2 技术实现方案：高精度、高交互的可视化架构

基于“解析引擎+模板驱动+交互优化”的技术思路，构建分层可视化架构，覆盖模型解析、模板管理、渲染优化、交互控制四大环节，解决多格式适配、大模型渲染卡顿、参数展示不直观等核心问题，确保 20 类模型的可视化需求落地。

模型解析引擎开发：集成 Netron 开源解析内核与自定义适配器，实现多格式模型结构数据的精准提取与标准化处理，为可视化渲染提供统一数据格式。

基础解析能力：基于 Netron 核心库解析 ONNX、PyTorch、TensorFlow 等格式模型，提取模型拓扑结构（层名称、类型、父层/子层关系）、算子参数（卷积核大小、步长、填充方式）、输入输出维度等基础数据，解析结果转换为 JSON 格式（包含“layers”数组与“connections”数组，layers 记录层信息，connections 记录层间连接关系），确保后续渲染模块可直接调用。

数据标准化处理：解析完成后对数据进行标准化（如统一参数命名格式，将“kernel-size”“filter-size”统一为“kernel”；统一维度表示方式，将“(1, 3, 224, 224)”标准化为“[batch, channel, height, width]”），避免因模型格式差异导致的渲染异常；针对大模型（如 1000 层以上的 GPT-2），对层数数据进行抽样预处理（默认保留每 10 层的详细参数，其余层仅保留核心信息），降低渲染数据量，预处理耗时 ≤ 2 秒（针对 30GB NLP 大模型）。

模板驱动渲染系统：采用“模板+数据”分离的渲染模式，为不同类型模型设计可视化模板，通过模板定义渲染规则（布局方式、颜色配置、参数展示逻辑），实现差异化展示，同时提升渲染灵活性与可扩展性。

模板结构设计：可视化模板采用 JSON 格式定义，包含“layout”（布局规则）、“style”（样式配置）、“parameterDisplay”（参数展示逻辑）三大模块。例如，YOLO 模型模板的 layout 模块定义“分层折叠式”布局，style 模块定义 backbone 为蓝色、head 为橙色，parameterDisplay 模块定义卷积层需展示“输入通道→输出通道，kernel 尺寸，stride”参数。模板存储于 MySQL 的 visual-template 表，通过“模型类型”字段与模型关联（如“模型类型=目标检测-YOLO”关联 YOLO 模板），支持管理员在线编辑模板（修改颜色、调整参数展示项），修改后实时生效。

多引擎渲染适配：根据模型类型与数据规模选择合适的渲染引擎，确保渲染

效率与展示效果平衡。针对计算机视觉模型（层数 ≤ 200 层），采用 D3.js 渲染拓扑图，支持精细的节点交互与参数展示；针对自然语言大模型（层数 ≥ 100 层），采用 SVG 矢量图形渲染，减少内存占用，同时支持无限缩放；针对目标跟踪模型的流程可视化，采用 Mermaid.js 绘制流程图，结合 ECharts 绘制轨迹模拟动画。渲染过程采用“分层渲染”策略：优先渲染模型整体结构（仅显示层名称与连接关系），用户交互时（如展开模块、点击节点）再动态加载详细参数与图形（如注意力热力图），初始渲染时间 ≤ 1 秒（针对 200 层以内模型），交互响应时间 $\leq 100\text{ms}$ 。

交互控制模块开发：基于前端事件驱动机制，开发丰富的交互功能，支持鼠标、键盘操作与触屏操作（适配试验场调试机的触屏显示器），提升操作便捷性。

基础交互功能：实现鼠标拖拽（拖动层节点调整位置）、滚轮缩放（缩放视图）、点击选中（选中层节点显示详情）、框选（选中多个节点对比参数）等基础操作；支持键盘快捷键（Ctrl+滚轮缩放、Ctrl+F 搜索层节点、Ctrl+S 导出可视化结果），提升操作效率；针对触屏设备，支持双指缩放、单指拖拽，确保多终端适配。

高级交互功能：开发“结构对比”功能，支持加载两个模型（如转换前后的模型），系统自动对齐层结构，用红色标识差异层，绿色标识一致层，点击差异层显示具体差异（如“转换前 Conv2d \rightarrow 转换后 GeMM”）；开发“参数筛选”功能，支持按参数条件筛选层（如“筛选所有卷积核尺寸为 3×3 的层”“筛选计算量 $>1\text{e}6$ FLOPs 的层”），筛选结果用黄色高亮标识，快速定位关键层；开发“动态模拟”功能，针对目标检测模型，通过 WebGL 加载测试图像，调用 ONNX Runtime 执行推理，实时获取各层特征图数据并渲染展示（特征图用热力图形式呈现），支持帧步进（逐帧查看特征图变化）与进度条拖动（快速定位至某一帧）。

性能优化与兼容性保障：针对大模型渲染卡顿、浏览器兼容性差等问题，采用多项优化技术，确保可视化功能在试验场调试机环境下稳定运行。

渲染性能优化：利用 WebWorker 在后台线程处理模型解析与数据预处理，避免阻塞主线程导致界面无响应；采用“数据分片加载”策略，大模型（如 1000 层以上）的层数据按模块分片加载（一次加载 10 个模块），滚动视图时动态加载后续分片；对大规模参数数据（如注意力权重矩阵）采用抽样展示（默认显示 10% 的数据，用户点击“查看全部”加载完整数据），降低数据传输与渲染压力。优化后，支持 3000 层以上模型的流畅可视化，界面操作无卡顿。

浏览器兼容性适配：针对 Chrome、Edge、Firefox 主流浏览器（版本 ≥ 90.0 ），通过 Babel 转译 ES6+ 语法，确保代码兼容；采用 CSS 前缀自动补全（如 -webkit-、-moz-），确保样式在不同浏览器一致。

异常处理与容错：开发异常监测模块，实时捕捉解析与渲染过程中的错误（如模型结构损坏导致解析失败、渲染数据格式错误），并弹出友好提示（如“模型文件损坏，建议重新上传”“当前浏览器版本过低，推荐升级至 Chrome 90+”），同时记录错误日志（包含错误类型、模型 ID、操作时间）至 MySQL 的

visual-error-log 表，便于开发人员排查问题。支持“降级渲染”（解析失败时自动切换至基础解析模式，仅展示层名称与连接关系，避免可视化功能完全不可用），容错率 $\geq 98\%$ 。

4.2.6 跨平台推理测试模块

跨平台推理测试模块作为模型转换效果的“验证关口”，需实现“多平台真机测试、多维度指标评估、自动化报告生成”，严格匹配“20类模型+4类平台”的测试覆盖要求，通过在华为昇腾、瑞芯微、英伟达、寒武纪真机环境下执行推理验证，精准评估转换后模型的功能完整性、性能达标率与精度稳定性，确保模型可直接用于试验场内场调试机的部署测试，避免因转换问题导致的部署失败。

4.2.6.1 功能描述：全维度推理验证能力

围绕“功能可用、性能达标、精度可控”核心目标，构建从“测试用例管理-多平台推理执行-指标评估-报告生成”的全流程测试体系，覆盖计算机视觉与自然语言处理模型的差异化测试需求，同时支持测试结果追溯与问题定位，为模型转换优化提供数据支撑。

测试用例全生命周期管理：支持测试数据的上传、分类、版本控制与复用，满足20类模型的测试数据需求。

多类型测试数据适配：兼容计算机视觉模型的图像数据（JPG/PNG/BMP 格式，支持单张/批量上传，批量上限1000张，单张图像分辨率支持 $320 \times 320 \sim 1280 \times 1280$ ）、自然语言模型的文本数据（TXT 格式，支持单条文本输入与批量数据集上传，批量数据集需按“输入文本+预期结果”格式组织，如GLUE数据集的SST-2任务格式）、目标跟踪模型的视频数据（MP4/AVI 格式，支持帧提取与关键帧标记）。上传时自动校验数据格式（如检测图像是否为RGB三通道、文本是否符合UTF-8编码），不兼容数据实时提示并提供格式转换工具（如将灰度图自动转为RGB图、ANSI编码文本转为UTF-8）。

测试用例分类与标签管理：按模型类型（目标检测/分类/跟踪/NLP大模型）与测试场景（功能验证/性能测试/精度对比）对测试用例打标签，如“目标检测-COCO验证集-精度测试”“BERT-GLUE-SST2-功能验证”。支持按标签快速筛选用例（如筛选“所有目标检测类精度测试用例”），同时支持用例版本控制（同一数据集更新时生成新版本，保留历史版本用于对比测试），用例复用率 $\geq 80\%$ ，减少重复数据上传工作量。

测试用例自动化生成：针对无现成测试数据的场景，提供自动化生成工具：目标检测模型支持生成含随机目标（如汽车、行人）的合成图像（可配置目标数量、背景类型）；自然语言模型支持基于模板生成文本（如“[主题]的[属性]是[值]”模板生成“智能模型的功能是格式转换”类文本）；生成数据自动关联

模型类型(如为 YOLOv8 生成 640×640 分辨率图像),生成数量与参数可配置(如生成 100 张合成图像,目标数量 2^5 个/张),生成耗时 ≤ 1 分钟/100 条数据。

多平台真机推理执行:支持在 4 类平台的真机测试节点上执行推理任务,覆盖不同硬件配置(如华为 Atlas 800 服务器、瑞芯微 RK3588 开发板、英伟达 Jetson AGX Orin、寒武纪思元 370 服务器),实现“一键提交-分布式执行-实时监控”。

测试节点选择与任务分配:用户提交推理任务时,可选择单平台单节点(如仅在瑞芯微 RK3588 节点测试)、单平台多节点(如在 2 台华为 Atlas 800 节点并行测试)、多平台多节点(如同时在英伟达 RTX 4090 与寒武纪思元 590 节点测试),系统自动检查所选节点在线状态与负载(CPU 利用率 $\leq 70\%$ 、显存/NPU 内存占用 $\leq 60\%$),若节点离线或过载则提示更换节点或等待。任务分配采用“负载均衡”策略,将批量测试任务(如 1000 张图像推理)拆分至多个节点并行执行,执行效率提升至单节点的 3^5 倍(取决于节点数量)。

推理参数配置:提供精细化参数配置功能,匹配不同模型与平台的测试需求:

推理精度模式:支持与模型转换一致的精度模式(FP32/FP16/INT8),确保测试环境与实际部署环境一致;

推理批次大小:支持批量推理(batch size=1/2/4/8,最大支持 batch size=16,取决于节点内存大小),用于测试模型的吞吐量性能;

结果保存策略:可选择保存所有推理结果(如每张图像的检测框坐标、文本的分类概率)或仅保存异常结果(如推理失败、精度超差的结果),减少存储占用;

日志输出级别:支持 WARN/ERROR 二级日志。

推理过程实时监控:任务执行过程中,实时展示任务进度(如“已完成 80/100 张图像推理,进度 80%”)、各节点执行状态(如“节点 1: 处理中,剩余 30s; 节点 2: 已完成”)、关键指标实时数据(如当前推理延迟、吞吐量)。支持任务暂停(暂停后保留已完成进度,恢复时从断点继续)与取消(取消后释放节点资源,删除临时结果),暂停/取消操作响应时间 $\leq 100\text{ms}$ 。

多维度推理指标评估:从功能、性能、精度三个核心维度评估推理效果,生成量化指标,全面验证模型转换质量。

功能验证指标:评估模型在目标平台的推理可用性,核心指标包括推理成功率(成功推理的测试用例数/总测试用例数,目标 $\geq 99.5\%$)、模型加载成功率(成功加载模型的节点数/总节点数,目标 100%)、无异常输出率(输出结果格式合法、无空值/NaN 值的测试用例数/总测试用例数,目标 $\geq 99\%$)。若推理失败(如模型加载报错、推理超时),自动记录失败原因(如“显存不足”“算子执行错误”)与错误日志,便于定位转换或硬件问题。

性能测试指标:评估模型在目标平台的运行效率,核心指标包括:

推理延迟:单条测试用例从输入到输出的耗时,取 100 轮推理的平均值(排除首轮加载耗时),单位 ms,不同平台延迟阈值不同(如英伟达 RTX 4090 目标检测模型延迟 $\leq 300\text{ms}$,瑞芯微 RK3588 延迟 $\leq 300\text{ms}$);

吞吐量 (QPS)：单位时间内处理的测试用例数，计算公式为“batch size/平均推理延迟 × 1000”，用于评估模型的并发处理能力；

硬件资源占用：推理过程中 CPU 利用率（目标 ≤ 80%）、显存/NPU 内存占用峰值（目标 ≤ 70%）、硬件功耗（仅边缘节点如 RK3588/Jetson Orin，目标 ≤ 15W），通过平台专用工具（nvidia-smi、npu-smi、rk-smi）实时采集。

精度对比指标：以英伟达 RTX 4090 平台（FP32 精度）的推理结果为基准，计算其他平台的精度损失，核心指标按模型类型差异化设计：

目标检测模型：mAP (mean Average Precision, 目标损失 ≤ 10%)、Precision（精确率，目标损失 ≤ 10%）、Recall（召回率，目标损失 ≤ 10%），采用 COCO 评估标准计算；

目标分类模型：Top-1 Accuracy（目标损失 ≤ 10%）、Top-5 Accuracy（目标损失 ≤ 10%），采用 ImageNet 评估标准计算；

目标跟踪模型：MOTA（多目标跟踪精度，目标损失 ≤ 10%）、ID Switch（身份切换次数，目标 ≤ 10 次/100 帧），采用 MOT17 评估标准计算；

自然语言大模型：GLUE Score（含 CoLA、SST-2 等子任务，目标损失 ≤ 5%）、BLEU Score（生成式模型，目标损失 ≤ 10%），采用 Hugging Face 评估库计算。

若精度损失超过预设阈值，自动标记为“精度异常”，并分析异常原因（如“INT8 量化导致小目标检测精度下降”“算子替换导致注意力权重偏移”），提供优化建议（如“降低量化精度至 FP16”“更换算子映射规则”）。注意，本系统重点是设置精度损失阈值和提示，不包括精度损失性能指标提升。

自动化测试报告生成与追溯：推理测试完成后，自动生成标准化测试报告，支持在线查看、下载与问题追溯，为模型转换优化与验收提供依据。

报告内容与格式：报告包含测试概况（测试模型、目标平台、测试时间、测试人员、用例数量）、功能验证结果（各指标数值与达标情况，如“推理成功率 99.8%，达标”）、性能数据图表（不同平台的推理延迟对比图、吞吐量趋势图、资源占用热力图）、精度对比详情（各指标基准值、测试值、损失率，异常指标标红）、问题汇总（推理失败/精度异常的用例列表、失败原因、优化建议）。报告支持 HTML/PDF/Excel 三种格式，HTML 格式支持交互操作（如点击图表查看具体数值），PDF 格式用于存档与验收，Excel 格式用于数据导出与二次分析。

报告追溯与版本关联：报告自动关联模型版本（如“YOLOv8-V2.0”）、转换任务 ID（如“task-20240520-001”）与测试用例版本，用户可通过模型版本或任务 ID 快速查询历史测试报告（默认保留 90 天），对比不同版本模型的测试结果（如“V2.0 比 V1.0 在瑞芯微平台推理延迟降低 20ms，mAP 提升 2%”）。报告中所有数据均关联原始测试日志，点击“查看日志”可跳转至对应的推理日志页面，查看详细操作记录与错误信息，便于问题定位。

4.2.6.2 技术实现方案：高可靠、自动化的测试架构

围绕“4类平台真机验证、20类模型全量测试”核心需求，采用“分布式调度+插件化适配+自动化评估”技术架构，从测试用例管理、推理执行、指标评估、异常处理四大维度构建技术体系，解决多平台环境差异、大规模测试效率低、精度评估复杂等问题，确保推理测试结果精准匹配试验场内场调试机的部署需求。

采用“分布式调度+平台插件化+自动化评估”架构，解决多平台适配与大规模测试效率问题，确保验证结果精准可靠。

测试用例管理实现：采用“MinIO存储+MySQL索引”架构，测试数据按“模型类型/测试场景”路径存储于MinIO，元数据（用例ID、存储路径、标签）存入MySQL并建立联合索引，查询耗时 $\leq 200\text{ms}$ 。自动生成工具基于OpenCV（图像）、Jinja2模板（文本）开发，支持后台异步生成，生成后自动上传并创建索引。

推理执行架构：以RabbitMQ构建调度中心，手动选择节点分配任务。为4类平台开发专用插件：华为昇腾基于CANN SDK封装OM模型推理逻辑，瑞芯微调用RKNN Toolkit实现RKNN模型加载与推理，英伟达通过TensorRT API优化执行效率，寒武纪依托MagicMind SDK适配CNModel格式，插件支持动态加载，新增平台无需修改核心代码。

指标评估与报告生成：集成行业标准评估算法（如COCO Evaluator计算mAP、Hugging Face库计算GLUE Score），自动对比基准数据计算精度损失。报告基于模板驱动生成，调用ECharts绘制性能对比图表，支持在线查看与导出，关键数据关联原始推理日志，点击可跳转查看详细错误信息（如算子执行报错日志）。

4.2.7 任务监控模块

任务监控模块作为系统运行状态的“可视化中枢”，需实现“全类型任务跟踪、多维度状态监控、异常实时告警、操作可追溯”，覆盖模型上传、格式转换、推理测试等全流程任务，帮助试验场技术人员实时掌握任务进度、快速定位异常问题，确保系统在多任务并发场景下稳定可控。

4.2.7.1 功能描述

围绕“实时性、全面性、可操作性”核心目标，构建任务全生命周期监控体系，覆盖任务可视化、状态追踪、异常告警、操作管理四大核心能力，适配试验场多用户、多任务并行的使用场景。

全类型任务可视化展示：支持模型上传、格式转换、推理测试等所有系统任务的统一展示，以列表视图呈现任务核心信息，包括任务ID（UUID唯一标识）、任务名称（如“YOLOv8-华为昇腾转换”）、任务类型（模型上传/格式转换/推理测试）、提交人、提交时间、当前状态（等待中/处理中/成功/失败）、关联

资源（如涉及的模型 ID、目标平台）。支持按任务类型（如仅查看“格式转换”任务）、状态（如筛选“失败”任务）、时间范围（今日/昨日/近 7 天/自定义）、提交人多维度筛选，列表支持分页（每页 10/20/50 条可选）与排序（按提交时间降序/进度升序），单页加载耗时 $\leq 300\text{ms}$ 。

任务状态实时追踪：针对处理中任务，实时更新进度与执行细节：进度通过“进度条+百分比”直观展示，同时标注当前执行阶段（如格式转换任务标注“ONNX 格式解析中”“目标平台插件加载中”）；点击任务条目可展开详情面板，查看任务参数（如转换任务的目标平台、精度模式，推理测试的用例数量）、执行日志（按时间顺序记录关键操作，如“2024-05-20 14: 30: 00 模型解析完成，耗时 15s”“2024-05-20 14: 32: 00 开始 INT8 量化”）、资源占用（如 CPU/内存/显存实时利用率，仅推理/转换任务展示），日志支持关键词搜索（如搜索“错误”快速定位异常节点）。

异常实时告警与处理：当任务出现异常（如转换失败、推理超时、硬件节点离线）时，系统立即触发告警：前端界面弹出红色告警弹窗，显示任务 ID、异常类型（如“算子不兼容”“显存溢出”）与简要原因；同时在任务列表中用红色标识异常任务，标注“异常”标签。提供快速处理入口：点击告警弹窗的“查看详情”可跳转至任务详情面板，点击“重试任务”可沿用原参数重新提交（失败任务专属），告警响应时间 $\leq 100\text{ms}$ 。

任务操作与管理：支持任务全生命周期操作，满足灵活管控需求：处理中任务支持“暂停”（保留当前进度，恢复时从断点继续）与“取消”（释放占用资源，删除临时文件），暂停/取消操作实时生效；成功任务支持“查看结果”（如跳转至转换后模型下载页、推理测试报告页）与“重新执行”（基于原参数生成新任务）；失败任务支持“查看日志”（定位失败原因）与“参数修改后重试”（跳转至参数配置页，保留原配置便于调整）。同时支持批量操作（如选中多个等待中任务批量取消、批量导出任务日志），提升管理效率。

4.2.7.2 技术实现方案

基于“前端实时交互+后端状态同步+分布式日志采集”技术架构，实现任务监控的实时性、准确性与可扩展性，适配系统多模块协同与多平台部署需求。

前端交互层实现：采用 Vue3+Element Plus 构建响应式监控界面，通过 WebSocket 或者前段轮询方式实现任务状态实时更新，避免页面刷新。任务列表基于虚拟滚动列表实现，支持万级任务数据流畅加载，解决大规模任务场景下的界面卡顿问题；进度条与状态标签采用动态样式绑定，处理中任务进度条展示动画效果，异常任务标签自动闪烁提示；详情面板采用折叠面板组件，默认折叠核心信息，点击展开详细日志与资源监控图表（基于 ECharts 绘制 CPU/内存利用率趋势图），日志展示支持代码高亮（突出错误信息）与分页加载（每 50 条日志一页），提升可读性。

后端状态管理与同步：基于常见 Rest 框架+Redis+MySQL 构建任务状态管理体系：MySQL 的“task-status”表存储任务基础状态（如等待中/成功/失败）与核心参数，建立“任务 ID”主键索引与“状态+提交时间”联合索引，确保状态查询耗时 $\leq 200\text{ms}$ ；Redis 缓存处理中任务的实时进度（如“task:123:progress”存储进度百分比）与执行阶段（如“task:123:step”存储当前阶段），采用 Hash 数据结构存储多维度状态信息，支持单独更新某一字段（如仅更新进度，无需重新写入完整状态）；后端服务通过 RabbitMQ 接收各业务模块（模型管理、转换系统、推理测试）的任务状态变更通知，同步更新 MySQL 与 Redis 数据，并通过 WebSocket 向前端推送状态更新消息，状态同步延迟 $\leq 3000\text{ms}$ 。

分布式日志采集与展示：采用 ELK（Elasticsearch+Logstash+Kibana）架构实现任务日志的分布式采集与检索：各业务模块（如转换系统、推理模块）将任务执行日志通过 Logstash 输出至 Elasticsearch，按“任务 ID”分片存储，建立“任务 ID+时间戳”索引，确保日志查询耗时 ≤ 1 秒；前端日志查看功能通过调用 Elasticsearch REST API 实现日志检索，支持关键词模糊查询、时间范围筛选与日志级别过滤（DEBUG/INFO/WARN/ERROR），日志展示采用分页加载与滚动加载结合方式，避免大量日志数据一次性渲染导致的界面卡顿。

异常告警机制实现：基于“规则触发+多渠道通知”设计异常告警系统：在 MySQL 中配置告警规则（如“推理超时 $> 300\text{s}$ 触发二级告警”“硬件节点离线触发一级告警”），后端定时任务（基于 XXL-Job）轮询任务状态与硬件节点状态，匹配告警规则后生成告警事件；一级告警通过 JavaMail 组件发送邮件通知（调用试验场邮件服务器，包含任务详情与处理建议链接），二级告警通过 WebSocket 推送至前端界面；所有告警事件存储于 MySQL 的“alarm-record”表，记录告警 ID、任务 ID、告警类型、处理状态（未处理/已处理）、处理人、处理时间，支持告警历史查询与统计分析，便于系统运维优化。

4.3 运行环境

结合“跨平台格式转化、模型可视化”功能指标及“20 类模型+4 类平台”性能指标要求，按全周期顺序划分为如下工作项：

1. 需求深度调研与确认

对接试验场甲方，明确调试机硬件环境（如操作系统、算力配置）、需适配的 20 类智能模型清单（目标检测/分类/跟踪/NLP 大模型具体型号）及 4 类平台（华为/瑞星微/英伟达/寒武纪）的技术参数；确认模型可视化需求细节（如是否需展示层参数、拓扑图交互功能），输出《需求规格说明书》并经甲方签字确认。

2. 资源与环境搭建

租赁 4 类平台硬件设备（华为昇腾、瑞星微、英伟达 Jetson、寒武纪思元），搭建真机测试环境；收集 20 类模型原生文件（例如 ONNX/TensorFlow/PyTorch

格式)及标准测试数据集(例如COCO/ImageNet等),建立“模型测试库”;部署开发工具链(华为昇腾CANN SDK等)。

3. 软件架构与模块开发

设计“数据输入层-核心处理层-输出与交互层”三层架构,完成模块拆分与接口定义;

数据输入层开发:实现模型文件解析引擎(支持多格式导入)、平台参数配置功能,确保可读取20类模型原生格式;

核心处理层开发:

格式转化模块:开发跨平台转化核心,针对4类平台分别开发插件,实现“原生格式→ONNX→目标平台格式”转化,适配20类模型的算子映射与量化(INT8/FP16);

可视化模块:开发模型拓扑解析逻辑,客户端可视化界面、D3.js实现网页端交互(支持节点缩放/拖拽、层参数查看),确保20类模型网络结构可展示;

输出与交互层开发:实现转化后模型多格式导出(适配4类平台原生格式)、日志记录(数据库存储)、客户端集成。

4. 20类模型与4类平台适配调试

针对每类模型(如YOLO、TensorFlow等),调试转化模板,确保转化成功率≥90%;在4类平台真机上测试转化后模型的部署兼容性,优化算子调用效率,确保模型可正常运行。

5. 功能与性能测试、问题修复与优化

验证格式转化(20类模型+4类平台全量测试)、可视化(网络结构完整性、参数展示准确性)功能;测试模型转化时间、可视化响应时间、转化后模型精度损失,输出《测试报告》;

针对测试中发现的问题,集中修复并更新《问题修复清单》;根据甲方预验收反馈,优化软件交互体验。

4.4 售后服务

在试验场内场调试机上安装部署软件,逐台验证功能(模型导入-转化-可视化全流程);交付最终软件安装包、全套文档(《用户手册》《测试报告》等)。

开展现场培训(操作培训:软件使用+常见问题处理;维护培训:日志分析+版本更新),覆盖甲方技术人员。

一、服务期限:自软件验收合格之日起,提供为期1年的售后服务(若双方签订的合同对服务期限另有约定,以合同为准)。

二、核心服务内容:

技术支持:我方实行首接负责制,当场可解决的问题即时答复,疑难问题2小时内回电告知进展。

故障处理:当甲方提出需求时,我方将严格遵守时效承诺:确保在72小时

内对需求作出明确响应,并在 72 小时内抵达甲方指定场所,开展现场服务工作。

系统维护:系统功能性升级及安全更新,提前 1 周通知,升级后 24 小时内跟进运行情况。

三、服务保障:我方将组建专门面向甲方的服务团队,并提供团队联系人及完整联系方式;若团队人员或联系方式发生变动,我方将第一时间通知甲方,确保沟通渠道始终畅通。

4.5 进度安排

序号	项目阶段	里程碑节点	备注
1	需求调研与分析，方案设计	2025.10.11	完成需求分析和方案设计
2	核心开发	2025.11.10	1. 数据输入层：完成模型解析引擎开发，支持ONNX、TensorFlow、PyTorch原生格式解析；2. 核心处理层：格式转化模块：完成核心适配+4类平台插件开发，完成20类模型转化模板调试；可视化模块：完成拓扑解析逻辑+界面开发，完成交互功能优化（如节点拖拽、参数查看）；3. 输出与交互层：完成多格式导出接口开发，完成日志模块开发与客户端整体集成
3	测试与优化	2025.11.15	1. 完成兼容性测试（20类模型全量测试+4类平台真机测试）、性能测试（转化时间/可视化响应/精度损失），输出初版《测试报告》；2. 针对测试问题集中修复（如模型转化失败、可视化参数缺失）；3. 开展发布版本测试+甲方预验收测试，根据验收反馈完成最终优化
4	交付与培训	2025.11.30	1. 完成试验场内场调试机软件安装部署，逐台验证功能（如模型导入、转化、可视化）；2. 组织现场培训（操作培训：含软件操作、常见问题处理；维护培训：含日志查看、版本更新），覆盖甲方技术人员，输出《培训纪要》；3. 交付最终软件安装包、全套文档（《用户手册》《测试报告》等），组织项目终验收，签署《项目交付确认单》

4.6 交付成果清单

本项目交付成果清单包括：

序号	成果名称	数量
1	跨平台智能算法模型转化工具（含可执行程序、可二次开发源代码）	1 套
2	跨平台智能算法模型转化工具技术方案	1 份
3	跨平台智能算法模型转化工具设计说明	1 份
4	跨平台智能算法模型转化工具自测试报告	1 份
5	跨平台智能算法模型转化工具使用说明书	1 份
6	跨平台智能算法模型转化工具自主可控测试报告	1 份
7	跨平台智能算法模型转化工具软件授权书	1 份

4.7 项目团队

序号	姓名	性别	身份证号	学历	职称等级	单位
1	李敏	女	320322199009016544	博士研究生	讲师	西南交通大学计算机与人工智能学院
2	龚勋	男	433127198009140056	博士研究生	教授	西南交通大学计算机与人工智能学院
3	徐天赐	男	620421200003094117	博士		西南交通大学计算机与人工智能学院
4	林晓兵	男	511324199508014737	博士		西南交通大学计算机与人工智能学院
5	李家辉	男	410526199908033856	硕士		西南交通大学计算机与人工智能学院

专业技术职务任职资格
证书

(交大职证字第 [2021]A027 号)

根据国家有关文件规定以及国家颁发的各类专业技术职务（试行）条例中的相应任职条件，经上级领导机关授权的评审委员会评审，本证持有者已具有相应专业技术职务任职资格。
特发此证。



Qualification Certificate for a Professional Title
(Certificate No. [2021]A027)

In accordance with State regulations and their corresponding qualification status, the bearer of this certificate is officially authorized to hold the professional title presented herein.

Southwest Jiaotong University
Date 3 Month 2021 Year

姓名 Name	龚勋			
性别 Sex	男			
出生日期 Date of Birth	1980 年 5 月 14 日	Y	M	D
身份证号码 ID Number	433127198009140056			
工作单位 Work Department	信息科学与技术学院			
从事专业 Speciality	计算机科学与技术			
任职资格 Professional Title	教授			
评审机构 Approved By	西南交通大学专业技术职务评审委员会			
评审通过时间 Date of Approval	2020 年 12 月 31 日 Y M D			

西南交通大学

西交人技干（2024）103 号



关于 **李敏** 同志具备 **讲师** 任职资格的 通 知

经考核，**李敏** 同志从 二〇二四年五月二十八日 起具备
讲师 专业技术职务任职资格。

特此通知。



主送：本人、计算机与人工智能学院

抄送：档案、人力资源部

教育部学籍在线验证报告

更新日期：2025年09月17日

姓名	徐天赐
性别	男
出生日期	2000年03月09日
民族	汉族
学校名称	西南交通大学
层次	博士研究生
专业	计算机科学与技术
学制	4 年
学历类别	普通高等教育
学习形式	全日制
分院	
系所	计算机与人工智能学院
入学日期	2024年09月01日
学籍状态	在籍（注册学籍）
预计毕业日期	2028年06月15日



在线验证码 **ASYVELBN76XJLONG**

- ①验证报告在线查验网址：<https://www.chsi.com.cn/xlcx/bgcx.jsp>
- ②使用学信网App扫描二维码验证

注意事项：

- 1、《学籍在线验证报告》是教育部学籍电子注册备案的查询结果。
- 2、报告内容如有修改，请以最新在线验证的内容为准。
- 3、未经学籍信息权属人同意，不得将报告用于违背权属人意愿之用途。
- 4、报告在线验证有效期由报告权属人设置（1~6个月），其在报告验证到期前可再次延长验证有效期。



教育部学籍在线验证报告

更新日期：2025年04月30日

姓名	林晓兵
性别	男
出生日期	1995年08月01日
民族	汉族
学校名称	西南交通大学
层次	博士研究生
专业	计算机科学与技术
学制	4 年
学历类别	普通高等教育
学习形式	全日制
分院	
系所	计算机与人工智能学院
入学日期	2021年09月01日
学籍状态	在籍（注册学籍）
预计毕业日期	2025年12月15日



在线验证码 **AHZYFUPY4GWX32VD**

①验证报告在线查验网址：<https://www.chsi.com.cn/xlcx/bgcx.jsp>

②使用学信网App扫描二维码验证

注意事项：

- 1、《学籍在线验证报告》是教育部学籍电子注册备案的查询结果。
- 2、报告内容如有修改，请以最新在线验证的内容为准。
- 3、未经学籍信息权属人同意，不得将报告用于违背权属人意愿之用途。
- 4、报告在线验证有效期由报告权属人设置（1~6个月），其在报告验证到期前可再次延长验证有效期。



教育部学籍在线验证报告

更新日期：2025年09月17日

姓名	李家辉
性别	男
出生日期	1999年08月03日
民族	汉族
学校名称	西南交通大学
层次	硕士研究生
专业	计算机技术
学制	3年
学历类别	普通高等教育
学习形式	全日制
分院	计算机与人工智能学院
系所	计算机与人工智能学院
入学日期	2023年09月01日
学籍状态	在籍（注册学籍）
预计毕业日期	2026年06月15日



在线验证码 **ARPEXXD56SAGDXNY**

- ①验证报告在线查验网址：<https://www.chsi.com.cn/xlcx/bgcx.jsp>
②使用学信网App扫描二维码验证

注意事项：

- 1、《学籍在线验证报告》是教育部学籍电子注册备案的查询结果。
- 2、报告内容如有修改，请以最新在线验证的内容为准。
- 3、未经学籍信息权属人同意，不得将报告用于违背权属人意愿之用途。
- 4、报告在线验证有效期由报告权属人设置（1~6个月），其在报告验证到期前可再次延长验证有效期。



投标人名称：西南交通大学

法定代表人或授权代表：

年 月 日

4.8 配套能力

申请团队一直从事智能模型的机理理论和技术应用工作，近年来承担了大量基于大模型的垂直领域工程实践项目，项目覆盖空地多机协同决策、战术级策略推演、电网智能巡检、隧道智能建造等诸多行业和工程领域。构建了基于多视角迁移的视觉语言大模型，实现了异视角下的空地多机联合侦查；构建了基于知识图谱和检索增强的策略推演大模型，实现了兵棋推演环境下的智能策略生成；构建了虚实结合的机器人感知与决策基座大模型，实现了电网环境下的四足机器人智能巡检；构建了基于多智能体和扩展知识库的隧道智能建造大模型，为隧道全生命周期管理提供了智能化解决方案。团队积累了大量智能模型行业应用的知识和技术，可为项目开展的跨平台智能算法模型转化工具提供扎实的技术基础。

此外，申请人团队还开展了大量技术创新工作，在 **AAAI** 等国内外知名期刊和会议上发表了大模型相关的高水平学术论文 **10** 余篇。这些工作填补了智能大模型在具身智能和多智能体协同决策方面的理论和技术空白，积累了大量大模型在领域应用中的核心技术，为项目的有效开展提供了有力的创新基础。研究了基于大模型的无人设备群体智能系统，证明了长时序思维链在多机器人协同决策方面的卓越性能；研究了基于强化反馈的多模态游戏 **AI** 大模型，证明了基于人类反馈的强化学习对于多模态大模型的强力支持作用；研究了基于场景语义理解的室内机器人导航大模型，证明了三维场景语义图对于决策智能大模型的有效支撑。

投标人名称：西南交通大学

法定代表人或授权代表：

年 月 日

五、 资质文件

5.1 事业单位法人证书副本



事业单位法人证书

统一社会信用代码 12100000450752090P

法定代表人 闫学东

经费来源 财政补助、上级补助、事业、经营、
捐赠收入

开办资金 ¥71470万元

举办单位 教育部

登记机关 机关

有效期 自2024年07月24日至2029年07月23日
请于每年3月31日前向登记机关报送上一年度的年度报告

仅用于招标投标使用

国家事业单位登记管理局监制

名称 西南交通大学

宗旨和业务范围 培养高级专门人才，实施科技文化发展工程，开展科学研究、技术开发、成果转化、学术交流、咨询服务、资产管理、后勤保障、社会服务和国际交流与合作。管理学校资产和财务，开展学校后勤服务和后勤社会化改革，开展学校后勤服务社会化改革，开展学校后勤服务社会化改革。

住所 四川省成都市二环路北一段

登记机关 机关

登记机关 机关

5.2 GJB 质量管理体系证书

有纸质版

5.3 2022 年至 2024 年资产负债表

资产负债表							
编制单位：西南交通大学							
2024年度							
行次	项 目	年初余额	年末余额	行次	项 目	年初余额	年末余额
1	2	3	4	39	流动负债：	—	—
1	流动资产：	—	—	40	短期借款	—	—
2	库存现金	191,721.99	1,374,000.41	41	应交增值税	1,654,654.96	7,614,625.61
3	银行存款	4,234,473,859.19	4,684,137,483.26	42	其他应交税费	9,435,538.98	14,267,425.76
4	其他货币资金	—	—	43	应缴财政款	—	—
5	短期投资	—	—	44	应付职工薪酬	12,518,688.50	12,660,101.42
6	财政应返还额度	1,921,359.16	55,540,121.67	45	应付票据	—	—
7	应收票据	59,563,587.07	51,224,897.58	46	应付账款	5,352,527.59	5,190,036.41
8	应收账款净额	1,537,956.37	1,746,026.49	47	应付利息	—	—
9	预付账款	327,475,145.60	305,206,030.98	48	预收账款	2,859,995.00	527,740.09
10	应收股利	—	—	49	其他应付款	599,945,665.30	1,155,217,724.78
11	应收利息	—	—	50	预提费用	—	545,401.33
12	其他应收款净额	119,011,458.77	102,703,564.91	51	一年内到期的非流动负债	—	—
13	存货	5,673,344.19	5,744,074.74	52	其他流动负债	—	—
14	待摊费用	—	—	53	流动负债合计	631,767,070.33	1,196,023,055.40
15	一年内到期的非流动资产	—	—	54	非流动负债：	—	—
16	其他流动资产	—	—	55	长期借款	1,074,363,498.94	1,070,871,231.69
17	流动资产合计	4,749,848,432.34	5,207,676,200.04	56	长期应付款	234,804.64	588,663.56
18	非流动资产：	—	—	57	预计负债	—	—
19	长期股权投资	808,921,558.05	941,808,047.99	58	其他非流动负债	—	—
20	长期债券投资	—	—	59	非流动负债合计	1,074,598,303.58	1,071,459,899.25
21	固定投资	8,397,367,685.75	8,666,851,539.29	60	受托代理负债	—	—
22	减：固定投资累计折旧	4,035,719,434.69	4,354,499,380.82	61	负债合计	1,715,228,619.37	2,277,136,197.15
23	固定投资净值	4,361,648,251.06	4,312,352,158.47	62	净资产：	—	—
24	工程物资	—	—	63	累计盈余	9,011,899,174.12	9,294,974,111.70
25	在建工程	678,706,588.79	1,006,060,208.33	64	专用基金：	1,935,852.80	1,943,172.80
26	无形资产原值	135,762,196.05	154,432,339.23	65	职工福利基金	—	1,560.00
27	减：无形资产累计摊销	73,544,505.12	94,929,904.29	66	学生奖励基金	—	—
28	无形资产净值	62,217,690.93	59,502,434.94	67	其他专用基金	1,935,852.80	1,941,612.80
29	研发支出	—	—	68	权益法调整	-58,857,879.66	-37,001,185.38
30	长期待摊费用	—	—	69	净资产合计	8,954,977,147.26	9,259,916,099.12
31	待处理财产损益	—	—	70	—	—	—
32	其他非流动资产	—	—	71	—	—	—
33	非流动资产合计	5,911,494,088.83	6,319,722,849.73	72	—	—	—
34	受托代理资产	8,863,245.46	9,653,246.50	73	—	—	—
35	—	—	—	74	—	—	—
36	—	—	—	75	—	—	—
37	—	—	—	76	—	—	—
38	资产总计	10,670,205,766.63	11,537,052,296.27	76	负债和净资产总计	10,670,205,766.63	11,537,052,296.27

资产负债表							
编制单位：西南交通大学							
2023年度							
行次	项 目	年初余额	年末余额	行次	项 目	年初余额	年末余额
1	2	3	4	39	流动负债：	—	—
1	流动资产：	—	—	40	短期借款	0.00	0.00
2	库存现金	170,347.24	191,721.99	41	应交增值税	4,956,260.99	1,654,654.96
3	银行存款	4,009,986,029.17	4,234,473,859.19	42	其他应交税费	8,873,354.05	9,435,538.98
4	其他货币资金	0.00	0.00	43	应缴财政款	0.00	0.00
5	短期投资	0.00	0.00	44	应付职工薪酬	12,376,588.11	12,518,688.50
6	财政应返还额度	29,474,727.20	1,921,359.16	45	应付票据	0.00	0.00
7	应收票据	74,555,077.39	59,563,587.07	46	应付账款	6,652,085.57	5,352,527.59
8	应收账款净额	2,514,902.65	1,537,956.37	47	应付利息	0.00	0.00
9	预付账款	310,745,170.65	327,475,145.60	48	预收账款	486,216.25	2,859,995.00
10	应收股利	0.00	0.00	49	其他应付款	652,399,242.57	599,945,665.30
11	应收利息	0.00	0.00	50	预提费用	0.00	0.00
12	其他应收款净额	191,396,819.37	119,011,458.77	51	一年内到期的非流动负债	0.00	0.00
13	存货	1,273,773.61	5,673,344.19	52	其他流动负债	0.00	0.00
14	待摊费用	0.00	0.00	53	流动负债合计	685,743,747.54	631,767,070.33
15	一年内到期的非流动资产	—	—	54	非流动负债：	—	—
16	其他流动资产	0.00	0.00	55	长期借款	1,076,229,709.28	1,074,363,498.94
17	流动资产合计	4,620,116,847.28	4,749,848,432.34	56	长期应付款	92,695.46	234,804.64
18	非流动资产：	—	—	57	预计负债	0.00	0.00
19	长期股权投资	971,174,693.43	808,921,558.05	58	其他非流动负债	0.00	0.00
20	长期债券投资	0.00	0.00	59	非流动负债合计	1,076,322,404.74	1,074,598,303.58
21	固定投资	7,521,757,964.76	8,397,367,685.75	60	受托代理负债	9,325,563.75	8,863,245.46
22	减：固定投资累计折旧	3,652,276,447.87	4,035,719,434.69	61	负债合计	1,771,391,716.03	1,715,228,619.37
23	固定投资净值	3,869,481,516.89	4,361,648,251.06	62	净资产：	—	—
24	工程物资	0.00	0.00	63	累计盈余	8,389,729,422.34	9,011,899,174.12
25	在建工程	605,671,655.62	678,706,588.79	64	专用基金：	1,935,852.80	1,935,852.80
26	无形资产原值	96,381,856.75	135,762,196.05	65	职工福利基金	0.00	0.00
27	减：无形资产累计摊销	52,585,296.22	73,544,505.12	66	学生奖励基金	0.00	0.00
28	无形资产净值	43,796,560.53	62,217,690.93	67	其他专用基金	1,935,852.80	1,941,612.80
29	研发支出	0.00	0.00	68	权益法调整	-43,490,153.67	-58,857,879.66
30	长期待摊费用	0.00	0.00	69	净资产合计	8,348,175,121.47	8,954,977,147.26
31	待处理财产损益	0.00	0.00	70	—	—	—
32	其他非流动资产	0.00	0.00	71	—	—	—
33	非流动资产合计	5,490,124,426.47	5,911,494,088.83	72	—	—	—
34	受托代理资产	9,325,563.75	8,863,245.46	73	—	—	—
35	—	—	—	74	—	—	—
36	—	—	—	75	—	—	—
37	—	—	—	76	—	—	—
38	资产总计	10,119,566,837.50	10,670,205,766.63	76	负债和净资产总计	10,119,566,837.50	10,670,205,766.63



资产负债表

财基06表

编制单位：南京航空航天大学

2022年度

金额单位：元

行次	项目	年初余额	年末余额	行次	项目	年初余额	年末余额
1	2	3	4	39	40	41	42
1	流动资产：			39	流动负债：		
2	货币资金	270,921.91	170,347.24	40	短期借款	0.00	0.00
3	衍生金融资产	3,147,331,657.77	4,009,986,029.17	41	应付票据	5,491,153.75	4,956,260.99
4	其他货币资金	0.00	0.00	42	其他应付款	11,957,078.75	8,873,354.05
5	交易性金融资产	0.00	0.00	43	应付账款	0.00	0.00
6	应收账款	9,951,276.01	29,474,727.20	44	应付职工薪酬	2,984,156.91	12,376,588.11
7	应收票据	56,777,550.19	74,555,077.39	45	应付利息	0.00	0.00
8	应收账款净额	1,401,516.94	2,514,902.65	46	应付股利	6,576,651.36	6,652,085.57
9	预付账款	110,328,366.87	310,745,170.65	47	其他应付款	0.00	0.00
10	应收股利	0.00	0.00	48	预收账款	109,786.45	486,216.25
11	应收利息	0.00	0.00	49	其他应付款	804,274,457.11	652,399,242.57
12	其他应收款净额	387,109,526.48	191,396,819.37	50	预提费用	0.00	0.00
13	存货	5,172,731.78	1,273,773.61	51	一年内到期的非流动负债	0.00	0.00
14	待摊费用	0.00	0.00	52	其他流动负债	0.00	0.00
15	一年内到期的非流动资产	—	—	53	流动负债合计	831,393,284.33	685,743,747.54
16	其他流动资产	0.00	0.00	54			
17	流动资产合计	3,718,343,547.95	4,620,116,847.28	55	非流动负债：		
18				56	长期借款	359,840,000.00	1,076,229,709.28
19	非流动资产：			57	长期应付款	46,019.79	92,695.45
20	长期股权投资	606,220,590.07	971,174,693.43	58	预计负债	0.00	0.00
21	债权投资	0.00	0.00	59	其他非流动负债	0.00	0.00
22	固定资产原值	7,327,701,892.73	7,521,757,964.76	60	非流动负债合计	359,886,019.79	1,076,322,404.74
23	减：固定资产累计折旧	3,355,611,876.71	3,652,276,447.87	61			
24	固定资产净值	3,972,090,016.02	3,869,481,516.89	62	受托代理负债	7,352,995.86	9,325,563.75
25	工程物资	0.00	0.00	63			
26	在建工程	568,980,298.14	605,671,655.62	64	负债合计	1,198,632,299.98	1,771,391,716.03
27	无形资产原值	84,497,493.49	96,381,856.75	65			
28	减：无形资产累计摊销	35,529,717.46	52,585,296.22	66	净资产：		
29	无形资产净值	48,967,776.03	43,796,560.53	67	累计盈余	7,898,586,112.36	8,389,729,422.34
30	研发支出	0.00	0.00	68	专用基金：	1,935,852.80	1,935,852.80
31	长期待摊费用	0.00	0.00	69	职工福利基金	0.00	0.00
32	待处理财产损益	172,439,041.04	0.00	70	学生奖励基金	0.00	0.00
33	其他非流动资产	0.00	0.00	71	其他专用基金	1,935,852.80	1,935,852.80
34	非流动资产合计	5,368,697,721.30	5,490,124,426.47	72	权益法调整	-4,760,000.03	-43,490,153.67
35				73	净资产合计	7,895,761,965.13	8,348,175,121.47
36	受托代理资产	7,352,995.86	9,325,563.75	74			
37				75			
38	资产总计	9,094,394,265.11	10,119,566,837.50	76	负债和净资产总计	9,094,394,265.11	10,119,566,837.50

5.4 2022 年至 2024 年收入支出表

预算收入支出表									
编制单位：西南交通大学									
2024年度									
行次	项 目 栏 次	本年数	上年数	行次	项 目 栏 次	本年数	上年数	财基01表	
		1	2			3	4	金额单位：元	
1	一、本年预算收入	4,970,433,427.79	5,032,369,668.85	44	二、本年预算支出	5,050,452,519.69	4,901,709,363.60		
2	(一) 财政拨款预算收入	1,871,935,400.00	1,642,385,000.00	45	(一) 财政拨款支出	1,818,308,837.49	1,669,950,170.41		
3	2. 教育级数预算收入	1,639,767,200.00	1,435,421,200.00	46	1. 教育事业支出	1,616,125,234.94	1,357,763,199.47		
4	(1) 基本支出	1,200,995,900.00	1,181,640,600.00	47	2. 科研事业支出	75,065,827.72	84,392,135.10		
5	(2) 项目支出	438,770,300.00	253,780,600.00	48	3. 行政事业支出	107,735,569.02	183,985,620.03		
6	中央高校改革基本办学条件经费	109,517,800.00	103,111,300.00	49	4. 后勤保障支出	19,106,907.08	27,533,573.53		
7	中央高校教育综合改革支持经费	19,650,000.00	19,300,000.00	50	5. 离退休支出	275,298.73	16,274,642.28		
8	中央高校基本科研业务费	49,050,000.00	46,160,000.00	51	6. 其他事业支出				
9	中央高校捐赠配比资金	6,110,000.00	1,510,000.00	52	(二) 非财政拨款支出	3,149,276,189.15	3,012,643,955.40		
10	中央高校管理改革等绩效拨款	6,032,500.00	16,051,300.00	53	1. 教育事业支出	1,393,412,856.30	1,349,048,273.08		
11	中央高校建设世界一流大学(学科)和特色发展引导专项	69,300,000.00	64,100,000.00	54	2. 科研事业支出	1,295,000,771.51	1,240,189,247.64		
12	基本建设经费	42,380,000.00		55	3. 行政管理支出	206,373,969.43	104,578,481.97		
13	其他专项资金	136,730,000.00	3,548,000.00	56	4. 后勤保障支出	208,341,839.37	281,233,307.15		
14	2. 科研拨款预算收入	10,200,000.00	7,000,000.00	57	5. 离退休支出	46,146,782.34	37,594,645.56		
15	(1) 基本支出	1,200,000.00		58	6. 其他事业支出				
16	(2) 项目支出	9,000,000.00	7,000,000.00	59	(三) 经营支出				
17	3. 其他拨款预算收入	221,968,200.00	199,963,800.00	60	(四) 上缴上级支出				
18	(1) 基本支出	221,390,700.00	199,386,300.00	61	(五) 对附属单位补助支出				
19	(2) 项目支出	577,500.00	577,500.00	62	(六) 投资支出				
20	(二) 事业收入	2,531,157,750.43	2,051,986,270.14	63	(七) 债务还本支出	53,212,267.25	187,324,653.34		
21	1. 教育事业预算收入	1,102,656,638.62	800,000,000.00	64	(八) 其他支出	29,655,225.80	31,790,584.45		
22	(1) 财政专户管理资金	1,102,656,638.62	800,000,000.00	65	其中：利息支出	29,655,225.80	31,790,584.45		
23	(2) 其他教育事业预算收入			66	捐赠支出				
24	2. 科研事业预算收入	1,428,501,151.81	1,251,986,270.14	67	三、本年预算收支差额	-80,019,091.90	130,660,305.23		
25	(1) 非同级财政拨款	638,818,606.22	542,446,386.22						
26	其中：中央科研经费拨款	488,782,618.56	388,633,177.90						
27	地方科研经费拨款	150,035,987.66	153,813,208.32						
28	(2) 其他科研事业预算收入	789,682,545.59	709,539,883.92						
29	3. 其他事业预算收入								
30	(三) 上级补助预算收入								
31	(四) 附属单位上缴预算收入								
32	(五) 经营预算收入								
33	(六) 债务预算收入	49,720,000.00	185,458,443.00						
34	(七) 非同级财政拨款预算收入	175,282,167.00	436,004,203.53						
35	(1) 中央拨款	29,797,664.71	51,049,040.17						
36	(2) 地方拨款	145,484,502.29	384,955,163.36						
37	(八) 投资预算收益		120,000,000.00						
38	(九) 其他预算收入	342,338,070.36	596,535,782.18						
39	1. 租金预算收入	2,388,732.75	3,762,207.40						
40	2. 捐赠预算收入	25,397,793.55	18,161,238.09						
41	3. 利息预算收入	61,108,511.30	86,896,417.46						
42	4. 后勤保障单位净预算收入								
43	5. 其他	253,443,032.76	487,715,889.23						
补充资料：									
一、全日制本专科生(不含网络和成人教育学生)学费：		本年应收	316,047,168.11 元	本年实收：	311,148,437.00 元	欠 费 率：	1.55 (%)		
二、全日制研究生学费：		本年应收	182,046,538.98 元	本年实收：	175,365,431.00 元	欠 费 率：	3.67 (%)		

预算收入支出表									
编制单位：西南交通大学									
2023年度									
行次	项 目 栏 次	本年数	上年数	行次	项 目 栏 次	本年数	上年数	财基01表	
		1	2			3	4	金额单位：元	
1	一、本年预算收入	5,032,369,668.85	5,090,324,339.04	44	二、本年预算支出	4,901,709,363.60	4,126,687,839.91		
2	(一) 财政拨款预算收入	1,642,385,000.00	1,672,056,600.00	45	(一) 财政拨款支出	1,669,950,170.41	1,652,375,084.02		
3	5. 教育级数预算收入	1,435,421,200.00	1,441,386,700.00	46	1. 教育事业支出	1,357,763,199.47	1,304,816,641.41		
4	(1) 基本支出	1,181,640,600.00	1,156,782,200.00	47	2. 科研事业支出	84,392,135.10	133,851,563.33		
5	(2) 项目支出	253,780,600.00	284,604,500.00	48	3. 行政事业支出	183,985,620.03	182,472,308.00		
6	中央高校改革基本办学条件经费	103,111,300.00	105,089,700.00	49	4. 后勤保障支出	27,533,573.53	8,950,016.08		
7	中央高校教育综合改革支持经费	19,300,000.00	19,890,000.00	50	5. 离退休支出	16,274,642.28	22,341,555.17		
8	中央高校基本科研业务费	46,160,000.00	45,880,000.00	51	6. 其他事业支出	0.00	0.00		
9	中央高校捐赠配比资金	1,510,000.00	1,780,000.00	52	(二) 非财政拨款支出	3,012,643,955.40	2,433,862,231.20		
10	中央高校管理改革等绩效拨款	16,051,300.00	17,046,200.00	53	1. 教育事业支出	1,349,048,273.08	1,225,598,295.16		
11	中央高校建设世界一流大学(学科)和特色发展引导专项	64,100,000.00	62,200,000.00	54	2. 科研事业支出	1,240,189,247.64	787,829,917.50		
12	基本建设经费	0.00	30,410,000.00	55	3. 行政管理支出	104,578,481.97	109,174,769.03		
13	其他专项资金	3,548,000.00	2,308,600.00	56	4. 后勤保障支出	281,233,307.15	274,254,458.96		
14	2. 科研拨款预算收入	7,000,000.00	19,540,000.00	57	5. 离退休支出	37,594,645.56	37,004,790.53		
15	(1) 基本支出	0.00	0.00	58	6. 其他事业支出	0.00	0.00		
16	(2) 项目支出	7,000,000.00	19,540,000.00	59	(三) 经营支出	0.00	0.00		
17	3. 其他拨款预算收入	199,963,800.00	211,129,900.00	60	(四) 上缴上级支出	0.00	0.00		
18	(1) 基本支出	199,386,300.00	210,552,400.00	61	(五) 对附属单位补助支出	0.00	0.00		
19	(2) 项目支出	577,500.00	577,500.00	62	(六) 投资支出	0.00	0.00		
20	(二) 事业收入	2,051,986,270.14	1,814,580,102.93	63	(七) 债务还本支出	187,324,653.34	24,860,000.00		
21	1. 教育事业预算收入	800,000,000.00	729,594,355.36	64	(八) 其他支出	31,790,584.45	15,590,524.65		
22	(1) 财政专户管理资金	800,000,000.00	729,594,355.36	65	其中：利息支出	31,790,584.45	15,590,524.65		
23	(2) 其他教育事业预算收入	0.00	0.00	66	捐赠支出	0.00	0.00		
24	2. 科研事业预算收入	1,251,986,270.14	1,084,985,747.07	67	三、本年预算收支差额	130,660,305.23	963,636,499.13		
25	(1) 非同级财政拨款	542,446,386.22	438,614,135.73						
26	其中：中央科研经费拨款	388,633,177.90	321,285,580.88						
27	地方科研经费拨款	153,813,208.32	107,328,554.85						
28	(2) 其他科研事业预算收入	709,539,883.92	656,371,611.84						
29	3. 其他事业预算收入	0.00	0.00						
30	(三) 上级补助预算收入	0.00	0.00						
31	(四) 附属单位上缴预算收入	0.00	0.00						
32	(五) 经营预算收入	0.00	0.00						
33	(六) 债务预算收入	185,458,443.00	741,249,709.28						
34	(七) 非同级财政拨款预算收入	436,004,203.53	282,361,603.92						
35	(1) 中央拨款	51,049,040.17	38,974,800.27						
36	(2) 地方拨款	384,955,163.36	243,386,803.65						
37	(八) 投资预算收益	120,000,000.00	21,100,400.00						
38	(九) 其他预算收入	596,535,782.18	558,975,522.51						
39	1. 租金预算收入	3,762,207.40	1,774,143.78						
40	2. 捐赠预算收入	18,161,238.09	16,981,131.00						
41	3. 利息预算收入	86,896,417.46	77,275,582.44						
42	4. 后勤保障单位净预算收入	0.00	0.00						
43	5. 其他	487,715,889.23	492,945,065.69						



预算收入支出表

财基01表
金额单位：元

2022年度		2021年度		2020年度			
行次	项 目	本年数	上年数	行次	项 目	本年数	上年数
1	一、本年预算收入	5,090,324,339.04	4,172,184,687.16	44	二、本年预算支出	4,126,687,839.91	4,228,155,205.38
2	(一) 财政拨款预算收入	1,672,056,600.00	1,672,934,000.00	45	(一) 财政拨款支出	1,652,375,084.03	1,665,179,683.39
3	1. 教育拨款预算收入	1,441,386,700.00	1,526,117,600.00	46	1. 教育事业支出	1,304,816,641.41	1,314,927,937.87
4	(1) 基本支出	1,156,782,200.00	1,187,280,100.00	47	2. 科研事业支出	133,851,563.33	134,888,807.03
5	(2) 项目支出	284,604,500.00	338,837,500.00	48	3. 行政管理支出	182,472,308.03	183,886,324.03
6	中央高校改善基本办学条件经费	105,089,700.00	124,170,000.00	49	4. 后勤保障支出	8,893,016.08	8,961,929.92
7	中央高校教育教学改革支持经费	19,890,000.00	21,310,000.00	50	5. 离退休支出	22,341,555.17	22,514,684.54
8	中央高校基本科研业务费	45,880,000.00	47,140,000.00	51	6. 其他事业支出	0.00	0.00
9	中央高校捐赠配比资金	1,780,000.00	9,290,000.00	52	(二) 非财政拨款支出	2,433,862,231.20	2,472,083,857.99
10	中央高校管理改革等绩效拨款	17,046,200.00	25,630,000.00	53	1. 教育事业支出	1,225,598,295.16	1,198,252,458.75
11	中央高校建设世界一流大学(学科)和特色发展引导专项	62,200,000.00	56,000,000.00	54	2. 科研事业支出	787,829,917.50	872,930,451.62
12	基本建设经费	30,410,000.00	53,000,000.00	55	3. 行政管理支出	109,174,769.03	101,033,723.14
13	其他专项经费	2,308,600.00	2,297,500.00	56	4. 后勤保障支出	274,254,458.98	286,571,323.86
14	2. 科研拨款预算收入	19,540,000.00	40,310,000.00	57	5. 离退休支出	37,004,790.53	13,295,900.62
15	(1) 基本支出	0.00	0.00	58	6. 其他事业支出	0.00	0.00
16	(2) 项目支出	19,540,000.00	40,310,000.00	59	(三) 经营支出	0.00	0.00
17	3. 其他拨款预算收入	211,129,900.00	106,506,400.00	60	(四) 上缴上级支出	0.00	0.00
18	(1) 基本支出	210,552,400.00	106,176,400.00	61	(五) 对附属单位补助支出	0.00	0.00
19	(2) 项目支出	577,500.00	330,000.00	62	(六) 投资支出	0.00	0.00
20	(二) 事业收入	1,814,580,102.93	1,748,700,910.56	63	(七) 债务还本支出	24,860,000.00	75,240,000.00
21	1. 教育事业预算收入	729,594,355.36	769,980,880.52	64	(八) 其他支出	15,590,524.69	15,651,664.00
22	(1) 财政专户管理资金	729,594,355.36	769,980,880.52	65	其中：利息支出	15,590,524.69	15,651,664.00
23	(2) 其他教育事业预算收入	0.00	0.00	66	捐赠支出	0.00	0.00
24	2. 科研事业预算收入	1,084,985,747.57	978,720,030.04	67	三、本年预算收支差额	963,636,499.13	-55,970,518.22
25	(1) 非同级财政拨款	428,614,135.73	301,732,401.10				
26	其中：中央科研经费拨款	321,285,580.88	187,254,332.70				
27	地方科研经费拨款	107,328,554.85	114,478,068.40				
28	(2) 其他科研事业预算收入	656,371,611.84	676,987,628.94				
29	3. 其他事业预算收入	0.00	0.00				
30	(三) 上级补助预算收入	0.00	0.00				
31	(四) 附属单位上缴预算收入	0.00	0.00				
32	(五) 经营预算收入	0.00	0.00				
33	(六) 债务预算收入	741,249,709.28	75,240,000.00				
34	(七) 非同级财政拨款预算收入	282,361,603.92	273,437,104.30				
35	(1) 中央拨款	38,974,800.27	46,525,497.73				
36	(2) 地方拨款	243,386,803.65	226,911,606.57				
37	(八) 投资预算收入	21,100,400.00	12,879,980.04				
38	(九) 其他预算收入	558,975,922.91	388,992,692.26				
39	1. 租金预算收入	1,774,143.78	1,086,030.88				
40	2. 捐赠预算收入	16,981,131.00	26,348,291.86				
41	3. 利息预算收入	77,275,582.44	75,237,555.67				
42	4. 后勤保障单位净预算收入	0.00	0.00				
43	5. 其他	462,945,065.69	286,320,813.85				

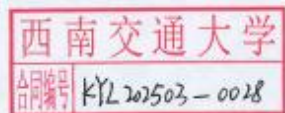
补充资料：

一、全日制本专科生(不含网络和成人教育学生)学费： 本年应收 248,631,867.77 元， 本年实收： 245,460,798.31 元，欠费率： 1.28 (%)
二、全日制研究生学费： 本年应收 146,501,100.00 元， 本年实收： 142,724,286.99 元，欠费率： 2.58 (%)

5.5 三年内承担过同类或类似项目

近年的类似项目情况表

货物名称	模型 AI 开发平台
规格和型号	/
项目名称	思特奇 AI 开发平台建设项目
买方名称	北京思特奇信息技术股份有限公司
买方联系人及电话	吴奕霏 028-85973556
合同价格	80 万元
项目概况及供应商履约情况	提供图像识别、算法改进、数据分析模型构建等大数据与 AI 相关技术服务
备注	/



合同登记编号:

技术开发合同书

项目名称: 思特奇 AI 开发平台建设项目

甲方(委托方): 北京思特奇信息技术股份有限公司

乙方(开发方): 西南交通大学

签订地点: 四川省成都市

签订时间: 2024 年 12 月

中华人民共和国科学技术部印制

依据《中华人民共和国民法典》《中华人民共和国政府采购法》等相关法律法规的规定，合同双方就思特奇运营商 AI 开发平台建设项目项目的技术开发（该项目属技术创新与智能化升级计划*），经平等协商，在真实、自愿、充分地表达各自意愿的基础上，达成如下协议，并由双方共同恪守。

一、★本合同研究开发项目的要求：

（一）技术目标：根据甲方客户需求，战略规划及市场竞争需求，在既定的开发目标和要求下，乙方对甲方的核心数据，业务流程按统一规范进行开发，流程充分考虑甲方的实际需求。

（二）技术内容：包括但不限于图像识别、算法改进、数据分析模型构建等大数据与 AI 相关技术服务。

（三）技术方法和路线：1. 数据准备与预处理。2. 模型训练与优化。3. 模型部署与监控。

二、乙方应在本合同生效后 100 日内向甲方提交研究开发计划。

（一）总体计划：要求：1. 立足项目整体目标，划分开发任务，明确各功能模块的技术实现路径。2. 确定系统开发时间表、资源配置方案及交付节点。3. 确保系统在项目期内按时、按质、按量交付。

（二）阶段性计划：签订合同后 60 天内编制阶段性计划（具体详见附件一）

（三）提交时间和方式：2025 年 6 月

三、发票事宜

3.1 乙方需在双方约定的支付时间前 14 日，向甲方提交符合要求的增值税发票。

3.2 乙方需开具类目 [研发和技术服务·科研经费] 的 [增值税专用发票]，甲方收到发票后，将按约定时间，分期支付。

四、费用事宜

根据双方认可工作成果报告为验收付款要件；共分 4 期完成任务，乙方每期交付的开发要件，必须通过本阶段评审，才可发起下一轮支付流程。

甲方按以下方式支付研究开发经费和报酬：

1、本项目研究开发经费及报酬含税总额为人民币：800,000.00 元（大写：捌拾万元整），其中不含税价款金额 776700 元（大写：柒拾柒万陆仟柒佰元整），税额 23300 元（大写：贰万叁仟叁佰元整）。

2、经费和报酬支付方式及时限（采用以下第 2 种）：

①、一次总付： / 万元，时间： / 。


②、分期支付：贰拾 万元，时间：2025 年 6 月 30 日

二十三、*其他（含中介方的权利、义务、服务费及其支付方式；定金、财产抵押、担保等上述条款未尽事宜）： 无。

二十四、本合同自双方法定代表人或授权代表签字并加盖公章后立即生效。本合同一式四份，甲方执 两 份，乙方执 两 份，具有同等效力。

委 托 方 （ 甲 方 ）	单位名称	北京思特奇信息技术股份有限公司			技术合同专用章	
	法 定 代表人	吴飞舟				
	委 托 代理人	吴奕霏				
	联系人 或经办人	吴奕霏				
	住 所 （通讯地址）	北京市海淀区中关村南大街6号中电信息大厦16层	邮政编 码	100080		2025年3月21日
	电 话	028-85973556	传真	/		
	开户银行	北京银行中关村科技园区支行				
	银行帐号	01090879400120105028968				
	组织机构代码	91110108633062121U				
开 发 方 （ 乙 方 ）	单位名称	西南交通大学			技术合同专用章	
	法定代表人或 委托代理人	何学东				
	项目负责人	黄勤				
	住 所 （通讯地址）	四川省成都市二环路北一段111号	邮政编 码	610031		2025年3月21日
	电 话		传真			


5.6 近一年中任意一个月依法纳税的凭证



中华人民共和国
税 收 完 税 证 明

No. 351015250600215481

填发日期: 2025 年 6 月 20 日 税务机关: 国家税务总局成都市金牛区税务局

纳税人识别号	12100000450752090P		纳税人名称	西南交通大学	
原凭证号	税 种	品 目 名 称	税款所属时期	入(退)库日期	实缴(退)金额
351016250600184793	地方教育附加	增值税地方教育附加	2025-05-01 至 2025-05-31	2025-06-12	18,386.96
351016250600184793	教育费附加	增值税教育费附加	2025-05-01 至 2025-05-31	2025-06-12	27,580.44
351016250600184793	城市维护建设税	市区	2025-05-01 至 2025-05-31	2025-06-12	64,354.37
351016250600184793	增值税	鉴证咨询服务	2025-05-01 至 2025-05-31	2025-06-12	663,946.75
351016250600184793	增值税	食品加工制造业	2025-05-01 至 2025-05-31	2025-06-12	255,401.39
金额合计	(大写)人民币壹佰零贰万玖仟陆佰陆拾玖元玖角壹分				¥1029669.91
		填 票 人 电子税务局	备注: 一般申报 正税 主管税务所(科、分局): 国家税务总局成都市金牛区税务局第二税务所		

妥善保管

收
据
联

交
纳
税
人
作
完
税
证
明

5.7 近一年中任意一个月缴纳社会保障资金的凭证

ICBC 中国工商银行

电子缴税付款
成都西南交大新区支行
业务专用章
6C54C2461006

转账日期: 2025-07-07 凭证字号: 2025070810446378

纳税人全称及纳税人识别号: 西南交通大学 12100000450752090P

付款人全称: 西南交通大学
付款人账号: 4402088509100000675
付款人开户银行: 工行成都西南交大新区支行
小写(合计)金额: RMB13,370.475.84
大写(合计)金额: 壹仟叁佰叁拾柒万零肆佰柒拾伍元捌角肆分

征收机关名称: 国家税务总局成都市税务局第二税务分局
收款国库(银行)名称: 国家金库成都市中心支库
缴款书交易流水号: 12635028
税票号码: 451016250700186166

税(费)种名称
机关事业单位基本养老保险费

所属日期
20250701-20250731

实缴金额
13,370.475.84

第 1 次打印 打印时间: 2025-09-04

5.8 “信用中国”信息查询截图



信用中国
WWW.CREDITCHINA.GOV.CN

报告编号：20250917102929521880M3
生成时间：2025年09月17日 10:29:29

扫一扫

核验码

公共信用信息概览

西南交通大学

登记注册基本信息

基础信息

统一社会信用代码	12100000450752090P	法定代表人	闫学东
举办单位	教育部	审批机关	国家事业单位登记管理局
地址	四川省成都市二环路北一段		

信用信息概要

行政管理	71条	诚实守信	0条
严重失信	0条	经营异常	0条
信用承诺	0条	信用评价	0条
司法判决	0条	其他	0条

报告生成日期	2025年09月17日	报告出具单位	国家公共信用和地理空间信息中心
--------	-------------	--------	-----------------

5.9 不存在控股、管理等关联关系的承诺书

不存在控股、管理等关联关系的承诺书

中国船舶集团有限公司第七一六研究所：

我单位参加贵单位组织的 跨平台智能算法模型转化工具 采购项目，郑重承诺：

在参加本次采购活动不存在与参加本项目的其它供应商单位负责人为同一人或者存在直接控股管理关系。本单位对上述承诺的真实性负责。如有虚假，将依法承担相应责任。

特此说明！

投标人名称：西南交通大学

法定代表人或授权代表：

年 月 日

5.10 不出具第三方审计报告情况说明

不出具第三方审计报告情况说明

中国船舶集团有限公司第七一六研究所：

由于投标人(西南交通大学)为事业单位，无法出具第三方审计报告，故在财务报表方面本单位目前只有资产负债表和预算收入支出表(现金流量表)，暂时无利润表、所有者权益变动表、财务情况说明书，因此涉及的计算均由资产负债表和预算收入支出表数据完成，希望招标单位和评阅专家知晓，如若后续仍需提供特定的报表，本单位将努力进行协商，争取解决。

特此说明！

投标人名称：西南交通大学

法定代表人或授权代表：

年 月 日