GRADING RUBRIC

人工智能与图像识别
课程评分标准

产教融合项目式课程 · 十五组七十五人考核细则（满分 100 分）

课程代码：46050107 · 深度学习与图像识别

主讲：张洪进 · 工程学院

授课班级：2024 机器人 1 班 / 2024 机器人 2 班（15 组 · 75 人）

2026—2027 学年第一学期

目　　录

第一章　总则.....1

1.1　适用范围与总评构成.....1

1.2　评分通则.....1

第二章　平时成绩（20 分）：十二讲课后习题“十二选二”.....3

2.1　评分办法.....3

2.2　十二讲习题与达标标准（验收线）一览.....3

第三章　实验成绩（30 分）：实验二、三、四四档制评分标准.....7

3.1　实验安排与四档制通用判据.....7

3.2　第 1 组　六目拼接 · 匹配决策组.....8

3.3　第 2 组　六目拼接 · 特征质量组.....10

3.4　第 3 组　六目拼接 · 缝合融合组.....11

3.5　第 4 组　六目拼接 · 数据与标签组.....13

3.6　第 5 组　六目拼接 · 测试集与评测组.....15

3.7　第 6 组　六目拼接 · 端到端数据组.....17

3.8　第 7 组　六目拼接 · 端到端训练组.....19

3.9　第 8 组　六目拼接 · 端到端优化组.....20

3.10　第 9 组　水下双目 · 数据组.....22

3.11　第 10 组　水下双目 · 部署组.....24

3.12　第 11 组　水下双目 · 训练组.....26

3.13　第 12 组　水下双目 · 评测组.....28

3.14　第 13 组　数据基座 · 数据结构化组.....29

3.15　第 14 组　数据基座 · 智能问答与图检索组.....31
3.16　第 15 组　数据基座 · 虚假内容识别与对抗训练组.....33
第四章　课程大作业成绩（50 分）：期末项目成果评分标准.....36
4.1　评分通则.....36
4.2　各组满分标准表.....36
4.3　第 1 组　六目拼接 · 匹配决策组.....36
4.4　第 2 组　六目拼接 · 特征质量组.....37
4.5　第 3 组　六目拼接 · 缝合融合组.....37
4.6　第 4 组　六目拼接 · 数据与标签组.....38
4.7　第 5 组　六目拼接 · 测试集与评测组.....38
4.8　第 6 组　六目拼接 · 端到端数据组.....39
4.9　第 7 组　六目拼接 · 端到端训练组.....39
4.10　第 8 组　六目拼接 · 端到端优化组.....40
4.11　第 9 组　水下双目 · 数据组.....40
4.12　第 10 组　水下双目 · 部署组.....41
4.13　第 11 组　水下双目 · 训练组.....41
4.14　第 12 组　水下双目 · 评测组.....41
4.15　第 13 组　数据基座 · 数据结构化组.....42
4.16　第 14 组　数据基座 · 智能问答与图检索组.....42
4.17　第 15 组　数据基座 · 虚假内容识别与对抗训练组.....43
附录　成绩登记表.....44

注：本目录由域代码生成。编辑文档后请在目录上右键选择“更新域”，以保证页码准确。

第一章　总则

1.1　适用范围与总评构成

本标准适用于《深度学习与图像识别》（课程代码 46050107，工程学院，专业选修，2 学分 32 学时）2026—2027 学年第一学期全部教学班（2024 机器人 1 班、2024 机器人 2 班），覆盖 15 个项目小组、75 名学生。评定依据为《人工智能与图像识别——产教融合项目式课程讲义》第一篇（十二讲课后习题）、第三篇（15 份任务书）与《教师指导书》第三章；评定只认证据包（命令全文+完整日志+产物路径三件套）与双台账，不认形容词。

总评满分 100 分，由三部分构成：

表 1-1　总评构成（满分 100 分）

考核环节	分值	评分方式	对应章节
平时成绩	20	第一篇第 1–12 讲课后编程题“十二选二”，每题 10 分达标制（达标即记 10 分）；可提交多题，取最高 2 题计分，上限 20 分	第二章
实验成绩	30	实验二、三、四各 10 分，按“优秀 10 / 良好 9 / 中等 8 / 及格 7”四档制评分；实验一（10 月 12 日）为入场达标实验不计分，必须通过验收方可进入后续实验	第三章
课程大作业	50	期末项目成果（第 19 周周五 2027-1-22 提交），按各组任务书的满分标准表逐项打分：功能/成果主体 20（或按组细分）、证据完整 10、报告质量、协作与台账、加分	第四章
合计	**100**	——	——

1.2　评分通则

1. 证据文化：凡声称“做出来了”，必须同时拿出命令全文、完整日志、产物路径三件套；三件缺任何一件，按“无证据=未完成”处理。所有节点验收以各组维护的双台账与证据包为唯一依据。
2. 如实条款一票否决：虚构训练精度、伪造或拼接证据、改日志时间戳、选择性上报，该节点（该部分）成绩作废并按讲义规则处理，全课程无弹性。

3. 四档制与保底：实验分按四档制执行，“基本不是特别不像话都有 7 分”——通用判据不满足合格线的缓议并补做，补做通过按合格档（7 分）记；大作业成果分按 rubric 严格执行，不设保底。
4. 组内区分：每组报告附分工声明与组内互评，个人得分可在组分基础上浮动 ±10%（主讲核定），防止搭便车。
5. 评分流程与申诉：组内自评（按 rubric 打草分+证据索引）→ 助教初评（逐条写依据）→ 主讲复核（抽 30% 组全查，其余抽查）→ 公示 48 小时 → 申诉窗口（申诉必须指到具体条目与证据，主讲终裁）。

第二章　平时成绩（20 分）：十二讲课后习题“十二选二”

2.1　评分办法

1. 题源：第一篇第 1–12 讲每讲末的“课后编程题（用 AI 实现）”，共 12 道；学生任选 2 道完成（十二选二）。各讲交付物以该讲“交付物”栏为准，普遍为：可运行代码文件＋运行截图（或图件）＋一段不少于 50 字的“我改了 AI 代码里的什么、为什么”说明。
2. 计分：每题按达标制计分——达到该讲“验收线”的全部要求记 10 分，未达标记 0 分，不设中间分。
3. 多交限高：可提交多于 2 道题，按成绩最高的 2 道计入平时分，上限 20 分。
4. 抽问为达标的必要环节：随机抽问须讲清代码原理与自己的修改——“代码是 AI 写的没关系，讲不清楚就是没做”；抽问未通过视为未达标，可另约时间再次验收。
5. 提交与补交：随讲次进度滚动提交，截止时间以课堂通知为准；未达标的题目可按反馈修改后再次申请验收，以最终达标记录为准。

2.2　十二讲习题与达标标准（验收线）一览

下表“达标标准”即各讲课后题的验收线，全部满足记 10 分。

表 2-1　十二讲课后编程题与达标标准

讲次	课后编程题（用 AI 实现）	达标标准（验收线，全部满足记 10 分）
第 1 讲 感知器、线性单元与学习规则	指挥 AI 辅助实现感知器分类器：在 sklearn 鸢尾花数据集上完成两类分类（只用花瓣长度、花瓣宽度两个特征），训练后打印准确率，用 matplotlib 画出样本散点与决策边界；再把数据换成另外两类（versicolor 与 virginica）重跑，观察并解释结果为什么变了。	代码在本机能跑通；随机抽问能不看书讲清 w、b、η 各自的作用和“什么时候更新”；至少有一处自己的修改（换特征、换学习率、换类别都算）。

讲次	课后编程题（用 AI 实现）	达标标准（验收线，全部满足记 10 分）
第 2 讲 BP 神经网络、反向传播与梯度检查	用 numpy 手写 2-2-1 的 MLP（2 输入、2 个 tanh 隐藏神经元、1 个线性输出）拟合 XOR 四个样本：先做数值梯度检查（相对误差小于 1e-5）再开训；平方损失加全批量梯度下降训练到均方误差低于 0.01；画“训练轮数 - 损失”曲线存 png；打印四个输入的预测值。	梯度检查通过且截图在；损失曲线单调下降并收敛到 0.01 以下；四个预测值与真值之差全部小于 0.1；抽问能讲清 δ 如何从输出层倒推回隐藏层；至少一处自己的修改（换激活函数、换学习率、加一个隐藏神经元都算）。
第 3 讲 超参数、权重初始化、模型训练与评估	超参数对比实验：在双月牙（或 XOR）数据集上，用完全相同的设置训练同一个网络三次，每次只改一个超参数（三个学习率如 0.001/0.1/10，或三种初始化如全 0/方差过大/Xavier）；三条 loss 曲线画进同一张图（带图例）并写不少于 50 字结论。	代码能跑，三条曲线对比清晰、结论与曲线对得上（如“lr=10 震荡甚至 nan，lr=0.001 龟速未收敛”）；抽问能讲清 batch、epoch、iteration 三者区别与 batch size 的权衡；至少一处自己的修改（换三种初始化对比、加早停、改 batch size 都算）。
第 4 讲 自编码器	在 sklearn 手写数字集 load_digits（8×8 灰度图）上训练 64→32→16→32→64 自编码器；交付“原图 vs 重建”对比图与全部样本 16 维隐码的 TSNE 二维散点图（按 0–9 数字着色），并写结论：16 维隐码够不够装下数字的“本质”、哪些数字的簇易黏连。	两张图都生成且能讲清——重建图能指出丢了什么细节、保住了什么笔画；TSNE 图上至少能辨认出 2–3 个明显分开的簇并解释 4 和 9、3 和 8 这类数字为何易粘连；能用自己的话说出瓶颈的作用；至少一处自己的修改。
第 5 讲 卷积神经网络原理与前向计算	“卷积核体验器”：对任意一张自选图片（先转灰度）分别应用锐化、模糊、边缘检测（Laplace）三块 3×3 核；滑动窗口必须自己写双重 for 循环实现（禁用 cv2.filter2D 等现成卷积 API），画“原图+三张结果”2×2 四宫格存 png。	四宫格生成且三张结果肉眼可辨（锐化更“脆”、模糊更“糊”、边缘图上轮廓发亮）；能对着图讲清“某个亮点=模板与该处图像块的点积得分高”；能当场回答两问——3×3 核滑 1000×1000 的图要点多少次积、为什么工程实现不用 Python 循环；至少一处自己的修改。

讲次	课后编程题（用 AI 实现）	达标标准（验收线，全部满足记 10 分）
第 6 讲 卷积神经网络的训练与计算模式	实现 LeNet 风格的 CNN（两层卷积－池化堆叠+全连接收尾）在 MNIST 上训练，测试准确率达到 98% 以上；然后自己改动一个超参数（学习率、batch size、是否加数据增强三选一），用完全相同的方式再跑一遍，把两次的训练曲线各画一张图，对比差异并解释原因。	代码在本机跑通且最终准确率不低于 98%；随机抽问能不看资料讲清训练循环五步各自干了什么、model.train() 与 model.eval() 的差别；至少一处自己的修改（换超参数、加 BN、加增强都算）。
第 7 讲 卷积神经网络的演化：加宽、加深与残差	在 CIFAR-10 上做“有无残差”对照实验：搭两个参数量接近的 6 层卷积网络——A 版普通串接（plain），B 版每两层卷积加一条恒等捷径（shortcut）；两者用完全相同的训练设置各训 20 个 epoch，两条训练 loss 曲线与两条测试准确率曲线画在同一张图上对比，写不少于 50 字结论。	两个网络都训完 20 个 epoch 且训练设置完全一致；能不看资料向同学解释“捷径相加为什么能让深层网络训练误差降下来”；至少一处自己的修改（加深到 8 层、改学习率、换 batch size 都算）。
第 8 讲 轻量化网络、全卷积网络与卷积操作模式设计	“模型体检报告”：取 torchvision 预训练的 resnet18 与 mobilenet_v2，在猫狗二分类数据集上（自备几百张或用 CIFAR-10 猫/狗子集）替换分类头各微调 2 个 epoch，测四项指标——参数量（百万）、权重文件大小（MB）、CPU 单张平均推理耗时（测 20 张取平均）、测试集准确率——整理成对比表并写不少于 50 字选型建议。	四项指标齐全且来自自己电脑的真实运行；能不看资料解释“MobileNetV2 为什么更小更快、代价是什么”；至少一处自己的修改（换第三个模型、换输入尺寸、换数据集都算）。
第 9 讲 循环神经网络 RNN	最小 RNN 时序实验（二选一）：(A) 字符级文字接龙——找 5000 字以上自选中文语料训练字符级 RNN（隐层 64–128），训练完从种子字“今”开始采样生成 100 个字；(B) 正弦波预测——前 20 步预测第 21 步，训练后从序列中段递推预测 50 步，画“预测 vs 真值”曲线。代码须注释标清 h 在哪一步更新、“按时间展开”发生在哪里。	代码能跑、loss 曲线整体下降；能对同学讲清 ht=tanh(Wx·xt+Wh·h(t-1)+b) 里每个符号是什么、“展开”是什么意思；能当场回答两问——“RNN 为什么慢”“为什么记不住 50 步前的字”；至少一处自己的修改（换语料、改窗口长度、改隐层大小都算）。

讲次	课后编程题（用 AI 实现）	达标标准（验收线，全部满足记 10 分）
第 10 讲 长短时记忆网络 LSTM 与 GRU	RNN 与 LSTM 对比实验：生成 600 点正弦序列加 8% 噪声，滑窗“前 40 步预测第 41 步”；隐层同为 64、学习率同为 0.001、训练轮数相同，分别训练 nn.RNN 与 nn.LSTM 各一次；各自从序列中段递推预测 60 步，画“真值 vs RNN vs LSTM”对比曲线一张、两条训练 loss 曲线一张，写不少于 100 字结论（参数量与耗时各打印出来对比）。	两张 png 生成且现象与讲义一致（LSTM 贴合真值、RNN 漂移或幅度跑偏）；能对着打印出的参数量回答“LSTM 多在哪”（三扇门多出的那几套矩阵）；能对同学讲清“递推预测”与“单步预测”的区别；至少一处自己的修改（窗口加长、换 GRU 三方对比、去噪声都算）。
第 11 讲 注意力模型	“注意力可视化器”：对句子“小明把书放回书架因为它看完了”，用 transformers 加载 bert-base-chinese，抽取某一层某几个头的自注意力权重画成热图（行列标注分词结果；没有网络条件时可用 AI 生成的 6×6 模拟注意力矩阵代替），找出“它”那一行最亮的格子并解释这一格为什么亮、符合什么语言直觉（指代消解）。	代码能跑出热图；能向同学讲清楚 Q、K、V 在代码里各对应哪个张量、热图的一行和一列各是什么意思；至少有一处自己的修改（换句子、换层、把最亮格子框出来）。
第 12 讲 生成对抗模型	用 AI 辅助训练一个 DCGAN，在 MNIST 上生成手写数字（50 个 epoch，CPU 半小时内可跑完；实在没有条件可改用 diffusers 库调用预训练扩散管线生成数字，交付要求相同）；出 5×10 生成样本网格图与生成器/判别器双 loss 曲线，重点观察是否出现模式崩塌并写观察记录。	代码能跑完并出图；能向同学讲清楚“谁在学谁”——D 在学分辨真假、G 在学 D 的鉴定标准；至少一处自己的修改（噪声维度、加深一层、换学习率）并说出看到的变化。

第三章　实验成绩（30 分）：实验二、三、四四档制评分标准

3.1　实验安排与四档制通用判据

四次实验全班同频：实验一“大模型 agent 模式配置及应用”（2026 年 10 月 12 日，入场达标实验，不计分，必须通过验收方可进入后续实验）；实验二“匹配合适程序并执行本地部署”（11 月 2 日）；实验三“深度网络的模型训练实践”（11 月 23 日）；实验四“模型优化训练实践”（12 月 21 日，合并联调）。实验二、三、四按四档制评分，每场 10 分，共 30 分；档位判据分两层——先按下表通用判据定档，再按分项目实测要点在同档内微调（10 与 9、9 与 8 之间靠它区分）。

表 3-1　实验分四档制通用判据（各实验 10 分）

档位	判据（全部满足）
优秀 10	验收证据三件套齐全且当场可复现；当次实验的全部验收线达成；有至少一处超出命令级步骤的有效自主改进（改参数并给出对比、自写小工具、发现并绕过一个坑）；能向提问者讲清每一步为什么
良好 9	三件套齐全可复现；全部验收线达成；能讲清主要步骤的原理；无自主改进或有但价值有限
中等 8	三件套齐全；验收线达成但有轻微瑕疵（如一项证据格式不规范、一处解释含糊）；当场能复现主要结果
合格 7	核心验收线达成、三件套基本齐全（缺次要件并当场补上）；能讲清“做了什么”，对“为什么”回答部分正确
不合格（缓议）	核心验收线未达成、或证据缺失、或讲不清自己做了什么——补做通过后按合格档（7 分）记

表 3-2　分项目实测要点（判档微调用）

实验	实测要点	助教怎么测
实验二（部署）	六目：九行 [Provider] 与探针 12/12 真实性；水下：50 对冒烟三项数字是否亲手所测；数据基座：本地模型确为本地（拔网线测试）	现场抽问 2 人重跑关键命令；对日志时间戳
实验三（训练）	先 --smoke/先基线纪律；loss 曲线真实性（有无波动、epoch 数与时长是否自洽）；闸门与时段记录	让学生在自己终端重放最后一步；对比曲线与日志数字
实验四（联调）	接口对账签字是否双向；A/B 与台账数字是否一致；遗留问题是否如实登记	抽查一组的三件套与对方组确认单；比对台账

判档顺序：先看通用判据定档，再用实测要点微调；通用判据不满足“合格”线的缓议并补做，补做通过按合格档（7 分）记，再往下压不提倡。实验四当日只判“完成/未完成”，档位评分课后按证据包进行。以下 3.2–3.16 按组给出实验二、三、四各自“验收线＋四档”评分标准，共 45 张小表；每表第一行为该组该实验的验收线（达成门槛），其余四行为档位判据。

3.2 第 1 组 六目拼接 · 匹配决策组

表 3-3 第 1 组 · 实验二（11 月 2 日 · 本地部署）：本机部署，亲眼看见九个头上线（满分 10 分）

档位	评分标准
验收线	① 编译成功末段日志（VS2022 切 Release 生成 360stitcher.sln） ② 九行 [Provider] 加载日志（--benchmark-stitch 基准运行） ③ 模型探针 PASS (12/12)（nine_head_probe.py） （助教测评方式：现场抽问 2 人重跑关键命令、对日志时间戳）
优秀（10 分）	证据三件套（命令全文+完整日志+产物路径）齐全且当场可复现；本实验验收线全部达成；有至少一处超出命令级步骤的有效自主改进（改参数并给出对比、自写小工具、发现并绕过一个坑）；能向提问者讲清每一步为什么；经现场抽问重跑关键命令、核对日志时间戳无异常。
良好（9 分）	三件套齐全且当场可复现；本实验验收线全部达成；能讲清主要步骤的原理；无自主改进，或有但价值有限。
中等（8 分）	三件套齐全；验收线达成但有轻微瑕疵（如一项证据格式不规范、一处解释含糊）；当场能复现主要结果。
及格（7 分）	核心验收线达成；三件套基本齐全（缺次要件并当场补上）；能讲清“做了什么”，对“为什么”的回答部分正确。

表 3-4 第 1 组 · 实验三（11 月 23 日 · 离线训练）：mc 头真训练、真交付、真对比（满分 10 分）

档位	评分标准
验收线	① --smoke 与 --telemetry 两段完整训练日志 ② mc.metrics.json ③ 交付后重跑探针仍 PASS (12/12) ④ A/B 对比结果（ab_compare.py） （助教测评方式：让学生在自己终端重放最后一步，对比曲线与日志数字）

档位	评分标准
优秀（10 分）	证据三件套（命令全文+完整日志+产物路径）齐全且当场可复现；本实验验收线全部达成；有至少一处超出命令级步骤的有效自主改进（改参数并给出对比、自写小工具、发现并绕过一个坑）；能向提问者讲清每一步为什么；先 --smoke/先基线纪律与 loss 曲线真实性经查无异常，能在自己终端重放最后一步。
良好（9 分）	三件套齐全且当场可复现；本实验验收线全部达成；能讲清主要步骤的原理；无自主改进，或有但价值有限。
中等（8 分）	三件套齐全；验收线达成但有轻微瑕疵（如一项证据格式不规范、一处解释含糊）；当场能复现主要结果。
及格（7 分）	核心验收线达成；三件套基本齐全（缺次要件并当场补上）；能讲清“做了什么”，对“为什么”的回答部分正确。

表 3-5　第 1 组 · 实验四（12 月 21 日 · 合并联调）：补齐两头，并用新标签复训 mc（满分 10 分）

档位	评分标准
验收线	① geometry、policy 两头训练日志与 metrics.json ② 交付后探针 PASS (12/12) ③ 三版 mc 对比表（基线/公开仿真标签版/真机标签版）与不少于 100 字结论 （助教测评方式：抽查证据包与对方组确认单，比对台账）
优秀（10 分）	证据三件套（命令全文+完整日志+产物路径）齐全且当场可复现；本实验验收线全部达成；有至少一处超出命令级步骤的有效自主改进（改参数并给出对比、自写小工具、发现并绕过一个坑）；能向提问者讲清每一步为什么；接口对账签字双向、A/B 与台账数字一致，遗留问题如实登记。
良好（9 分）	三件套齐全且当场可复现；本实验验收线全部达成；能讲清主要步骤的原理；无自主改进，或有但价值有限。
中等（8 分）	三件套齐全；验收线达成但有轻微瑕疵（如一项证据格式不规范、一处解释含糊）；当场能复现主要结果。
及格（7 分）	核心验收线达成；三件套基本齐全（缺次要件并当场补上）；能讲清“做了什么”，对“为什么”的回答部分正确。

3.3 第 2 组 六目拼接 · 特征质量组

表 3-6 第 2 组 · 实验二（11 月 2 日 · 本地部署）：本机部署，外加一次“占位模型”对照（满分 10 分）

档位	评分标准
验收线	① 编译、九行 [Provider]、探针三段日志 ② 装占位模型前、后两份 wired 值对照记录（应当完全一致） （助教测评方式：现场抽问 2 人重跑关键命令、对日志时间戳）
优秀（10 分）	证据三件套（命令全文+完整日志+产物路径）齐全且当场可复现；本实验验收线全部达成；有至少一处超出命令级步骤的有效自主改进（改参数并给出对比、自写小工具、发现并绕过一个坑）；能向提问者讲清每一步为什么；经现场抽问重跑关键命令、核对日志时间戳无异常。
良好（9 分）	三件套齐全且当场可复现；本实验验收线全部达成；能讲清主要步骤的原理；无自主改进，或有但价值有限。
中等（8 分）	三件套齐全；验收线达成但有轻微瑕疵（如一项证据格式不规范、一处解释含糊）；当场能复现主要结果。
及格（7 分）	核心验收线达成；三件套基本齐全（缺次要件并当场补上）；能讲清“做了什么”，对“为什么”的回答部分正确。

表 3-7 第 2 组 · 实验三（11 月 23 日 · 离线训练）：temporal_state 真训练交付，图像头交设计（满分 10 分）

档位	评分标准
验收线	① temporal_state 训练日志、metrics.json、探针 PASS、A/B 结果 ② 两个图像头设计文档各一页 ③ 两段 --smoke 完整日志（公开 patch 先行训练） （助教测评方式：让学生在自己终端重放最后一步，对比曲线与日志数字）
优秀（10 分）	证据三件套（命令全文+完整日志+产物路径）齐全且当场可复现；本实验验收线全部达成；有至少一处超出命令级步骤的有效自主改进（改参数并给出对比、自写小工具、发现并绕过一个坑）；能向提问者讲清每一步为什么；先 --smoke/先基线纪律与 loss 曲线真实性经查无异常，能在自己终端重放最后一步。
良好（9 分）	三件套齐全且当场可复现；本实验验收线全部达成；能讲清主要步骤的原理；无自主改进，或有但价值有限。
中等（8 分）	三件套齐全；验收线达成但有轻微瑕疵（如一项证据格式不规范、一处解释含糊）；当场能复现主要结果。

档位	评分标准
及格（7 分）	核心验收线达成；三件套基本齐全（缺次要件并当场补上）；能讲清“做了什么”，对“为什么”的回答部分正确。

表 3-8　第 2 组 · 实验四（12 月 21 日 · 合并联调）：真机 patch 联调，图像头迁移转真（满分 10 分）

档位	评分标准
验收线	① 抽检 20 张真机 patch 的记录 ② 两个图像头训练日志、metrics.json ③ 交付后探针 PASS (12/12) ④ 公开版/真机版 A/B 结果 （助教测评方式：抽查证据包与对方组确认单，比对台账）
优秀（10 分）	证据三件套（命令全文+完整日志+产物路径）齐全且当场可复现；本实验验收线全部达成；有至少一处超出命令级步骤的有效自主改进（改参数并给出对比、自写小工具、发现并绕过一个坑）；能向提问者讲清每一步为什么；接口对账签字双向、A/B 与台账数字一致，遗留问题如实登记。
良好（9 分）	三件套齐全且当场可复现；本实验验收线全部达成；能讲清主要步骤的原理；无自主改进，或有但价值有限。
中等（8 分）	三件套齐全；验收线达成但有轻微瑕疵（如一项证据格式不规范、一处解释含糊）；当场能复现主要结果。
及格（7 分）	核心验收线达成；三件套基本齐全（缺次要件并当场补上）；能讲清“做了什么”，对“为什么”的回答部分正确。

3.4　第 3 组　六目拼接 · 缝合融合组

表 3-9　第 3 组 · 实验二（11 月 2 日 · 本地部署）：本机部署，顺手记下基线数字（满分 10 分）

档位	评分标准
验收线	① 编译、九行 [Provider]、探针三段日志 ② 三遍基线计时记录表（luminance 增益补偿段与 blend 融合段，同一基准至少跑三遍取稳定读数） （助教测评方式：现场抽问 2 人重跑关键命令、对日志时间戳）

档位	评分标准
优秀（10 分）	证据三件套（命令全文+完整日志+产物路径）齐全且当场可复现；本实验验收线全部达成；有至少一处超出命令级步骤的有效自主改进（改参数并给出对比、自写小工具、发现并绕过一个坑）；能向提问者讲清每一步为什么；经现场抽问重跑关键命令、核对日志时间戳无异常。
良好（9 分）	三件套齐全且当场可复现；本实验验收线全部达成；能讲清主要步骤的原理；无自主改进，或有但价值有限。
中等（8 分）	三件套齐全；验收线达成但有轻微瑕疵（如一项证据格式不规范、一处解释含糊）；当场能复现主要结果。
及格（7 分）	核心验收线达成；三件套基本齐全（缺次要件并当场补上）；能讲清“做了什么”，对“为什么”的回答部分正确。

表 3-10　第 3 组 · 实验三（11 月 23 日 · 离线训练）：gain_field 真训练交付，seam 头交设计（满分 10 分）

档位	评分标准
验收线	① gain_field 训练日志、metrics.json、探针 PASS、A/B 结果（与实验二基线对照） ② seam_uncertainty 设计文档一页 ③ seam 全链路 --smoke 日志（随机数数据，目录留证） （助教测评方式：让学生在自己终端重放最后一步，对比曲线与日志数字）
优秀（10 分）	证据三件套（命令全文+完整日志+产物路径）齐全且当场可复现；本实验验收线全部达成；有至少一处超出命令级步骤的有效自主改进（改参数并给出对比、自写小工具、发现并绕过一个坑）；能向提问者讲清每一步为什么；先 --smoke/先基线纪律与 loss 曲线真实性经查无异常，能在自己终端重放最后一步。
良好（9 分）	三件套齐全且当场可复现；本实验验收线全部达成；能讲清主要步骤的原理；无自主改进，或有但价值有限。
中等（8 分）	三件套齐全；验收线达成但有轻微瑕疵（如一项证据格式不规范、一处解释含糊）；当场能复现主要结果。
及格（7 分）	核心验收线达成；三件套基本齐全（缺次要件并当场补上）；能讲清“做了什么”，对“为什么”的回答部分正确。

表 3-11　第 3 组 · 实验四（12 月 21 日 · 合并联调）：八头齐，fusion 正式版压轴上场（合并联调）（满分 10 分）

档位	评分标准
验收线	① 上游（第 1、2 组）交付确认记录 ② 放行后 fusion 训练日志、metrics.json ③ 交付后探针 PASS (12/12) ④ A/B 结果与彩排版/正式版两版记录 （助教测评方式：抽查证据包与对方组确认单，比对台账）
优秀（10 分）	证据三件套（命令全文+完整日志+产物路径）齐全且当场可复现；本实验验收线全部达成；有至少一处超出命令级步骤的有效自主改进（改参数并给出对比、自写小工具、发现并绕过一个坑）；能向提问者讲清每一步为什么；接口对账签字双向、A/B 与台账数字一致，遗留问题如实登记。
良好（9 分）	三件套齐全且当场可复现；本实验验收线全部达成；能讲清主要步骤的原理；无自主改进，或有但价值有限。
中等（8 分）	三件套齐全；验收线达成但有轻微瑕疵（如一项证据格式不规范、一处解释含糊）；当场能复现主要结果。
及格（7 分）	核心验收线达成；三件套基本齐全（缺次要件并当场补上）；能讲清“做了什么”，对“为什么”的回答部分正确。

3.5　第 4 组　六目拼接 · 数据与标签组

表 3-12　第 4 组 · 实验二（11 月 2 日 · 本地部署）：先在基准图上跑通全流程（满分 10 分）

档位	评分标准
验收线	① 遥测体检 RESULT: PASS 日志（check_telemetry.py） ② 五张基准图跑遥测的完整日志 ③ 三段标签文件与校验输出（UDIS-D 公开仿真标签 match/geometry 各不少于 200 行并过校验） （助教测评方式：现场抽问 2 人重跑关键命令、对日志时间戳）
优秀（10 分）	证据三件套（命令全文+完整日志+产物路径）齐全且当场可复现；本实验验收线全部达成；有至少一处超出命令级步骤的有效自主改进（改参数并给出对比、自写小工具、发现并绕过一个坑）；能向提问者讲清每一步为什么；经现场抽问重跑关键命令、核对日志时间戳无异常。
良好（9 分）	三件套齐全且当场可复现；本实验验收线全部达成；能讲清主要步骤的原理；无自主改进，或有但价值有限。

档位	评分标准
中等（8 分）	三件套齐全；验收线达成但有轻微瑕疵（如一项证据格式不规范、一处解释含糊）；当场能复现主要结果。
及格（7 分）	核心验收线达成；三件套基本齐全（缺次要件并当场补上）；能讲清“做了什么”，对“为什么”的回答部分正确。

表 3-13　第 4 组 · 实验三（11 月 23 日 · 离线训练）：真机场地采集，交出第一条验收线（满分 10 分）

档位	评分标准
验收线	① 视频清单（场景/时长）与抽帧统计 ② 遥测校验 PASS 日志 ③ 三段标签各不少于 200 行并过校验 （助教测评方式：让学生在自己终端重放最后一步，对比曲线与日志数字）
优秀（10 分）	证据三件套（命令全文+完整日志+产物路径）齐全且当场可复现；本实验验收线全部达成；有至少一处超出命令级步骤的有效自主改进（改参数并给出对比、自写小工具、发现并绕过一个坑）；能向提问者讲清每一步为什么；先 --smoke/先基线纪律与 loss 曲线真实性经查无异常，能在自己终端重放最后一步。
良好（9 分）	三件套齐全且当场可复现；本实验验收线全部达成；能讲清主要步骤的原理；无自主改进，或有但价值有限。
中等（8 分）	三件套齐全；验收线达成但有轻微瑕疵（如一项证据格式不规范、一处解释含糊）；当场能复现主要结果。
及格（7 分）	核心验收线达成；三件套基本齐全（缺次要件并当场补上）；能讲清“做了什么”，对“为什么”的回答部分正确。

表 3-14　第 4 组 · 实验四（12 月 21 日 · 合并联调）：扩量与质检，交出最终数据家底（满分 10 分）

档位	评分标准
验收线	① 真机标签各不少于 500 行的校验通过日志 ② 字段分布图与异常样本处理清单（逐条有名有姓有理由） ③ 数据质量报告 PDF 与台账登记 ④ 公开仿真标签与真机标签对比说明 （助教测评方式：抽查证据包与对方组确认单，比对台账）

档位	评分标准
优秀（10 分）	证据三件套（命令全文+完整日志+产物路径）齐全且当场可复现；本实验验收线全部达成；有至少一处超出命令级步骤的有效自主改进（改参数并给出对比、自写小工具、发现并绕过一个坑）；能向提问者讲清每一步为什么；接口对账签字双向、A/B 与台账数字一致，遗留问题如实登记。
良好（9 分）	三件套齐全且当场可复现；本实验验收线全部达成；能讲清主要步骤的原理；无自主改进，或有但价值有限。
中等（8 分）	三件套齐全；验收线达成但有轻微瑕疵（如一项证据格式不规范、一处解释含糊）；当场能复现主要结果。
及格（7 分）	核心验收线达成；三件套基本齐全（缺次要件并当场补上）；能讲清“做了什么”，对“为什么”的回答部分正确。

3.6 第 5 组 六目拼接 · 测试集与评测组

表 3-15 第 5 组 · 实验二（11 月 2 日 · 本地部署）：两条路线都部署，两种探针都点亮（满分 10 分）

档位	评分标准
验收线	① 两条路线部署日志各一段（360stitcher 与 360stitcher_e2e） ② 九头探针 PASS (12/12) ③ 引擎探针 PASS (5/5) （助教测评方式：现场抽问 2 人重跑关键命令、对日志时间戳）
优秀（10 分）	证据三件套（命令全文+完整日志+产物路径）齐全且当场可复现；本实验验收线全部达成；有至少一处超出命令级步骤的有效自主改进（改参数并给出对比、自写小工具、发现并绕过一个坑）；能向提问者讲清每一步为什么；经现场抽问重跑关键命令、核对日志时间戳无异常。
良好（9 分）	三件套齐全且当场可复现；本实验验收线全部达成；能讲清主要步骤的原理；无自主改进，或有但价值有限。
中等（8 分）	三件套齐全；验收线达成但有轻微瑕疵（如一项证据格式不规范、一处解释含糊）；当场能复现主要结果。
及格（7 分）	核心验收线达成；三件套基本齐全（缺次要件并当场补上）；能讲清“做了什么”，对“为什么”的回答部分正确。

表 3-16　第 5 组 · 实验三（11 月 23 日 · 离线训练）：冻结测试集 v1，基线入账，回归全绿（满分 10 分）

档位	评分标准
验收线	① 测试集 v1 的 SHA-256 哈希清单 ② 基线指标台账条目（初始化耗时、稳态均值、稳态 P95） ③ run_regression.ps1 全绿一遍的完整日志 （助教测评方式：让学生在自己终端重放最后一步，对比曲线与日志数字）
优秀（10 分）	证据三件套（命令全文+完整日志+产物路径）齐全且当场可复现；本实验验收线全部达成；有至少一处超出命令级步骤的有效自主改进（改参数并给出对比、自写小工具、发现并绕过一个坑）；能向提问者讲清每一步为什么；先 --smoke/先基线纪律与 loss 曲线真实性经查无异常，能在自己终端重放最后一步。
良好（9 分）	三件套齐全且当场可复现；本实验验收线全部达成；能讲清主要步骤的原理；无自主改进，或有但价值有限。
中等（8 分）	三件套齐全；验收线达成但有轻微瑕疵（如一项证据格式不规范、一处解释含糊）；当场能复现主要结果。
及格（7 分）	核心验收线达成；三件套基本齐全（缺次要件并当场补上）；能讲清“做了什么”，对“为什么”的回答部分正确。

表 3-17　第 5 组 · 实验四（12 月 21 日 · 合并联调）：台账 v2 与困难样本，裁判的终场哨（满分 10 分）

档位	评分标准
验收线	① 台账 v2 全表（逐行核对记录） ② 测试集 v2 哈希清单（易/中/难分桶） ③ 各组难桶补测记录 ④ 评分汇总表定稿（交叉复核） （助教测评方式：抽查证据包与对方组确认单，比对台账）
优秀（10 分）	证据三件套（命令全文+完整日志+产物路径）齐全且当场可复现；本实验验收线全部达成；有至少一处超出命令级步骤的有效自主改进（改参数并给出对比、自写小工具、发现并绕过一个坑）；能向提问者讲清每一步为什么；接口对账签字双向、A/B 与台账数字一致，遗留问题如实登记。
良好（9 分）	三件套齐全且当场可复现；本实验验收线全部达成；能讲清主要步骤的原理；无自主改进，或有但价值有限。
中等（8 分）	三件套齐全；验收线达成但有轻微瑕疵（如一项证据格式不规范、一处解释含糊）；当场能复现主要结果。

档位	评分标准
及格（7 分）	核心验收线达成；三件套基本齐全（缺次要件并当场补上）；能讲清“做了什么”，对“为什么”的回答部分正确。

3.7 第 6 组 六目拼接 · 端到端数据组

表 3-18 第 6 组 · 实验二（11 月 2 日 · 本地部署）：合成数据全流程，32+8 对先干跑一遍（满分 10 分）

档位	评分标准
验收线	① 32+8 对生成日志与目录清单（make_demo_data.py） ② infer_overlay.jpg 推理自测可视化 ③ 真实测试批次说明（来源与用途，只放 val） （助教测评方式：现场抽问 2 人重跑关键命令、对日志时间戳）
优秀（10 分）	证据三件套（命令全文+完整日志+产物路径）齐全且当场可复现；本实验验收线全部达成；有至少一处超出命令级步骤的有效自主改进（改参数并给出对比、自写小工具、发现并绕过一个坑）；能向提问者讲清每一步为什么；经现场抽问重跑关键命令、核对日志时间戳无异常。
良好（9 分）	三件套齐全且当场可复现；本实验验收线全部达成；能讲清主要步骤的原理；无自主改进，或有但价值有限。
中等（8 分）	三件套齐全；验收线达成但有轻微瑕疵（如一项证据格式不规范、一处解释含糊）；当场能复现主要结果。
及格（7 分）	核心验收线达成；三件套基本齐全（缺次要件并当场补上）；能讲清“做了什么”，对“为什么”的回答部分正确。

表 3-19 第 6 组 · 实验三（11 月 23 日 · 离线训练）：首批公开数据集（UDIS-D），50 对起步（满分 10 分）

档位	评分标准
验收线	① 不少于 50 对、覆盖不少于 2 个场景的清单与统计（build_dataset.py） ② 数据集卡片一页（如实标注“公开数据版”） ③ 10 对抽检可视化与交接组 7 的记录 （助教测评方式：让学生在自己终端重放最后一步，对比曲线与日志数字）

档位	评分标准
优秀（10 分）	证据三件套（命令全文+完整日志+产物路径）齐全且当场可复现；本实验验收线全部达成；有至少一处超出命令级步骤的有效自主改进（改参数并给出对比、自写小工具、发现并绕过一个坑）；能向提问者讲清每一步为什么；先 --smoke/先基线纪律与 loss 曲线真实性经查无异常，能在自己终端重放最后一步。
良好（9 分）	三件套齐全且当场可复现；本实验验收线全部达成；能讲清主要步骤的原理；无自主改进，或有但价值有限。
中等（8 分）	三件套齐全；验收线达成但有轻微瑕疵（如一项证据格式不规范、一处解释含糊）；当场能复现主要结果。
及格（7 分）	核心验收线达成；三件套基本齐全（缺次要件并当场补上）；能讲清“做了什么”，对“为什么”的回答部分正确。

表 3-20　第 6 组 · 实验四（12 月 21 日 · 合并联调）：数据扩量与 benchmark 初版，双线收官（满分 10 分）

档位	评分标准
验收线	① 扩量不少于 150 对、不少于 3 个场景的统计与 v2 卡片 ② benchmark_report.json（逐对明细+均值/P95） ③ benchmark.py 的 README（命令与期望输出） （助教测评方式：抽查证据包与对方组确认单，比对台账）
优秀（10 分）	证据三件套（命令全文+完整日志+产物路径）齐全且当场可复现；本实验验收线全部达成；有至少一处超出命令级步骤的有效自主改进（改参数并给出对比、自写小工具、发现并绕过一个坑）；能向提问者讲清每一步为什么；接口对账签字双向、A/B 与台账数字一致，遗留问题如实登记。
良好（9 分）	三件套齐全且当场可复现；本实验验收线全部达成；能讲清主要步骤的原理；无自主改进，或有但价值有限。
中等（8 分）	三件套齐全；验收线达成但有轻微瑕疵（如一项证据格式不规范、一处解释含糊）；当场能复现主要结果。
及格（7 分）	核心验收线达成；三件套基本齐全（缺次要件并当场补上）；能讲清“做了什么”，对“为什么”的回答部分正确。

3.8 第 7 组 六目拼接 · 端到端训练组

表 3-21 第 7 组 · 实验二（11 月 2 日 · 本地部署）：本机 CPU 冒烟，三十分钟走通全链路（满分 10 分）

档位	评分标准
验收线	① CPU 冒烟训练日志（--epochs 2 --batch 4） ② 导出自检输出（数值自检误差远小于 1e-3） ③ 交付 [OK] 提示（deliver_model.bat） ④ 引擎探针 PASS (5/5) （助教测评方式：现场抽问 2 人重跑关键命令、对日志时间戳）
优秀（10 分）	证据三件套（命令全文+完整日志+产物路径）齐全且当场可复现；本实验验收线全部达成；有至少一处超出命令级步骤的有效自主改进（改参数并给出对比、自写小工具、发现并绕过一个坑）；能向提问者讲清每一步为什么；经现场抽问重跑关键命令、核对日志时间戳无异常。
良好（9 分）	三件套齐全且当场可复现；本实验验收线全部达成；能讲清主要步骤的原理；无自主改进，或有但价值有限。
中等（8 分）	三件套齐全；验收线达成但有轻微瑕疵（如一项证据格式不规范、一处解释含糊）；当场能复现主要结果。
及格（7 分）	核心验收线达成；三件套基本齐全（缺次要件并当场补上）；能讲清“做了什么”，对“为什么”的回答部分正确。

表 3-22 第 7 组 · 实验三（11 月 23 日 · 离线训练）：4080 真训练，全书唯一的 200 epoch（满分 10 分）

档位	评分标准
验收线	① 200 epoch 完整训练日志（含运行者与起止时间） ② ropstitch.metrics.json ③ 交付后探针 PASS (5/5) ④ 基线 vs v1 的 A/B 对比（组 5 冻结测试集） （助教测评方式：让学生在自己终端重放最后一步，对比曲线与日志数字）
优秀（10 分）	证据三件套（命令全文+完整日志+产物路径）齐全且当场可复现；本实验验收线全部达成；有至少一处超出命令级步骤的有效自主改进（改参数并给出对比、自写小工具、发现并绕过一个坑）；能向提问者讲清每一步为什么；先 --smoke/先基线纪律与 loss 曲线真实性经查无异常，能在自己终端重放最后一步。
良好（9 分）	三件套齐全且当场可复现；本实验验收线全部达成；能讲清主要步骤的原理；无自主改进，或有但价值有限。

档位	评分标准
中等（8 分）	三件套齐全；验收线达成但有轻微瑕疵（如一项证据格式不规范、一处解释含糊）；当场能复现主要结果。
及格（7 分）	核心验收线达成；三件套基本齐全（缺次要件并当场补上）；能讲清“做了什么”，对“为什么”的回答部分正确。

表 3-23　第 7 组 · 实验四（12 月 21 日 · 合并联调）：全量真实数据微调，v2 收官（满分 10 分）

档位	评分标准
验收线	① v2 训练日志与 ropstitch.metrics.json ② 交付后探针 PASS (5/5) ③ 基线/v1/v2 三列 A/B 对比表与不少于 100 字结论（v2 不如 v1 时如实回退并记录） （助教测评方式：抽查证据包与对方组确认单，比对台账）
优秀（10 分）	证据三件套（命令全文+完整日志+产物路径）齐全且当场可复现；本实验验收线全部达成；有至少一处超出命令级步骤的有效自主改进（改参数并给出对比、自写小工具、发现并绕过一个坑）；能向提问者讲清每一步为什么；接口对账签字双向、A/B 与台账数字一致，遗留问题如实登记。
良好（9 分）	三件套齐全且当场可复现；本实验验收线全部达成；能讲清主要步骤的原理；无自主改进，或有但价值有限。
中等（8 分）	三件套齐全；验收线达成但有轻微瑕疵（如一项证据格式不规范、一处解释含糊）；当场能复现主要结果。
及格（7 分）	核心验收线达成；三件套基本齐全（缺次要件并当场补上）；能讲清“做了什么”，对“为什么”的回答部分正确。

3.9　第 8 组　六目拼接 · 端到端优化组

表 3-24　第 8 组 · 实验二（11 月 2 日 · 本地部署）：引擎模式开起来，逐帧计时基线立起来（满分 10 分）

档位	评分标准
验收线	① 完整 [BenchmarkFrame] 逐帧日志（30 帧以上） ② 逐帧计时基线表（均值+P95，分模型加载/预处理/引擎推理/后处理渲染） ③ 与组 6 benchmark.py 交叉验证记录 （助教测评方式：现场抽问 2 人重跑关键命令、对日志时间戳）

档位	评分标准
优秀（10 分）	证据三件套（命令全文+完整日志+产物路径）齐全且当场可复现；本实验验收线全部达成；有至少一处超出命令级步骤的有效自主改进（改参数并给出对比、自写小工具、发现并绕过一个坑）；能向提问者讲清每一步为什么；经现场抽问重跑关键命令、核对日志时间戳无异常。
良好（9 分）	三件套齐全且当场可复现；本实验验收线全部达成；能讲清主要步骤的原理；无自主改进，或有但价值有限。
中等（8 分）	三件套齐全；验收线达成但有轻微瑕疵（如一项证据格式不规范、一处解释含糊）；当场能复现主要结果。
及格（7 分）	核心验收线达成；三件套基本齐全（缺次要件并当场补上）；能讲清“做了什么”，对“为什么”的回答部分正确。

表 3-25　第 8 组 · 实验三（11 月 23 日 · 离线训练）：复现组 7 公开版 v1，指标对得上账（满分 10 分）

档位	评分标准
验收线	① 探针 PASS (5/5) ② 复现对照表（自测值 vs metrics.json，偏差在 1e-3 量级以内） ③ 复现报告一页（命令、输出、对照） （助教测评方式：让学生在自己终端重放最后一步，对比曲线与日志数字）
优秀（10 分）	证据三件套（命令全文+完整日志+产物路径）齐全且当场可复现；本实验验收线全部达成；有至少一处超出命令级步骤的有效自主改进（改参数并给出对比、自写小工具、发现并绕过一个坑）；能向提问者讲清每一步为什么；先 --smoke/先基线纪律与 loss 曲线真实性经查无异常，能在自己终端重放最后一步。
良好（9 分）	三件套齐全且当场可复现；本实验验收线全部达成；能讲清主要步骤的原理；无自主改进，或有但价值有限。
中等（8 分）	三件套齐全；验收线达成但有轻微瑕疵（如一项证据格式不规范、一处解释含糊）；当场能复现主要结果。
及格（7 分）	核心验收线达成；三件套基本齐全（缺次要件并当场补上）；能讲清“做了什么”，对“为什么”的回答部分正确。

表 3-26　第 8 组 · 实验四（12 月 21 日 · 合并联调）：三档权衡表与 batch 方案，招牌交付收官（满分 10 分）

档位	评分标准
验收线	① fp32/fp16/int8 三档权衡表（含原始数据，同一冻结测试集同一输入） ② batch 推理方案文档 ③ 量化版交付与探针记录（或不达标不交付的结论说明，劣化<10% 为硬线） ④ 性能预算书复盘 （助教测评方式：抽查证据包与对方组确认单，比对台账）
优秀（10 分）	证据三件套（命令全文+完整日志+产物路径）齐全且当场可复现；本实验验收线全部达成；有至少一处超出命令级步骤的有效自主改进（改参数并给出对比、自写小工具、发现并绕过一个坑）；能向提问者讲清每一步为什么；接口对账签字双向、A/B 与台账数字一致，遗留问题如实登记。
良好（9 分）	三件套齐全且当场可复现；本实验验收线全部达成；能讲清主要步骤的原理；无自主改进，或有但价值有限。
中等（8 分）	三件套齐全；验收线达成但有轻微瑕疵（如一项证据格式不规范、一处解释含糊）；当场能复现主要结果。
及格（7 分）	核心验收线达成；三件套基本齐全（缺次要件并当场补上）；能讲清“做了什么”，对“为什么”的回答部分正确。

3.10　第 9 组　水下双目 · 数据组

表 3-27　第 9 组 · 实验二（11 月 2 日 · 本地部署）：抽帧去冗一条龙，切分防住泄漏（满分 10 分）

档位	评分标准
验收线	① 工具链三条版本号输出（ffmpeg / scenedetect / cv2） ② 抽帧与去冗日志（抽了多少、去掉多少、留了多少） ③ 按场景/日期切分的训练/验证划分清单与数据卡（同一天同一场景只进一边） （助教测评方式：现场抽问 2 人重跑关键命令、对日志时间戳）
优秀（10 分）	证据三件套（命令全文+完整日志+产物路径）齐全且当场可复现；本实验验收线全部达成；有至少一处超出命令级步骤的有效自主改进（改参数并给出对比、自写小工具、发现并绕过一个坑）；能向提问者讲清每一步为什么；经现场抽问重跑关键命令、核对日志时间戳无异常。

档位	评分标准
良好（9 分）	三件套齐全且当场可复现；本实验验收线全部达成；能讲清主要步骤的原理；无自主改进，或有但价值有限。
中等（8 分）	三件套齐全；验收线达成但有轻微瑕疵（如一项证据格式不规范、一处解释含糊）；当场能复现主要结果。
及格（7 分）	核心验收线达成；三件套基本齐全（缺次要件并当场补上）；能讲清“做了什么”，对“为什么”的回答部分正确。

表 3-28　第 9 组 · 实验三（11 月 23 日 · 离线训练）：数据组也要懂训练——亲眼看 loss 下降（满分 10 分）

档位	评分标准
验收线	① 两段训练日志（1–2 个 epoch 小微调，含每个 epoch 的 loss 数字） ② 两条 loss 曲线（不同数据子集对比） ③ 数据子集差异说明 （助教测评方式：让学生在自己终端重放最后一步，对比曲线与日志数字）
优秀（10 分）	证据三件套（命令全文+完整日志+产物路径）齐全且当场可复现；本实验验收线全部达成；有至少一处超出命令级步骤的有效自主改进（改参数并给出对比、自写小工具、发现并绕过一个坑）；能向提问者讲清每一步为什么；先 --smoke/先基线纪律与 loss 曲线真实性经查无异常，能在自己终端重放最后一步。
良好（9 分）	三件套齐全且当场可复现；本实验验收线全部达成；能讲清主要步骤的原理；无自主改进，或有但价值有限。
中等（8 分）	三件套齐全；验收线达成但有轻微瑕疵（如一项证据格式不规范、一处解释含糊）；当场能复现主要结果。
及格（7 分）	核心验收线达成；三件套基本齐全（缺次要件并当场补上）；能讲清“做了什么”，对“为什么”的回答部分正确。

表 3-29　第 9 组 · 实验四（12 月 21 日 · 合并联调）：换配比、调阈值，用验证 EPE 选出更好的一口粮（满分 10 分）

档位	评分标准
验收线	① 配比设计说明（2–3 种配比，验证集固定） ② 各配比训练日志（训练参数与实验三完全一致） ③ 验证 EPE 对比表（配比/帧数/EPE/耗时）与不少于 100 字推荐结论 （助教测评方式：抽查证据包与对方组确认单，比对台账）

档位	评分标准
优秀（10 分）	证据三件套（命令全文+完整日志+产物路径）齐全且当场可复现；本实验验收线全部达成；有至少一处超出命令级步骤的有效自主改进（改参数并给出对比、自写小工具、发现并绕过一个坑）；能向提问者讲清每一步为什么；接口对账签字双向、A/B 与台账数字一致，遗留问题如实登记。
良好（9 分）	三件套齐全且当场可复现；本实验验收线全部达成；能讲清主要步骤的原理；无自主改进，或有但价值有限。
中等（8 分）	三件套齐全；验收线达成但有轻微瑕疵（如一项证据格式不规范、一处解释含糊）；当场能复现主要结果。
及格（7 分）	核心验收线达成；三件套基本齐全（缺次要件并当场补上）；能讲清“做了什么”，对“为什么”的回答部分正确。

3.11 第 10 组 水下双目 · 部署组

表 3-30 第 10 组 · 实验二（11 月 2 日 · 本地部署）：装上 4080，跑通第一对图（本组主力）（满分 10 分）

档位	评分标准
验收线	① run_demo.py 成功日志（看到视差图输出） ② UWStereo 50 对零样本冒烟三项数字表（EPE/单对耗时/显存峰值） ③ 降级链处置记录（如有：vitl→vits→IGEV++，每级记卡点与数字） （助教测评方式：现场抽问 2 人重跑关键命令、对日志时间戳）
优秀（10 分）	证据三件套（命令全文+完整日志+产物路径）齐全且当场可复现；本实验验收线全部达成；有至少一处超出命令级步骤的有效自主改进（改参数并给出对比、自写小工具、发现并绕过一个坑）；能向提问者讲清每一步为什么；经现场抽问重跑关键命令、核对日志时间戳无异常。
良好（9 分）	三件套齐全且当场可复现；本实验验收线全部达成；能讲清主要步骤的原理；无自主改进，或有但价值有限。
中等（8 分）	三件套齐全；验收线达成但有轻微瑕疵（如一项证据格式不规范、一处解释含糊）；当场能复现主要结果。
及格（7 分）	核心验收线达成；三件套基本齐全（缺次要件并当场补上）；能讲清“做了什么”，对“为什么”的回答部分正确。

表 3-31　第 10 组 · 实验三（11 月 23 日 · 离线训练）：微调护航——你的环境要能“载训练”（满分 10 分）

档位	评分标准
验收线	① 小微调成功日志（1–2 个 epoch） ② 标注峰值的显存曲线（按秒采样） ③ 吞吐数字（对/分钟）与一周可吃数据量的排班建议 （助教测评方式：让学生在自己终端重放最后一步，对比曲线与日志数字）
优秀（10 分）	证据三件套（命令全文+完整日志+产物路径）齐全且当场可复现；本实验验收线全部达成；有至少一处超出命令级步骤的有效自主改进（改参数并给出对比、自写小工具、发现并绕过一个坑）；能向提问者讲清每一步为什么；先 --smoke/先基线纪律与 loss 曲线真实性经查无异常，能在自己终端重放最后一步。
良好（9 分）	三件套齐全且当场可复现；本实验验收线全部达成；能讲清主要步骤的原理；无自主改进，或有但价值有限。
中等（8 分）	三件套齐全；验收线达成但有轻微瑕疵（如一项证据格式不规范、一处解释含糊）；当场能复现主要结果。
及格（7 分）	核心验收线达成；三件套基本齐全（缺次要件并当场补上）；能讲清“做了什么”，对“为什么”的回答部分正确。

表 3-32　第 10 组 · 实验四（12 月 21 日 · 合并联调）：三组旋钮+ONNX 初探，给出权衡建议（满分 10 分）

档位	评分标准
验收线	① 三组对比表（批量 1/2/4、分辨率 1.0x/0.5x、FP32/FP16，各记 EPE+耗时+显存） ② ONNX 导出记录（能导则测速，导不出则写清卡点与报错原文） ③ “精度－速度”权衡建议一页（每个推荐都有数字） （助教测评方式：抽查证据包与对方组确认单，比对台账）
优秀（10 分）	证据三件套（命令全文+完整日志+产物路径）齐全且当场可复现；本实验验收线全部达成；有至少一处超出命令级步骤的有效自主改进（改参数并给出对比、自写小工具、发现并绕过一个坑）；能向提问者讲清每一步为什么；接口对账签字双向、A/B 与台账数字一致，遗留问题如实登记。
良好（9 分）	三件套齐全且当场可复现；本实验验收线全部达成；能讲清主要步骤的原理；无自主改进，或有但价值有限。
中等（8 分）	三件套齐全；验收线达成但有轻微瑕疵（如一项证据格式不规范、一处解释含糊）；当场能复现主要结果。

档位	评分标准
及格（7 分）	核心验收线达成；三件套基本齐全（缺次要件并当场补上）；能讲清“做了什么”，对“为什么”的回答部分正确。

3.12 第 11 组 水下双目 · 训练组

表 3-33 第 11 组 · 实验二（11 月 2 日 · 本地部署）：让一个 batch 走完全程（满分 10 分）

档位	评分标准
验收线	① batch 形状打印（左目/右目/视差真值的形状与数值范围） ② 初始 loss 数字 ③ 反向与优化器更新成功日志 ④ 显存快照（nvidia-smi） （助教测评方式：现场抽问 2 人重跑关键命令、对日志时间戳）
优秀（10 分）	证据三件套（命令全文+完整日志+产物路径）齐全且当场可复现；本实验验收线全部达成；有至少一处超出命令级步骤的有效自主改进（改参数并给出对比、自写小工具、发现并绕过一个坑）；能向提问者讲清每一步为什么；经现场抽问重跑关键命令、核对日志时间戳无异常。
良好（9 分）	三件套齐全且当场可复现；本实验验收线全部达成；能讲清主要步骤的原理；无自主改进，或有但价值有限。
中等（8 分）	三件套齐全；验收线达成但有轻微瑕疵（如一项证据格式不规范、一处解释含糊）；当场能复现主要结果。
及格（7 分）	核心验收线达成；三件套基本齐全（缺次要件并当场补上）；能讲清“做了什么”，对“为什么”的回答部分正确。

表 3-34 第 11 组 · 实验三（11 月 23 日 · 离线训练）：先立基线，再开微调（本组主力）（满分 10 分）

档位	评分标准
验收线	① 零样本基线 EPE 数字（含初步模型 v0 参照） ② 闸门申请与批准记录（epoch 数不超过 5、时段、配置快照） ③ 完整训练日志与逐 epoch checkpoint 存档 ④ 基线/微调对比表、loss 与验证 EPE 双曲线（标出最优 checkpoint） （助教测评方式：让学生在自己终端重放最后一步，对比曲线与日志数字）

档位	评分标准
优秀（10 分）	证据三件套（命令全文+完整日志+产物路径）齐全且当场可复现；本实验验收线全部达成；有至少一处超出命令级步骤的有效自主改进（改参数并给出对比、自写小工具、发现并绕过一个坑）；能向提问者讲清每一步为什么；先 --smoke/先基线纪律与 loss 曲线真实性经查无异常，能在自己终端重放最后一步。
良好（9 分）	三件套齐全且当场可复现；本实验验收线全部达成；能讲清主要步骤的原理；无自主改进，或有但价值有限。
中等（8 分）	三件套齐全；验收线达成但有轻微瑕疵（如一项证据格式不规范、一处解释含糊）；当场能复现主要结果。
及格（7 分）	核心验收线达成；三件套基本齐全（缺次要件并当场补上）；能讲清“做了什么”，对“为什么”的回答部分正确。

表 3-35　第 11 组 · 实验四（12 月 21 日 · 合并联调）：一次只动一个旋钮（满分 10 分）

档位	评分标准
验收线	① 各组配置快照与训练日志（固定随机种子，其余参数与实验三完全一致） ② EPE 对比表（配置差异/基线 EPE/各组 EPE/训练耗时） ③ 不少于 100 字结论段 （助教测评方式：抽查证据包与对方组确认单，比对台账）
优秀（10 分）	证据三件套（命令全文+完整日志+产物路径）齐全且当场可复现；本实验验收线全部达成；有至少一处超出命令级步骤的有效自主改进（改参数并给出对比、自写小工具、发现并绕过一个坑）；能向提问者讲清每一步为什么；接口对账签字双向、A/B 与台账数字一致，遗留问题如实登记。
良好（9 分）	三件套齐全且当场可复现；本实验验收线全部达成；能讲清主要步骤的原理；无自主改进，或有但价值有限。
中等（8 分）	三件套齐全；验收线达成但有轻微瑕疵（如一项证据格式不规范、一处解释含糊）；当场能复现主要结果。
及格（7 分）	核心验收线达成；三件套基本齐全（缺次要件并当场补上）；能讲清“做了什么”，对“为什么”的回答部分正确。

3.13 第 12 组 水下双目 · 评测组

表 3-36 第 12 组 · 实验二（11 月 2 日 · 本地部署）：一键打通“推理→视差→指标”（满分 10 分）

档位	评分标准
验收线	① 一键脚本（tasks/u12_eval/，推理→保存→算分） ② 单对与 10 对批量日志 ③ 视差可视化色图样例（float 存盘防截断） （助教测评方式：现场抽问 2 人重跑关键命令、对日志时间戳）
优秀（10 分）	证据三件套（命令全文+完整日志+产物路径）齐全且当场可复现；本实验验收线全部达成；有至少一处超出命令级步骤的有效自主改进（改参数并给出对比、自写小工具、发现并绕过一个坑）；能向提问者讲清每一步为什么；经现场抽问重跑关键命令、核对日志时间戳无异常。
良好（9 分）	三件套齐全且当场可复现；本实验验收线全部达成；能讲清主要步骤的原理；无自主改进，或有但价值有限。
中等（8 分）	三件套齐全；验收线达成但有轻微瑕疵（如一项证据格式不规范、一处解释含糊）；当场能复现主要结果。
及格（7 分）	核心验收线达成；三件套基本齐全（缺次要件并当场补上）；能讲清“做了什么”，对“为什么”的回答部分正确。

表 3-37 第 12 组 · 实验三（11 月 23 日 · 离线训练）：两版权重同场比武，每模型不少于 50 对（满分 10 分）

档位	评分标准
验收线	① 评测配置（分辨率/对齐方式/子集名单/随机种子，随表交出） ② 两模型批量日志（每模型不少于 50 对，EPE/bad-1/2/4px/D1） ③ 基线数字表（总分+分桶+每版权重一句“体检结论”） （助教测评方式：让学生在自己终端重放最后一步，对比曲线与日志数字）
优秀（10 分）	证据三件套（命令全文+完整日志+产物路径）齐全且当场可复现；本实验验收线全部达成；有至少一处超出命令级步骤的有效自主改进（改参数并给出对比、自写小工具、发现并绕过一个坑）；能向提问者讲清每一步为什么；先 --smoke/先基线纪律与 loss 曲线真实性经查无异常，能在自己终端重放最后一步。
良好（9 分）	三件套齐全且当场可复现；本实验验收线全部达成；能讲清主要步骤的原理；无自主改进，或有但价值有限。

档位	评分标准
中等（8 分）	三件套齐全；验收线达成但有轻微瑕疵（如一项证据格式不规范、一处解释含糊）；当场能复现主要结果。
及格（7 分）	核心验收线达成；三件套基本齐全（缺次要件并当场补上）；能讲清“做了什么”，对“为什么”的回答部分正确。

表 3-38　第 12 组 · 实验四（12 月 21 日 · 合并联调）：扫阈值、找最差桶、当个好裁判（满分 10 分）

档位	评分标准
验收线	① 阈值 τ 扫描表与保留率 - 精度曲线（含推荐 τ 与理由） ② 按浑浊度/视差分桶的数字表（找出最差桶） ③ 不少于 20 张误差可视化图的人工归因记录（AI 不可代看） （助教测评方式：抽查证据包与对方组确认单，比对台账）
优秀（10 分）	证据三件套（命令全文+完整日志+产物路径）齐全且当场可复现；本实验验收线全部达成；有至少一处超出命令级步骤的有效自主改进（改参数并给出对比、自写小工具、发现并绕过一个坑）；能向提问者讲清每一步为什么；接口对账签字双向、A/B 与台账数字一致，遗留问题如实登记。
良好（9 分）	三件套齐全且当场可复现；本实验验收线全部达成；能讲清主要步骤的原理；无自主改进，或有但价值有限。
中等（8 分）	三件套齐全；验收线达成但有轻微瑕疵（如一项证据格式不规范、一处解释含糊）；当场能复现主要结果。
及格（7 分）	核心验收线达成；三件套基本齐全（缺次要件并当场补上）；能讲清“做了什么”，对“为什么”的回答部分正确。

3.14　第 13 组　数据基座 · 数据结构化组

表 3-39　第 13 组 · 实验二（11 月 2 日 · 本地部署）：解析环境装通，10 页流水线跑通（满分 10 分）

档位	评分标准
验收线	① 两个解析库版本号输出（PyMuPDF/pdfplumber 三选二） ② 10 页“抽取 - 锚点 - 登记”流水线完整终端日志 ③ 结构化 JSON 产物路径与统计（几张表、几段公式、几条锚点、几个回退区域） （助教测评方式：现场抽问 2 人重跑关键命令、对日志时间戳）

档位	评分标准
优秀（10 分）	证据三件套（命令全文+完整日志+产物路径）齐全且当场可复现；本实验验收线全部达成；有至少一处超出命令级步骤的有效自主改进（改参数并给出对比、自写小工具、发现并绕过一个坑）；能向提问者讲清每一步为什么；经现场抽问重跑关键命令、核对日志时间戳无异常。
良好（9 分）	三件套齐全且当场可复现；本实验验收线全部达成；能讲清主要步骤的原理；无自主改进，或有但价值有限。
中等（8 分）	三件套齐全；验收线达成但有轻微瑕疵（如一项证据格式不规范、一处解释含糊）；当场能复现主要结果。
及格（7 分）	核心验收线达成；三件套基本齐全（缺次要件并当场补上）；能讲清“做了什么”，对“为什么”的回答部分正确。

表 3-40　第 13 组 · 实验三（11 月 23 日 · 离线训练）：自己造数据，训一个“认版面”的小模型（满分 10 分）

档位	评分标准
验收线	① 标注集（不少于 200 条，三类都有像样样本）与标注规范一页（AI 预标、人逐条终审留痕） ② 训练日志与 loss 曲线（训练/验证按页切分） ③ 混淆矩阵与宏平均分析（哪两类易混） （助教测评方式：让学生在自己终端重放最后一步，对比曲线与日志数字）
优秀（10 分）	证据三件套（命令全文+完整日志+产物路径）齐全且当场可复现；本实验验收线全部达成；有至少一处超出命令级步骤的有效自主改进（改参数并给出对比、自写小工具、发现并绕过一个坑）；能向提问者讲清每一步为什么；先 --smoke/先基线纪律与 loss 曲线真实性经查无异常，能在自己终端重放最后一步。
良好（9 分）	三件套齐全且当场可复现；本实验验收线全部达成；能讲清主要步骤的原理；无自主改进，或有但价值有限。
中等（8 分）	三件套齐全；验收线达成但有轻微瑕疵（如一项证据格式不规范、一处解释含糊）；当场能复现主要结果。
及格（7 分）	核心验收线达成；三件套基本齐全（缺次要件并当场补上）；能讲清“做了什么”，对“为什么”的回答部分正确。

表 3-41　第 13 组 · 实验四（12 月 21 日 · 合并联调）：在线大模型救疑难表格，规则法当对照组（满分 10 分）

档位	评分标准
验收线	① 疑难表格清单（不少于 5 张，登记“难在哪”） ② 提示词全文与 few-shot 样例、API 调用记录（次数与花费账单） ③ 规则法 vs 在线增强法对比表（列名对/数值对/合并关系对、耗时、花费） ④ 不少于 100 字结论段 （助教测评方式：抽查证据包与对方组确认单，比对台账）
优秀（10 分）	证据三件套（命令全文+完整日志+产物路径）齐全且当场可复现；本实验验收线全部达成；有至少一处超出命令级步骤的有效自主改进（改参数并给出对比、自写小工具、发现并绕过一个坑）；能向提问者讲清每一步为什么；接口对账签字双向、A/B 与台账数字一致，遗留问题如实登记。
良好（9 分）	三件套齐全且当场可复现；本实验验收线全部达成；能讲清主要步骤的原理；无自主改进，或有但价值有限。
中等（8 分）	三件套齐全；验收线达成但有轻微瑕疵（如一项证据格式不规范、一处解释含糊）；当场能复现主要结果。
及格（7 分）	核心验收线达成；三件套基本齐全（缺次要件并当场补上）；能讲清“做了什么”，对“为什么”的回答部分正确。

3.15　第 14 组　数据基座 · 智能问答与图检索组

表 3-42　第 14 组 · 实验二（11 月 2 日 · 本地部署）：本地小模型上岗，一句话问数跑通（满分 10 分）

档位	评分标准
验收线	① ollama list 输出（模型下载完成再调用） ② 四个精简 CSV 建库行数统计 ③ 三个样例问答录屏（明细/聚合/列名说明各一，含 SQL 与查询结果；白名单与只读两道闸、自设注入测试通过） （助教测评方式：现场抽问 2 人重跑关键命令、对日志时间戳）
优秀（10 分）	证据三件套（命令全文+完整日志+产物路径）齐全且当场可复现；本实验验收线全部达成；有至少一处超出命令级步骤的有效自主改进（改参数并给出对比、自写小工具、发现并绕过一个坑）；能向提问者讲清每一步为什么；经现场抽问重跑关键命令、核对日志时间戳无异常。

档位	评分标准
良好（9 分）	三件套齐全且当场可复现；本实验验收线全部达成；能讲清主要步骤的原理；无自主改进，或有但价值有限。
中等（8 分）	三件套齐全；验收线达成但有轻微瑕疵（如一项证据格式不规范、一处解释含糊）；当场能复现主要结果。
及格（7 分）	核心验收线达成；三件套基本齐全（缺次要件并当场补上）；能讲清“做了什么”，对“为什么”的回答部分正确。

表 3-43　第 14 组 · 实验三（11 月 23 日 · 离线训练）：小图上真训练——两层 R-GCN 做链路预测（满分 10 分）

档位	评分标准
验收线	① 建图统计（50–100 节点、四类实体、至少三类关系，登记节点数/边数） ② 训练日志与 loss 曲线（切边口径写清、特征归一化） ③ 验证集 AUC 或 Hit@10 数字及其含义解释 （助教测评方式：让学生在自己终端重放最后一步，对比曲线与日志数字）
优秀（10 分）	证据三件套（命令全文+完整日志+产物路径）齐全且当场可复现；本实验验收线全部达成；有至少一处超出命令级步骤的有效自主改进（改参数并给出对比、自写小工具、发现并绕过一个坑）；能向提问者讲清每一步为什么；先 --smoke/先基线纪律与 loss 曲线真实性经查无异常，能在自己终端重放最后一步。
良好（9 分）	三件套齐全且当场可复现；本实验验收线全部达成；能讲清主要步骤的原理；无自主改进，或有但价值有限。
中等（8 分）	三件套齐全；验收线达成但有轻微瑕疵（如一项证据格式不规范、一处解释含糊）；当场能复现主要结果。
及格（7 分）	核心验收线达成；三件套基本齐全（缺次要件并当场补上）；能讲清“做了什么”，对“为什么”的回答部分正确。

表 3-44　第 14 组 · 实验四（12 月 21 日 · 合并联调）：本地 vs 在线——30 题上一张对比表（满分 10 分）

档位	评分标准
验收线	① 固定 30 题逐题明细表（标准答案独立编写，至少 10 题原创边界题；本地与在线各一份） ② 三列对比表（准确率/中位延迟/API 花费） ③ 不少于 100 字建议段与花费账单 （助教测评方式：抽查证据包与对方组确认单，比对台账）

档位	评分标准
优秀（10 分）	证据三件套（命令全文+完整日志+产物路径）齐全且当场可复现；本实验验收线全部达成；有至少一处超出命令级步骤的有效自主改进（改参数并给出对比、自写小工具、发现并绕过一个坑）；能向提问者讲清每一步为什么；接口对账签字双向、A/B 与台账数字一致，遗留问题如实登记。
良好（9 分）	三件套齐全且当场可复现；本实验验收线全部达成；能讲清主要步骤的原理；无自主改进，或有但价值有限。
中等（8 分）	三件套齐全；验收线达成但有轻微瑕疵（如一项证据格式不规范、一处解释含糊）；当场能复现主要结果。
及格（7 分）	核心验收线达成；三件套基本齐全（缺次要件并当场补上）；能讲清“做了什么”，对“为什么”的回答部分正确。

3.16 第 15 组 数据基座 · 虚假内容识别与对抗训练组

表 3-45 第 15 组 · 实验二（11 月 2 日 · 本地部署）：规则当第一道闸，词袋小模型当基线（满分 10 分）

档位	评分标准
验收线	① 三条规则的可配置检查代码（数值越界/单位错/时间戳矛盾） ② 造假样本不少于 100 条（AI 按配额造假并记账） ③ 规则法与 TF-IDF+逻辑回归两版基线指标表（准确率/查准/查全/F1 同表） （助教测评方式：现场抽问 2 人重跑关键命令、对日志时间戳）
优秀（10 分）	证据三件套（命令全文+完整日志+产物路径）齐全且当场可复现；本实验验收线全部达成；有至少一处超出命令级步骤的有效自主改进（改参数并给出对比、自写小工具、发现并绕过一个坑）；能向提问者讲清每一步为什么；经现场抽问重跑关键命令、核对日志时间戳无异常。
良好（9 分）	三件套齐全且当场可复现；本实验验收线全部达成；能讲清主要步骤的原理；无自主改进，或有但价值有限。
中等（8 分）	三件套齐全；验收线达成但有轻微瑕疵（如一项证据格式不规范、一处解释含糊）；当场能复现主要结果。
及格（7 分）	核心验收线达成；三件套基本齐全（缺次要件并当场补上）；能讲清“做了什么”，对“为什么”的回答部分正确。

表 3-46　第 15 组 · 实验三（11 月 23 日 · 离线训练）：主力会战——500 条数据集、两个模型、一份自查（满分 10 分）

档位	评分标准
验收线	① 数据集 v1（CSV，不少于 500 条：假 250 易/中/难分层+真 250，分层随机洗牌，双金标逐条填好） ② TF-IDF 基线与 embedding 小网络两版训练日志及 loss/F1 曲线（报宏观 F1） ③ 泄漏自查重叠率数字（近似重复与事实级重叠两条口径） ④ 对照 0.88→0.74 的心得段（不少于 100 字） （助教测评方式：让学生在自己终端重放最后一步，对比曲线与日志数字）
优秀（10 分）	证据三件套（命令全文+完整日志+产物路径）齐全且当场可复现；本实验验收线全部达成；有至少一处超出命令级步骤的有效自主改进（改参数并给出对比、自写小工具、发现并绕过一个坑）；能向提问者讲清每一步为什么；先 --smoke/先基线纪律与 loss 曲线真实性经查无异常，能在自己终端重放最后一步。
良好（9 分）	三件套齐全且当场可复现；本实验验收线全部达成；能讲清主要步骤的原理；无自主改进，或有但价值有限。
中等（8 分）	三件套齐全；验收线达成但有轻微瑕疵（如一项证据格式不规范、一处解释含糊）；当场能复现主要结果。
及格（7 分）	核心验收线达成；三件套基本齐全（缺次要件并当场补上）；能讲清“做了什么”，对“为什么”的回答部分正确。

表 3-47　第 15 组 · 实验四（12 月 21 日 · 合并联调）：在线判官复核，蒸馏回流再上一分（满分 10 分）

档位	评分标准
验收线	① 考卷封存记录（先封存 50 条后训练，时间戳为证） ② 纯本地/纯在线/混合三种判定的宏观 F1 与成本表 ③ 难样本 3 倍权重回流前后对比 ④ 不少于 100 字结论段与 API 账单 （助教测评方式：抽查证据包与对方组确认单，比对台账）
优秀（10 分）	证据三件套（命令全文+完整日志+产物路径）齐全且当场可复现；本实验验收线全部达成；有至少一处超出命令级步骤的有效自主改进（改参数并给出对比、自写小工具、发现并绕过一个坑）；能向提问者讲清每一步为什么；接口对账签字双向、A/B 与台账数字一致，遗留问题如实登记。

档位	评分标准
良好（9 分）	三件套齐全且当场可复现；本实验验收线全部达成；能讲清主要步骤的原理；无自主改进，或有但价值有限。
中等（8 分）	三件套齐全；验收线达成但有轻微瑕疵（如一项证据格式不规范、一处解释含糊）；当场能复现主要结果。
及格（7 分）	核心验收线达成；三件套基本齐全（缺次要件并当场补上）；能讲清“做了什么”，对“为什么”的回答部分正确。

第四章　课程大作业成绩（50 分）：期末项目成果评分标准

4.1　评分通则

课程大作业即期末项目成果，第 19 周周五（2027 年 1 月 22 日）完全提交离场。评分严格按讲义第三篇各组任务书“防糊弄指南与满分标准”一节的满分标准表逐条打分，不打印象分：助教按表逐行写“给了几分、依据是什么”，主讲复核。三条执行纪律：

1. rubric 逐条打分：功能达成、证据完整、报告质量、协作与台账逐项给分，加分项据实给分；每组满分均为 50 分（各组栏目划分略有差异，以各组下表为准）。
2. 如实条款一票否决：虚构训练精度、伪造证据，该部分计零并按讲义规则处理。
3. 组内区分：报告附分工声明与组内互评，个人得分可在组分基础上浮动 ±10%（主讲核定），防止搭便车。

提交物四件（15 组通用）：项目报告（PDF，8–15 页，摘要/背景/方法/实验/结论/分工声明六段）、代码包（zip 含 README，别人拿到能跑）、证据包（按四个实验节点分四包，命令+日志+产物三件套）、AI 使用声明一页；各组另有组内特有交付物（如六目各组的三头 ONNX 模型），以各组任务书提交清单为准。

4.2　各组满分标准表

以下 15 张表与各组任务书中的满分标准表一致，为第四章评分的执行依据。

4.3　第 1 组　六目拼接 · 匹配决策组

表 4-1　第 1 组成果满分标准（50 分制）

分值	项目	细分标准（怎么拿满）
20	功能达成	实验三交付 mc 头并完成 A/B（8 分）；实验四交付 geometry、policy 两头（7 分）；每次交付后重跑探针仍 PASS (12/12)（5 分）
10	证据完整	四次实验证据包三件套齐全、可复现；缺一件扣 5 分
8	报告质量	组报告按模板 8–15 页，方法讲得清、图表齐全、结论有 A/B 数据支撑

分值	项目	细分标准（怎么拿满）
7	协作与台账	排班执行有记录；组 5 双台账中本组条目无缺失；求助记录规范
5	加分	如：为 mc 设计新的统计特征并用 A/B 证明提升；对三个头做联合分析并给出数据

4.4 第 2 组 六目拼接 · 特征质量组

表 4-2 第 2 组成果满分标准（50 分制）

分值	项目	细分标准（怎么拿满）
20	功能达成	实验三交付 temporal_state 并完成 A/B（8 分）；实验四交付两个图像头（7 分）；每次交付后探针仍 PASS (12/12)（5 分）
10	证据完整	四次实验证据包齐全可复现，含规格说明书与两页设计文档；缺一件扣 5 分
8	报告质量	组报告按模板 8–15 页；规格说明书与设计文档被助教判定“能照着实现”
7	协作与台账	排班执行有记录；组 5 双台账中本组条目无缺失；求助记录规范
5	加分	如：时序窗口长度 T 的消融对比有数据；给软化头设计新的评测角度并用 A/B 验证

4.5 第 3 组 六目拼接 · 缝合融合组

表 4-3 第 3 组成果满分标准（50 分制）

分值	项目	细分标准（怎么拿满）
20	功能达成	实验三交付 gain_field（公开仿真标签版）并完成 A/B（8 分）；实验四放行后交付 fusion 正式版，附彩排版与正式版两版记录（7 分）；每次交付后探针仍 PASS (12/12)（5 分）
10	证据完整	四次实验证据包齐全可复现，基线计时三遍记录在包；缺一件扣 5 分
8	报告质量	组报告按模板 8–15 页；seam 设计文档达到“能照着实现”的水平
7	协作与台账	排班执行有记录；对第 1、2 组交付进度的跟踪记录完整；组 5 台账无缺失
5	加分	如：seam 设计文档被总体组采纳进实现；对 fusion 做八头联动分析并给出数据

4.6 第 4 组 六目拼接 · 数据与标签组

表 4-4 第 4 组成果满分标准（50 分制）

分值	项目	细分标准（怎么拿满）
20	功能达成	实验三达成验收线：match/geometry 各不少于 200 行并过校验（10 分）；实验四扩量至各不少于 500 行且质量报告合格（10 分）
10	证据完整	采集证据包齐全可追溯：视频清单、抽帧统计、两道校验的 PASS 日志；缺一件扣 5 分
8	报告质量	组报告按模板 8–15 页；质量报告字段分布讲得清、异常处理有据
7	协作与台账	排班与现场分工有记录；向第 1–3 组交付有交接记录；组 5 数据量台账无缺失
5	加分	如：质量报告发现并追溯出遥测或标签管线的 bug 且被总体组采纳修复；给下游提供“按场景切片”的标签使用建议

4.7 第 5 组 六目拼接 · 测试集与评测组

表 4-5 第 5 组成果满分标准（50 分制）

分值	项目	细分标准（怎么拿满）
20	功能达成	实验二两条路线部署验收双双通过（8 分）；实验三测试集 v1 冻结+基线入账+回归全绿（7 分）；实验四台账 v2 全覆盖+困难样本测试集 v2（5 分）
10	证据完整	四次实验证据包齐全：哈希清单、部署与探针日志、全绿回归日志、A/B 归档；缺一件扣 5 分
8	报告质量	组报告按模板 8–15 页；评测指标说明书讲得清，评分表可回溯、有交叉复核记录
7	协作与台账	排班有记录；各组交付登记无缺失、无迟到；评分汇总按时产出并用于离场验收会
5	加分	如：设计“易/中/难”分桶并给出各组在难桶上的差距分析；为回归脚本补充新检查项且被采纳

4.8 第 6 组 六目拼接 · 端到端数据组

表 4-6 第 6 组成果满分标准（50 分制）

分值	项目	细分标准（怎么拿满）
20	功能达成	实验二合成 32+8 对全流程跑通（5 分）；实验三真实数据不少于 50 对+数据集卡片合格（8 分）；实验四扩量不少于 150 对且 benchmark.py 对 ropstitch.onnx 出数（7 分）
10	证据完整	四次实验证据包齐全：生成日志、推理自测可视化、清单统计、benchmark_report.json；缺一件扣 5 分
8	报告质量	组报告按模板 8–15 页；卡片四要素齐全，benchmark 报告均值与 P95 并列、结论有数据
7	协作与台账	向组 7/组 8 交接有记录；数据量按时入组 5 双台账；求助记录规范
5	加分	如：用数据增强扩量并用 A/B 证明不伤指标；产出“困难样本建议清单”交组 5 进难桶

4.9 第 7 组 六目拼接 · 端到端训练组

表 4-7 第 7 组成果满分标准（50 分制）

分值	项目	细分标准（怎么拿满）
20	功能达成	实验二 CPU 冒烟全链路+探针 PASS (5/5)（5 分）；实验三 4080 真训练交付 v1 且 A/B 优于基线（10 分）；实验四 v2 微调完成并给出对比结论或如实回退（5 分）
10	证据完整	三段训练日志、metrics.json、交付后探针 PASS、A/B 报告，缺一件扣 5 分
8	报告质量	组报告按模板 8–15 页；训练曲线、超参数表、A/B 表齐全，结论有数据支撑
7	协作与台账	GPU 时段预约与使用登记无违规；交付按时入组 5 台账；与组 6/组 8 交接有记录
5	加分	如：学习率/batch 小实验（2–3 组对比曲线有结论）；对损失函数做改进且 A/B 有证据

4.10 第 8 组 六目拼接 · 端到端优化组

表 4-8 第 8 组成果满分标准（50 分制）

分值	项目	细分标准（怎么拿满）
20	功能达成	实验二逐帧计时基线完整（5 分）；实验三 v1 复现对账通过（5 分）；实验四三档权衡表+batch 方案合格（10 分）
10	证据完整	四次实验证据包齐全：预算书、逐帧日志、复现对账记录、三档原始数据；缺一件扣 5 分
8	报告质量	组报告按模板 8–15 页；权衡表讲得清、结论有数据、预算复盘诚实
7	协作与台账	GPU 时段申请与使用有记录；与组 7/组 6 交接顺畅；结论按时入组 5 台账
5	加分	如：batch 方案做出原型并给出实测收益；单次引擎调用实测压进 300 ms；发现并反馈引擎管线的耗时 bug

4.11 第 9 组 水下双目 · 数据组

表 4-9 第 9 组成果满分标准（50 分制）

分值	项目	细分标准（怎么拿满）
20	功能达成	实验一体检报告+30 帧人工核对记录（5 分）；实验二抽帧去冗全流程+防泄漏划分清单（5 分）；实验三资产包数据小微调、亲眼看 loss 下降（5 分）；实验四 2–3 种配比对比+推荐配比（5 分）
10	证据完整	四次实验证据包三件套齐全、可复现；缺一件扣 5 分
8	报告质量	组报告按模板 8–15 页，数据卡清晰、配比结论有 EPE 数据支撑
7	协作与台账	排班执行有记录；4080 时段预约与用量记录无缺失；求助记录规范
5	加分	如：自研更聪明的去冗指标并用数字证明有效；为浑浊/低照度场景设计专门过滤规则并被训练组采用；把体检做成一键复跑的工具

4.12 第 10 组 水下双目 · 部署组

表 4-10 第 10 组成果满分标准（50 分制）

分值	项目	细分标准（怎么拿满）
20	功能达成	实验一配置清单经逐条核实（3 分）；实验二 run_demo.py 跑通 +50 对冒烟三项数字齐全（10 分）；实验三微调护航显存曲线与吞吐（2 分）；实验四三组对比+权衡建议（5 分）
10	证据完整	四次实验证据包三件套齐全、可复现；缺一件扣 5 分
8	报告质量	组报告按模板 8–15 页，部署手册可复现、权衡建议有数据支撑
7	协作与台账	4080 时段台账无缺失；值守与代跑记录规范；求助记录规范
5	加分	如：ONNX 导出真跑通并给出可用性结论；写一键环境校验脚本被三组采用；FP16 精度损失的定量分析

4.13 第 11 组 水下双目 · 训练组

表 4-11 第 11 组成果满分标准（50 分制）

分值	项目	细分标准（怎么拿满）
20	功能达成	实验一配置说明书+逐项核对（3 分）；实验二数据加载+前向/反向冒烟（5 分）；实验三零样本基线+闸门内短程微调+对照曲线（10 分）；实验四 1–2 项超参 2–3 组对比（2 分）
10	证据完整	四次实验证据包三件套齐全、可复现；缺一件扣 5 分
8	报告质量	组报告按模板 8–15 页，对照结论有基线数据支撑、曲线图齐全
7	协作与台账	闸门申请与批准记录齐全；4080 时段台账无缺失；求助记录规范
5	加分	如：数据增强消融实验；按浑浊度分层训练子集并给出 EPE 对比；冻结主干只调局部的对照

4.14 第 12 组 水下双目 · 评测组

表 4-12 第 12 组成果满分标准（50 分制）

分值	项目	细分标准（怎么拿满）
20	功能达成	实验一指标白皮书+手工验算对账成功（5 分）；实验二“推理→视差→指标”一键链路（5 分）；实验三每模型不少于 50 对的基线数字表（含分桶）（8 分）；实验四 τ 扫描曲线+最差桶归因（2 分）

分值	项目	细分标准（怎么拿满）
10	证据完整	四次实验证据包三件套齐全、可复现；缺一件扣 5 分
8	报告质量	组报告按模板 8–15 页，评测口径交代清楚、结论有基线与分桶数据支撑
7	协作与台账	为三组提供统一评测口径并有记录；4080 时段台账无缺失；求助记录规范
5	加分	如：自动浑浊度分桶的量化指标被数据组采用；评测封装成一条命令被三组复用；提出并论证一个新指标

4.15 第 13 组 数据基座 · 数据结构化组

表 4-13 第 13 组成果满分标准（50 分制）

分值	项目	细分标准（怎么拿满）
15	流水线功能	实验二 10 页流水线跑通且产物字段齐全（8 分）；实验四疑难表格二次抽取并出对比报告（7 分）
10	质量报告与如实度	每个数字与产物对得上账；回退、存疑、人工核对计数如实登记，无一件虚报
8	小模型训练	标注不少于 200 条且终审留痕；loss 曲线齐全；混淆矩阵与宏平均有分析
7	报告质量	组报告按模板 8–15 页，方法讲得清、图表齐全、结论有数据支撑
5	协作与台账	排班执行有记录；标注分工与终审留痕；求助记录规范
5	加分	如：跨页表格自动合并逻辑并用数字证明有效；为扫描件接 OCR 兜底并对比效果；把流水线做成一键复跑工具

4.16 第 14 组 数据基座 · 智能问答与图检索组

表 4-14 第 14 组成果满分标准（50 分制）

分值	项目	细分标准（怎么拿满）
15	问数 demo	实验二建库+一句话问数跑通并录屏（8 分）；白名单工具与只读检查齐备、通过自设注入测试（7 分）
10	R-GCN 实验	建图规范、训练真实完成、loss 曲线齐全，报出 AUC 或 Hit@10 并有分析
8	测试集与评估	30 题独立标准答案+至少 10 题原创边界题；每题对错明细可对账

分值	项目	细分标准（怎么拿满）
7	报告质量	组报告按模板 8–15 页，方法讲得清、图表齐全、结论有数据支撑
5	协作与台账	排班执行有记录；录屏与明细归档完整；求助记录规范
5	加分	如：给 demo 加“缺条件追问”并演示；加语义检索提升表名/列名命中率；对 R-GCN 做关系类型消融实验

4.17 第 15 组 数据基座 · 虚假内容识别与对抗训练组

表 4-15 第 15 组成果满分标准（50 分制）

分值	项目	细分标准（怎么拿满）
13	数据集与配方	不少于 500 条分层齐全、洗牌可验证、双金标一致（8 分）；配方不少于 10 个家族且每家族有可复现示例（5 分）
12	分类器与指标	规则+TF-IDF 基线（5 分）；embedding 小网络与回流对比（7 分）；宏观 F1 口径正确、曲线齐全
8	泄漏自查与如实度	给出重叠率数字与检查方法；抽检留痕（抽了哪 100 条、判对多少）；“人工确认计数”如实登记——虚报一处，本栏清零
7	报告质量	组报告按模板 8–15 页，方法讲得清、图表齐全、结论有数据支撑
5	协作与台账	排班执行有记录；API 账本无缺失；求助记录规范
5	加分	如：自创造假家族并被配方收录；对纯文字造假做专门评测并给出盲区分析；抽检升级为双人独立判+仲裁

附录　成绩登记表

全班 15 组、每组 5 人、共 75 人。本表供登记使用：平时分按第二章（上限 20）、实验分按第三章（上限 30）、大作业分按第四章（上限 50）录入，总评为三项之和（满分 100 分）；大作业部分个人得分可在组分基础上浮动 ±10%（见 1.2）。

表 A-1　十五组成绩登记表（满分 100 分）

组号	组名	组员（姓名 / 学号）	平时(20 分)	实验(30 分)	大作业(50 分)	总评(100 分)
1	六目拼接 · 匹配决策组	1. _____ 2. _____ 3. _____ 4. _____ 5. _____				
2	六目拼接 · 特征质量组	1. _____ 2. _____ 3. _____ 4. _____ 5. _____				
3	六目拼接 · 缝合融合组	1. _____ 2. _____ 3. _____ 4. _____ 5. _____				
4	六目拼接 · 数据与标签组	1. _____ 2. _____ 3. _____ 4. _____ 5. _____				
5	六目拼接 · 测试集与评测组	1. _____ 2. _____ 3. _____ 4. _____ 5. _____				

组号	组名	组员（姓名 / 学号）	平时(20 分)	实验(30 分)	大作业 (50 分)	总评 (100 分)
6	六目拼接 · 端到端数据组	1. _____ 2. _____ 3. _____ 4. _____ 5. _____				
7	六目拼接 · 端到端训练组	1. _____ 2. _____ 3. _____ 4. _____ 5. _____				
8	六目拼接 · 端到端优化组	1. _____ 2. _____ 3. _____ 4. _____ 5. _____				
9	水下双目 · 数据组	1. _____ 2. _____ 3. _____ 4. _____ 5. _____				
10	水下双目 · 部署组	1. _____ 2. _____ 3. _____ 4. _____ 5. _____				
11	水下双目 · 训练组	1. _____ 2. _____ 3. _____ 4. _____ 5. _____				
12	水下双目 · 评测组	1. _____ 2. _____ 3. _____ 4. _____ 5. _____				

组号	组名	组员（姓名 / 学号）	平时(20 分)	实验(30 分)	大作业 (50 分)	总评 (100 分)
13	数据基座 · 数据结构化组	1. _____ 2. _____ 3. _____ 4. _____ 5. _____				
14	数据基座 · 智能问答与图检索组	1. _____ 2. _____ 3. _____ 4. _____ 5. _____				
15	数据基座 · 虚假内容识别与对抗训练组	1. _____ 2. _____ 3. _____ 4. _____ 5. _____				