**按 Excel 划分 Train/Test/Validation 的文件梳理 GUI — 用户使用手册**

版本：v1.0｜更新日期：2025-09-26

# 1. 简介

本工具用于依据 Excel 统计表中的划分列（train/test/validation），将 root 目录下的 case 目录批量梳理到目标目录 dst 的子文件夹中。

程序采用 tkinter + pandas 实现：支持严格匹配与宽松匹配两种 ID 对齐策略，提供“预览梳理”以在执行前检查冲突、缺失与孤儿目录，并支持复制/移动两种输出模式。

# 2. 环境与依赖

• 操作系统：Windows / Linux / macOS（建议 Windows 10/11）

• Python：3.8 及以上

• 依赖：pandas、openpyxl、tkinter（标准库自带）

安装命令：pip install pandas openpyxl

# 3. 目录与命名约定

• root 目录：其第一层子目录即视作各个 caseID（例如 root/1001、root/1002…）。

• dst 目录：执行后将在 dst 下创建 train、test、validation 三个子目录，并将对应的 case 目录复制/移动至其中。

• 允许的划分标签（不区分大小写）：train、test、validation。其余值视为未知并跳过。

# 4. 界面与参数

|  |  |  |
| --- | --- | --- |
| 控件/参数 | 说明 | 备注/示例 |
| 数据根目录 root | 选择包含多个 caseID 子目录的根路径。 | 示例：D:/dataset/cases |
| 输出目录 dst | 结果输出路径，程序将创建 train/test/validation 子目录。 | 示例：D:/dataset/splits |
| 统计表 Excel | 包含 ID 与划分列的 Excel 文件。 | 支持 .xlsx/.xls |
| ID 列名 | Excel 中 case 标识列名。 | 如：ID、case、病例编号 |
| 划分列名 | Excel 中划分列名。 | 如：split、subset、partition |
| 严格匹配 | 勾选后严格区分大小写与空格；否则宽松匹配（去首尾空格、压缩多空格、统一小写）。 | 建议优先严格匹配，必要时关闭 |
| 只处理表中出现的ID | 启用后，root 中未出现在 Excel 的“孤儿”目录将不处理。 | 默认开启 |
| 复制/移动 | 选择输出模式：复制（推荐）或移动。 | 复制可合并覆盖同名文件；移动遇已存在目录将跳过 |
| 扫描 root | 扫描 root 第一层子目录并建立 case 列表。 |  |
| 浏览Excel并统计 | 读取 Excel 并统计 train/test/validation 计数、未知值、冲突。 |  |
| 预览梳理 | 生成计划（plan），统计将处理的数量以及跳过原因。 | 强烈建议先预览 |
| 开始梳理 | 根据计划执行复制/移动，带进度条与日志。 |  |

# 5. 操作流程（Step-by-Step）

1. 选择【root】与【dst】路径。

2. 点击【扫描 root】：日志将显示在 root 下发现的 case 数量与部分示例。

3. 设置【统计表 Excel】【ID 列名】【划分列名】，点击【浏览Excel并统计】：统计各划分计数；标出未知/空值与冲突；并与 root 对比“未找到/孤儿”。

4. 确认【严格匹配】与【只处理表中出现的ID】；选择【复制】或【移动】模式。

5. 点击【预览梳理】：生成 plan，给出 train/test/validation 数量与跳过统计。

6. 点击【开始梳理】：后台线程执行，进度条实时更新；日志记录每个 case 的 COPY/MOVE/ERR/DONE。

# 6. 匹配策略与冲突处理

• 严格匹配：Excel 的 ID 必须与 root 第一层目录名完全一致（区分大小写与空格）。

• 宽松匹配：对 Excel ID 与目录名做 normalize（去首尾空格、多空格压缩为一个、统一小写）后匹配。

• 冲突：同一 ID 在 Excel 中对应多个不同划分，记为冲突并跳过该 ID。

• 未知划分：非 train/test/validation（大小写不敏感）视为未知并跳过。

• 未找到：Excel 中的 ID 在 root 中找不到目录也将跳过。

• 孤儿目录：root 中存在但 Excel 未出现的目录，若启用“只处理表中出现的ID”，则不处理。

# 7. 复制与移动的差异

• 复制（推荐）：允许目标目录已存在时合并复制并覆盖同名文件；不改变 root 原始数据，便于回滚。

• 移动：若目标目录已存在，将跳过该 case；root 中相应目录将被移走，需谨慎执行并做好备份。

# 8. 日志与进度条

• 日志标签示例：[INFO]、[PREVIEW]、[COPY]、[COPY-MERGE]、[MOVE]、[ERR]、[DONE]。

• 进度条显示“已处理/总数”。完成后弹窗汇总“成功/失败”。

# 9. Excel 规范与示例

• 至少包含两列：ID 与划分列；列名需与界面填写一致。

• 划分列允许值：train / test / validation（大小写不敏感）。

|  |  |
| --- | --- |
| ID | split |
| 1001 | train |
| 1002 | test |
| 1003 | validation |
| 1004 | Train |
| 1005 | TEST |

# 10. 常见问题（FAQ）

Q：扫描为 0 个 case？

A：确认 root 第一层是否为 case 目录；检查权限/路径可达性。

Q：冲突/未知/未找到较多？

A：检查 Excel：ID 是否唯一、split 是否为 train/test/validation；必要时关闭“严格匹配”排查。

Q：复制是否覆盖？

A：复制模式会合并并覆盖同名文件；移动模式遇已存在目录则跳过。

Q：能否按文件级过滤？

A：当前以“目录”为单位，不支持目录内按文件过滤。

Q：报错如何排查？

A：查看 [ERR] 信息，常见原因：权限不足、磁盘空间不足、路径过长等。

# 11. 安全建议

• 先“预览”，再小样本试运行，最后全量执行。

• 优先使用“复制”模式，保留原始数据。

• 跨盘/网络盘操作注意权限与空间；使用“移动”前请备份。

# 12. 版本信息

• 程序：文件区分梳理工具（按 Excel 划分 train/test/validation）

• 主要特性：严格/宽松匹配、预览与统计、复制/移动、进度条与详细日志、冲突/未知/孤儿处理

• 兼容性：Windows / Linux / macOS；支持中文路径