

# pptx-wzq 使用手册与技术分析

PPT 多模态教学知识库自动化构建 · 版本 2.5.0 · 使用手册（第一部分）+ 技术分析报告  
（第二部分）

作者：吴振谦 · wuzhenqian@nbu.edu.cn · QQ: 38328063

## 第一部分 · 使用手册

### 零、关于本工具与作者

作者：吴振谦 · wuzhenqian@nbu.edu.cn · QQ: 38328063

开发原因：高校教师在课程建设与教材建设中，长期面临「课件里的大量图片、公式、文本散落各处，难以整理为规范、可复用、图文并茂的教学资源」的痛点——手工整理一份课程的图文知识库往往要耗费数周。本工具把 PPT 自动转化为图文并茂的多模态教学知识库，让教师从机械整理中解放出来。

版本：当前 2.5.0（历经理念重构：v1.5 六步管线 → v2.0 十六规则落地 → v2.3 整篇分章分节 → v2.5 文档体系完备）。

### 一、它能做什么（工作流程总览）

输入一份教学 PPT (.pptx)，自动完成 8 环节流水线：

- ① **blocks 图块提取**：组合(grpSp)即图块，整体保留原生 XML 段到 sources/；像素图/Visio/矢量按规则拆块；PowerPoint 渲染块图。
- ② **text 文本提取**：逐页正文（排除页眉页脚），组内文字标「图块内文本」，表格输出 Markdown。
- ③ **formula 公式提取**：非组合公式三路径（OMML/MTEF/OCR）→ LaTeX。
- ④ **caption 图块解读**：按 sources/ 顺序——XML 段→DeepSeek 读 XML（公式转 LaTeX）；像素图→qwen 兜底。
- ⑤ **related 相关性过滤**：剔除 logo/作者/装饰块（审计 json）。
- ⑥ **author 教材文案**：整篇视为一部教材，DeepSeek 自主分章分节（一节可含多页），每页标注章节；500 字为限（不足扩写、超出直出整理）。
- ⑦ **blocks\_json 组装**：visual\_blocks.json (v2.0 schema) + DeepSeek 语义增强 + 跨模态关系 + binding 导出。
- ⑧ **输出**：归位 images/ + sources/ + 各 JSON/MD；成功即清理过程文件；中断断点续传。

## 二、环境要求与下载

### 系统要求

- Windows 10/11（强烈建议；渲染通道仅支持 PowerPoint COM）
- Python 3.9 及以上
- Microsoft Office PowerPoint（必装：块渲染走 PowerPoint 原生，无 Office 则无渲染图、优雅降级）
- API Key：**DEEPSEEK\_API\_KEY**（语义解读/教材文案）、**DASHSCOPE\_API\_KEY**（qwen VLM 兜底）

### 下载方式

1. pip 安装发行版：`pip install pptx-wzq`
2. GitHub 源码：<https://github.com/wuzhenqian6611/pptx-wzq>
3. 安装后自带文档：`python -c "from pptx_wzq import docs; print(docs.__path__)"`

### 三、快速开始

```
# 一条命令跑完 8 环节（推荐）
pptx-paser "C:\课件\战略管理.pptx" -o "C:\输出\战略管理知识库"

# 只跑图块+文本+公式+解读（跳过相关性与文案，省 Token）
pptx-paser "C:\课件\战略管理.pptx" -o out --skip related,author

# 断点续传（中断后重跑同命令即可自动接续）
pptx-paser "C:\课件\战略管理.pptx" -o out

# 单步命令（叶子工具）
pptx-blocks out --pptx "战略管理.pptx" --texts out\战略管理_texts.md --no-vlm
pptx-author --texts out\战略管理_texts.md --formulas out\战略管理_formulas.md --captions out\战略管理_
```

### 四、命令一览（10 个 console\_script）

| 命令                                              | 功能                                                                        |
|-------------------------------------------------|---------------------------------------------------------------------------|
| <code>pptx-paser</code>                         | 总编排器：8 环节一条命令跑完，断点续传/成功清理                                                 |
| <code>pptx-blocks</code>                        | 图块提取+解构+渲染； <code>--caption-sources</code> 模式按 <code>sources/</code> 顺序解读 |
| <code>pptx-text</code>                          | 文本提取（ <code>in_group</code> /表格 Markdown）                                 |
| <code>pptx-formula</code>                       | 公式提取 → LaTeX                                                              |
| <code>pptx-caption</code>                       | 图片/图块解读（模型路由）                                                             |
| <code>pptx-related</code>                       | 块相关性过滤 + 审计                                                               |
| <code>pptx-author</code>                        | 整篇分章分节教材文案（500 字为限）                                                       |
| <code>pptx-bind</code>                          | 图文绑定 JSON                                                                 |
| <code>pptx-html</code> / <code>pptx-deck</code> | 教材 HTML / 教学 Deck 导出                                                      |

## 五、输入 PPTX 预处理要求

核心思想：用 PPT 自带工具给解析器打「块边界」标注。组合的数量 = 该页图块数量的上限基准，可据此验收。  
预处理不是必须的（无组合也能跑），但组合能让图块提取完全确定、可回归。

| 对象          | 预处理操作                     | 解析器行为                                          |
|-------------|---------------------------|------------------------------------------------|
| 逻辑图/示意图     | Ctrl+G 把底图+文字+箭头合成一个组合    | 整体 = 1 个 group 块，XML 段导出 sources/，渲染图存 images/ |
| 嵌套组合        | 有意为之才用；想分开就移出外层           | 并入外层 children，不单独成块                            |
| Visio/vsdx  | 不要组合进其他形状；保持独立 OLE        | 独立 visio 块：可剥离 .vsdx / 不可剥离 XML 段              |
| 公式          | 行内公式保持独立；想并入图块就移进组合       | 组合内随块转 LaTeX；非组合进 formulas.md                  |
| 首页/尾页       | 无需操作（默认跳过图块）              | 封面/致谢不产块（文本/公式仍提取）                             |
| 小像素图        | <20% 页面默认舍弃；想保留 → 组合进相邻图形 | 组合内豁免；非组合丢弃入审计                                 |
| 表格          | 无需操作                      | Markdown 表格输出（texts.md）                        |
| srcRect 裁剪图 | 无需操作（自动一致）                | 元数据/资源/渲染/描述全程按裁剪                              |

## 六、输出文档说明

```
<名>_result/
├─ sources/ # 图块源（解读唯一输入）：grpSp XML 段 / .vsdx / 像素图原图 / rldimg/ 资源图
├─ images/ # 块渲染 PNG（PowerPoint，仅供人阅览）
├─ <名>_visual_blocks.json # 核心结构化（v2.0）
├─ <名>_visualBlock_text_binding.json # 图文关联（v1.0）
├─ <名>_textbook.md # 教材文案（篇→章→节→页，每页标注章节）
├─ <名>_captions.md # 块解读（绑定 sources/ 文件名+通道）
```

## 七、常见问题

| 问题                    | 说明                                                                           |
|-----------------------|------------------------------------------------------------------------------|
| 没有 Office 会怎样？        | 渲染降级：无块渲染图（images/ 为空）；XML/JSON/文案等其余产物正常。                                   |
| 没有 API Key？           | 跳过对应解读/文案步骤： <code>PPTX_PASER_NO_VLM=1</code> 可跳过 VLM 全流程 0 Token。           |
| 中途中断？                 | 重跑同命令自动接续（state.json + doc_md5 换源检测）。                                        |
| DeepSeek 对某 XML 段空响应？ | 自动降级 qwen 读渲染图 → 规则模板兜底，保证每块有解读。                                             |
| 如何查安装版本与文档？           | <code>pptx-paser --version</code> ； <code>from pptx_wzq import docs</code> 。 |

# 第二部分 · 技术分析报告

## 1. 系统定位与核心能力

### 定位

把任意 PPT 课件解构为「多模态教学知识库」：文本/公式/图块三路抽取 + AI 解读 + 教材级文案，输出 JSON/Markdown/渲染图三件套。

### 技术栈

OOXML 原生解析 + DeepSeek-V4-Flash（语义/文案）+ qwen3.7-plus（VLM 兜底）+ PowerPoint COM 渲染。

### 交付物

sources/ + images/ + visual\_blocks.json + binding + textbook.md + captions.md。

图块优先：组合即图块，整体保留而非拆散。

源语言级解构：grpSp 原生 XML 段、Visio 原生文件、公式 LaTeX。

模型按块路由：DeepSeek 读 XML / qwen 像素图兜底。

教材级文案：整篇自主分章分节、每页标注章节、500 字为限。

生命周期自管：成功即清理、中断续传。

## 2. 技术架构

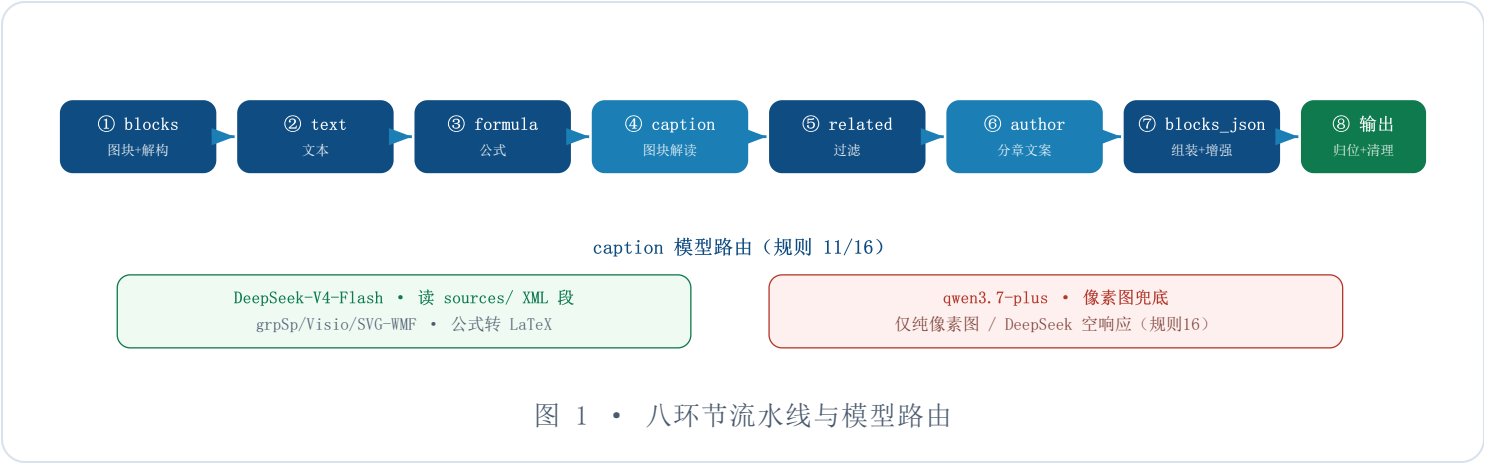
### 2.1 模块划分

| 模块                                  | 职责                                                                        | Token         |
|-------------------------------------|---------------------------------------------------------------------------|---------------|
| <code>extract_pptx_images.py</code> | OOXML 原子对象提取（图片/形状/grpSp/表格/公式/图表）；XML 段原生提取；srcRect 裁剪；PowerPoint COM 渲染 | 0             |
| <code>extract_texts.py</code>       | 页面文本（in_group 标记）、表格 Markdown                                             | 0             |
| <code>visual_blocks.py</code>       | 单阶段确定性拆块、块渲染、语义增强、跨模态关系                                                   | DeepSeek      |
| <code>cli_blocks.py</code>          | XML 段导出、rldimg 复制、caption 路由、binding、captions.md                          | DeepSeek+qwen |
| <code>cli_author.py</code>          | 整篇分章分节文案、500 字为限、自动分批                                                     | DeepSeek      |
| <code>cli_related.py</code>         | 相关性过滤 + 审计                                                                | DeepSeek      |
| <code>cli_paser.py</code>           | 总编排、断点续传、归位、成功清理                                                          | —             |

### 2.2 命令体系

```
pptx-paser / pptx-blocks / pptx-text / pptx-formula / pptx-caption
pptx-related / pptx-author / pptx-bind / pptx-html / pptx-deck
```

## 3. 工作流程（8 环节）



## 4. 图块识别规则（16 条）

### 4.1 识别层（1-9）——「组合即图块声明」

| # | 规则                                | 实现                          |
|---|-----------------------------------|-----------------------------|
| 1 | 一个 grpSp 组合即是一个图块，组合内一切内容读取为该图块内容 | kind="group", children 递归收编 |
| 2 | 嵌套组合不单独提取                         | 嵌套仅作外层 children             |
| 3 | OLE Visio/vsdx 独立成块（除非在组合内）       | kind="visio" 分支             |
| 4 | 非组合像素图单独成块；重叠文本并入                 | raster 独立块 + 重叠文本并入         |
| 5 | 组合内公式作块内容；非组合公式独立提取               | formula 排除 in_group         |
| 6 | 首页/尾页图块舍弃                         | skip_cover_pages 整页跳过       |
| 7 | 非组合像素图 < 整页 20% 舍弃                | raster_min_area_ratio=0.20  |
| 8 | 表格仍读取为表格（Markdown）                | 表格移交 text 步骤                |
| 9 | grpSp 内 srcRect 须全程一致             | children 携带 src_rect + 裁剪落盘 |

### 4.2 解构/解读层（10-16）

| #  | 规则                                              | 实现                                |
|----|-------------------------------------------------|-----------------------------------|
| 10 | grpSp 保留整段 XML，页标记，每组合一个独立 .xml 存 sources/      | sources/slide_{页}_{块id}_grp.xml   |
| 11 | grpSp 块 caption 用 DeepSeek 读 XML；公式转 LaTeX      | _ds_read_xml + 超长压缩 + 重试          |
| 12 | grpSp 块 PowerPoint 渲染 PNG 存 images/；PNG 不送 qwen | COM ExportAsFixedFormat + PyMuPDF |
| 13 | Visio 可剥离 → .vsdx 存 sources/ + PNG 存 images/    | 原生剥离 + 渲染                         |
| 14 | Visio 不可剥离 → XML 段，同 grpSp                      | sources/slide_{页}_{块id}_ole.xml   |
| 15 | SVG/WMF 同 Visio：尽量 XML 段，不行才 PNG                | sources/slide_{页}_{块id}_vec.xml   |
| 16 | 仅无法 DeepSeek 解读的块才送 qwen                        | 模型路由 + 空响应降级链                     |

## 5. 输出文档体系与内部格式

### 5.1 结果目录结构

```
<名>_result/
├─ sources/ # 图块源（解读唯一输入）： grpSp XML 段 / .vsdx / 像素图原图 / rldimg/
├─ images/ # 块渲染 PNG（PowerPoint）
├─ <名>_visual_blocks.json # pptx_multimodal_slide_v2.0
├─ <名>_visualBlock_text_binding.json # pptx_visual_block_text_binding_v1.0
├─ <名>_textbook.md # 篇→章→节→页（每页标注章节）
├─ <名>_captions.md # 绑定 sources/ 文件名 + 通道
├─ <名>_texts.md / _text_entries.json / _formulas.md / _related_filter.json
（成功即清理，无过程文件）
```

### 5.2 VisualBlock 内部格式

```
{
  "block_id": "blk_01", "page": 7, "block_type": "战略管理概念框架",
```

```
"bbox": {"x": 142.3, "y": 226.4, "w": 1026.2, "h": 400.2},
"assets": {"xml_source": "./sources/slide_07_blk_01_grp.xml",
           "raster_png": null,
           "rldimg": ["../sources/rldimg/slide_07_blk_01_image4.png", "..."]},
"internal_structure": {"nodes": [...], "edges": [...]},
"semantic_description": {
  "block_type": "战略管理概念框架",
  "expression_goal": "展示战略管理概念框架的核心逻辑",
  "expression_role": "将抽象概念通过战略哲学/商道/天道/人道具象化...",
  "expression_features": ["概念框架", "层次结构", "关系图"],
  "vlm_caption": "该图块以“战略哲学”为中心...",
  "teaching_use": "教学辅助图示",
  "formula_latex": "", "caption_source": "deepseek_xml"
},
"member_obj_ids": [...], "vector_resources": []
}
```

## 6. 生命周期与可靠性

### 6.1 成功即清理

成功后中间产物（by\_page/atomic/manifest）全部删除，只留交付物。  
删除失败打印警告而非静默。

### 6.2 中断续传

- 中断保留 work + state.json（步骤状态机）。
- 再运行跳过已完成、自动接续缺失：doc\_md5 换源 → 全量重跑。

## 7. 版本演进

| 版本    | 里程碑                                                     |
|-------|---------------------------------------------------------|
| 1.5.0 | 六步管线重构、三件套交付物、版本统一                                      |
| 2.0.0 | 16 条规则落地、组合即块、XML 段导出、DeepSeek caption 路由、PowerPoint 渲染 |
| 2.1.0 | captions 绑定 sources/ + 通道标注；每页扩写 500 字                  |
| 2.2.0 | textbook 500 字为限（不足扩写、超出直出整理）                           |
| 2.3.0 | Author 整篇自主分章分节，每页标注章节                                  |

|       |                           |
|-------|---------------------------|
| 2.4.0 | 技术分析报告（html/pdf/md）随安装包分发 |
| 2.5.0 | 使用手册+技术分析合并文档、README 全量更新 |

附注：本手册与技术分析的所有字段与格式均来自 pptx-wzq 真实产物（visual\_blocks.json / binding / textbook.md / captions.md / sources/），非虚构。