1.实现对题目图片中各类信息的精确捕捉与结构化处理。这包括对印刷体、手写体文字、复杂公式和特殊符号的精准识别。我们期望字符级别的识别准确率在处理标准印刷体时能达到99.5%以上，对于书写工整的手写体也能维持在95%左右。特别是在公式识别方面，我们要求系统能将图片中的公式完整且结构正确地转换为LaTeX代码，其准确率需达到98%以上，确保覆盖中小学阶段所有常见的公式类型。

此外，识别范围需扩展至各类图表信息，如数学几何图形、化学实验装置图与方程式、以及物理电路图等，确保重要信息无任何遗漏。同时，该技术方案必须具备强大的鲁棒性，能够有效应对真实拍摄场景中的各种干扰，如光影变化、纸张褶皱、图片模糊、拍摄角度倾斜等。对于小面积的污渍或内容遮挡，我们期望算法能展现出一定的智能容错能力，结合视觉上下文进行识别或标记。

我们期望输入的信息在文本公式识别的精确度上能超越Pix2text的基线水平，同时在图像内容的结构化理解上，其效果也应显著优于直接将原始图片信息传递给通义千问VL-Max等多模态模型。

2.版面分析与内容结构化能力。当处理包含多道题目的整页图片时，系统必须能够智能地进行单题分割，准确框定出每一道独立题目的完整区域。衡量标准：在进行单题区域检测时，其结果与标准答案的交并比（IoU）大于0.9的比例必须达到95%以上。

在成功分割出单题后，系统还需要进一步解析其内部结构，准确识别并标注出题干文本、选择题选项、配图、表格以及独立的公式块等不同类型的元素。对于这些元素类型的识别，我们要求其综合评价指标F1-Score（基于准确率和召回率计算）同样需达到95%以上。

另外，由于我们的产品兼具解题与批改两大核心功能，因此，算法必须能够精准地识别出手写笔迹，并将其作为“手写答案”元素与对应的印刷体题目进行正确关联。

3.数据输出格式，我们要求所有识别与分析过程的最终产物是一个结构化的JSON对象，而非简单的文本流。这个JSON对象应逻辑清晰，便于我们进行后续的程序化处理和传递给大模型。具体而言，它需要包含完整的题目分割结果，即每道题目的边界框坐标。在每个题目单元内部，应进一步列出其包含的所有元素的详细信息，包括该元素的类型（如题干、配图、手写答案等）、精确的边界框坐标，以及识别出的具体内容，如文本字符串或LaTeX代码。这种结构化的输出将作为我们连接图像识别前端与多模态大模型后端的标准数据接口。