* 输入图像，首先由形状自适应裁剪模块(Shape-Adaptive Cropping Module)，将文档中不同布局的内容裁剪出来
* 然后，将裁剪出的子图像同时通过视觉编码器(visual Encoder)和视觉抽象器(visual Abstractor)
* 最后，为了使大语言模型能够关联多个裁剪的子图像，应用裁剪位置编码模块(crop Postion Encoding)来引入跨子图像的空间信息, 送入LLM中

带有文本的图像具有各种宽高比和多种分辨率，简单地将图像大小调整为MLLM 的原始分辨率会导致文本模糊、扭曲和无法识别。 因此，提出了形状自适应裁剪模块。

如何自适应呢？

* 首先，预先定义各种形状的网格，分别表示网格的行数和列数，网格的最小格子长宽为
* 要为形状为的图像选择合适的网格，应遵循两个规则：

为此，采用Iou算法，来获得最佳网格布局

* 1. 网格应尽可能保留图像的分辨率
  2. 网格应适合输入图像的长宽比

**大语言模型裁剪图像建模**

MLLM 主要使用单个图像作为输入进行训练。 由于裁剪模块，需要将多个图像的视觉特征输入到语言模型中。 大语言模型的一维位置嵌入不能反映每个子图像的空间位置，这对于关联局部图像至关重要。

因此，结合二维裁剪位置编码来帮助语言模型理解裁剪图像之间的空间关系。

具体操作如下：

* 为所选网格的每个单元分配一个位置索引
* 通过两个辅助嵌入层获取它们的行嵌入和列嵌入
* 通过沿着可学习查询的维度进行广播，将嵌入添加到语言空间中每个单元的视觉特征中
* 然后，将视觉特征重塑为一张图，由此产生的空间感知视觉特征和输入句子的词嵌入在序列维度上连接并发送到大型语言模型。
* 为了增强语言模型有效建模多个图像的能力，同时保持较低的训练成本，冻结了原始语言模型并采用低秩适应方法进行微调。