生成网络在第一阶段利用全局句子向量生成低分辨率图像。在接下来的步骤中，它使用每个子区域中的图像向量，通过使用注意层来形成单词上下文向量来查询单词向量。

然后，该模型将区域图像向量和相应的单词上下文向量相结合，形成多模态上下文向量，并在此基础上在周围的子区域生成新的图像特征。这有效地产生了更高分辨率的图片，每个阶段都有更多细节。