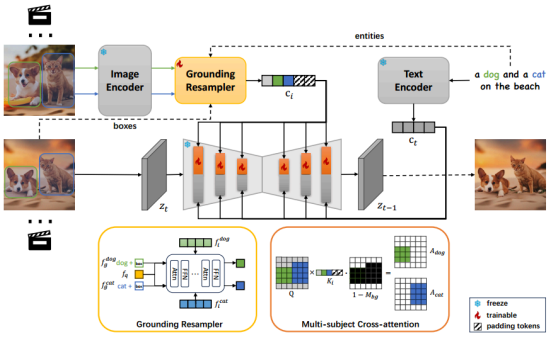
MS-Diffusion

Tags

研究解决的问题：

1.根据文本描述准确地保持每个指定主体的细节；
2.多主体生成同时保持
局部细节的准确性（每个主体的表现）和**整体布局的协调性**（主体与主体之间、主体与背景之间的关系）



核心思想-布局引导多主体生成

MS-Diffusion通过布局引导和交叉注意力机制解决了多主体生成中的条件混淆和空间冲突问题，并首次支持在冻结基础模型的条件下进行多主体的个性化图像生成。

模型架构

1. **Grounding Resampler**

利用可学习的查询标记从图像特征中提取细节信息，同时结合文本嵌入和位置信息增强生成的语义和空间准确性。

（随机丢弃标记训练，以防模型过度依赖布局信息）

2. **Multi-subject Cross-attention**

通过注意力掩码将生成的图像分区，并限制特定主体在指定区域内表示，以避免主体间冲突或相互覆盖。

引入虚拟标记用于背景部分的表示，确保未标记区域仅受文本控制。

工作流程

1. 数据准备

输入：

从文本中提取语义（主体）信息和场景描述。

利用辅助模块生成布局引导（如边界框）。

生成初始条件：

Grounding Tokens：

使用文本描述生成每个主体的语义嵌入。

将边界框信息通过 Fourier 嵌入生成空间先验。

2. 条件生成与分配

Grounding Resampler：

输入图像特征和布局引导。

输出每个主体的独立条件向量。

交叉注意力计算：

使用多主体掩码（将每个条件绑定到特定的空间区域。

背景 token 控制未明确分配的区域。

3. 图像生成

U-Net 扩散生成：

条件向量输入到 U-Net 中，通过扩散过程逐步生成图像。

交叉注意力在每个扩散步中持续作用，确保生成内容符合条件。

多主体分离：

掩码机制确保每个主体的生成独立。

背景 token 保证场景完整性。

4. 输出优化

在生成过程中，通过注意力权重动态调整每个主体的细节。

最终输出符合文本描述、布局引导且细节丰富的多主体图像。