**一、研究题目**

**基于剪枝的高效视觉自回归模型推理研究**

**二、研究背景**

随着人工智能的发展，视觉生成技术已经在图像与视频合成、创意设计、虚拟现实等领域展现出巨大的潜力。传统的自回归（Autoregressive, AR）模型通常采用“逐像素”或“逐token”预测方式，在高分辨率图像生成中存在推理步骤冗长、计算复杂度过高的问题。近年来提出的视觉自回归（Visual Autoregressive, VAR）建模采用“逐尺度”预测（next-scale prediction），能够在较少步骤内完成高质量图像生成，为高效视觉生成提供了新的思路。然而，现有VAR模型在分辨率扩展时，仍然面临显著的计算与存储开销瓶颈，尤其在大尺度阶段，其运行时间和显存消耗急剧上升。因此，如何在保证生成质量的前提下降低计算开销、实现高分辨率可扩展生成，成为一个重要的研究课题。

**三、研究现状**

现有研究主要集中在以下几个方面：

1. **扩散模型的加速**：已有大量工作通过蒸馏、量化、剪枝和缓存等手段加速扩散模型，但这些方法大多依赖扩散网络结构，不适用于VAR。
2. **自回归模型的并行解码**：在文本与图像生成任务中，部分工作探索了推测解码（speculative decoding）、并行解码等方法来减少生成步骤，但这些方法针对的是“逐token”预测，不适用于VAR的“逐尺度”建模。
3. **VAR加速的探索**：已有如 **Infinity**、**HART** 等代表性VAR模型，在高分辨率生成上表现突出，但其推理复杂度仍随分辨率平方级或更高增长，难以在消费级GPU上实现2K级别图像生成。近期提出的 **FastVAR** 方法，通过“缓存令牌剪枝”（Cached Token Pruning），在大尺度阶段仅保留关键信息token，并利用缓存恢复被剪枝的token，有效降低了推理复杂度，在保证质量的同时实现推理加速。

综上，虽然已有方法在扩散模型或VAR框架下取得一定进展，但仍存在适配性不足、效率有限等问题，为研究提供了很好的切入点。

**四、研究内容**

本课题拟在 **FastVAR 框架** 的基础上，开展以下研究：

1. **理论分析与模型复现**

* 学习并理解VAR模型的“逐尺度预测”机制；熟悉HART与Infinity代码。
* 掌握FastVAR中的核心思想，包括“关键token选择（PTS）”和“缓存token恢复（CTR）”；
* 复现论文中的核心实验，验证在HART与Infinity模型上的加速效果。

1. **方法改进与实验探索**

* 分析不同剪枝比例对生成质量与推理速度的影响，寻找项目可实现的“效率-质量平衡点”；
* 探索FastVAR与其他加速技术如（CoDe: Collaborative Decoding Makes Visual Auto-Regressive Modeling Efficient ）的方法结合，研究在协同解码和剪枝策略共同作用下的推理加速潜力；
* 设计对比实验，系统评估不同方法组合的加速比、生成质量以及显存消耗，形成完整的效率-性能权衡分析。