Future Research Directions

Efficient Inference on Edge/Mobile Devices

- **Why it matters:**Many VLMs are too large and compute-intensive for deployment on phones, AR glasses, or IoT devices.
- **Challenges:**
  - High memory, compute, and energy usage
  - Real-time latency constraints
- **Solutions:**
  - Model quantization, distillation, hardware-specific optimization
  - Lightweight architectures (e.g., MobileViT, MobileSAM)
- **Use Cases:**AR assistants, mobile translators, smart wearables, offline captioning tools.