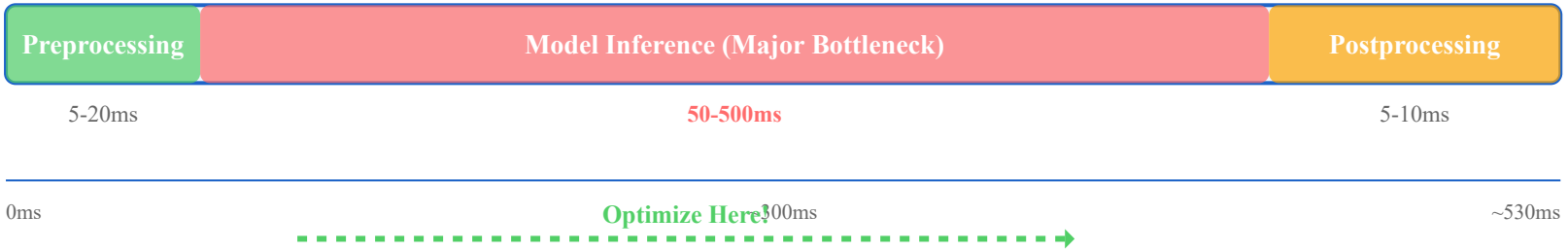
Latency Optimization

Latency Optimization

Inference speed improvement techniques for real-time medical applications

Inference Latency Breakdown



Latency Components

Data Preprocessing
Image resizing, normalization
5-20ms

Model Inference
Largest proportion
50-500ms

Postprocessing
Result interpretation, visualization
5-10ms

Optimization Techniques

Model Compression

- Quantization
- Pruning
- Distillation

2-10x speedup

Inference Engine

- TensorRT
- ONNX Runtime
- TFLite

1.5-3x speedup

Hardware Acceleration

- GPU
- NPU/TPU
- FPGA

5-100x speedup

Real-time Processing Standards

Soft Real-time

< 100ms

General diagnostic assistance

Hard Real-time

< 10ms

Surgical robots, emergency systems