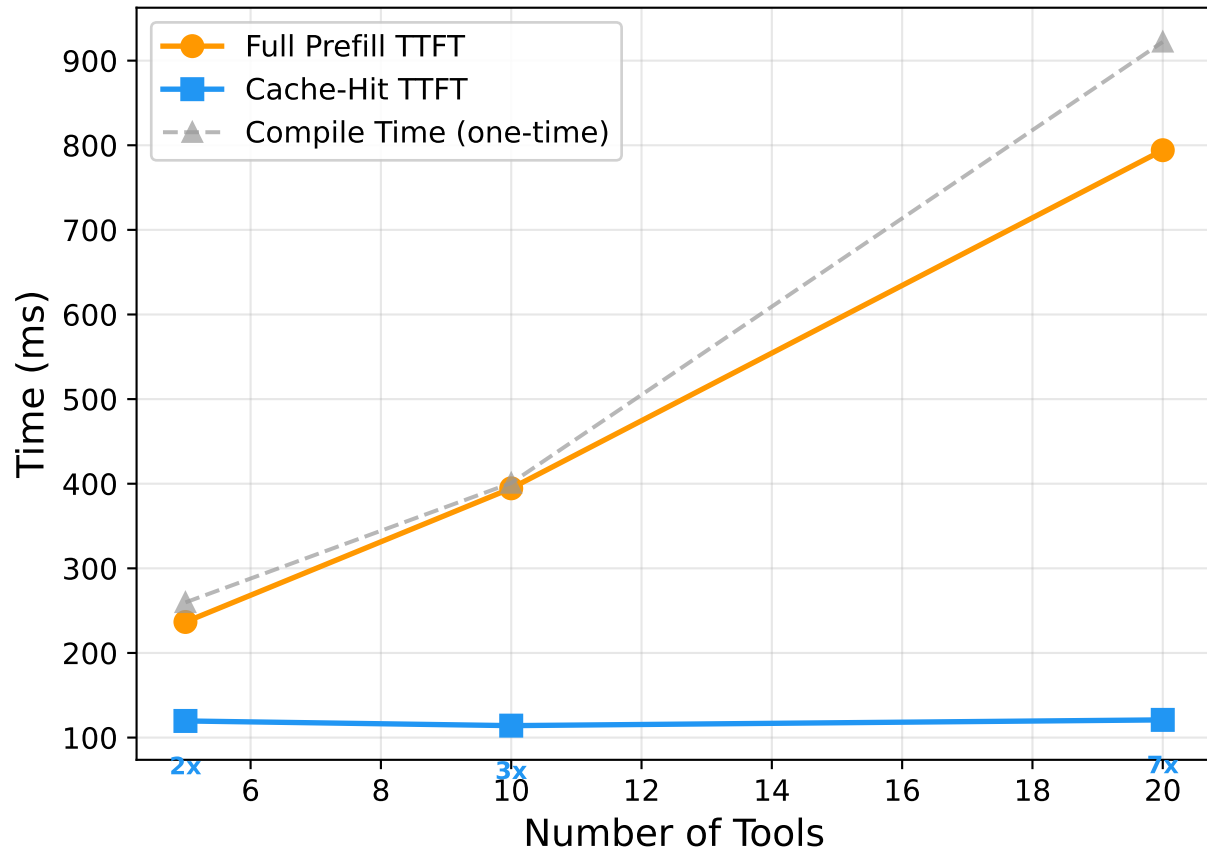


TTFT Scaling



GPU Memory Cost

