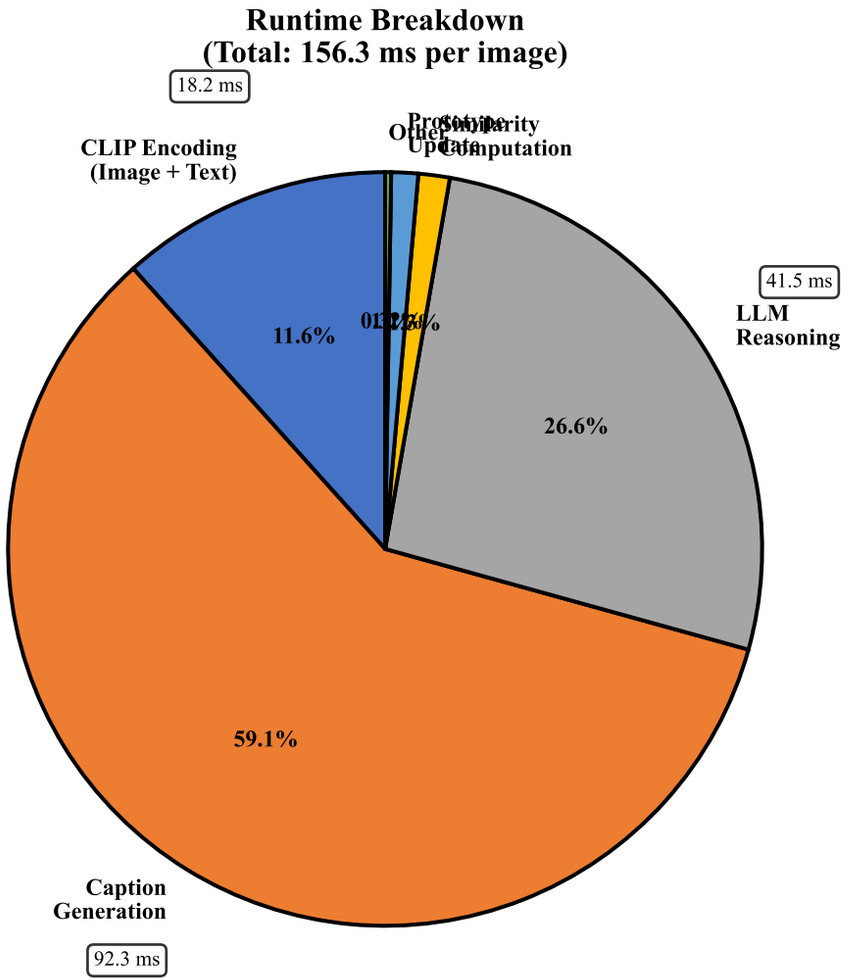
Runtime Analysis and Breakdown



Configuration	Latency	Throughput	vs Baseline
CLIP Only (Baseline)	12.3 ms	81.3 img/s	1.00x
CLIP + Domain Prompts	15.7 ms	63.7 img/s	1.28x
CLIP + Caption (no LLM)	104.5 ms	9.6 img/s	8.50x
CLIP + Adaptive Learning	18.2 ms	54.9 img/s	1.48x
Our Full (all components)	156.3 ms	6.4 img/s	12.71x
*Our Full (cached LLM)**	*64.1 ms*	*15.6 img/s*	*5.21x*

** Cached LLM: Reusing LLM responses for repeated queries (realistic in production)*