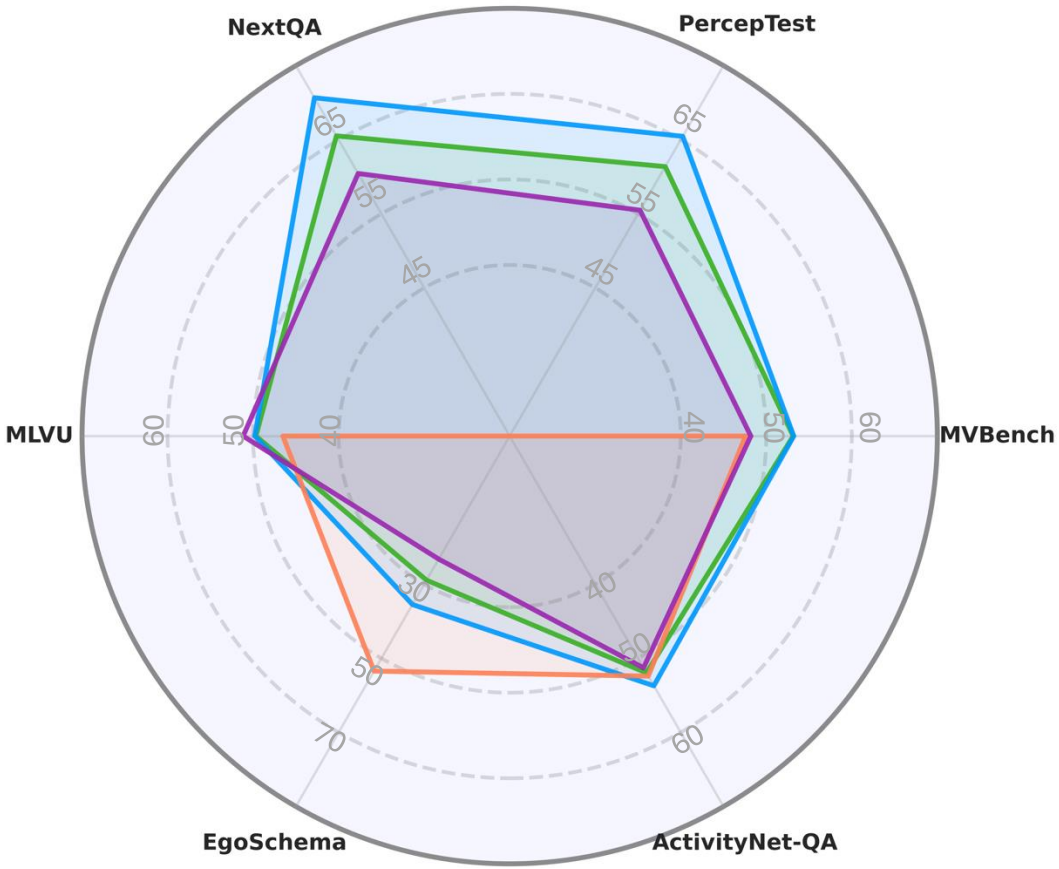


| Model | Total params | Throughput (Tokens/second) | |
|---|---|---|---|
| | | Jetson Orin Nano | RTX A6000 |
| Mobile-VideoGPT-0.5B | **0.6B** | **7.3** | **45.9** |
| Mobile-VideoGPT-1.5B | 1.6B | 6.1 | 41.0 |
| LLaVA-one-vision-0.5B | 1.0B | 3.4 | 22.7 |
| LLaVA-Mini-8B | 8.4B | NA | 7.6 |