VLM Performance Metrics (Higher is Better)

Model	Size Efficiency	Memory Efficiency	Loading Speed	Inference Speed	Response Quality
Qwen2.5-VL-7B	1.00	1.00	0.79	0.97	1.00
Gemma-3-4b-it	0.57	0.54	0.87	0.93	1.00
InternVL2.5-4B	0.57	0.48	0.91	0.91	1.00
Moondream2-2B	0.28	0.65	0.96	0.96	1.00
Qwen2.5-VL-3B	0.43	0.40	0.89	0.94	1.00
DeepSeek-VL-7B	1.00	0.84	0.81	0.94	0.00
Qwen2-VL-2B	0.29	0.26	0.94	0.95	1.00
InternVL3-2B	0.29	0.30	0.94	0.89	1.00
GLM-Edge-V-2B	0.29	0.65	0.95	0.52	1.00
DeepSeek-VL-1.3B	0.19	0.28	0.95	0.95	1.00
SmolVLM2-500M	0.07	0.33	0.97	0.98	1.00
Florence-2-base	0.03	0.09	0.80	0.99	1.00
Phi-3.5-vision-instruct	0.59	0.41	0.89	0.00	1.00
InternVL3-1B	0.14	0.15	0.97	0.90	0.50
PaliGemma-3B	0.43	0.39	0.82	0.99	0.00
SmolVLM2-2.2B	0.31	0.25	0.90	0.95	0.00
Florence-2-large	0.11	0.28	0.00	0.99	1.00
SmolVLM2-256M	0.04	0.18	0.99	0.97	0.00