Table 2. Performance on general visual understanding across various VLM backbones.

Backbone	GQA	MMBench	MMStar	TextVQA	DocVQA	InfoVQA	ChartQA	AI2D	OCRBench	MMMU	LogicVista	POPE
Qwen2.5-VL-7B [4]	60.4	82.2	56.7	83.2	93.3	78.6	83.3	78.7	84.3	50.1	39.1	86.1
+Ours	62.1 ↑1.7	83.7 ↑1.5	61.6 ↑4.9	85.0 ↑1.8	95.2 ↑1.9	80.5 ↑1.9	87.6 ↑4.3	82.6 ↑3.9	86.7 ↑2.4	51.0 ↑0.9	42.4 ↑3.3	87.9 ↑1.8
LLaVA-NeXT-7B [19]	63.0	64.9	33.5	65.4	66.1	30.7	53.5	63.4	52.4	35.8	26.3	86.1
+Ours	64.6 ↑1.6	67.4 ↑2.5	36.5 ↑3.0	67.8 ↑2.4	70.2 ↑4.1	35.8 ↑5.1	55.7 ↑2.2	65.2 ↑1.8	53.7 ↑1.3	37.4 ↑1.6	29.2 ↑2.9	88.2 ↑2.1
Qwen3-VL-8B [3]	61.8	83.8	61.1	82.7	94.3	80.8	84.6	82.9	85.8	54.0	35.3	87.3
+Ours	63.0 ↑1.2	86.4 ↑2.6	65.5 ↑4.4	84.6 ↑1.9	95.8 ↑1.5	83.6 ↑2.8	89.9 ↑5.3	84.3 ↑1.4	88.6 ↑2.8	55.7 ↑1.7	39.5 ↑4.2	88.5 ↑1.2
InternVL3.5-8B [37]	58.3	81.1	64.6	74.6	90.2	73.6	84.5	81.0	83.7	58.4	37.3	89.7
+Ours	61.1 ↑2.8	83.2 ↑2.1	66.6 ↑2.0	77.8 ↑3.2	92.5 ↑2.3	75.5 ↑1.9	87.5 ↑3.0	83.7 ↑2.7	86.0 ↑2.3	61.9 ↑3.5	40.0 ↑2.7	91.0 ↑1.3
Qwen2.5-VL-3B [4]	54.6	77.8	55.8	79.9	93.0	74.2	82.5	76.2	79.7	44.2	32.8	86.5
+Ours	56.7 ↑2.1	79.3 ↑1.5	58.4 ↑2.6	80.8 ↑0.9	93.2 ↑0.2	76.9 ↑2.7	83.8 ↑1.3	78.3 ↑2.1	82.7 ↑3.0	45.6 ↑1.4	34.4 ↑1.6	88.3 ↑1.8
LLaVA-OV-1.5-4B [1]	55.7	82.1	60.5	76.8	90.4	70.1	84.5	82.2	78.7	51.7	39.1	86.0
+Ours	57.0 ↑1.3	83.7 ↑1.6	63.6 ↑3.1	79.0 ↑2.2	92.1 ↑1.7	72.7 ↑2.6	86.1 ↑1.6	83.0 ↑0.8	81.2 ↑2.5	53.2 ↑1.5	42.4 ↑3.3	87.2 ↑1.2
LLaVA-OV-2-8B [2]	58.5	84.7	65.5	76.9	93.0	72.8	85.1	81.9	77.3	55.0	41.7	88.0
+Ours	60.0 ↑1.5	86.7 ↑2.0	68.9 ↑3.4	79.7 ↑2.8	93.5 ↑0.5	74.3 ↑1.5	87.4 ↑2.3	83.5 ↑1.6	79.2 ↑1.9	56.7 ↑1.7	43.1 ↑1.4	89.4 ↑1.4
Gemma-3-12B [31]	47.8	77.0	49.9	72.4	82.9	52.0	58.6	78.3	70.7	46.1	33.3	90.2
+Ours	48.7 ↑0.9	78.8 ↑1.8	52.3 ↑2.4	77.7 ↑5.3	87.1 ↑4.2	57.2 ↑5.2	60.0 ↑1.4	80.8 ↑2.5	72.6 ↑1.9	48.3 ↑2.2	35.3 ↑2.0	90.7 ↑0.5