Table 1. Model comparison on 22 Benchmarks. We provide model variants with different Vision Encoders: SigLIP-SO400M-Patch14-384 (FUSION) and SigLIP2-Giant-OPT-Patch16-384 (FUSION-X). FUSION-L is built upon FUSION-X by employing interpolation on image tokens during inference. FUSION-X achieves superior performance with fewer tokens.

(a) Results on general multimodal benchmarks.

Model		General Benchmarks										
Method	# Vis tok.	MMB^{EN}	MMB^{CN}	VizWiz	POPE	MM-Vet	MME^{P}	MME^{C}	Seed-Image	HallB	LLaVA^{W}	MMStar
<=4B Model Comparison												
Qwen2.5VL 3B	-	79.1	**78.1**	-	85.9	**61.4**	1592.4	**607.5**	74.0	46.6	-	**56.3**
InternVL2 4B	-	78.5	73.9	-	84.6	50.5	1532.8	531.8	73.2	42.4	-	53.9
DeepSeek-VL2-Tiny	-	74.6	72.1	-	-	52.5	1548.3	357.1	72.3	39.6	-	45.9
MM1.5 3B	-	-	-	-	88.1	41.0	1478.4	319.6	72.4	-	73.0	-
Phi 3.5-Vision	-	75.5	64.2	58.2	82.2	46.5	1473.4	412.1	69.9	53.3	68.8	49.0
Florence-VL 3B	576	71.6	60.8	59.1	88.3	51.0	1498.7	403.9	70.6	**58.1**	71.1	44.9
FUSION 3B (ours)	780	79.5	71.7	64.6	**88.9**	57.2	**1595.9**	416.5	74.6	51.4	84.7	52.4
FUSION-X 3B (ours)	620	**80.3**	74.8	**66.1**	88.7	60.3	1582.1	440.0	**75.3**	51.9	**85.2**	50.9
FUSION-L 3B (ours)	308	77.6	70.8	65.3	88.3	56.7	1573.7	406.8	74.1	48.7	77.6	47.7
>=7B Model Comparison												
Qwen2VL 7B	-	**83.0**	80.5	-	88.4	**62.0**	**1639.2**	**637.1**	76.0	50.6	-	**60.7**
InternVL2 8B	-	81.7	**81.2**	-	86.9	54.2	1639.7	575.3	75.4	45.2	-	61.5
LLaVA-OneVision 8B	-	81.7	78.0	-	87.2	58.8	1626.0	483.0	74.8	47.5	86.9	60.9
MM1.5 7B	-	-	-	-	88.6	42.2	1514.9	346.4	73.4	-	74.2	-
Cambrain 8B	576	75.9	67.9	-	87.4	48.0	1547.1	-	74.7	48.7	71.0	50.0
Florence-VL 8B	576	76.2	69.5	59.1	**89.9**	56.3	1560.0	381.1	74.9	**57.3**	74.2	50.0
Eagle 8B	1024	75.9	-	-	-	-	1559.0	-	76.3	-	-	-
FUSION 8B (ours)	780	80.5	74.9	59.5	89.3	60.0	1592.3	396.1	77.2	52.6	86.9	52.4
FUSION-X 8B (ours)	620	82.0	76.2	**62.9**	88.8	60.0	1607.5	337.2	78.2	51.4	88.0	52.7
FUSION-L 8B (ours)	308	80.0	73.6	59.9	88.5	57.3	1601.7	338.9	75.9	46.7	82.1	49.3

(b) Results on Vision centric, Knowledge based, and OCR & Chart benchmarks.

Model		Vision centric				Knowledge based				OCR & Chart			
Method	# Vis tok.	MME-RW	RWQA	CV-Bench	MMVP	AI2D	MathVista	MMMU	SQA	TextVQA	OCRBench	ChartQA	DocVQA
<= 4B Model Comparison													
Qwen2.5VL 3B	-	**53.1**	**65.5**	-	-	**81.4**	**61.2**	**51.2**	79.3	-	**82.8**	**84.0**	**93.9**
InternVL2 4B	-	52.1	60.5	-	-	79.0	58.5	48.3	**96.0**	74.7	78.4	81.5	89.2
DeepSeek-VL2-Tiny	-	-	64.2	-	-	71.6	53.6	40.7	-	**80.7**	80.5	81.0	86.9
MM1.5 3B	-	-	56.9	-	-	65.7	44.4	37.1	85.8	76.5	65.7	74.2	87.5
Phi 3.5 Vision	-	-	53.5	69.3	67.7	77.4	-	43.3	89.0	61.1	59.8	72.0	75.9
Florence-VL 3B	576	-	60.4	70.2	64.7	73.8	52.2	41.8	84.6	69.1	63.0	70.7	-
FUSION 3B (ours)	780	41.5	65.1	76.4	76.0	78.9	54.3	44.7	87.1	71.8	60.0	75.7	70.9
FUSION-X 3B (ours)	620	41.7	63.7	**78.3**	**78.1**	79.2	54.9	44.2	87.3	73.9	63.7	75.8	71.1
FUSION-L 3B (ours)	308	39.5	61.8	76.2	77.0	77.3	48.6	43.4	85.6	71.4	56.9	67.7	63.5
>= 7B Model Comparison													
Qwen2VL 7B	-	**57.4**	**70.1**	-	-	**83.0**	**58.2**	**54.1**	85.5	**84.3**	**86.6**	83.0	**94.5**
InternVL2 8B	-	53.5	64.4	-	-	83.6	58.3	52.6	96.3	77.4	79.4	**83.3**	91.6
LLaVA-OneVision 7B	-	57.5	65.5	-	-	81.6	56.1	47.7	**96.6**	78.5	69.7	78.8	87.5
MM1.5 7B	-	-	62.5	-	-	72.2	47.6	41.8	89.6	76.5	63.5	88.1	78.2
Cambrian 8B	576	-	64.2	72.2	51.3	73.0	49.0	42.7	80.4	71.7	62.4	73.3	77.8
Florence-VL 8B	576	-	64.2	73.4	73.3	74.2	55.5	43.7	85.9	74.2	63.4	74.7	-
Eagle 8B	1024	-	66.5	-	71.6	76.1	52.7	43.8	84.3	77.1	62.6	80.1	86.6
FUSION 8B (ours)	780	46.0	65.2	78.7	78.7	80.4	56.6	43.1	89.2	77.3	63.8	80.3	78.6
FUSION-X 8B (ours)	620	44.7	66.1	**79.2**	**79.7**	81.4	59.4	42.2	90.3	74.7	66.6	79.8	77.8
FUSION-L 8B (ours)	308	42.3	65.1	78.2	76.7	79.2	55.2	41.8	88.3	72.8	59.5	73.0	66.0