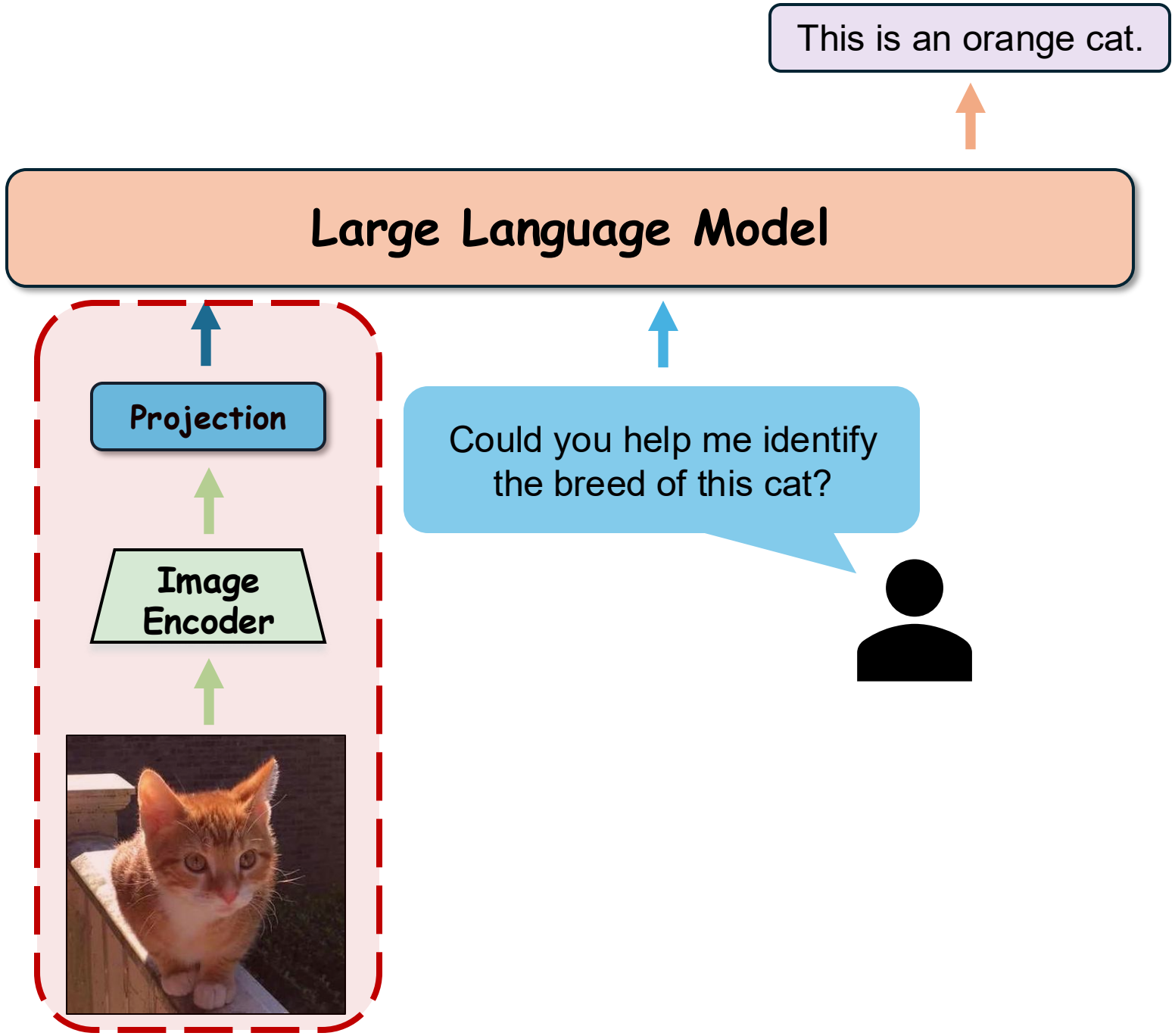
OpenVision

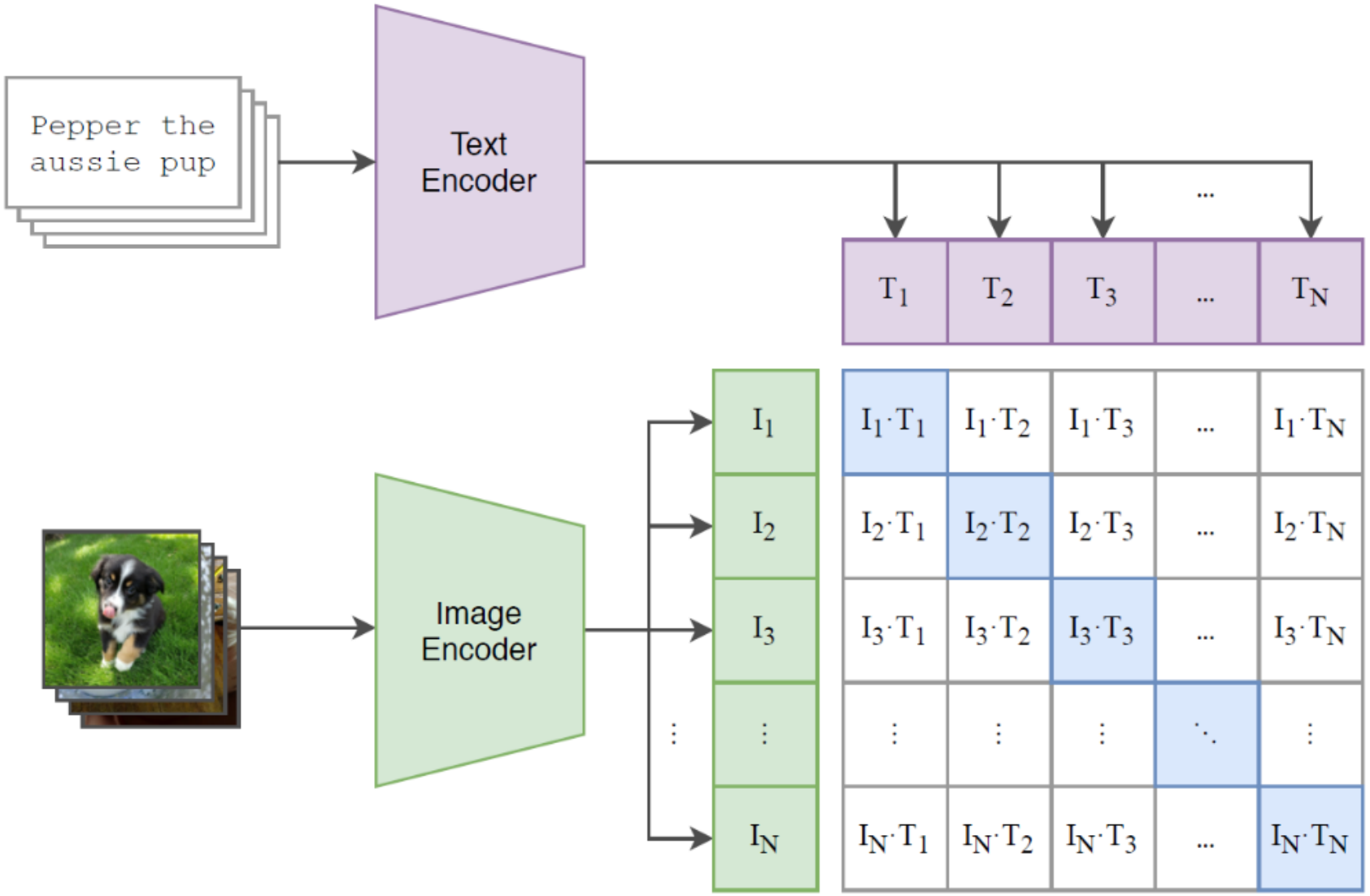
A Fully-Open & Cost-Effective Family of Vision Encoder For MultiModal Learning

Xianhang Li, Yanqing Liu, Haoqin Tu, Cihang Xie

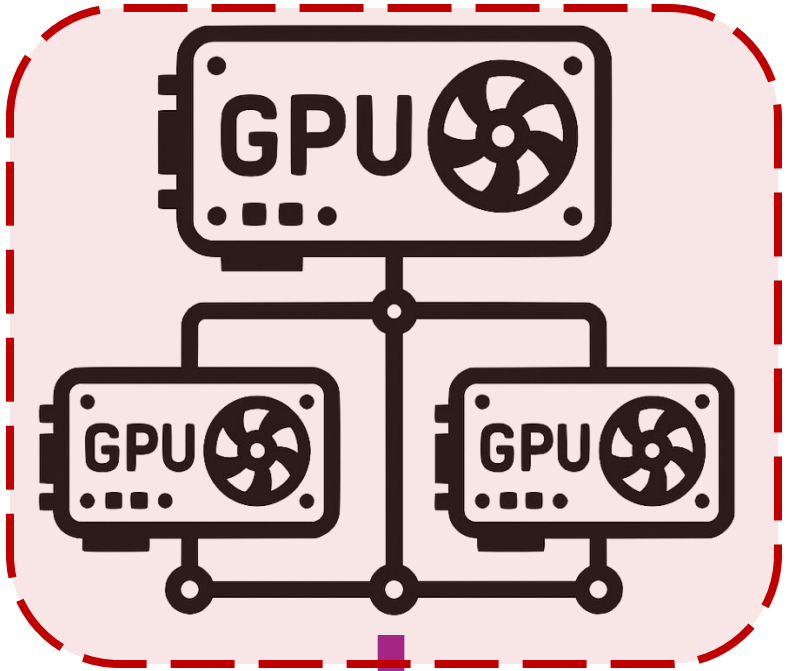
UNIVERSITY OF CALIFORNIA SANTA CRUZ

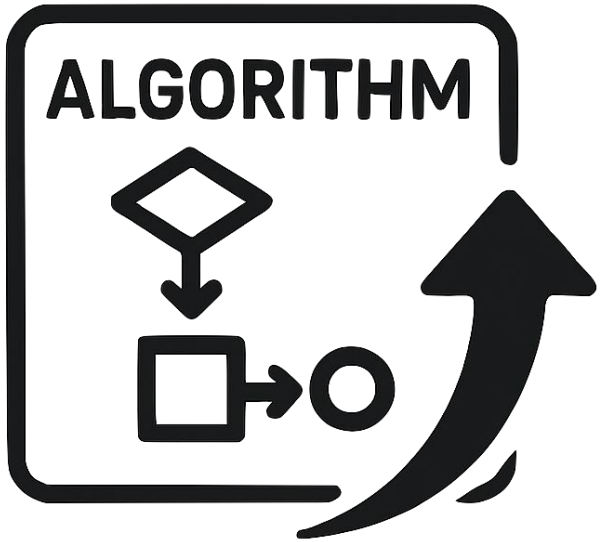

This is an orange cat.
Large Language Model
Projection
Image Encoder
Could you help me identify the breed of this cat?

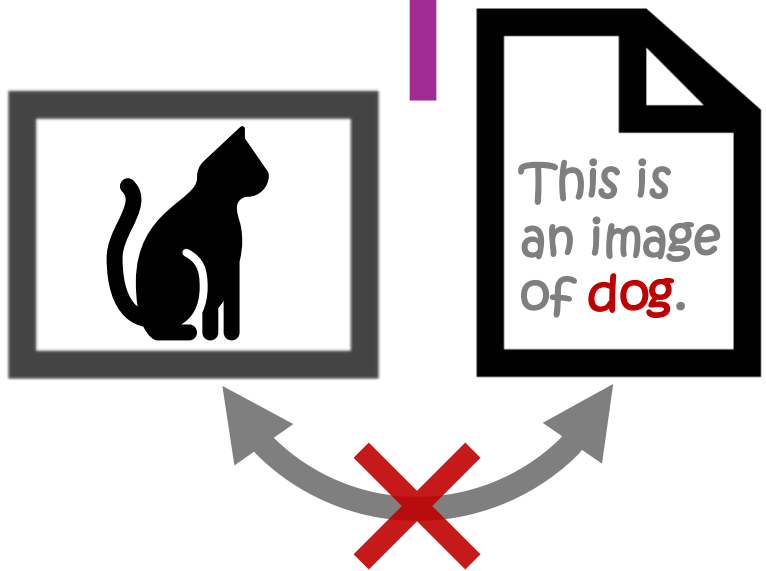
Contrastive Language-Image Pre-Training



Model	Training data	Resolution	# of samples seen	ImageNet zero-shot acc.
ConvNext-Base	LAION-2B	256px	13B	71.5%
ConvNext-Large	LAION-2B	320px	29B	76.9%
ConvNext-XXLarge	LAION-2B	256px	34B	79.5%
ViT-B/32	DataComp-1B	256px	34B	72.8%
ViT-B/16	DataComp-1B	224px	13B	73.5%
ViT-L/14	LAION-2B	224px	32B	75.3%
ViT-H/14	LAION-2B	224px	32B	78.0%
ViT-L/14	DataComp-1B	224px	13B	79.2%
ViT-G/14	LAION-2B	224px	34B	80.1%
ViT-L/14-quickgelu (Original CLIP)	WIT	224px	13B	75.5%
ViT-SO400M/14 (SigLIP)	WebLI	224px	45B	82.0%
ViT-L/14 (DFN)	DFN-2B	224px	39B	82.2%
ViT-SO400M-14-SigLIP-384 (SigLIP)	WebLI	384px	45B	83.1%
ViT-H/14-quickgelu (DFN)	DFN-5B	224px	39B	83.4%
ViT-H-14-378-quickgelu (DFN)	DFN-5B	378px	44B	84.4%

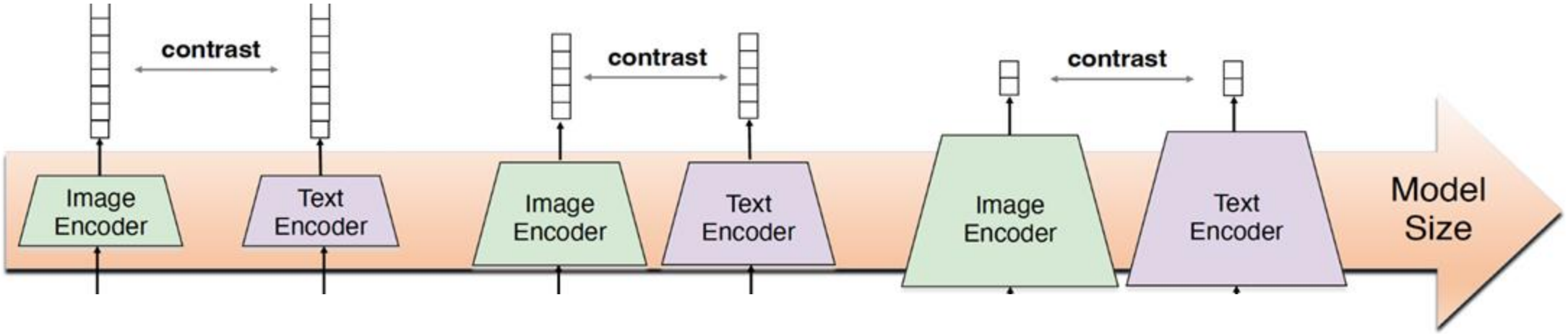




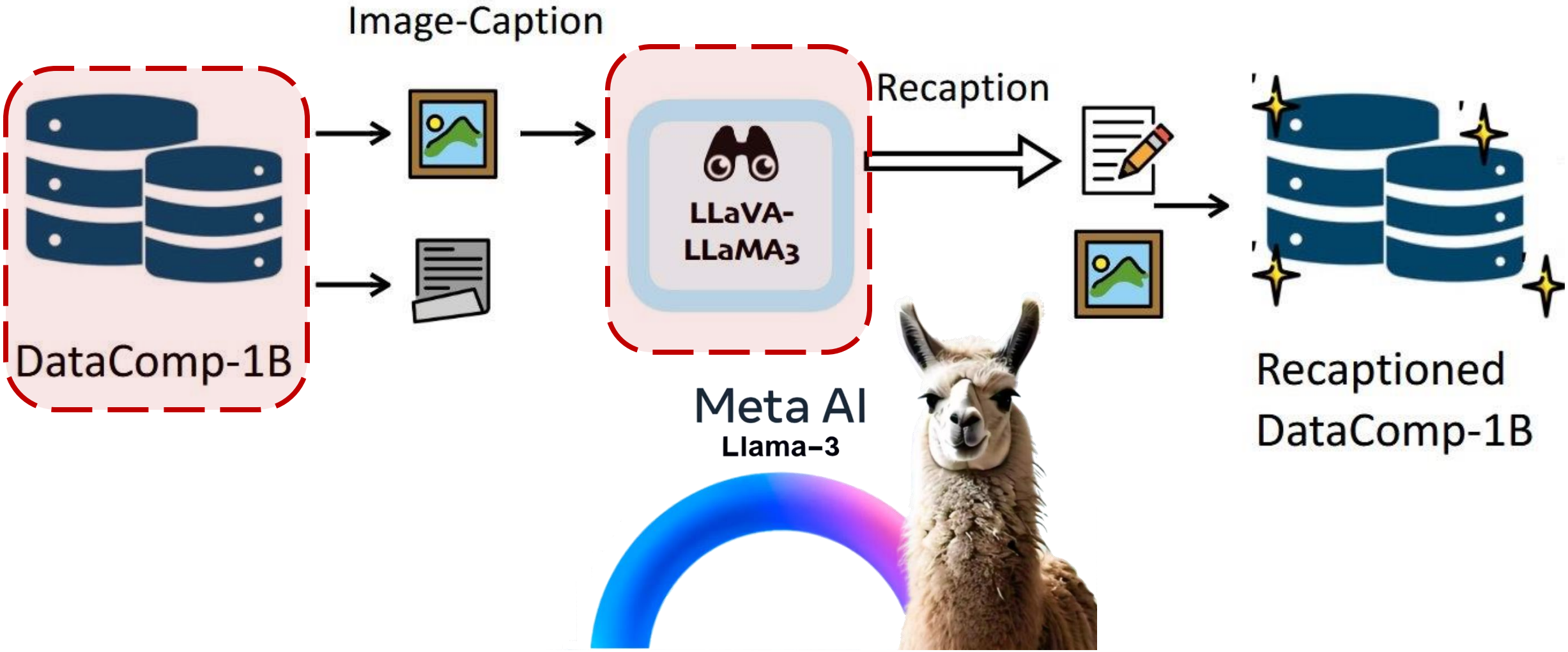


Challenges

An Inverse Scaling Law for CLIP Training



Our Recaption Pipeline



Synthetic Caption (Full)

The image features an adorable orange tabby kitten with a compact and fluffy body perched on a wooden ledge. Its fur is a warm orange color, paired with a white chin, chest, and paws, creating a striking contrast. The kitten‘s eyes are large, round, and a soft shade of orange or brown, giving it an inquisitive expression. The ears of the kitten are perked up, indicating alertness, and they catch the sunlight, making them appear slightly translucent. The wooden ledge is weathered and light-colored, adding to the overall charm of the scene.

Re-caption

CLIPS

Text Decoder

Image Encoder

contrast

Text Encoder

Synthetic Caption (partial)

The image features an adorable orange tabby kitten with a compact and fluffy body perched on a wooden ledge. Its fur is a warm orange color, paired with a white chin, chest, and paws, creating a striking contrast.

Web Caption

An orange cat that is sitting on a ledge.

MLLM Benchmarks

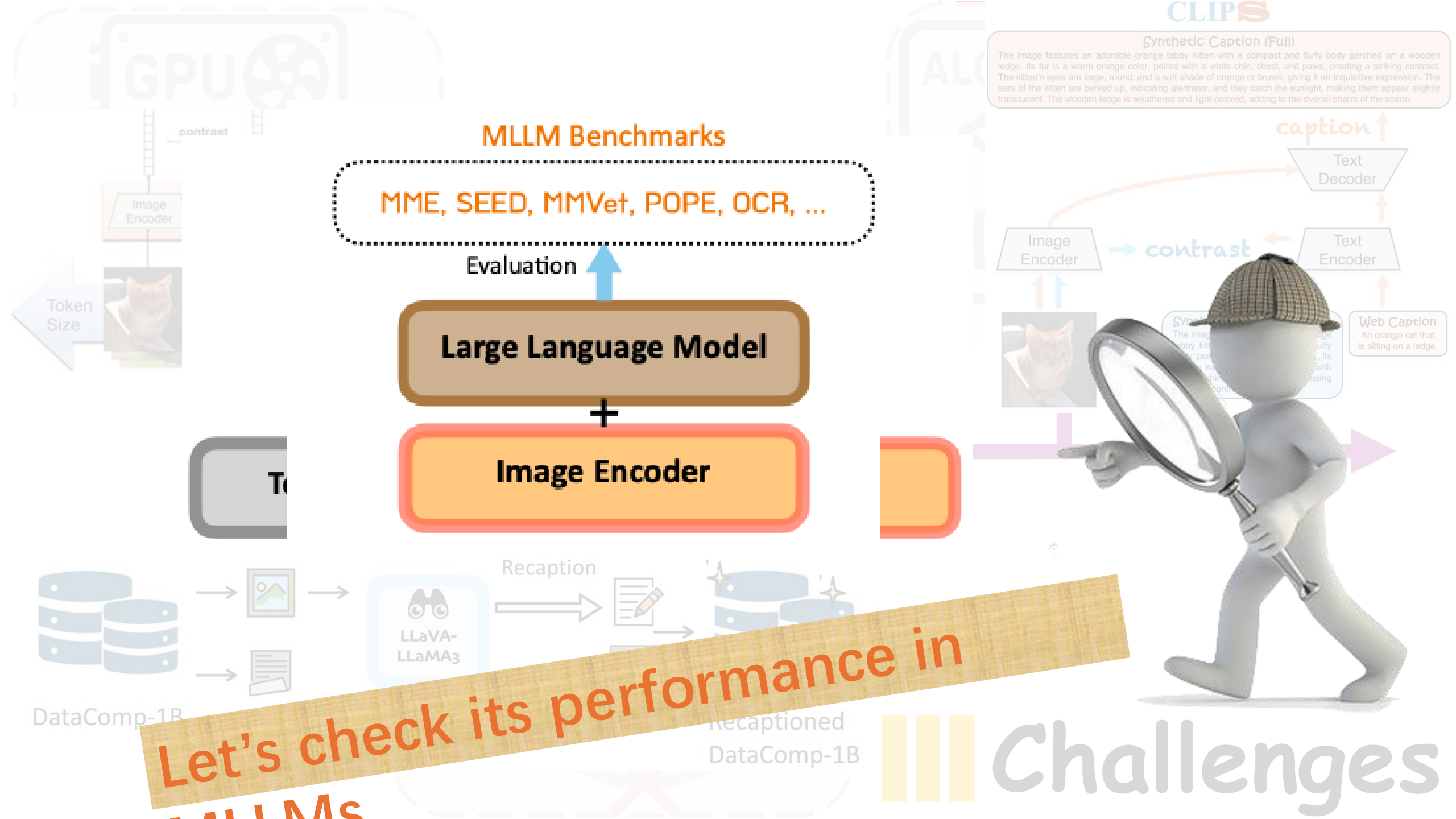
MME, SEED, MMVet, POPE, OCR, ...

Evaluation

Large Language Model

+

Image Encoder



Let's check its performance in MLLMs

Challenges

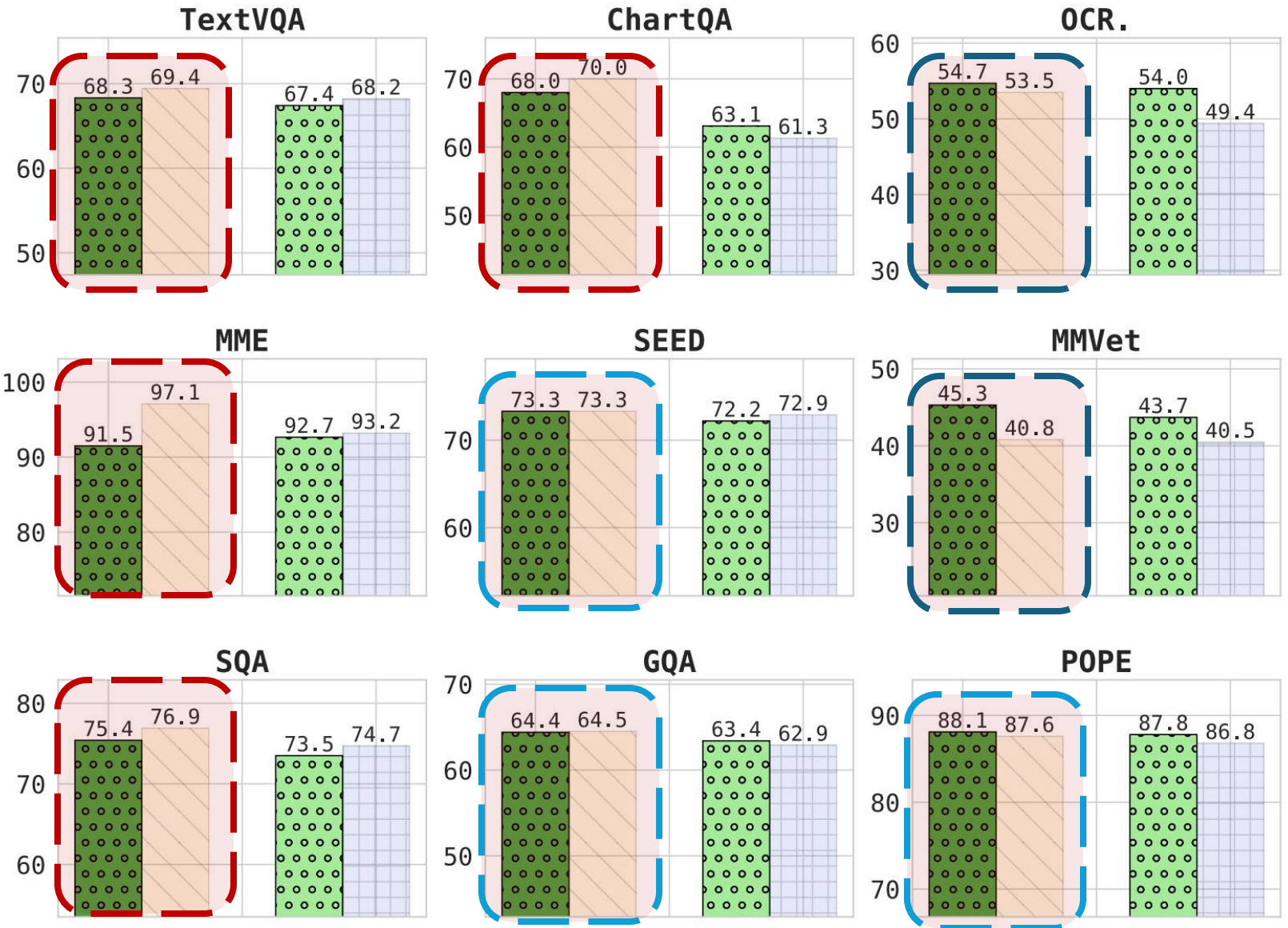
LLaVA-NeXT: Open Large Multimodal Models

llava video paper | llava onevision paper | llava next blog

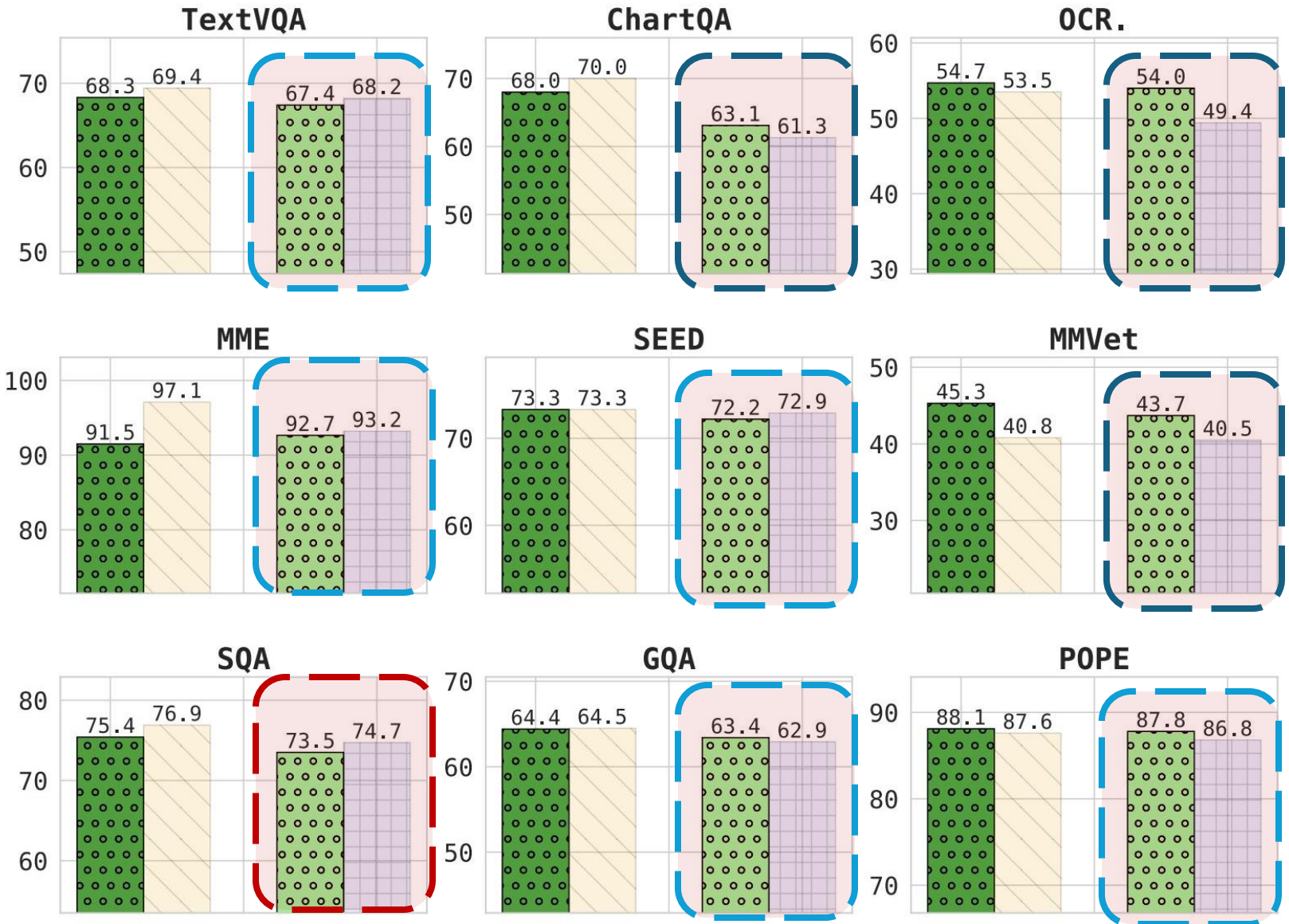
llava onevision demo | llava video demo | llava next interleave demo | Demo OpenBayes贝式计算

llava video checkpoints | llava onevision checkpoints | llava next interleave checkpoints | llava next image checkpoints

VS OpenAI's CLIP-L/14



Google's SigLIP-SoViT/400M




FULLY
OPEN

Ablation: Visual Encoder Sizes

Table 6: Performance of `OpenVision` encoders at different scales with Llama3-8B under LLaVA-1.5.

Vision Encoder	# Res.	# Params.	CLIP-Bench		Text VQA	Chart QA	OCR.	MME	SEED	MMVet	SQA	GQA	POPE
			Cls.	Retr.									
OpenAI-CLIP-L/14	224	303.7M	75.5	36.5/56.3	56.1	13.2	177	1443/306	66.0	32.8	73.4	60.8	85.0
L/14	224	303.7M	78.4	55.3/75.2	57.7	13.9	315	1487/317	69.5	35.2	73.6	62.9	86.4
H/14	224	632.1M	80.4	57.4/77.0	57.9	13.6	330	1501/308	69.3	35.8	75.9	61.9	87.0
B/16	224	87.4M	73.7	51.1/71.6	54.1	11.8	262	1496/293	68.2	30.9	74.4	61.6	86.6
S/16	224	22.4M	65.9	43.6/64.5	51.8	11.0	202	1348/264	65.5	24.6	71.8	60.1	84.6
Ti/16	224	5.9M	49.6	50.0/30.4	48.9	11.7	128	1273/282	59.9	21.8	71.8	57.4	82.0

Ablation: Patchification Size

Table 5: Impact of different patch sizes in LLaVA-1.5. Smaller patch sizes generally improve performance.

Vision Encoder	Patch Size	Text VQA	Chart QA	OCR.	MME	SEED	MMVet	SQA	GQA	POPE
Ti	16	50.2	11.6	139	1329/280	62.0	21.4	73.1	58.0	82.8
Ti	8	54.6	12.9	223	1383/310	66.3	25.1	73.1	59.7	85.3
S	16	54.3	12.0	235	1393/343	67.5	28.8	73.2	61.6	85.7
S	8	59.3	15.9	310	1449/303	70.3	32.5	74.7	62.0	87.1
B	16	57.9	14.5	293	1432/333	69.8	33.2	73.5	62.8	87.8
B	8	61.2	17.2	345	1545/299	71.8	35.5	74.0	63.0	87.0

Ablation: SmolLM-135M

More Stage 2 Data

More & Better Stage 3 Data

Much Higher Resolution

Stage 2	Res.	Stage 3 Data Scale	TextVQA	ChartQA	OCR-VQA	MME	SEED-Bench	MMVet	SQA	GQA	POPE
(1) Scale Stage 2 Data: ×1, ×2, ×4, ×6, ×8 (fix resolution=384, Stage 3=LLaVA (665K))											
×1	384	LLaVA (665K)	33.2	10.3	194	743/212	48.8	15.8	38.2	54.2	85.0
×2	384		34.2	10.6	200	785/204	50.0	16.4	37.0	54.3	85.1
×4	384		34.7	10.2	204	760/210	48.2	16.3	33.9	54.4	84.7
×6	384		34.7	10.1	223	806/201	47.4	15.8	37.5	53.9	84.6
×8	384		35.4	10.8	234	788/215	45.1	16.4	35.6	54.2	84.7
(2) Scale Stage 3 Data: LLaVA (665K), LLaVA-Next (1M), LLaVA-One (3M) (fix Stage 2=×8, Res=384)											
×8	384	LLaVA-Next (1M)	34.5	26.1	284	869/219	50.8	16.4	39.0	53.9	84.5
×8	384	LLaVA-OneVision (3M)	36.3	31.3	319	1051/248	41.6	20.7	37.6	53.3	84.6
(3) Scale Input Resolution: 384→448→512→672→768 (fix Stage 2=×8, Stage 3=LLaVA-OneVision (3M))											
×8	448	LLaVA-OneVision (3M)	37.0	34.9	333	907/246	41.3	18.1	36.8	53.5	85.0
×8	512		38.2	37.2	347	886/226	39.3	20.8	39.0	53.9	86.0
×8	672		38.3	43.2	355	1126/203	46.6	18.8	43.7	53.3	85.5
×8	768		40.6	44.7	382	1080/242	45.8	22.0	39.5	53.2	86.3