InternVL Family

- InternVL (CVPR 2024)
- InternVL3 (2025)

Georgia Tech

Pauline Legendre

- MSCS grad student
- Interests: Computer Vision, Object Detection, VLM
- International student from France

Georgia Tech

Outline

- Problem Statement
- Related Works
- Approach
- Experiments & Results
- Limitations, Societal Implications
- Summary of Strengths, Weaknesses, Relationship to Other Papers

Problem Statement

- Visual encoders in VLLM much less parameters than LLM (~1B vs ~1000B)

VLLM	Visual Encoder (params)	LLM (params)
BLIP-2 (2023)	ViT-L/14 (~0.4B)	Flan-T5-XXL (11B) or OPT (6.7B)
LLaVA-1.0 (2023)	CLIP ViT-L/14 (0.4B)	LLaMA (13B)
MiniGPT-4 (2023)	CLIP ViT-L/14 (0.4B)	Vicuna-13B (LLaMA 13B base)

Problem Statement

- Train visual encoder and language model separately

→ Visual encoder produces tokens that don’t naturally align with LLM

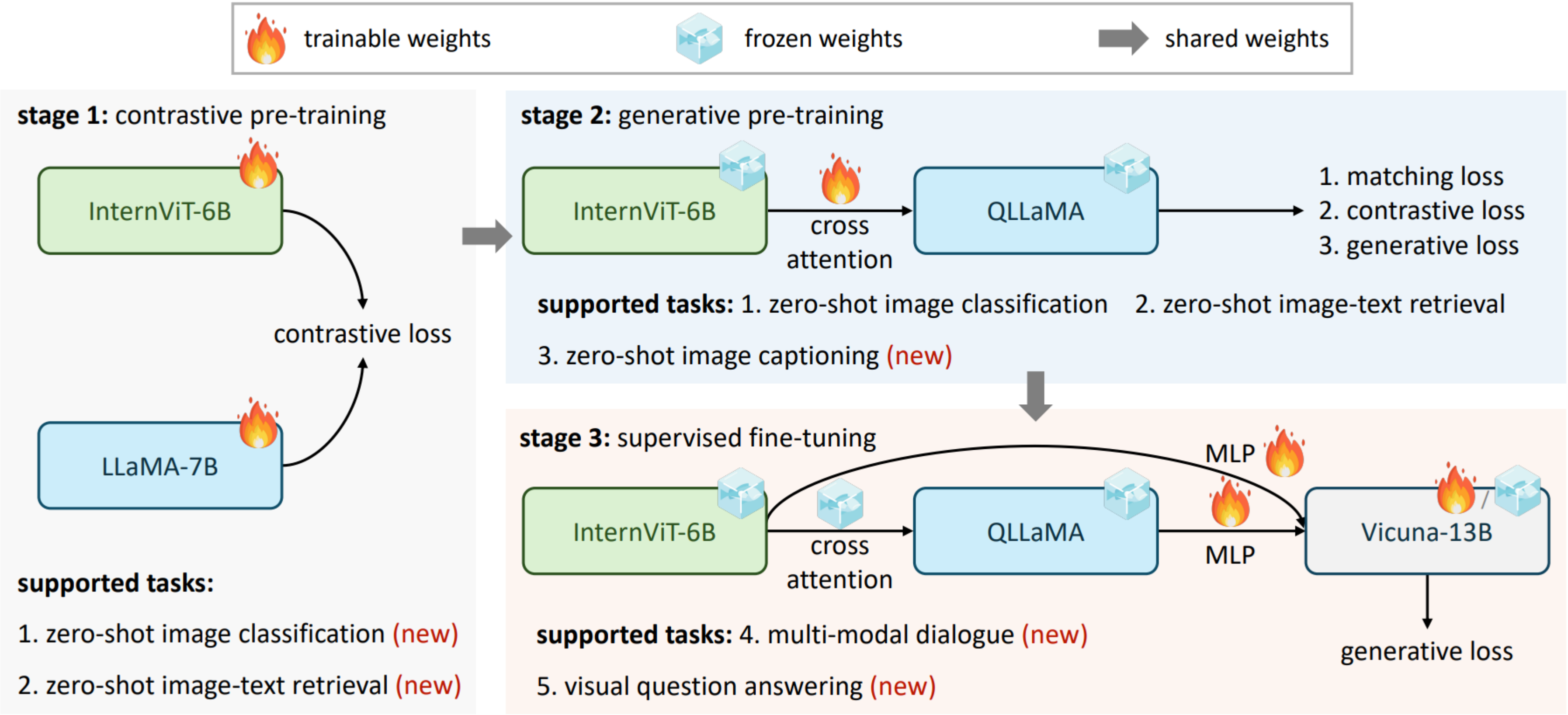
Problem Statement

- “Glue Layer”: module connecting
- Too simple (linear projection) → lose information, poor alignment
- Too heavy (transformers) → adds computational overhead

Georgia Tech

InternVL

Approach


trainable weights
frozen weights
shared weights
stage 1: contrastive pre-training
InternViT-6B
contrastive loss
LLaMA-7B
supported tasks:
1. zero-shot image classification (new)
2. zero-shot image-text retrieval (new)
stage 2: generative pre-training
InternViT-6B
cross attention
QLLaMA
1. matching loss
2. contrastive loss
3. generative loss
supported tasks: 1. zero-shot image classification 2. zero-shot image-text retrieval
3. zero-shot image captioning (new)
stage 3: supervised fine-tuning
MLP
InternViT-6B
cross attention
QLLaMA
MLP
Vicuna-13B
generative loss
supported tasks: 4. multi-modal dialogue (new)
5. visual question answering (new)

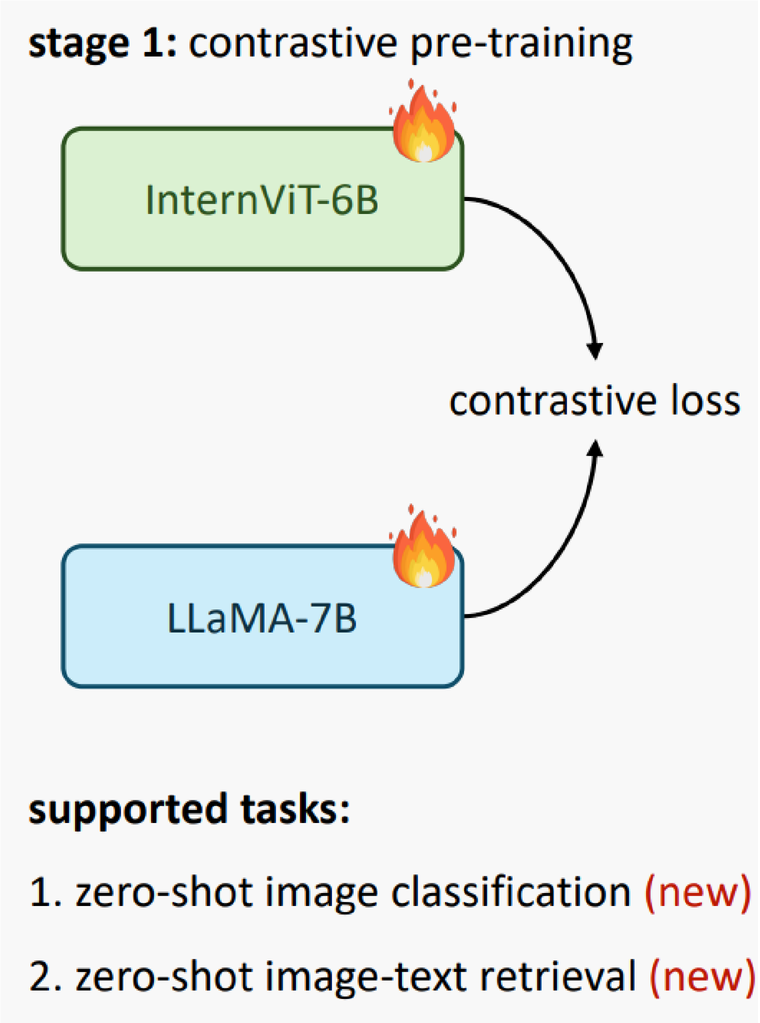
Training Data for stage 1 and 2

dataset	characteristics		stage 1		stage 2	
	language	original	cleaned	remain	cleaned	remain
LAION-en [120]	English	2.3B	1.94B	84.3%	91M	4.0%
LAION-COCO [121]	English	663M	550M	83.0%	550M	83.0%
COYO [14]	English	747M	535M	71.6%	200M	26.8%
CC12M [20]	English	12.4M	11.1M	89.5%	11.1M	89.5%
CC3M [124]	English	3.0M	2.6M	86.7%	2.6M	86.7%
SBU [112]	English	1.0M	1.0M	100%	1.0M	100%
Wukong [55]	Chinese	100M	69.4M	69.4%	69.4M	69.4%
LAION-multi [120]	Multi	2.2B	1.87B	85.0%	100M	4.5%
Total	Multi	6.03B	4.98B	82.6%	1.03B	17.0%

- Publicly available
- Multilingual content
- Combination of datasets and filter out low quality data

Georgia Tech

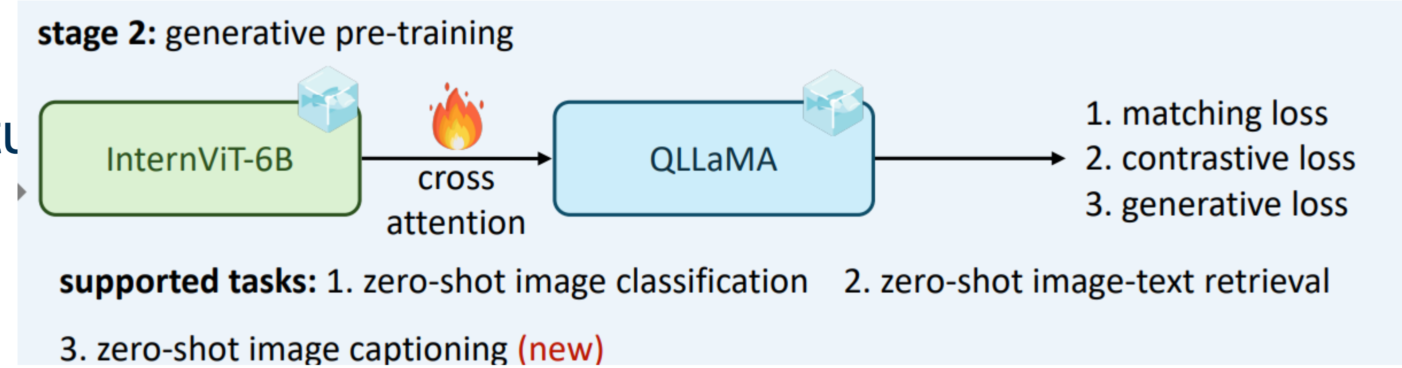
Stage 1: constrastive pretraining



- InternViT-6B and LLaMA-7B trained with contrastive loss
- Match image embeddings and textual embeddings
- Align visual and textual feature spaces

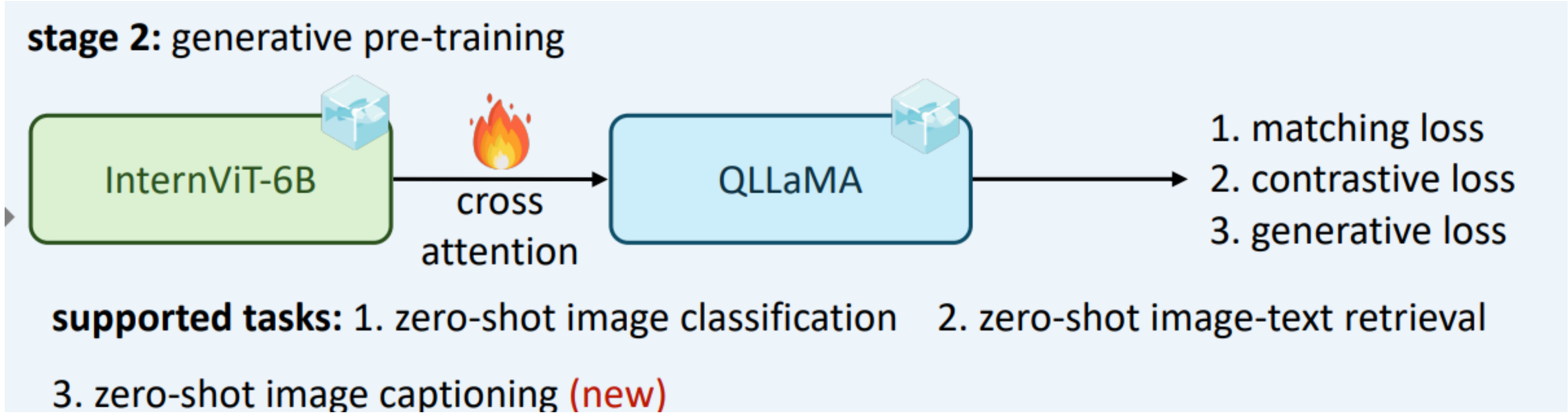
Stage 2: Generative pretraining

- InternViT-6B and QLLaMA frozen
- QLLaMA inherits LLaMA-7B weights from stage 1, InternViT-6B inherits weights as well
- Cross-attention to connect vision featu into LLM
- image-text contrastive (ITC) loss
- image-text matching (ITM) loss
- image-grounded text generation (ITG) loss
- Extract powerful visual representations (further alignment with LLM)



QLLaMA (Query LLaMA)

- Bridges gap between vision encoder and LLM (makes visual features into "acceptable" tokens for LLM)
- Inherited weights from LLaMA → already "speaks" LLM
- Cross-attention → query tokens attend to vision features

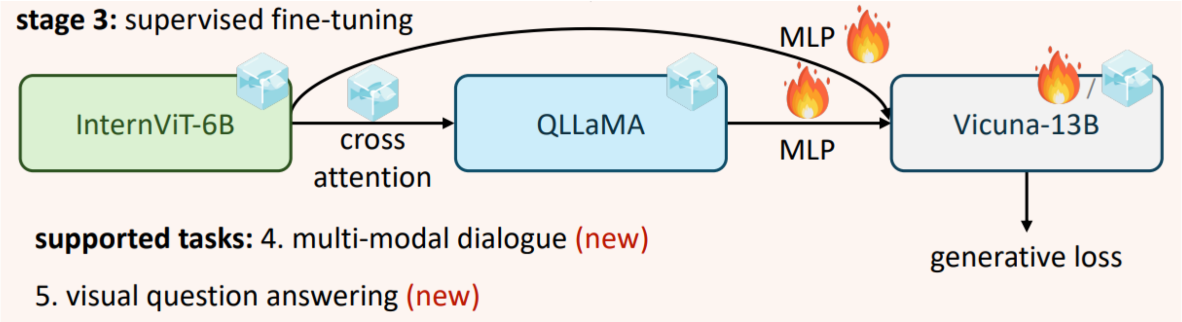


Training Data for stage 3

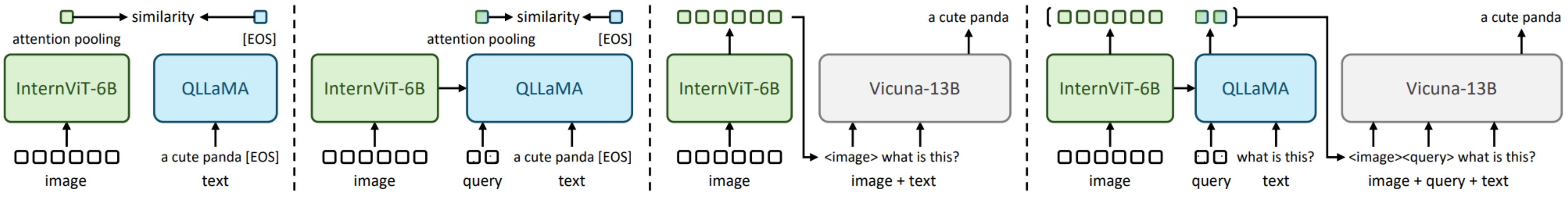
task	#samples	dataset
Captioning	588K	COCO Caption [22], TextCaps [126]
VQA	1.1M	VQAv2 [54], OKVQA [104], A-OKVQA [122], IconQA [99], AI2D [71], GQA [64]
OCR	294K	OCR-VQA [107], ChartQA [105], DocVQA [29], ST-VQA [12], EST-VQA [150], InfoVQA [106], LLaVAR [182]
Grounding	323K	RefCOCO/+/g [103, 170], Toloka [140]
Grounded Cap.	284K	RefCOCO/+/g [103, 170]
Conversation	1.4M	LLaVA-150K [92], SVIT [183], VisDial [36], LRV-Instruction [90], LLaVA-Mix-665K [91]

- High quality instruction data

Stage 3: Supervised Fine-Tuning



- InternViT-6B and QLLaMA frozen
- Vicuna-13B: instruction-tuned LLaMA, partially trainable via MLP adapters
- Train with supervised fine-tuning



(a) InternVL-C (b) InternVL-G (c) InternVL-Chat (w/o QLLaMA) (d) InternVL-Chat (w/ QLLaMA)

- (a) contrastive (stage 1): zero-shot classification, retrieval
- (b) generative (stage 2): captioning, retrieval, zero-shot image-text tasks
- (c) vision encoder outputs fed directly into Vicuna-13B
- (d) full system for multimodal dialogue (stage 3): InternVL-Chat

Linear evaluation for image classification

- Significant improvement over previous SOTA

method	#param	IN-1K	IN-ReaL	IN-V2	IN-A	IN-R	IN-Ske	avg.
OpenCLIP-H [67]	0.6B	84.4	88.4	75.5	–	–	–	–
OpenCLIP-G [67]	1.8B	86.2	89.4	77.2	63.8	87.8	66.4	78.5
DINOv2-g [111]	1.1B	86.5	89.6	78.4	75.9	78.8	62.5	78.6
EVA-01-CLIP-g [46]	1.1B	86.5	89.3	77.4	70.5	87.7	63.1	79.1
MAWS-ViT-6.5B [128]	6.5B	87.8	–	–	–	–	–	–
ViT-22B* [37]	21.7B	89.5	90.9	83.2	83.8	87.4	–	–
InternViT-6B (ours)	5.9B	**88.2**	**90.4**	**79.9**	**77.5**	**89.8**	**69.1**	**82.5**

Semantic segmentation on ADE20K

- Few-shot: fine-tuning backbone with linear head on limited dataset
- InternViT-6B consistently outperforms ViT-22B

method	#param	crop size	1/16	1/8	1/4	1/2	1
ViT-L [137]	0.3B	504^2	36.1	41.3	45.6	48.4	51.9
ViT-G [173]	1.8B	504^2	42.4	47.0	50.2	52.4	55.6
ViT-22B [37]	21.7B	504^2	44.7	47.2	50.6	52.5	54.9
InternViT-6B (ours)	5.9B	504^2	**46.5**	**50.0**	**53.3**	**55.8**	**57.2**

(a) Few-shot semantic segmentation with limited training data. Following ViT-22B [37], we fine-tune the InternViT-6B with a linear classifier.

method	decoder	#param (train/total)	crop size	mIoU
OpenCLIP-G$_{frozen}$ [67]	Linear	0.3M / 1.8B	512^2	39.3
ViT-22B$_{frozen}$ [37]	Linear	0.9M / 21.7B	504^2	34.6
InternViT-6B$_{frozen}$ (ours)	Linear	0.5M / 5.9B	504^2	**47.2**
ViT-22B$_{frozen}$ [37]	UperNet	0.8B / 22.5B	504^2	52.7
InternViT-6B$_{frozen}$ (ours)	UperNet	0.4B / 6.3B	504^2	**54.9**
ViT-22B [37]	UperNet	22.5B / 22.5B	504^2	55.3
InternViT-6B (ours)	UperNet	6.3B / 6.3B	504^2	**58.9**

(b) Semantic segmentation performance in three different settings, from top to bottom: linear probing, head tuning, and full-parameter tuning.

Georgia Tech

Zero-shot image classification

method	IN-1K	IN-A	IN-R	IN-V2	IN-Sketch	ObjectNet	Δ↓	avg.
OpenCLIP-H [67]	78.0	59.3	89.3	70.9	66.6	69.7	5.7	72.3
OpenCLIP-g [67]	78.5	60.8	90.2	71.7	67.5	69.2	5.5	73.0
OpenAI CLIP-L+ [117]	76.6	77.5	89.0	70.9	61.0	72.0	2.1	74.5
EVA-01-CLIP-g [130]	78.5	73.6	92.5	71.5	67.3	72.3	2.5	76.0
OpenCLIP-G [67]	80.1	69.3	92.1	73.6	68.9	73.0	3.9	76.2
EVA-01-CLIP-g+ [130]	79.3	74.1	92.5	72.1	68.1	75.3	2.4	76.9
MAWS-ViT-2B [128]	81.9	–	–	–	–	–	–	–
EVA-02-CLIP-E+ [130]	82.0	82.1	94.5	75.7	71.6	79.6	1.1	80.9
CoCa* [169]	86.3	90.2	96.5	80.7	77.6	82.7	0.6	85.7
LiT-22B* [37, 174]	85.9	90.1	96.0	80.9	–	87.6	–	–
InternVL-C (ours)	**83.2**	**83.8**	**95.5**	**77.3**	**73.9**	**80.6**	**0.8**	**82.4**

(a) ImageNet variants [38, 60, 61, 119, 141] and ObjectNet [8].

method	EN	ZH	JP	AR	IT	avg.
M-CLIP [16]	–	–	–	–	20.2	–
CLIP-Italian [11]	–	–	–	–	22.1	–
Japanese-CLIP-ViT-B [102]	–	–	54.6	–	–	–
Taiyi-CLIP-ViT-H [176]	–	54.4	–	–	–	–
WuKong-ViT-L-G [55]	–	57.5	–	–	–	–
CN-CLIP-ViT-H [162]	–	59.6	–	–	–	–
AltCLIP-ViT-L [26]	74.5	59.6	–	–	–	–
EVA-02-CLIP-E+ [130]	82.0	3.6	5.0	0.2	41.2	–
OpenCLIP-XLM-R-B [67]	62.3	42.7	37.9	26.5	43.7	42.6
OpenCLIP-XLM-R-H [67]	77.0	55.7	53.1	37.0	56.8	55.9
InternVL-C (ours)	**83.2**	**64.5**	**61.5**	**44.9**	**65.7**	**64.0**

(b) Multilingual ImageNet-1K [38, 76].

- Leading performance on various ImageNet variants
- Robust multilingual capabilities

Georgia Tech

Zero-shot image-text retrieval

method	multi-lingual	Flickr30K (English, 1K test set) [116]						COCO (English, 5K test set) [22]						avg.
		Image → Text			Text → Image			Image → Text			Text → Image			
		R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	
Florence [171]	×	90.9	99.1	–	76.7	93.6	–	64.7	85.9	–	47.2	71.4	–	–
ONE-PEACE [143]	×	90.9	98.8	99.8	77.2	93.5	96.2	64.7	86.0	91.9	48.0	71.5	79.6	83.2
OpenCLIP-H [67]	×	90.8	99.3	99.7	77.8	94.1	96.6	66.0	86.1	91.9	49.5	73.4	81.5	83.9
OpenCLIP-g [67]	×	91.4	99.2	99.6	77.7	94.1	96.9	66.4	86.0	91.8	48.8	73.3	81.5	83.9
OpenCLIP-XLM-R-H [67]	✓	91.8	99.4	99.8	77.8	94.1	96.5	65.9	86.2	92.2	49.3	73.2	81.5	84.0
EVA-01-CLIP-g+ [130]	×	91.6	99.3	99.8	78.9	94.5	96.9	68.2	87.5	92.5	50.3	74.0	82.1	84.6
CoCa [169]	×	92.5	99.5	99.9	80.4	95.7	97.7	66.3	86.2	91.8	51.2	74.2	82.0	84.8
OpenCLIP-G [67]	×	92.9	99.3	99.8	79.5	95.0	97.1	67.3	86.9	92.6	51.4	74.9	83.0	85.0
EVA-02-CLIP-E+ [130]	×	93.9	99.4	99.8	78.8	94.2	96.8	68.8	87.8	92.8	51.1	75.0	82.7	85.1
BLIP-2† [81]	×	97.6	100.0	100.0	89.7	98.1	98.9	–	–	–	–	–	–	–
InternVL-C (ours)	✓	94.7	99.6	99.9	81.7	96.0	98.2	70.6	89.0	93.5	54.1	77.3	84.6	86.6
InternVL-G (ours)	✓	**95.7**	**99.7**	**99.9**	**85.0**	**97.0**	**98.6**	**74.9**	**91.3**	**95.2**	**58.6**	**81.3**	**88.0**	**88.8**

method		Flickr30K-CN (Chinese, 1K test set) [77]						COCO-CN (Chinese, 1K test set) [84]						avg.
WuKong-ViT-L [55]	×	76.1	94.8	97.5	51.7	78.9	86.3	55.2	81.0	90.6	53.4	80.2	90.1	78.0
R2D2-ViT-L [159]	×	77.6	96.7	98.9	60.9	86.8	92.7	63.3	89.3	95.7	56.4	85.0	93.1	83.0
Taiyi-CLIP-ViT-H [176]	×	–	–	–	–	–	–	–	–	–	60.0	84.0	93.3	–
AltCLIP-ViT-H [26]	✓	88.9	98.5	99.5	74.5	92.0	95.5	–	–	–	–	–	–	–
CN-CLIP-ViT-H [162]	×	81.6	97.5	98.8	71.2	91.4	95.5	63.0	86.6	92.9	69.2	89.9	96.1	86.1
OpenCLIP-XLM-R-H [67]	✓	86.1	97.5	99.2	71.0	90.5	94.9	70.0	91.5	97.0	66.1	90.8	96.0	87.6
InternVL-C (ours)	✓	90.3	98.8	99.7	75.1	92.9	96.4	68.8	92.0	96.7	68.9	91.9	96.5	89.0
InternVL-G (ours)	✓	**92.9**	**99.4**	**99.8**	**77.7**	**94.8**	**97.3**	**71.4**	**93.9**	**97.7**	**73.8**	**94.4**	**98.1**	**90.9**

- Powerful multilingual image-text retrieval capabilities
- InternVL-G better results InternVL-C (thanks to language middleware QLLaMA)

Multi-Modal Dialogue

method	visual encoder	glue layer	LLM	Res.	PT	SFT	train. param	image captioning COCO	Flickr	NoCaps	visual question answering VQA^{v2}	GQA	VizWiz	VQA^{T}	dialogue MME	POPE
InstructBLIP [34]	EVA-g	QFormer	Vicuna-7B	224	129M	1.2M	188M	–	82.4	123.1	–	49.2	34.5	50.1	–	–
BLIP-2 [81]	EVA-g	QFormer	Vicuna-13B	224	129M	–	188M	–	71.6	103.9	41.0	41.0	19.6	42.5	1293.8	85.3
InstructBLIP [34]	EVA-g	QFormer	Vicuna-13B	224	129M	1.2M	188M	–	82.8	121.9	–	49.5	33.4	50.7	1212.8	78.9
InternVL-Chat (ours)	IViT-6B	QLLaMA	Vicuna-7B	224	1.0B	4.0M	64M	141.4^*	89.7	120.5	72.3^*	57.7^*	44.5	42.1	1298.5	85.2
InternVL-Chat (ours)	IViT-6B	QLLaMA	Vicuna-13B	224	1.0B	4.0M	90M	142.4^*	89.9	123.1	71.7^*	59.5^*	54.0	49.1	1317.2	85.4
Shikra [21]	CLIP-L	Linear	Vicuna-13B	224	600K	5.5M	7B	117.5^*	73.9	–	77.4^*	–	–	–	–	–
IDEFICS-80B [66]	CLIP-H	Cross-Attn	LLaMA-65B	224	1.6B	–	15B	91.8^*	53.7	65.0	60.0	45.2	36.0	30.9	–	–
IDEFICS-80B-I [66]	CLIP-H	Cross-Attn	LLaMA-65B	224	353M	6.7M	15B	117.2^*	65.3	104.5	37.4	–	26.0	–	–	–
Qwen-VL [5]	CLIP-G	VL-Adapter	Qwen-7B	448	$1.4B^\dagger$	$50M^\dagger$	9.6B	–	85.8	121.4	78.8^*	59.3^*	35.2	63.8	–	–
Qwen-VL-Chat [5]	CLIP-G	VL-Adapter	Qwen-7B	448	$1.4B^\dagger$	$50M^\dagger$	9.6B	–	81.0	120.2	78.2^*	57.5^*	38.9	**61.5**	1487.5	–
LLaVA-1.5 [91]	CLIP-L$_{336}$	MLP	Vicuna-7B	336	558K	665K	7B	–	–	–	78.5^*	62.0^*	50.0	58.2	1510.7	85.9
LLaVA-1.5 [91]	CLIP-L$_{336}$	MLP	Vicuna-13B	336	558K	665K	13B	–	–	–	80.0^*	63.3^*	53.6	61.3	1531.3	85.9
InternVL-Chat (ours)	IViT-6B	MLP	Vicuna-7B	336	558K	665K	7B	–	–	–	79.3^*	62.9^*	52.5	57.0	1525.1	86.4
InternVL-Chat (ours)	IViT-6B	MLP	Vicuna-13B	336	558K	665K	13B	–	–	–	80.2^*	63.9^*	54.6	58.7	1546.9	87.1
InternVL-Chat (ours)	IViT-6B	QLLaMA	Vicuna-13B	336	1.0B	4.0M	13B	**146.2***	**92.2**	**126.2**	**81.2***	**66.6***	**58.5**	**61.5**	**1586.4**	**87.6**

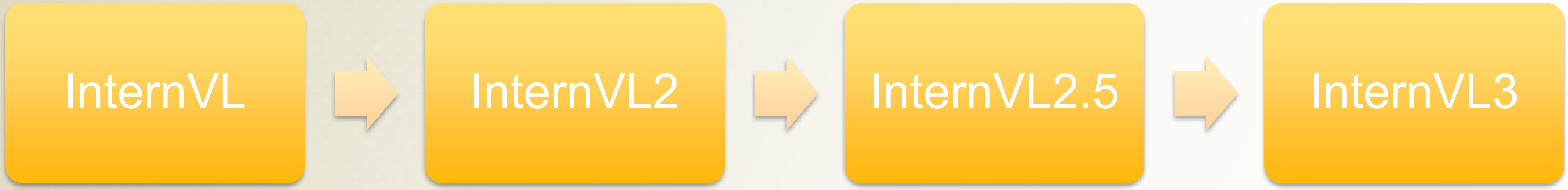
- MME 14 subtasks focused on perception and cognition abilities
- POPE evaluates object hallucination

Limitations

- Some feature misalignment
- Relied on noisy web data
- Can only handle limited resolution
- Weaknesses with abstract reasoning tasks

InternVL Family

InternVL2.5 (late 2024):
- Transition between InternVL2 and InternVL3
- Data quality filtering
- Optimize visual token compression



InternVL2 (07/2024):
- Bigger model family (1B – 108B)
- Dynamic resolution tiling
- Compression

Limitations:
- Not fully unified
- Still computationally heavy

Georgia Tech

InternVL3

Background

- InternVL2.5 still has not unified multimodal pretraining
- Reasoning strategies can still be improved

Variable Visual Position Encoding (V2PE)

- position encodings that can vary
- Better understanding of long multimodal without losing spatial coherence

Native Multimodal Pretraining (NMP)

- Interleave multimodal data with large scale textual corpora
- Model learns linguistic and vision-language alignment together, reducing mismatch between modalities

Post-Training

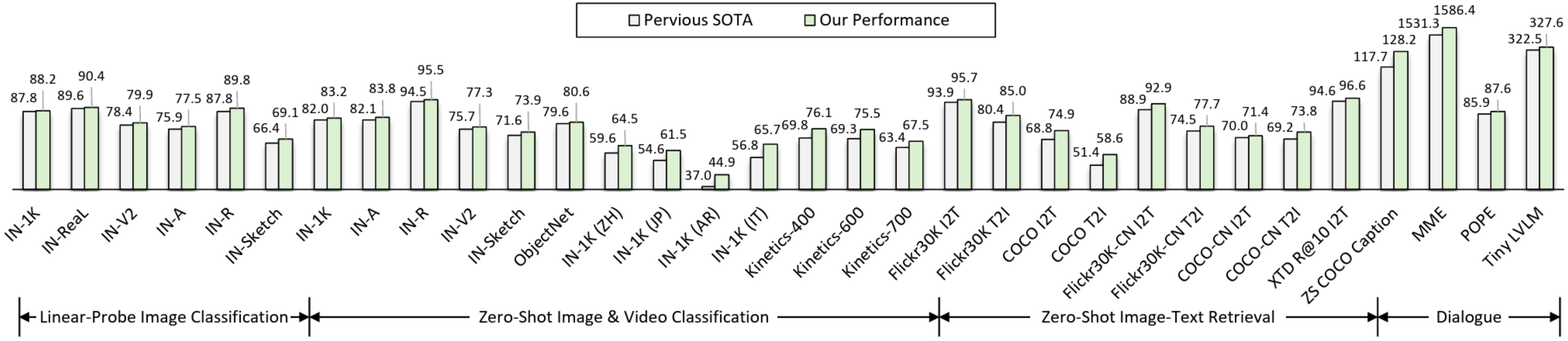
- Supervised Fine-Tuning: higher-quality and more diverse data than prior versions

- Mixed Preference Optimization (MPO): preference-based learning (positive vs negative samples) to align model outputs closer to what humans prefer

Test-Time Scaling

- "best-of-N": multiple responses are generated
- Critic model (VisualPRM-8B) picks the best
- Improves reasoning/math domain evaluations

Results on various generic visual-linguistic tasks



Best performance in all those tasks

Multimodal reasoning and mathematical performance

Model	MMMU	MathVista	MathVision	MathVerse	DynaMath	WeMath	LogicVista	Overall
GPT-4o-20241120 [97]	70.7	60.0	31.2	40.6	34.5	45.8	52.8	47.9
Claude-3.7-Sonnet [3]	75.0	66.8	41.9	46.7	39.7	49.3	58.2	53.9
Gemini-2.0-Flash [30]	72.6	70.4	43.6	47.8	42.1	47.4	52.3	53.7
Gemini-2.0-Pro [29]	69.9	71.3	48.1	67.3	43.3	56.5	53.2	58.5
LLaVA-OV-72B [60]	55.7	67.1	25.3	27.2	15.6	32.0	40.9	37.7
QvQ-72B-Preview [115]	70.3	70.3	34.9	48.2	30.7	39.0	58.2	50.2
Qwen2.5-VL-72B [7]	68.2	74.2	39.3	47.3	35.9	49.1	55.7	52.8
InternVL2.5-78B [18]	70.0	72.3	32.2	39.2	19.2	39.8	49.0	46.0
InternVL3-78B	72.2	79.0	43.1	51.0	35.1	46.1	55.9	54.6
w/ VisualPRM-Bo8 [125]	72.2	80.5	40.8	54.2	37.3	52.4	57.9	56.5

- Strong performance on all tested benchmarks

OCR, chart, and document understanding performance

Model Name	AI2D (w / wo M)	ChartQA (test avg)	TextVQA (val)	DocVQA (test)	InfoVQA (test)	OCR Bench	SEED-2 Plus	CharXiv (RQ / DQ)	VCR-EN-Easy (EM / Jaccard)	Overall
GPT-4V [97]	78.2 / 89.4	78.5	78.0	88.4	75.1	645	53.8	37.1 / 79.9	52.0 / 65.4	70.0
GPT-4o-20240513 [97]	84.6 / 94.2	85.7	77.4	92.8	79.2	736	72.0	47.1 / 84.5	91.6 / 96.4	81.6
Claude-3-Opus [3]	70.6 / 88.1	80.8	67.5	89.3	55.6	694	44.2	30.2 / 71.6	62.0 / 77.7	67.3
Claude-3.5-Sonnet [3]	81.2 / 94.7	90.8	74.1	95.2	74.3	788	71.7	60.2 / 84.3	63.9 / 74.7	78.7
Gemini-1.5-Pro [102]	79.1 / 94.4	87.2	78.8	93.1	81.0	754	–	43.3 / 72.0	62.7 / 77.7	–
LLaVA-OneVision-72B [60]	85.6 / –	83.7	80.5	91.3	74.9	741	–	–	–	–
NVLM-D-72B [28]	85.2 / 94.2	86.0	82.1	92.6	–	853	–	–	–	–
Molmo-72B [31]	– / 96.3	87.3	83.1	93.5	81.9	–	–	–	–	–
Qwen2-VL-72B [121]	88.1 / –	88.3	85.5	96.5	84.5	877	–	–	91.3 / 94.6	–
Qwen2.5-VL-72B [7]	88.7 / –	89.5	83.5	96.4	87.3	885	73.0	49.7 / 87.4	–	–
InternVL2-Llama3-76B [19]	87.6 / 94.8	88.4	84.4	94.1	82.0	839	69.7	38.9 / 75.2	83.2 / 91.3	81.1
InternVL2.5-78B [18]	89.1 / 95.7	88.3	83.4	95.1	84.1	854	71.3	42.4 / 82.3	95.7 / 94.5	83.9
InternVL3-78B	89.7 / 96.0	89.7	84.3	95.4	86.5	906	71.9	46.0 / 85.1	96.0 / 98.6	85.8

- “w/ VisualPRM-Bo8”: the model is evaluated with Best-of-8 settings
- Competitive performance

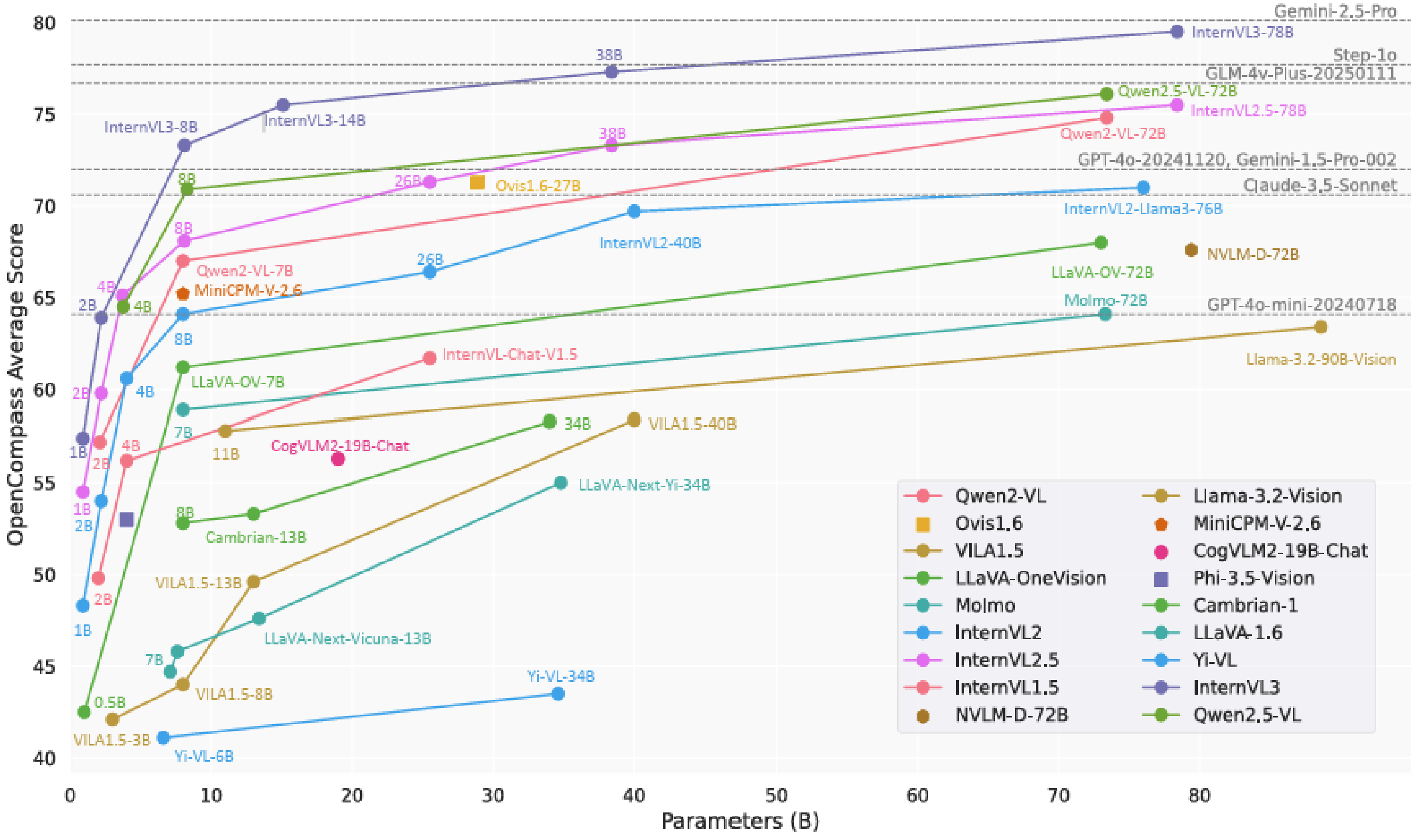
Multi-image and real-world understanding performance

Model Name	BLINK (val)	Mantis Eval	MMIU	Muir Bench	MMT (val)	MIRB (avg)	Overall	RealWorld QA	MME-RW (EN)	WildVision (win rate)	R-Bench (dis)	Overall
GPT-4V [97]	54.6	62.7	–	62.3	64.3	53.1	–	61.4	–	71.8	65.6	–
GPT-4o-20240513 [97]	68.0	–	55.7	68.0	65.4	–	–	75.4	45.2	80.6	77.7	69.7
Claude-3.5-Sonnet [3]	–	–	53.4	–	–	–	–	60.1	51.6	–	–	–
Gemini-1.5-Pro [102]	–	–	53.4	–	64.5	–	–	67.5	38.2	–	–	–
LLaVA-OneVision-72B [60]	55.4	77.6	–	54.8	–	–	–	71.9	–	–	–	–
Qwen2-VL-72B [121]	–	–	–	–	71.8	–	–	77.8	–	–	–	–
Qwen2.5-VL-72B [6]	64.4	–	–	70.7	–	–	–	75.7	63.2	–	–	–
InternVL2-Llama3-76B [19]	56.8	73.7	44.2	51.2	67.4	58.2	58.6	72.2	63.0	65.8	74.1	68.8
InternVL2.5-78B [18]	63.8	77.0	55.8	63.5	70.8	61.1	65.3	78.7	62.9	71.4	77.2	72.6
InternVL3-78B	66.3	79.3	60.4	64.5	73.2	64.3	68.0	78.0	65.4	73.6	77.4	73.6

- Multi-image: competitive results approximating GPT-4o
- Real-world comprehension:

Georgia Tech

Performance of various MLLMs



OpenCompass multimodal academic leaderboard: evaluates models across many tasks (math, OCR, reasoning, VQA, chart understanding…)

Georgia Tech

Limitations

- Huge compute cost
- Substantial memory cost
- Latency in inference (“best of N”)

InternVL3.5 (2025)

- Further results in reasoning abilities
- Better inference efficiency

Thanks!