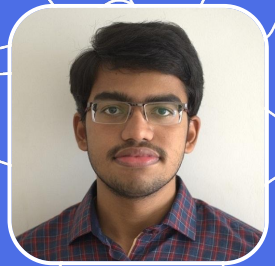

MULTIMODAL AI
SESSION #11
1st November 2024

About Us

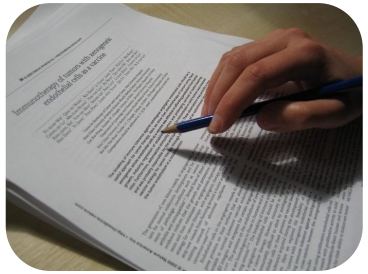
Henry Vo
@_lowkeyboi



Surya Guthikonda
@suryaguthikonda

What to expect?

#multimodal-ml - Channel
@multimodal - Role



Paper Reading Session

Guest Speaker Session

Coding Session

And More...

Second and Fourth
Friday 10 AM ET

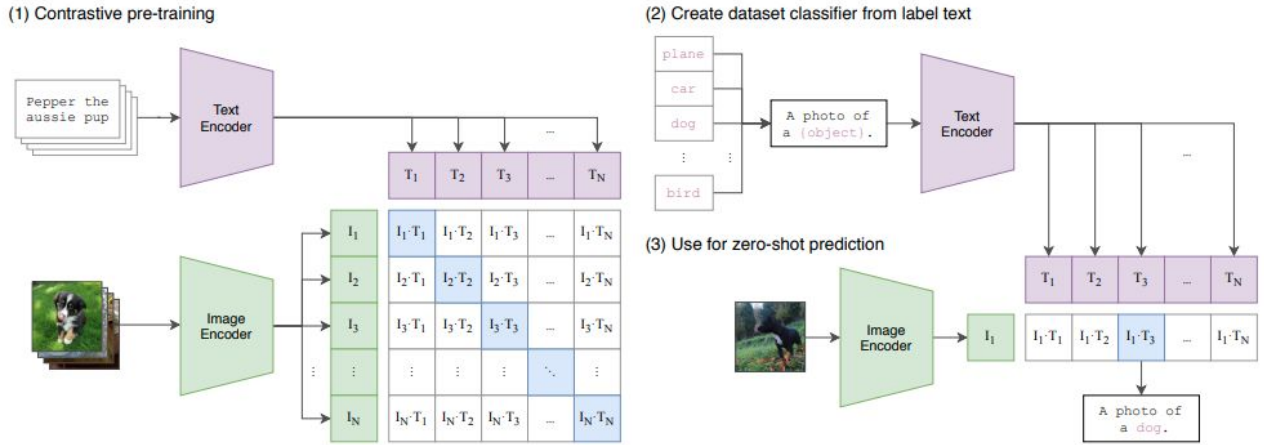
MM1: Methods, Analysis & Insights from Multimodal LLM Pre-training

Cohere For AI

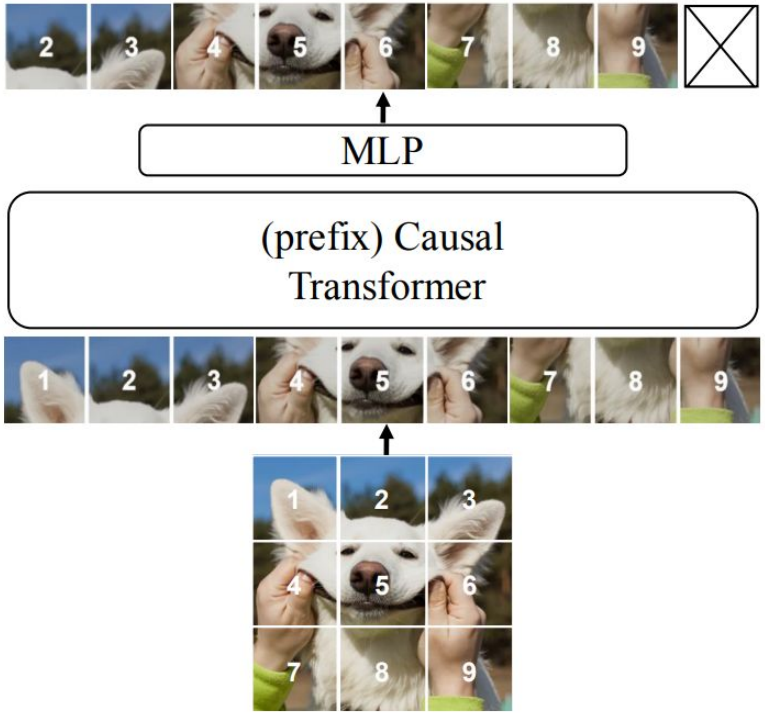
Introduction

- Little to no insights on algorithmic design choices (Open and Closed models)
- Performs ablation studies at small scale.
  - Data
  - Architecture
  - Training Procedure
- Mainly focuses on Pre-training and extends to SFT.
- (April 2024) Pre-trained MM1 model is SOTA and SFT model achieves competitive performance.

Recap: CLIP


(1) Contrastive pre-training
Pepper the aussie pup
Text Encoder
Image Encoder
(2) Create dataset classifier from label text
plane
car
dog
bird
A photo of a {object}.
Text Encoder
(3) Use for zero-shot prediction
Image Encoder
A photo of a dog.

Recap: AIM


2
3
4
5
6
7
8
9
MLP
(prefix) Causal
Transformer
1
2
3
4
5
6
7
8
9
1
2
3
4
5
6
7
8
9

Recap: Connector Types

Average Pooling

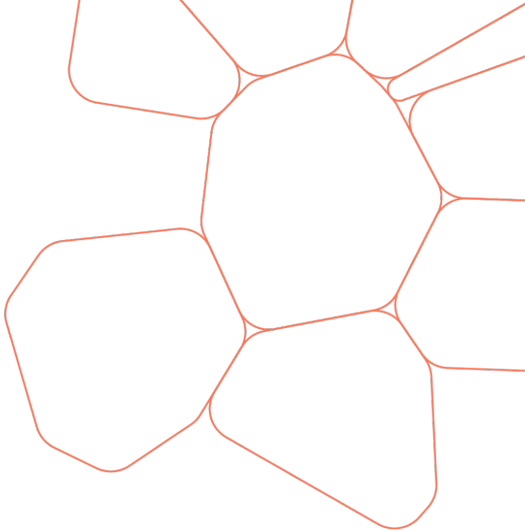
A simple and efficient feature-compressing connector that quickly reduces patch numbers without adding any parameters

Attention Pooling

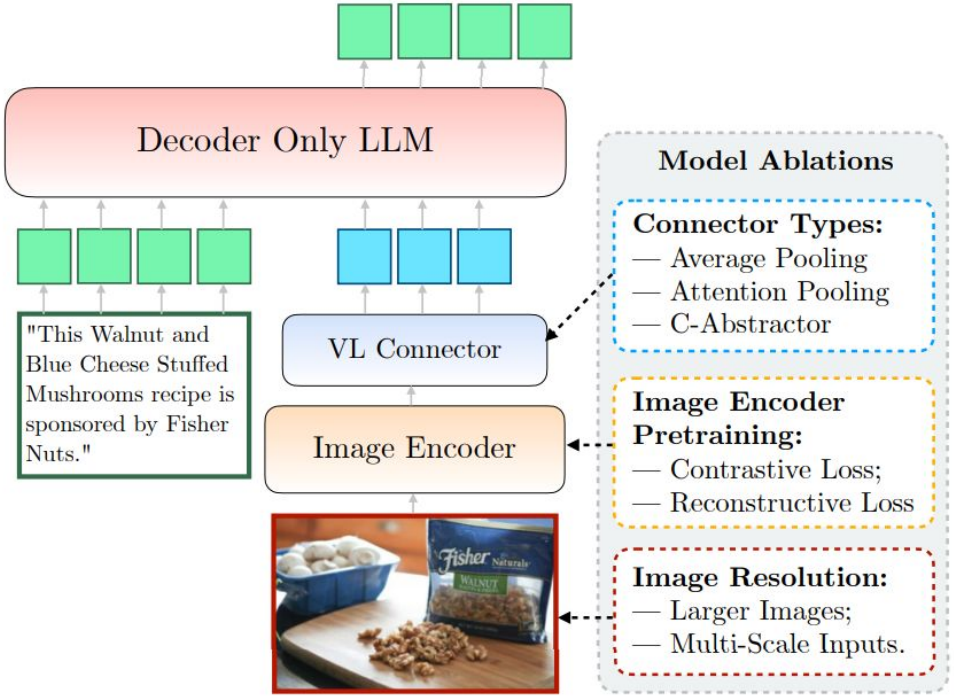
Uses cross-attention with learnable queries to interact with visual features, enabling global weighted pooling

C-Abstractor

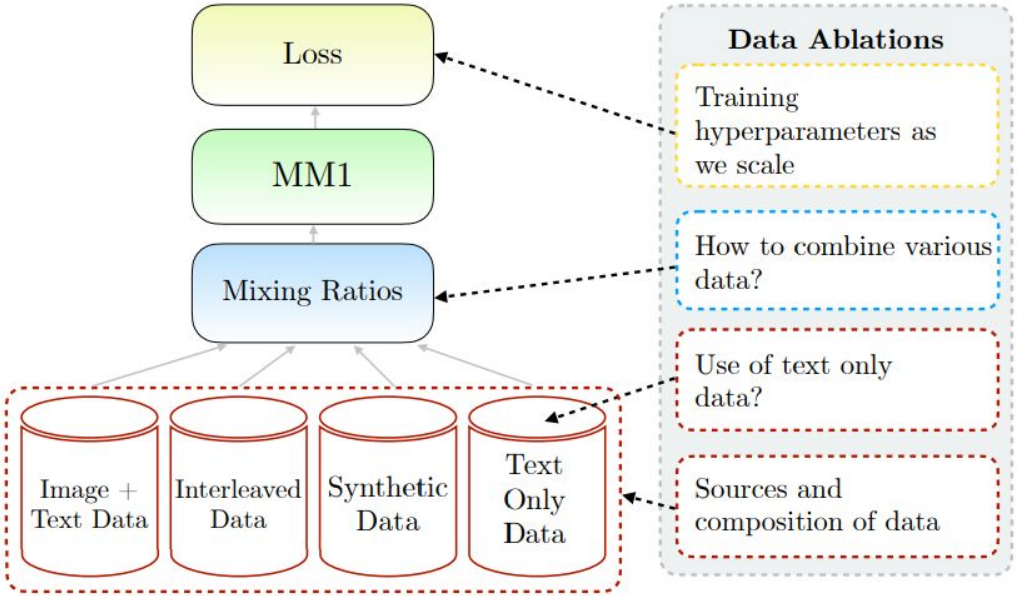
Uses a combination of convolutional layers and average pooling to reduce token numbers



Model Ablations



Data Ablations



Base Model Configuration and Evaluation

- **Image Encoder:**
  - ViT-L/14 (CLIP Loss on DFN-5B and VeCap-300M)
  - Image Size = 336x336
- **Vision-Language Connector:**
  - C-Abstractor with 144 Image Tokens
- **Pre-training Data**
  - Captioned Images (45%)
  - Interleaved Image-text documents (45%)
  - Text-Only (10%)
- **Language Model**
  - 1.2 B Transformer Decoder-only Language Model

To evaluate the different design decisions, used zero-shot and few-shot (4- and 8-shot) performance on a variety of captioning and VQA tasks: COCO Captioning, NoCaps, TextCaps, VQAv2, TextVQA, VizWiz, GQA, and OK-VQA.

Encoder Lesson

Image resolution has the highest impact, followed by model size and training data composition.

	Setup				Results		
	Model	Arch.	Image Res.	Data	0-shot	4-shot	8-shot
Recon.	AIM_{600M}	ViT/600M	224	DFN-2B	36.6	56.6	60.7
Recon.	AIM_{1B}	ViT/1B	224	DFN-2B	37.9	59.5	63.3
Recon.	AIM_{3B}	ViT/3B	224	DFN-2B	38.9	60.9	64.9
Contrastive	$CLIP_{DFN+VeCap}$	ViT-L	224	DFN-5B+VeCap	36.9	58.7	62.2
Contrastive	$CLIP_{DFN}$	ViT-H	224	DFN-5B	37.5	57.0	61.4
Contrastive	$CLIP_{DFN+VeCap}$	ViT-H	224	DFN-5B+VeCap	37.5	60.0	63.6
Contrastive	$CLIP_{DFN+VeCap}$	ViT-L	336	DFN-5B+VeCap	39.9	62.4	66.0
Contrastive	$CLIP_{DFN+VeCap}$	ViT-H	336	DFN-5B+VeCap	40.5	**62.6**	66.3
Contrastive	$CLIP_{OpenAI}$	ViT-L	336	ImageText-400M	39.3	62.2	66.1
Contrastive	$CLIP_{DFN}$	ViT-H	378	DFN-5B	**40.9**	62.5	**66.4**

Table 1: MM1 pre-training ablation across different image encoders (with 2.9B LLM).

- Image resolution from 224 to 336 - 3%
- ViT-L to ViT-H (double parameters) - <1%
- Addition of VeCap-300M Dataset (Few shot) - >1%
- Contrastive methods better than Reconstructive (Less Conclusive)

VL Connector Lesson

Number of visual tokens and image resolution matters most, while the type of VL connector has little effect.

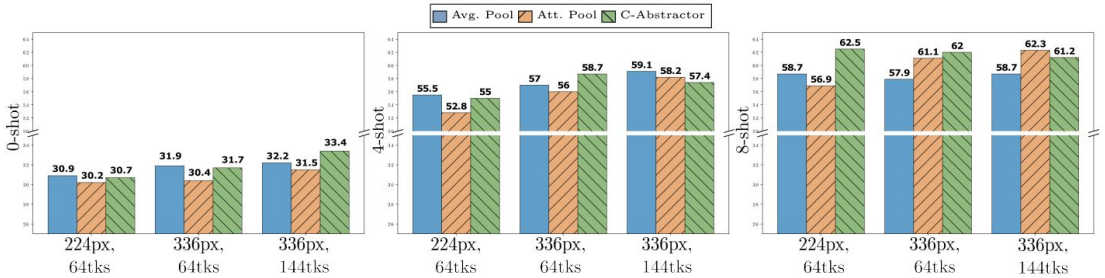


Fig. 4: 0-shot, 4-shot, and 8-shot ablations across different visual-language connectors for two image resolutions, and two image token sizes.

- Increase in Visual Tokens
- Increase in Image Resolution
- Type of Connector has little effect.

Pre-training Data Ablation

Short Text with high relevance to the Image

Data Type	Sources	Size
Captioned Images	CC3M [100], CC12M [13], HQIPT-204M [94], COYO [11], Web Image-Text-1B (Internal)	2B image-text pairs
Captioned Images (Synthetic)	VeCap [57]	300M image-text pairs
Interleaved Image-Text	OBELICS [58], Web Interleaved (Internal)	600M documents
Text-only	Webpages, Code, Social media, Books, Encyclopedic, Math	2T tokens

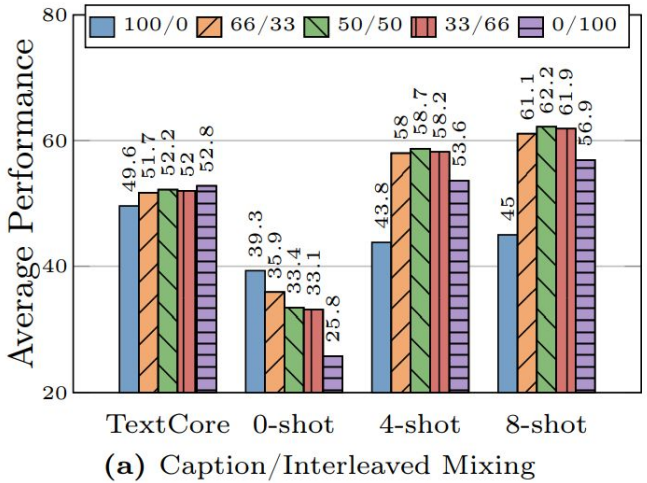
Table 2: List of datasets for pre-training multimodal large language models.

Longer and More Diverse Text with Less Relevance to Surrounding Images

- **Train for 200k steps**
- **Evaluates using TextCore: ARC, PIQA, LAMBADA, WinoGrande, HellaSWAG, SciQ, TriviaQA, and WebQS.**

Data Lesson 1

Interleaved data is instrumental for few-shot and text only performance, while captioning data lifts zero-shot performance.

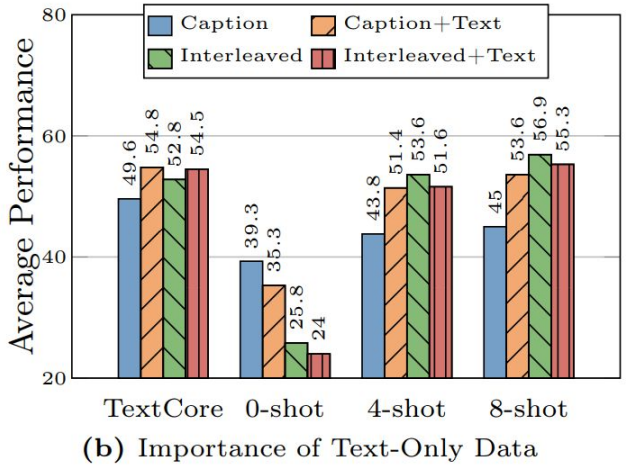


(a) Caption/Interleaved Mixing

- Captioned Data ↑ - 25.8% to 39.3% (Zero Shot) - 3 out of 8 benchmarks are captioning.
- At Least 50% of Interleaved data required for better 4-shot and 8-short performance.
- Use of Interleaved Data Boosts Performance on Captioning Benchmarks (Few-shot)
- Text-Only Performance benefits from Interleaved Data.

Data Lesson 2

Text-only data helps with few-shot and text-only performance.

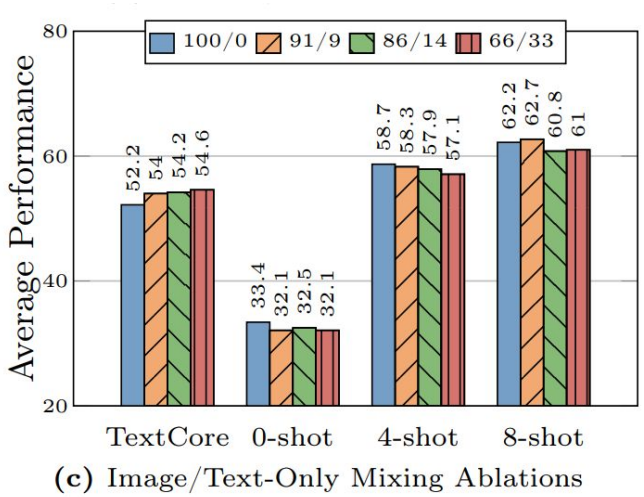


(b) Importance of Text-Only Data

- Text-Only + Captioned Data boost few shot performance.
- Text-Only + Interleaved Data drops in performance (minor).
- Both cases lead to boost in TextCore performance.

Data Lesson 3

Careful mixture of image and text data can yield optimal multimodal performance and retain strong text performance.



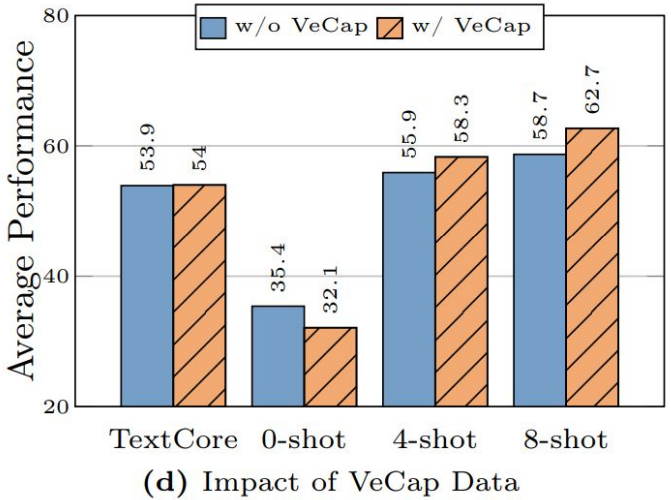
(c) Image/Text-Only Mixing Ablations

Mixing Ratios of Image (Caption and Interleaved) + Text-Only

- Caption/Interleaved/Text Ratio of 5:5:1 (91/9) Achieves good balance of Multimodal and Text-Only Performance.

Data Lesson 4

Synthetic data helps with few-shot learning



(d) Impact of VeCap Data

VeCap (High Quality but Relatively Small - 7% of Caption Data)

- Non-trivial boost of 2.4% to 4% in Few shot performance.

Final Model Configuration

- **Image Encoder:**
  - ViT-H (CLIP Loss on DFN-5B)
  - Image Size = 378x378
- **Vision-Language Connector:**
  - C-Abstractor with 144 Image Tokens
- **Pre-training Data**
  - Captioned Images (45%)
  - Interleaved Image-text documents (45%)
  - Text-Only (10%)
- **Language Model**
  - 3B, 7B and 30B Parameter Language Model

Pre-Training Configuration

- Train for 200k Steps (Approx 400B Tokens)
- Pre-trained Entire Model.
- Sequence Length 4096
- Up to 16 Images Per Sequence at 378X378 Resolution.
- Batch Size of 512 Sequences
- Trained using AXLearn Framework from Apple

Pre-Training Configuration

- **Model Scaling**
  - Performed grid search of learning rate at small scale 9M, 85M, 302M and 1.2B.

$$\eta = \exp(-0.4214 \ln(N) - 0.5535)$$

  - Simple rule of Scaling Weight Decay: $\lambda = 0.1\eta$

- **Scaling via Mixture-of-Experts (MoE)**
  - Two MoE Models
    - 3B-MoE with 64 Experts (64B)
    - 7B-MoE with 32 Experts (47B)

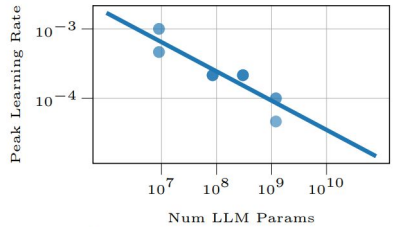


Fig. 6: Optimal peak learning rate as a function of model size. The data points represent experiments that achieved close-to-optimal 8-shot performance for their associated model size.

Pre-Training Results

Model	Shot	Captioning			Visual Question Answering			
		COCO	NoCaps	TextCaps	VQAv2	TextVQA	VizWiz	OKVQA
MM1-3B Model Comparisons								
Flamingo-3B [3]	0†	73.0	–	–	49.2	30.1	28.9	41.2
	8	90.6	–	–	55.4	32.4	38.4	44.6
MM1-3B	0	73.5	55.6	63.3	46.2	29.4	15.6	26.1
	8	**114.6**	**104.7**	**88.8**	**63.6**	**44.6**	**46.4**	**48.4**
MM1-7B Model Comparisons								
IDEFICS-9B [58]	0†	46.0*	36.8	25.4	50.9	25.9	35.5	38.4
	8	97.0*	86.8	63.2	56.4	27.5	40.4	47.7
Flamingo-9B [3]	0†	79.4	–	–	51.8	31.8	28.8	44.7
	8	99.0	–	–	58.0	33.6	39.4	50.0
Emu2-14B [105]	0†	–	–	–	52.9	–	34.4	42.8
	8	–	–	–	59.0	–	43.9	–
MM1-7B	0	76.3	61.0	64.2	47.8	28.8	15.6	22.6
	8	**116.3**	**106.6**	**88.2**	**63.6**	**46.3**	**45.3**	**51.4**
MM1-30B Model Comparisons								
IDEFICS-80B [58]	0†	91.8*	65.0	56.8	60.0	30.9	36.0	45.2
	8	114.3*	105.7	77.6	64.8	35.7	46.1	55.1
	16	116.6*	107.0	81.4	65.4	36.3	48.3	56.8
Flamingo-80B [3]	0†	84.3	–	–	56.3	35.0	31.6	50.6
	8	108.8	–	–	65.6	37.3	44.8	57.5
	16	110.5	–	–	66.8	37.6	48.4	57.8
Emu2-37B [105]	0	–	–	–	33.3	26.2	40.4	26.7
	8	–	–	–	67.8	49.3	54.7	54.1
	16	–	–	–	68.8	50.3	57.0	57.1
MM1-30B	0	70.3	54.6	64.9	48.9	28.2	14.5	24.1
	8	123.1	111.6	92.9	70.9	49.4	49.9	58.3
	16	**125.3**	**116.0**	**97.6**	**71.9**	**50.6**	**57.9**	**59.3**

- MM1 Outperforms all pre-trained MLLMs.
- Superior Performance at 30 B across captioning benchmarks and VizWiz-QA Benchmark
- Comparable to Emu2 on VQAv2, TextVQA and OKVQA.
- At Zero Shot comparable to Flamingo-3B and crushes at 8 shot at small scale.

SFT Data Mixture and Training

Datasets	Size	Prompting Strategy
Text-only SFT	13k	–
LLaVA-Conv [76]	57k	–
LLaVA-Complex [76]	77k	
ShareGPT-4V [15]	102k	
VQAv2 [38]	83k	"Answer the question using a single word or phrase."
GQA [46]	72k	
OKVQA [82]	9k	
OCRVQA [86]	80k	
DVQA [51]	200k	
ChartQA [83]	18k	
AI2D [52]	3k	
DocVQA [85]	39k	
InfoVQA [84]	24k	
A-OKVQA [98]	66k	"Answer with the option's letter from the given choices directly."
COCO Captions [18]	83k	Sample from a pre-generated prompt list, *e.g.*, "Provide a brief description of the given image."
TextCaps [103]	22k	
SynthDog-EN [53]	500k	Sample from a pre-generated prompt list, *e.g.*, "Please transcribe all the text in the picture."
Total	**1.45M**	–

Table 5: List of datasets used for supervised fine-tuning.

- All the datasets are mixed together and randomly sampled during training.
- Finetuned for 10k Steps
- Batch Size of 256
- Sequence Length of 2048
- AdaFactor Optimizer with peak learning of 1e-5 and Cosine decay to 0.
- Image Encoder and LLM are unfrozen.

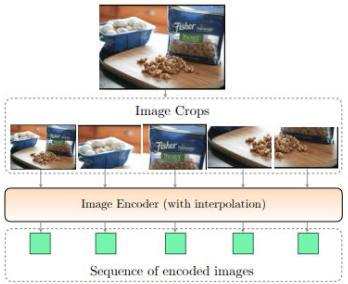
SFT Evaluation

- Used Greedy Decoding To Generate Responses
- Academic VL benchmarks
  - VQAv2
  - TextVQA
  - Image subset of ScienceQA
- MLLM Benchmarks
  - POPE
  - MME
  - MMBench
  - SEED-Bench
  - LLaVA-Bench-in-the-Wild
  - MM-Vet
  - MathVista
  - MMMU

GPT-4 Evaluation with average of 3 runs

Scaling to Higher Resolutions

- **Positional Embedding Interpolation**
  - Allows vision transformer backbone to be adapted to new image resolutions during fine-tuning
  - The method supports image resolutions of 448×448, 560×560, and 672×672
  - At 672×672 resolution with a 14×14 patch size, an image is represented by 2,304 tokens
- **Sub-Image Decomposition**
  - Addresses computational challenges of computing self-attention for images with over 2,000 tokens
  - Uses a strategy of constructing multiple images from a high-resolution input (e.g., 1344×1344)
  - Creates five images: one downsampled high-level representation (672×672) and four detailed sub-images (672×672 each)
  - By using positional embedding interpolation on sub-images, the method can support image resolutions up to 1792×1792 in experiments



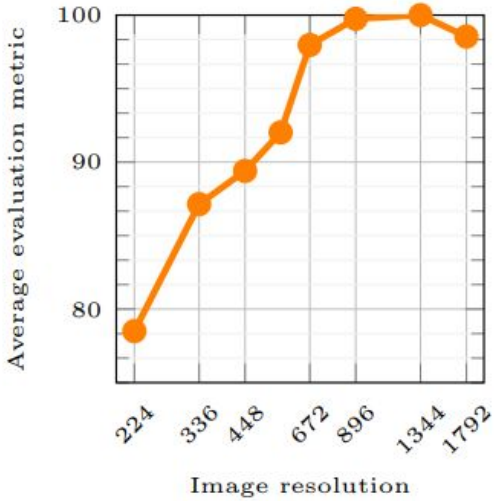
(a) High resolution image input processing.

SFT Results

Model	VQA^{v2}	VQA^{T}	SQA^{I}	MMMU	MathV	MME^{P}	MME^{C}	MMB	SEED	POPE	LLaVA^{W}	MM-Vet
3B Model Comparison												
MobileVLM [20]	–	47.5	61.0	–/–	–	1288.9	–	59.6	–/–	84.9	–	–
LLaVA-Phi [135]	71.4	48.6	68.4	–/–	–	1335.1	–	59.8	–/–	85.0	–	28.9
Imp-v1 [99]	79.45	59.38	69.96	–/–	–	1434.0	–	66.49	–	88.02	–	33.1
TinyLLaVA [133]	79.9	59.1	69.1	–/–	–	1464.9	–	66.9	–/–	86.4	75.8	32.0
Bunny [42]	79.8	–	70.9	38.2/33.0	–	1488.8	289.3	68.6	62.5/–	86.8	–	–
Gemini Nano-2 [106]	67.5	65.9	–	32.6/–	30.6	–	–	–	–	–	–	–
MM1-3B-Chat	82.0	71.9	69.4	33.9/33.7	32.0	1482.5	279.3	67.8	63.0/68.8	87.4	72.1	43.7
MM1-3B-MoE-Chat	82.5	72.9	76.1	38.6/35.7	32.6	1469.4	303.1	70.8	63.9/69.4	87.6	76.8	42.2
7B Model Comparison												
InstructBLIP-7B [24]	–	50.1	60.5	–/–	25.3	–	–	36.0	53.4/–	–	60.9	26.2
Qwen-VL-Chat-7B [5]	78.2	61.5	68.2	35.9/32.9	–	1487.5	360.7	60.6	58.2/65.4	–	–	–
LLaVA-1.5-7B [74]	78.5	58.2	66.8	–/–	–	1510.7	316.1	64.3	58.6/66.1	85.9	63.4	31.1
ShareGPT4V-7B [15]	80.6	60.4	68.4	–/–	–	1567.4	376.4	68.8	–/–	–	72.6	–
LVIS-Ins4V-7B [113]	79.6	58.7	68.3	–/–	–	1528.2	–	66.2	60.6/–	86.0	67.0	31.5
VILA-7B [71]	79.9	64.4	68.2	–/–	–	1531.3	–	68.9	61.1/–	85.5	69.7	34.9
SPHINX-Intern2 [36]	75.5	–	70.4	–/–	35.5	1260.4	294.6	57.9	68.8/–	86.9	57.6	36.5
LLaVA-NeXT-7B [75]	81.8	64.9	70.1	35.8/–	34.6	1519	332	67.4	–/70.2	86.53	81.6	43.9
MM1-7B-Chat	82.8	72.8	72.6	37.0/35.6	35.9	1529.3	328.9	72.3	64.0/69.9	86.6	81.5	42.1
MM1-7B-MoE-Chat	83.4	73.8	74.4	40.9/37.9	40.9	1597.4	394.6	72.7	65.5/70.9	87.8	84.7	45.2
30B Model Comparison												
Emu2-Chat-37B [105]	84.9	66.6	–	36.3/34.1	–	–	–	–	62.8/–	–	–	48.5
CogVLM-30B [114]	83.4	68.1	–	32.1/30.1	–	–	–	–	–	–	–	56.8
LLaVA-NeXT-34B [75]	83.7	69.5	81.8	51.1/44.7	46.5	1631	397	79.3	–/75.9	87.73	89.6	57.4
MM1-30B-Chat	83.7	73.5	81.0	44.7/40.3	39.4†	1637.6	431.4	75.1	65.9/72.1	87.6	89.3	48.7
Gemini Pro [106]	71.2	74.6	–	47.9/–	45.2	–	436.79	73.6	–/70.7	–	–	64.3
Gemini Ultra [106]	77.8	82.3	–	59.4/–	53.0	–	–	–	–	–	–	–
GPT4V [1]	77.2	78.0	–	56.8/55.7	49.9	–	517.14	75.8	67.3/69.1	–	–	67.6

- MM1-3B-Chat and MM1-7B-Chat outperforms all listed models of the same size
- MM1-3B-Chat and MM1-7B-Chat show particularly strong performance on VQAv2, TextVQA, ScienceQA, MMMU and MathVista
- MoE models performs better than dense counterparts.
- MM1-30B-Chat outperforms Emu2-Chat37B and CogVLM-30B on TextVQA, SEED, and MMMU
- Competitive Performance with LLaVA-NeXT

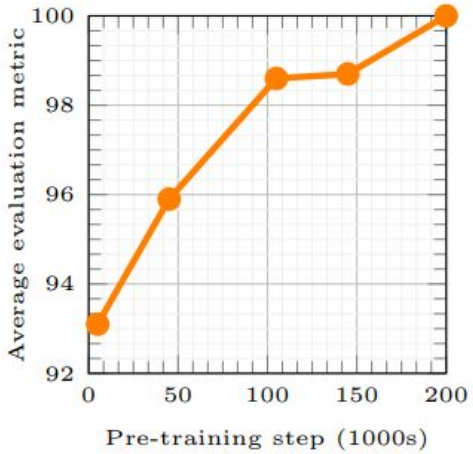
Impact of Image Resolution



(b) Impact of image resolution on SFT performance.

- Image resolution of 336x336 -> 1344x1344 results in 15% relative increase.
- For the largest image resolution of 1792×1792, average performance decreases slightly.

Impact of Pre-training



(c) Impact of pre-training on SFT performance.

- Model consistently improves as it has seen more pre-training data
- Large-scale multimodal pre-training enables strong in-context few-shot learning and multi-image reasoning capabilities

Conclusion

- Pre-training
  - Image resolution has the highest impact, followed by model size and training data composition.
  - Number of visual tokens and image resolution matters most, while the type of VL connector has little effect.
  - Interleaved data is instrumental for few-shot and text only performance, while captioning data lifts zero-shot performance.
  - Text-only data helps with few-shot and text-only performance.
  - Careful mixture of image and text data can yield optimal multimodal performance and retain strong text performance.
  - Synthetic data helps with few-shot learning
- SFT
  - With Increase in Image Resolution (upto certain point), the model performs better.
  - Pre-training the model with more data improves the model performance.

Next Session

MM1.5: Methods, Analysis & Insights from Multimodal LLM Fine-tuning

Haotian Zhang°, **Mingfei Gao**°, **Zhe Gan**°, **Philipp Dufter***, **Nina Wenzel***, **Forrest Huang***, **Dhruti Shah***, **Xianzhi Du***, **Bowen Zhang***, **Yanghao Li***, **Sam Dodge**, **Keen You**, **Zhen Yang**, **Aleksei Timofeev**, **Mingze Xu**, **Hong-You Chen**, **Jean-Philippe Fauconnier**, **Zhengfeng Lai**, **Haoxuan You**, **Zirui Wang**, **Afshin Dehghan**, **Peter Grasch***, **Yinfei Yang**†
Apple
{haotian.zhang2,mgao22,zhe.gan,yinfeiy}@apple.com
°First authors; *Core authors; †Project lead

Abstract

We present **MM1.5**, a new family of multimodal large language models (MLLMs) designed to enhance capabilities in text-rich image understanding, visual referring and grounding, and multi-image reasoning. Building upon the MM1 architecture, MM1.5 adopts a data-centric approach to model training, systematically exploring the impact of diverse data mixtures across the entire model training lifecycle. This includes high-quality OCR data and synthetic captions for continual pre-training, as well as an optimized visual instruction-tuning data mixture for supervised fine-tuning. Our models range from 1B to 30B parameters, encompassing both dense and mixture-of-experts (MoE) variants, and demonstrate that careful data curation and training strategies can yield strong performance even at small scales (1B and 3B). Additionally, we introduce two specialized variants: MM1.5-Video, designed for video understanding, and MM1.5-UI, tailored for mobile UI understanding. Through extensive empirical studies and ablations, we provide detailed insights into the training processes and decisions that inform our final designs, offering valuable guidance for future research in MLLM development.


THANK YOU :)