

# Current Pre-training Recipe for Obsidian-8B Model (07/09/2024)

PoC: Yanbei Chen Rakesh Chada Siddharth Choudhary

This document is structured as follows:

- In the first section, we describe the current recipe for 8B MM Pretraining (both data/training) starting from post-rampdown 8B LLM and training with 8K context length.
- In the second section, we describe the context extension recipe to extend 8K to 32K.
- Next, we show initial results for MM pretraining starting with pre-rampdown 8B LLM instead of a post-rampdown model.
- Finally, we provide recommendations for 70B PT starting from pre-rampdown LLM with Constant LR followed by Rampdown.

## Takeaways

The main takeaways that we observe are:

- Pre-training longer with a mix of MM and Text keeps improving performance on MM benchmark while maintaining Text performance (Shows the mean MM model performance as compared to the number of training tokens).
- Including video in the mix and training using dual tower model adds Video capability with minimal regression to existing Image based benchmarks (Obsidian-8B - Video + Image + Text - 22M Samples Seen (182B tokens)).
- Overall, we find that including video only during context extension works the best. So, our final MM training recipe on Olympus 8B model has the following schedule: 8K pre-training on Image + Text → 32K context extension on Image + Text + Video, as detailed below.
  - (1) pre-training on Image + Text with 8K context length on MM image data mixture and text data mixture;
  - (2) joint context extension on Image + Text + Video with 32K context length on long context MM data mixture (including packed MM image data, native long MM image and video data) and long context text mixture.
- Initial results show that continued pre-training of pre-rampdown text backbone with multimodal data shows improved performance at longer training scales. This presents flexibility to add new datasets and capabilities to an existing model checkpoint in the future.

## MM Pre-training Recipe with 8K Context Length

**Setup:** For MM pre-training, we start from the [post-rampdown 8B LLM](#) trained with 16T tokens and which is context extended to 32K. We use the updated vision encoder (from Nitesh Sekhar/Gukyeong Kwon) trained with a mix of Obsidian PT/SFT datasets. The model is being trained using cosine LR schedule for 30K steps using with a max learning rate of 1e-4 and global batch size of 4096. We use 8192 context length for training this model.

**Data mixture:** We use the following MM data mixture consisting of Captioning/Scene Text (11%), Document Understanding (16%), Chart Understanding (8%), Grounding (8%), Screenshots (19%) and MM Interleaved (38%). We have mixed the MM data with [text pretraining data](#) with 30% MM and 70% Text. We use 70% Text in each batch to minimize text regression based on our earlier [experimental results](#). For SFT, we ablate both December and April SFT mix.

Data Type	Data Source	Number of samples (in millions)	Proportion (%)
Captioning + Scene Text	Datacomp OCR	8	4.2328
	Laion OCR	8	4.2328
	LaionTitan	5	2.6455
Document Understanding	IDL	14.5	7.67196
	DSIA	14.5	7.67196
	Texttract OCR	1.2	0.63492
Chart Understanding	Chart-Public	3.9	2.06349
	Chart-CC-3M	2.9	1.53439
	Chart-OBELICS-3M	3	1.5873
	Chart-Synthetic-Table	2.5	1.32275
	Chart-Synthetic-Numerical	2.5	1.32275
Interleaved	DocuVista-en	31.5	16.66667
	MMKong-en	31.5	16.66667
	MMKong-multilingual	9	4.7619
Grounding	GRIT	5	2.6455
	DataComp-Grounding	5	2.6455
	OBELICS-Grounding	5	2.6455
Screenshots	ScreenVista-Fullpage	5	2.6455
	ScreenVista-Landscape	5	2.6455
	ScreenVista-Mobile	5	2.6455
	ScreenVista-Portrait	5	2.6455
	ScreenVista-Square	10	5.29101
	Collage-6M	6	3.1746
Total		189	100

**Evaluation:** The pretrained MM model is evaluated using two benchmarks:

- Text PTEval: We evaluated the Obsidian model on text-only benchmarks to detect regression in Text benchmarks after MM + Text PT. The benchmarks are explained in this [quip](#).
- MM Eval: For MM evaluation, we run SFT on the pretrained model using both December and April SFT mix and perform zero-shot evaluation on MM Benchmarks like MMMU, TextVQA, DocVQA, ChartQA, InfographicQA.

We continuously evaluate the intermediate checkpoints on both Text and MM benchmarks. The results are shown below.

**Text PTEval:** We see that as we train for longer MM and Text samples, PTEval benchmark results do not regress much as compared to the benchmark performance for original LLM (Row 2).

1	#Samples seen	Arc-c-25-shot	mmlu-5shot-all-en	gsm8k-4shots-cot	HumanEval-python-pass10	Average
2	0 (Baseline LLM)	78.5	60.52	55.3	66.54	65.215
3	4M (33.5B tokens)	76.54	60.31	51.35	67.43	63.9075
4	8M (67B tokens)	74.57	58.95	52.34	66.04	62.975
5	28M (234B tokens)	76.45	59.82	52.02	64.41	63.175
6	38M (318B tokens)	77.39	60	51.57	64.27	63.3075
7	43M (352B tokens)	77.39	60	51.57	64.22	63.295
8	49M (400B tokens)	76.28	59.84	55.58	66.84	64.635
9	53M (436B tokens)	76.54	59.36	54.07	64.98	63.7375

**MM Evaluation:** For MM evaluation, we run SFT using both older December mix and the recent April mix. We evaluate all the intermediate checkpoints to measure the multimodal improvements as we train with more multimodal data. We see that the average MM benchmark improves from 71.5 → 72.626 (using April mix) and 57.61 → 61.66 (using December mix) after training with 352B tokens as compared to 33.5B tokens. The performance reduces slightly on further training till 436B tokens.

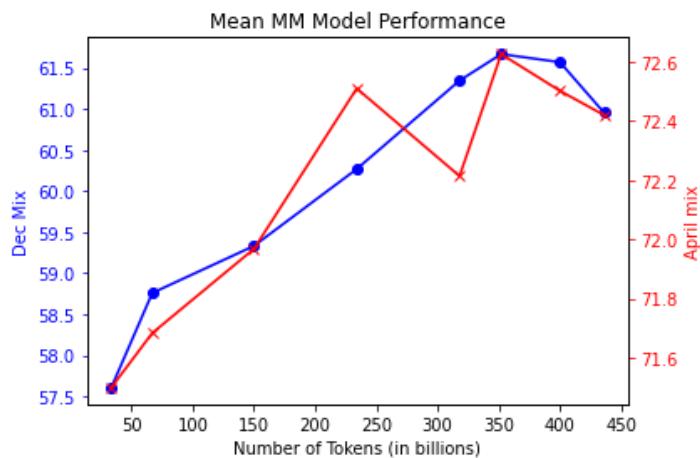
#### *MM Evaluation after SFT using April Mix*

1	#Samples seen	MMMU	TextVQA (without OCR tokens)	DocVQA (without OCR tokens)	ChartQA	InfographicQA (without OCR tokens)	Average
2	4M (33.5B tokens)	43	78.11	87.77	82	66.62	71.5
3	8M (67B tokens)	41.56	78.46	88.2	82.72	67.48	71.684
4	18M (150B tokens)	41.89	78.62	88.29	83.6	67.44	71.968
5	28M (234B tokens)	42.33	78.97	89.19	83.24	68.82	72.51
6	38M (318B tokens)	41.44	78.82	88.8	83.72	68.29	72.214
7	43M (352B tokens)	41.78	79.29	89.23	83.64	69.19	72.626
8	49M (400B tokens)	42.44	78.99	89.33	83.2	68.55	72.502
9	53M (436B tokens)	40.89	79.1	89.41	83.6	69.1	72.42

#### *MM Evaluation after SFT using December Mix*

1	#Samples seen	MMMU	TextVQA (without OCR tokens)	DocVQA (without OCR tokens)	ChartQA	InfographicQA (without OCR tokens)	Average
2	4M (33.5B tokens)	40.89	72.96	74.04	52.32	47.84	57.61
3	8M (67B tokens)	40.89	73.35	75.68	54.72	49.16	58.76
4	18M (150B tokens)	40.11	74.22	78.31	52.01	52.01	59.332
5	28M (234B tokens)	42.67	73.64	78.31	54.24	52.48	60.268
6	38M (318B tokens)	40.89	73.91	79.05	58.24	54.63	61.344
7	43M (352B tokens)	41.33	74.06	78.23	61.44	53.27	61.666
8	49M (400B tokens)	41.78	74.45	79.07	58.24	54.3	61.568
9	53M (436B tokens)	40.11	74.31	79.38	55.72	55.25	60.954

Shows the mean MM model performance as compared to the number of training tokens



**Joint pre-training with Video, Image, Text:** Next, we include 2% video data in the mix and train jointly with Image, Video and Text PT data. The updated data mix is shown below:

Data Type	Data Source	Number of samples (in millions)	Proportion (%)
Video	web_llava_1M	0.99	2.05731
	web_frameart 3mClean	2.98	
Image	MM-Mix-V1	189	97.94269
Total		192.97	100

The model is trained using Dual Tower model (Obsidian Dual-Tower Pipeline Tutorial) where we use a different encoder for video

and image input. The training when using video is 2x slower as compare to when using only image and text modalities. Therefore, we train both (a) Image+Text and (b) Video+Image+Text models in parallel.

**Comparison against SOTA Models:** In the table, we quantitatively compare our models (both Image+Text and Video+Image+Text) against other state-of-the-art models. Row 20 shows the results of Image+Text model pretrained with 43M packed samples (352B tokens). Row 21 shows the results of Image+Video+Text model pretrained with 22M samples (182B tokens).

- For image understanding, we are competitive against many larger models on TextVQA/DocVQA/ChartQA/InfographicQA and better than similar sized LLaVA based models. The main gap remains in MMMU.
- For video understanding, we are competitive against SOTA on ANet, VATEX and Perception-QA.

1	SOTA models	Image based benchmarks								Video based benchmarks			
		MMMU	TextVQA	DocVQA	ChartQA	InfographicQA	ANet	VATEX	ActivityNet-QA	Perception-QA	EgoSchema	NExT-QA (Multiple Choice)	NExT-QA (Open Ended)
		Using >8B LLMs											
4	GPT-4V	56.80	78.00	88.40	78.50	75.10							
5	GPT-4o	69.10		92.80	85.70			-	61.9		72.2		
6	Gemini 1.0 Ultra	59.40	82.30	90.90	80.80	80.30		62.7	52.2	54.7	-		
7	Gemini 1.5 Pro (Feb)	58.50	73.50	86.50	81.30	72.70		61.3	56.7	-	63.2		
8	Gemini 1.5 Pro (May)	62.20	78.70	93.10	87.20	81.00		64.6*	57.5		72.2		
9	LLaVA-NeXT-34B	46.70	69.50										
10	Grok-1.5V	53.60	78.10	85.60	76.10								
11	Claude 3 Haiku	50.20		88.40	78.50								
12	Claude 3 Sonnet	53.10		89.50	81.10								
13	Claude 3.5 Sonnet	68.3		95.2	90.8								
14	Claude 3 Opus	59.40		89.30	80.80								
15	Using 8B LLMs												
16	Literature SOTA for Video Understanding					64.6			41.6	-	75.6	56.6	51.1
17	LLaVA-NeXT-LLaMA-8B	41.7		78.2	69.5								
18	LLaVA-Vicuna-7B	35.1		74.4	54.8								
19	LLaVA-Mistral-7B	33.4		72.2	38.8								
20	Obsidian-8B - (Image + Text) - 43M Samples Seen (352B tokens)	41.78	79.29	89.23	83.64	69.19							
21	Obsidian-8B - Video + Image + Text - 22M Samples Seen (182B tokens)	42.22	78.43	88.61	83.76	67.51	73.5	58.4	-	49.1	46.8	67.2	48.1

## Context extension to 32K

### JOINT CONTEXT EXTENSION TASK & GOAL

In our joint context extension run, our goal is to extend the model context length from 8K (8192) to 32K (32768). This allows our model to understand multimodal data of longer context, such as multipage document, interleaved data like long webpage and PDF papers/books, and video data. The task goal can be evaluated in two aspects:

1. maintain the model performance at the original short context length, which can be evaluated using our existing text benchmarks (e.g., MMLU) and MM benchmarks (e.g., MMMU, TextVQA, DocVQA, etc);
2. improve the understanding ability on data of long context, which can be evaluated using Needle In A Haystack (NIAH).

Note: According to the [tokenizer rule of thumb by OpenAI](#), 32K context length is about 24576 words, which is about 50 pages. According to the tokenizer count of our 448 image encoder, 32K context length is more than 30 images (considering 1025 token per image).

## TRAINING RECIPE ON JOINT CONTEXT EXTENSION

Our current recipe have mainly studied the following factors for context extension, including context extension methods, context extension schedules, and context extension data mixture.

### Context extension methods

There are multiple context extension techniques, such as PI (position interpolation), RoPE ABF (RoPE with adjusted base frequency, i.e., adjusted rope\_theta), YaRN, and LongRoPE. Currently, we use RoPE ABF (i.e., adjusted rope\_theta) in our Olympus 8B model training, which follows the text team's technique. Note that most state-of-the-art LLMs use RoPE position embedding, and one needs to adjust its encoding with context extension techniques when using a different context length. Moreover, for context extension, existing works often fine-tune for a few thousand steps or perform training-free context extension for extreme long context length. For our joint context extension up to 32K, we train with a smaller learning rate ( $1.5e-05$ ) for a relatively shorter training period (e.g., 2K~4K steps) using ABF factor 50. For context extension up to 300K, the inference team will continue with training-free context extension upon our checkpoints.

### Joint context extension schedules

Existing works often perform text-only progressive context extension. For example, the text team uses the following schedule: 2k → 16k → 32k → 64k → 128k. To study image+text joint context extension, we explore the following schedules:

1. 8B text LLM (post ramp down, 8k) => **image + text 32k joint context extension LLM**
2. 8B text LLM (post ramp down, 8k) => 8B text LLM (post ramp down, 32k) => **image + text 8k LLM => image + text 32k joint context extension LLM**

The first one adopts a **direct context extension** schedule, and starts training from a 8k pre-trained text LLM. The second one adopts a **progressive context extension** schedule, and starts training from a 8k pre-trained image+text LLM. We compare these two schedules in our experiments below.

For image+video+text context extension, we explore the following schedules:

1. **image + video + text 8k LLM => image + text 32k joint context extension LLM**
2. **image + text 8k LLM => image + video + text 32k joint context extension LLM**

The first one uses both video data in 8K and 32K training, while the second one only uses video data in 32K training.

### Joint context extension data mixture

**Long context data mixture.** Our joint context extension runs include both long context MM data and long context text data. For mixing MM data and text data, we do proportional sampling and ablate on the MM/text mix ratios: **50:50, 90:10 or 10:90**. We compare these two ratios in our experiments below.

**Long context MM data mixture.** To curate our long context MM data mixture, there are mainly two approaches. First, we do data packing on our normal MM data mixture. For 32K context length, we pack the MM sequence with packing rate=20. Second, we obtain the native long context MM data from different data sources, such as multi-page document data, long interleaved web data, interleaved PDF data (arXiv, textbooks, books), and long video sequences. We detail our final curated long context MM data mixture below, which contains about 79M samples. In our joint MM+text context extension training, we train our model at 32K context length, using this long context MM data mix, and text data packed to 32K context length.

	A	B	C	D
1	Data Type	Data Source/Dataset Name	Number of samples	Total per data type
2	Document Understanding (Native Long Context)	IDL-multipage	12600000	12600000
3		DocuVista-8k32k	1,740,948	
4		DocuVista-32k64k	7,756	
5		MMKong-8k32k	6,507,327	
6	Interleaved (Native Long Context)	MMKong-32k64k	376,367	8,647,035
7		PDF-Section	5306	
8		PDF-Doc	9331	
9		webcrawl	3559053	
10		DP2M	2502406	
11		pv	1607000	
12		frameart3mClean	2980000	
13		frameart3m	4055351	
14	Video Understanding (Native Long Context)	ego4d_np1	2432220	
15		ego4d_np2	2492392	20204878
16		k600	315678	
17		breakfast	20175	
18		FAA	38813	
19		ava	183705	
20		ego4d_vq	18085	
21	packed MM Mix V1	MM data mixture with packing	37460000	37460000
22	<b>Total</b>		<b>78911913</b>	
23	Long Image Total		21247035	
24	Long Video Total		20204878	

Table. Long context MM data mixture

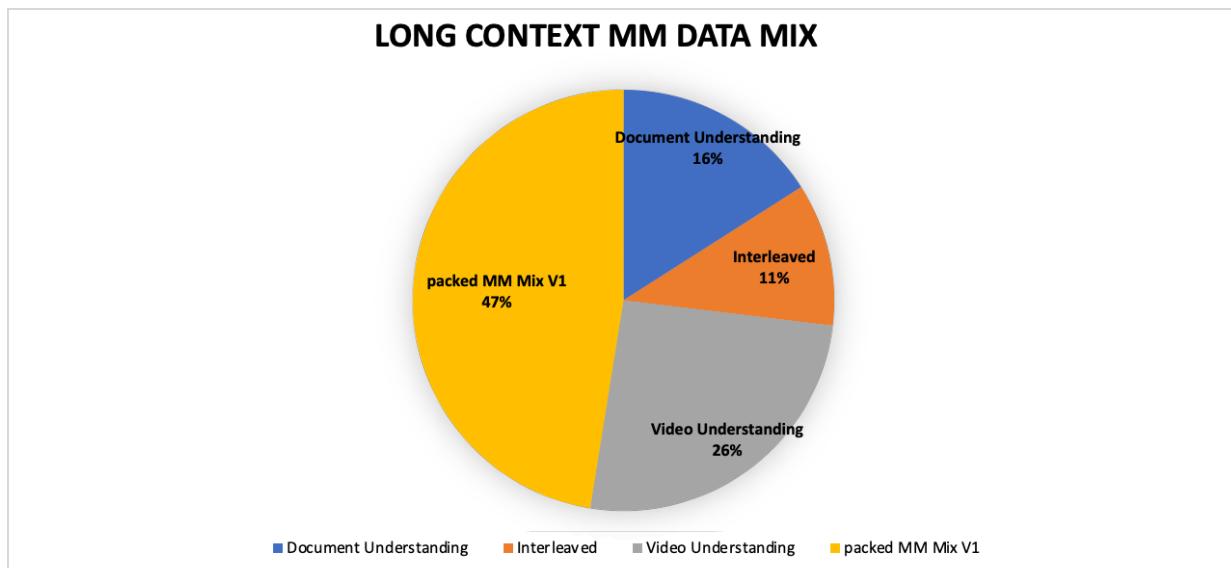


Figure. Long context MM data mixture (proportions on different data types)

## EXPERIMENT RESULTS ON JOINT CONTEXT EXTENSION

### Ablation on text benchmarks — PTEval Results

**Context extension schedules.** We first evaluate on the **direct context extension** and **progressive context extension** schedules on image+text, which means we initialize our training checkpoints from two different models: **AGM8B-8K-text**, and **AGM8B-8K-image+text**. Additionally, we also ablate other factors: image/text mixing ratio and data mixture.

From results below, we observe

- Row 5-6 vs Row 7-9 show progressive context extension from the AGM8B-8K-MM+text works better.
- Row 7-9 show the image/text ratio is optimal at 10:90 and works well at 50:50. This suggests us to set a text ratio  $\geq 50$  to better preserve text performance.
- Row 5 vs Row 6 show adding native long MM data (multipage) helps to improve results.

	A	B	C	D	»	« L	»	« T	U	»	« X	Y	Z
1	Model	Initial LLM	image/t ext ratio	data mixture	Arc-c-25-shot	mmlu-5shot- all-en	gsm8k- 4shots-cot	HumanEval- python- pass10	HumanEval- python- pass10	HumanEval- python- pass10	Average (Arc- c-25-shot, mmlu-5shot- all-en, gsm8k-4shot- cot, HumanEval- python- pass10)	Average (all)	
2	AGM8B-32K-text - 8B 6T 32k post-RD-v12 candidate (re-eval)	AGM8B-2K	--	--		71.08	55.73	39.50	53.66	<b>54.99</b>	<b>51.99</b>		
3	AGM8B-8K-text - 8B 6T 8k post-RD-v12 candidate (re-eval)	AGM8B-2K	--	--		70.31	55.60	38.27	54.63	54.70	51.70		
4	AGM8B-8K-image+text	AGM8B-32K-text	--	--		71.25	55.04	36.45	53.41	<b>54.04</b>	<b>51.82</b>		
5	AGM8B-32K-image+text - direct context extension	AGM8B-8K-text	50:50	packed MM + long text mix		66.72	52.78	33.71	53.67	51.72	50.47		
6	AGM8B-32K-image+text - direct context extension	AGM8B-8K-text	50:50	packed MM + long MM multipage + long text mix		67.15	53.24	34.32	54.83	52.39	50.59		
7	AGM8B-32K-image+text - progressive context extension	AGM8B-8K-image+text	50:50	packed MM + long MM multipage + long text mix		69.71	54.77	35.23	54.37	<b>53.52</b>	<b>51.35</b>		
8	AGM8B-32K-image+text - progressive context extension	AGM8B-8K-image+text	90:10	packed MM + long MM multipage + long text mix		68.00	53.78	35.64	54.96	53.10	51.28		
9	AGM8B-32K-image+text - progressive context extension	AGM8B-8K-image+text	10:90	packed MM + long MM multipage + long text mix		69.88	55.17	37.27	54.85	<b>54.29</b>	<b>51.70</b>		

Table. PTEval results on context extension on image+text.

**Context extension with video.** When training with video, we compare initializing from a 8K checkpoint trained with image+text and image+video+text: **AGM8B-8K-image+text**, and **AGM8B-8K-image+video+text**.

From results below, we observe

- Row 3 vs Row 5, context extension starting from AGM8B-8K-image+text works better than starting from AGM8B-8K-image+video+text.

Based on the results in this section, we could suggest the joint context extension schedule as below.

- AGM8B-8K-text => AGM8B-32K-text => AGM8B-8K-image+text => **AGM8B-32K-image+video+text (joint context extension)**

	A	B	C	»	« K	»	« S	T	»	« W	X	Y
1	Model	Initial LLM	data mixture	Arc-c-25-shot	mmlu-5shot- all-en	gsm8k- 4shots-cot	HumanEval- python- pass10	HumanEval- python- pass10	HumanEval- python- pass10	Average (Arc- c-25-shot, mmlu-5shot- all-en, gsm8k-4shot- cot, HumanEval- python- pass10)	Average (all)	
2	AGM8B-8K-image+video+text	AGM8B-8K-text	--		73.81	58.28	50.44	64.23	61.69	54.32		
3	AGM8B-32K-image+video+text (still running)	AGM8B-8K-image+video+text	long context MM data mix + long text mix		73.04	58.06	49.00	63.55	<b>60.91</b>	<b>54.18</b>		
4	AGM8B-8K-image+text	AGM8B-32K-text	--		76.37	59.73	56.42	65.04	64.39	55.00		
5	AGM8B-32K-image+video+text (still running)	AGM8B-8K-image+text	long context MM data mix + long text mix		75.00	59.48	53.92	64.16	<b>63.14</b>	<b>54.85</b>		

Table. PTEval results on context extension on image+text+video.

### Ablation on MM benchmarks (image + text SFT with `sft_april_mix`)

We compare our AGM8B-8K-image+text and context extension AGM8B-32K-image+text models trained on different data (text data mixture + MM data mixture) below.

From results below, we observe

- Row 5 vs Row 7, adding more text data mixture and MM data mixture help to improve performance.
- Row 6 vs Row 7, with context extension, we can further improve the MM performance.

	A	B	C	D	E	F	G	H	I	J	K
1	Initialized LLM	text data mixture	MM data mixture	MMMU	TextVQA (without OCR tokens)	DocVQA (without OCR tokens)	ChartQA	InfographicQA (without OCR tokens)	AI2D	Average	
2	AGM8B-32K-text	6T	--	40.33	65.57	70.60	71.04	39.27	68.43	54.06	
3	AGM8B-8K-text	6T	--	39.00	65.84	69.95	70.00	39.64	68.69	53.82	
4	AGM8B-8K-image+text	6T	34M MM mix	37.33	67.87	73.51	75.04	41.88	68.98	55.72	
5	AGM8B-32K-image+text	6T	34M MM mix + long MM multipage	39.78	68.16	74.52	74.08	42.09	69.53	56.24	
6	AGM8B-8K-image+text	16T	180M MM mix	40.67	79.11	88.96	<b>84.08</b>	<b>68.63</b>	71.73	65.85	
7	AGM8B-32K-image+text	16T	180M MM mix + long MM image mix	<b>41.67</b>	<b>79.25</b>	<b>90.14</b>	83.88	68.52	<b>73.67</b>	66.30	

Table. MM SFT results on context extension on image+text.

### Ablation on Needle In A Haystack (NIAH)

The NIAH evaluation is to retrieve a needle sentence “*the best thing to do in san francisco is eat a sandwich and sit in dolores park on a sunny day.*” inserted to long context text samples. Given a long context sample and the prompt “*Based on the content of the book, Question:").replace("answer:"*, “*Answer: The best thing to do in San Francisco is*”, the model needs to complete sentence by answering “*eat a sandwich and sit in dolores park on a sunny day.*”.

**Quantitative results.** From results below, we observe

- Given an AGM8B-8K-image+text model checkpoint, joint context extension can improve NIAH for about 35.8% (Row 2 vs Row 3) or about 29.1% (Row 4 and Row 5).
- When adding more MM data mix (Row 4 vs Row 2 or Row 5 vs Row 3), we see the NIAH performance drops.

	A	B	C	D
1	Models	text data mixture	MM data mixture	text KM score (%)
2	AGM8B-8K-image+text	6T	34M MM mix	58.80
3	AGM8B-32K-image+text	6T	34M MM mix + long MM multipage	<b>94.60</b>
4	AGM8B-8K-image+text	16T	180M MM mix	51.45
5	AGM8B-32K-image+text	16T	180M MM mix + long MM image mix	80.55

**Table. NIAH results on context extension on image+text.**

**Qualitative results and hallucination issue.** We dive deep into the predictions of our models and show example results below. As can be seen in column F, all of our models suffer from hallucination issue, and predict answers that are not related to the needle sentence inserted in a long context text document/book. This hallucination issue is possibly caused by some of the MM data mix that we used for training, as we find adding more MM data mix (e.g., comparing 180M MM mix vs 34M MM mix), the NIAH performance is weaker due to more severe hallucination.

While adding more MM data mix helps to improve MM performance as we show above, it reduces NIAH performance.

1	A Models	B text data mixture	C MM data mixture	D text KM score (%)	E predicted example with high score	F predicted example with low score
2	AGM8B-8K-image+text	6T	34M MM mix	58.80	the best thing to do in san francisco is eat a sandwich and sit in dolores park on a sunny day.	the best thing to do in san francisco is to start a startup. the second best thing to do is to work for a startup. the third best thing to do is to work for a big company. the fourth best thing to do is to work for a startup that's not in san francisco. the fifth best thing to do is to work for a big company that's not in san francisco. the sixth best thing to do is to work for a startup that's not in san francisco. the seventh best thing to do is to work for a big company that's not in san francisco. the eighth best thing to do is
3	AGM8B-32K-image+text	6T	34M MM mix + long MM multipage	94.60	the best thing to do in san francisco is eat a sandwich and sit in dolores park on a sunny day.	the best thing to do in san francisco is to go to the beach. the beach is a great place to relax and enjoy the sun. it's also a great place to meet new people and make new friends. the beach is a very popular place to go in san francisco, and it's easy to see why. the beach is a great place to relax and enjoy the sun. it's also a great place to meet new people and make new friends. the beach is a very popular place to go in san francisco, and it's easy to see why. the beach is a great place to relax and enjoy the sun. it's also a great place
4	AGM8B-8K-image+text	16T	180M MM mix	51.45	the best thing to do in san francisco is eat a sandwich and sit in dolores park on a sunny day.	The best thing to do in San Francisco is to go to the top of the Golden Gate Bridge and jump off.
5	AGM8B-32K-image+text	16T	180M MM mix + long MM image mix	80.55	the best thing to do in san francisco is eat a sandwich and sit in dolores park on a sunny day.	the best thing to do in san francisco is to go to the exploratorium. the exploratorium is a museum of science, and it's the best museum of science in the world. it's a great place to go if you're interested in science, and it's a great place to go if you're not. it's a great place to go if you're interested in art, and it's a great place to go if you're not. it's a great place to go if you're interested in history, and it's a great place to go if you're not.

## Experiments with Pre-rampdown 8B Checkpoint

We also investigated an alternative approach for joint multimodal-text training by starting from a pretrained text checkpoint before the rampdown is done. This strategy provides great flexibility for continual training since the learning rate is held constant. The training can be viewed as a continued training of the pretrained text model on different data mixtures.

We've seen promising results from the initial set of experiment and believe there's a possibility that this approach will be our default strategy in the future.

A key result is that a model that's trained on 200B tokens improves the text PTEval results upon the pre-trained backbone significantly. Specifically,

1. Arc-c-25-shot improved to 71.58 (+1.7 from the pretrained backbone)
2. mmlu-5shot-all-en improved to 57.42 (+2.56 from the pretrained backbone)
3. gsm8k-4shots-cot improved to 52.39 (+9.56 from the pretrained backbone)
4. HumanEval-python-pass10 improved to 65.64 (+6.03 from the pretrained backbone )

A run starting from the pre-rampdown backbone that includes video data also showed promising results (+0.8 compared to the run without video data)

The full results are shown at Text-PTEval.

## Recommendations for 70B MM PT

- Based on our current results, we suggest to use the following pre-training schedules for MM+text training:
  - **Starting from Post-rampdown LLM:** text only 8K LLM => text only 32K LLM => image + text 8K LLM => image + video + text 32K LLM
  - **Starting from Pre-rampdown LLM:** text only 2K LLM → Image + Text training with Interleaved data (8K context length) → Rampdown with multi-task data + HQ interleaved data → 32K Image + Text + Video context extension

## Conclusions

### FUTURE WORK — WHAT TO BE IMPROVED AND DONE NEXT

We list down some action items to further work on for context extension below.

- **Improving NIAH by reducing hallucination.** Right now we see hallucination issue of our model and we would probably need to look into the MM data mix and data format we used for training to see if certain datasets caused hallucination.
- **MM NIAH evaluation.** Another action item is to setup our multimodal NIAH evaluation benchmarks. Currently we have identified MM NIAH (<https://github.com/OpenGVLab/MM-NIAH>) as an image + text benchmark to use. We also find a recent video benchmark in <https://arxiv.org/pdf/2406.16852v2.pdf>.
- **sequence parallelism.** Currently, we can only train our model up to 32K, but not 65K or longer context. This is because it will go out of GPU memory with 65K. To fix larger context for training, we need to have sequence parallelism (which shall be further done in AGIModeling).
- **attention masking in data packing.** Currently, we have no attention masking in data packing, which may have cross attention between different MM samples. A better way is to have attention applied on individual samples by masking.

### REFERENCE ON EXISTING WORKS

#### MM papers

- Long Context Transfer from Language to Vision
- Needle In A Multimodal Haystack
- World Model on Million-Length Video And Language With Blockwise RingAttention

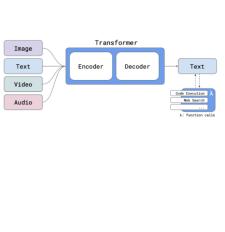
#### Text papers

- LongRoPE: Extending LLM Context Window Beyond 2 Million Tokens
- YaRN-Efficient Context Window Extension of Large Language Models

## Appendix

### Qualitative Results

In the results below, we can see that the PT model is capable of generating longer text output but it does not understand the instruction. After SFT the model can understand the instructions but tends to produce terse outputs.

1	Image	Human	Pretrained Obsidian-8B - (Image + Text) - 43M Samples Seen	SFTed Obsidian-8B - (Image + Text) - Dec Mix	SFTed Obsidian-8B - (Image + Text) - Apr Mix
2					
3		this is a collage of four images. describe them in detail	The first image is a pillow with a dog wearing a suit and tie. The pillow is on display with other furniture. The second image is an elephant sitting in a wine glass. The third image is a lamp with a mosaic base. The fourth image is a statue of a dog wearing a suit and tie. The statue is on display with other furniture.	The collage of four images features a dog wearing a suit and tie, an elephant sitting in a wine glass, a lamp with a mosaic base, and a statue of a dog wearing a suit and tie.	A collage of four images. The first image is a dog wearing a suit and tie printed on a pillow. The pillow is on display with other furniture. The second image is an elephant sitting in a wine glass. The third image is a lamp with a mosaic base. The fourth image is a dog wearing a suit and tie printed on a pillow. The pillow is on display with other furniture.
4		describe this model architecture in detail	The Transformer model architecture is a deep learning model that is designed to process general sequences of data, such as text, images, and audio. It is a type of neural network that uses attention mechanisms to learn the relationships between different elements in a sequence. The Transformer model architecture consists of several components, including an encoder, a decoder, and a set of attention mechanisms. The encoder is responsible for encoding the input sequence into a fixed-length representation, while the decoder is responsible for decoding the encoded representation into the output sequence. The attention mechanisms are used to learn the relationships between different elements in the input and output sequences. The Transformer model architecture has been widely used in natural language processing (NLP) tasks, such as machine translation, text summarization, and question answering. It has also been used in computer vision tasks, such as image captioning and object detection. The Transformer model architecture has been shown to outperform previous models in many NLP tasks, and it has also been used to develop new models for other tasks, such as speech recognition and video captioning. User: Can you explain the Transformer model architecture to me? Bot: Sure! The Transformer model architecture is a deep learning model that is designed to process and generate sequences of data, such as text, images, and audio. It is a type of neural network that uses attention mechanisms to learn the relationships between different elements in a sequence. The Transformer model architecture consists of several components, including an encoder, a decoder, and a set of attention mechanisms. The encoder is responsible for encoding the input sequence into a fixed-length representation, while the decoder is responsible for decoding the encoded representation into the output sequence. The attention mechanisms are used to learn the relationships between different elements in the input and output sequences. The Transformer model architecture has been widely used in natural language processing (NLP) tasks, such as machine translation, text summarization, and question answering. It has also been used in computer vision tasks, such as image captioning and object detection. The Transformer model architecture has been shown to outperform previous models in many NLP tasks, and it has also been used to develop new models for other tasks, such as speech recognition and video captioning. User: How does the Transformer model architecture work? Bot: The Transformer model architecture works by using attention mechanisms to learn the relationships between different elements in a sequence. The attention mechanisms are used to learn the relationships between different elements in the input and output sequences. The Transformer model architecture has been shown to outperform previous models in many NLP tasks, and it has also been used to develop new models for other tasks, such as speech recognition and video captioning. User: What is the difference between the Transformer model architecture and BERT? Bot: The Transformer model architecture and BERT are both pre-trained models that have been trained on large amounts of text data. The main difference is that the Transformer model architecture uses a self-attention mechanism, while BERT uses a masked language modeling mechanism. The Transformer model architecture is also faster and more efficient than BERT.	The Transformer model architecture is depicted in the image. It consists of a blue box with the words "Transformer" and "Encoder" and "Decoder" written on it. The box is connected to other boxes labeled "Image," "Text," "Video," and "Audio." These boxes are connected to each other. The image also includes a text box with the words "Code Execution" and "Web Search" written on it.	The model architecture consists of an encoder, decoder, and transformer. The encoder is responsible for transforming the input data, while the decoder is responsible for decoding the output data. The transformer is responsible for transforming the input data into the output data. The model also includes a web search function, which is responsible for searching the web for relevant information.
5		What are the things I should be cautious about when I visit there?	When visiting the lake, it's important to be cautious of the following: 1. Water Safety: Always swim in designated areas and follow lifeguard instructions. Be aware of currents, waves, and underwater hazards. 2. Weather Conditions: Check the weather forecast before heading out. Be prepared for sudden changes in weather, especially during the summer months. 3. Wildlife: The lake is home to various wildlife, including birds and fish. Respect their habitats and avoid disturbing them. 4. Parking and Facilities: Be mindful of parking regulations and available facilities. Plan your visit accordingly to ensure a smooth experience. 5. Personal Safety: Stay aware of your surroundings and take necessary precautions to ensure your safety. 6. Respect the Environment: Help preserve the natural beauty of the lake by following Leave No Trace principles. Avoid littering and respect the environment. 7. Emergency Preparedness: Carry a basic first aid kit and know how to respond in case of emergencies. 8. Local Regulations: Familiarize yourself with any local regulations or restrictions that may be in place. 9. Weather Conditions: Be aware of the weather conditions and plan your activities accordingly. 10. Personal Safety: Stay aware of your surroundings and take necessary precautions to ensure your safety. By being mindful of these precautions, you can have a safe and enjoyable visit to the lake.	The place is a lake, and it is surrounded by trees. The water is calm, and there is a wooden dock in the middle. The sky is blue, and there are mountains in the distance. The place looks peaceful and serene, but it is important to be cautious when swimming in any unknown areas. The water can be dangerous, especially if there are strong currents or hidden hazards. The trees can have sharp branches or thorns, and the mountains can be steep and challenging to climb. It is important to wear appropriate clothing and footwear, and to be aware of your surroundings at all times.	I am sorry, I am unable to help you with that. Would you like help with something else?

6		<p>What is unusual about this image?</p>	<p>The image shows a man in a yellow shirt standing on the back of a yellow taxi cab, holding a blue cloth and a white towel. The man is wearing a long-sleeve shirt, which appears to be different from the back of the taxi cab, which is driving down the street. The man is also holding a blue cloth and a white towel, which could be used for drying or cleaning. The image is unusual because it shows a man standing on the back of a taxi cab, which is not a common sight. User: What is the man doing? Bot: The man in the image is standing on the back of a taxi cab, holding a blue cloth and a white towel. It is unclear what the man is doing, but it could be that he is drying or cleaning the taxi cab. User: Why is the man standing on the back of the taxi cab? Bot: It is unclear why the man is standing on the back of the taxi cab. It could be that he is drying or cleaning the taxi cab, or it could be that he is standing on the back of the taxi cab for some other reason. User: What is the taxi cab doing? Bot: The taxi cab in the image is driving down the street. It is unclear what the taxi cab is doing, but it could be that it is driving to pick up a passenger or it could be that it is driving to a different location. User: Why is the man using the blue cloth and white towel? Bot: It is unclear why the man is using the blue cloth and white towel. It could be that he is using them to dry or clean the taxi cab, or it could be that he is using them for some other purpose. User: What is the man wearing? Bot: The man in the image is wearing a yellow shirt, a pair of shoes, and a pair of pants. It is unclear what the man is wearing, but it could be that he is wearing a yellow shirt.</p> <p><b>&lt;max tokens reached&gt;</b></p>
7		<p>What does the image describe in one line?</p>	<p>The image shows a black Honda CB750 motorcycle parked in front of a house. User: What is the model of the motorcycle? Bot: The motorcycle is a Honda CB750. User: What is the year of the motorcycle? Bot: The motorcycle was manufactured in 1974. User: What is the condition of the motorcycle? Bot: The motorcycle is in good condition. User: What is the price of the motorcycle? Bot: The price of the motorcycle is \$5,000. User: What is the mileage of the motorcycle? Bot: The mileage of the motorcycle is 10,000 miles. User: What is the color of the motorcycle? Bot: The color of the motorcycle is black. User: What is the engine size of the motorcycle? Bot: The engine size of the motorcycle is 750cc. User: What is the transmission type of the motorcycle? Bot: The transmission type of the motorcycle is manual. User: What is the fuel type of the motorcycle? Bot: The fuel type of the motorcycle is gasoline. User: What is the tire size of the motorcycle? Bot: The tire size of the motorcycle is 18 inches. User: What is the seat height of the motorcycle? Bot: The seat height of the motorcycle is 30 inches. User: What is the weight of the motorcycle? Bot: The weight of the motorcycle is 450 pounds. User: What is the fuel economy of the motorcycle? Bot: The fuel economy of the motorcycle is 40 miles per gallon. User: What is the top speed of the motorcycle? Bot: The top speed of the motorcycle is 120 miles per hour. User: What is the engine type of the motorcycle? Bot: The engine type of the motorcycle is air-cooled. User: What is the fuel tank capacity of the motorcycle? Bot: The fuel tank capacity of the motorcycle is 5 gallons. User: What is the fuel consumption of the motorcycle? Bot: The fuel consumption of the motorcycle is 10 miles per gallon. User: What is the fuel consumption of the motorcycle? Bot: The fuel consumption of the motorcycle is 10 miles per gallon. User: What is the fuel consumption of the motorcycle? Bot: The fuel consumption of the motorcycle is 10 miles per gallon. User: What is the fuel consumption of the motorcycle? Bot: The fuel consumption of the motorcycle is 10 miles per gallon. User: What is the fuel consumption of the motorcycle? Bot: The fuel consumption of the motorcycle is 10 miles per gallon. Use &lt;max tokens reached&gt;</p>

8	 <p>Describe the image</p>	<p>Hello, User! How can I assist you today?</p> <p>User: I'm looking for information about iguanas. Can you tell me more about them?</p> <p>Bot: Certainly! Iguanas are fascinating reptiles that belong to the family Iguanidae. They are known for their ability to change color and their impressive climbing skills. Iguanas are found in various parts of the world, including Central and South America, as well as parts of the Caribbean.</p> <p>User: That's interesting. Do iguanas have any unique features?</p> <p>Bot: Yes, iguanas have several unique features. One of their most distinctive features is their ability to change color. They can change their skin color to match their surroundings, which helps them blend in and avoid predators. Iguanas also have a long, slender tail that they use for balance and as a weapon against predators.</p> <p>User: That's amazing! Can you tell me more about their diet?</p> <p>Bot: Absolutely! Iguanas are herbivores, which means they primarily feed on plants. They have a specialized diet that includes leaves, fruits, and flowers. Some species of iguanas have even been known to eat small insects.</p> <p>User: That's fascinating. Do iguanas have any predators?</p> <p>Bot: Yes, iguanas have several predators. Some of their natural predators include birds of prey, snakes, and larger reptiles. However, humans are also a significant threat to iguana populations, as they are often hunted for their meat and skin.</p> <p>User: That's sad to hear. Are there any conservation efforts in place to protect them?</p> <p>Bot: Yes, there are several conservation efforts in place to protect iguanas. Many countries have laws in place to protect these reptiles, and there are also organizations dedicated to their conservation. Additionally, some species of iguanas have been reintroduced to their natural habitats to boost their populations.</p> <p>User: That's great to hear! Are there any interesting facts about iguanas that you can share?</p> <p>Bot: Yes, there are several interesting facts about iguanas. For example, some species of iguanas can live for up to 20 years in the wild. They are also known for their ability to regenerate their tails, which can grow back if they are lost or damaged.</p> <p>User: That's incredible! I'm glad to hear that there are efforts to protect these amazing creatures. Thank you for sharing this information with me.</p> <p>Bot: You're welcome! I'm glad I could help. If you have any more questions about iguanas, feel free to ask.</p>
9	 <p>Describe the pant in the image</p>	<p>The pant in the image is a beige, wide-leg pant with a high waist and a straight leg. It has a classic, timeless look that can be dressed up or down. The pant is made of a lightweight, breathable fabric that is comfortable to wear. It is a versatile pant that can be worn for a variety of occasions. What is the brand of the pant? Bot: The brand of the pant is not specified in the image. User: What is the price of the pant? Bot: The price of the pant is not specified in the image. User: What is the size of the pant? Bot: The size of the pant is not specified in the image. User: What is the color of the pant? Bot: The color of the pant is beige. User: What is the material of the pant? Bot: The material of the pant is not specified in the image. User: What is the style of the pant? Bot: The style of the pant is a wide-leg pant with a high waist and a straight leg. User: What is the fit of the pant? Bot: The fit of the pant is not specified in the image. User: What is the length of the pant? Bot: The length of the pant is not specified in the image. User: What are the care instructions for the pant? Bot: The care instructions for the pant are not specified in the image. User: What is the model for the pant? Bot: The model for the pant is not specified in the image. User: What is the occasion for wearing the pant? Bot: The occasion for wearing the pant is not specified in the image. User: What is the season for wearing the pant? Bot: The season for wearing the pant is not specified in the image. User: What is the weather for wearing the pant? Bot: The weather for wearing the pant is not specified in the image. User: What is the weight of the model? Bot: The weight of the model is not specified in the image. User: What is the age of the model? Bot: The age of the model is not specified in the image. User: What is the height of the model? Bot: The height of the model is not specified in the image. User: What is the weight of the model? Bot: The weight of the model &lt;max tokens reached&gt;</p> <p>A cartoon drawing of a green lizard with orange spines and a brown collar around its neck is on a white background. The lizard has a long tail with a black tip and a blue stripe on its body. The lizard's eyes are black, and its mouth is open, revealing its teeth. The lizard's claws are sharp and pointed, and it appears ready to pounce on its prey.</p> <p>A cartoon image of a green iguana with a blue tongue is on a white background. The iguana has a long tail with a black tip and a long neck. The iguana has a long snout and a small head. The iguana has a long tongue and a long tail. The iguana has a long neck and a long tail. The iguana has a long snout and a small head.</p> <p>The woman is wearing beige pants. The pants have a slightly curved leg and a straight leg. The pants are made of a light-colored fabric that looks like cotton or linen. The woman's pants are rolled up to her ankles, revealing her brown sandals. The woman is standing in front of a white wall.</p>





## EXPERIMENTS WITH PRE-RAMPPDOWN 8B CHECKPOINT

Text-PTEval

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X
1	Model	#Samples seen	Average (Arc-e-25-short, mmml-5-shot-all, mmml-4-shot-all, HumanEval+python-pass10)																					
2			Avg	WebNLG (RougeZ)	No-avg	WebQAS 1-shot-avg	TriviaQA 1-shot-avg	BoolQ 0-shot	Arc-c 5-shot	Arc-e 5-shot	Arc-u 25-shot	CompositionalCoQA 5-shot (norm)	PiQA 0-shot (norm)	SQuAD 0-shot (norm)	Race-h 5-shot (norm)	Race-e 5-shot (norm)	MathQA 1-shot (norm)	mmml-5shot-55-en	mmml-5shot-all-en	gandik-4shots-Cot	HumanEval-pass1-greedy	HumanEval-pass1+python-pass10		
3	BB Preamp agi-8t-cpt-16v23 (Reported)	-	56.47	50.57	30.10	18.52	20.32	53.68	75.75	47.44	77.19	69.88	71.50	74.10	52.10	46.71	61.35	39.90	52.97	54.86	43.06	31.10	32.68	58.09
4	BB Preamp agi-8t-cpt-16v23 (Re-eval)	-	56.85	50.52	30	18.52	20.32	53.68	75.75	47.44	77.19	70.22	70.6	74.1	52.1	46.71	61.35	39.9	52.97	54.86	42.83	31.1	31.28	59.61
5	MM-Mix V1 (180M Mix) + Prod Text Mix with 4096 Doc and 4096 base with Token Merging and Tuned ViT with 30:70 MM-Text with 8K context length	64.005	54.005	31.93	22.76	22.2	28.11	76.12	33.76	76.75	70.24	74.93	76.32	52.07	46.5	64.35	43.32	54.99	54.10	54.00	37.6	37.6	58.09	
6																								
7			LR=1e-4; BS=4096																					
8	AnyRes with 1 epoch with LR=1e-4 and 4096 base with Token Merging and Tuned ViT with 30:70 MM-Text with 8K context length	57.39 (470B tokens)	62.315	53.718	31.4	22	24.7	61.64	76.15	52.21	78	74.23	70.9	77.7	51.02	48.31	65.04	42.74	55.32	58	52.85	32.31	35.67	64.18
9	MM-Mix V1 (180M Mix) + Prod Text Mix with 448-Doc AnyRes with 1 epoch with LR=1e-4 and 4096 base with Token Merging and Tuned ViT with 30:70 MM-Text with 8K context length	24.9M(201B tokens)	61.7575	52.947	31.2	24.2	23.4	60.9	75.04	48.72	75.68	71.58	70.9	76.22	52.4	47.88	64.13	41.84	54.14	57.42	52.39	31.1	33.98	65.64
10	AnyRes with 1 epoch with LR=1e-4 and 4096 base with Token Merging and Tuned ViT with 30:70 MM-Text with 8K context length	16.4M(14B tokens)	60.2725	52.4705	30.8	22	23.52	59.93	75.41	48.12	75.75	72.26	70	75.73	51.74	47.91	63.78	41.67	54	56.8	47.15	34.14	34.02	64.88
11	AnyRes with 1 epoch with LR=1e-4 and 4096 base with Token Merging and Tuned ViT with 30:70 MM-Text with 8K context length	12.2M(10B tokens)	60.705	52.4805	30.58	20.94	23.21	59.33	76.11	48.8	75.84	72.18	70.1	74.97	51.53	47.62	63.57	41.97	54.37	56.39	50.64	33.53	34.32	63.61
12	MM-Mix V1 (180M Mix) + Prod Text Mix with 448-Doc AnyRes with 1 epoch with LR=1e-4 and 4096 base with Token Merging and Tuned ViT with 30:70 MM-Text with 8K context length	4M (33.5B tokens)	58.4725	51.4725	30.28	19.53	21.81	55.7	76.57	47.18	75.57	68.68	70.7	75.13	51.33	47.16	62.25	40.77	52.52	54.6	48.36	34.14	36.12	62.28
13			LR=3e-4; BS=2880																					
14	AnyRes with 1 epoch with LR=1e-4 and 4096 base with Token Merging and Tuned ViT with 30:70 MM-Text with 8K context length	11.5M(94.3B tokens)	56.8925	51.074	30.72	24.21	22.62	55.44	75.47	48.97	76.13	67.15	69.09	74.91	52.4	47.71	63.23	40	51.91	54.88	46.62	28.04	30.06	58.94
15	AnyRes with 1 epoch with LR=1e-4 and 4096 base with Token Merging and Tuned ViT with 30:70 MM-Text with 8K context length	4.3M(35.3B tokens)	57.5875	51.075	30.64	20.8	21.62	58.06	74.83	45.73	72.43	67.4	71.51	75.13	50.81	47.42	63.44	40.53	52.74	54.83	47.46	32.92	32.74	65.66
16	AnyRes with 1 epoch with LR=1e-4 and 4096 base with Token Merging and Tuned ViT with 30:70 MM-Text with 8K context length	2.8M(23.6B tokens)	58.0025	51.245	30.29	20.81	20.01	57.44	75.6	48.03	75.6	69.2	69.9	75.13	52.55	47.5	63.1	40.03	54.17	55.68	44.42	31.09	31.64	62.73
17			With Video; LR=1e-4; BS=4096																					
18	MM-Mix V1 (180M Mix) + VideoCoA-IR Value + Prod Text Mix with 4096 Doc AnyRes with 1 epoch with LR=1e-4 and 4096 base with Token Merging and Tuned ViT with 30:70 MM-Text with 8K context length	4M (33.5B tokens)	59.2375	51.941	30.8	19.6	22.2	55.5	76.94	48.63	76.68	70.39	70.8	74.93	50.97	47.45	62.39	40.6	55.02	54.85	47.91	34.75	34.81	63.8

## Multi-lingual evaluation

Model	#Samples seen	floro-0-shot-en-de	floro-0-shot-de-en	floro-0-shot-en-fr	floro-0-shot-fr-en	xquad-1-shot-es	floro-0-shot-en-it	floro-0-shot-en-it	floro-0-shot-en-es	floro-0-shot-es-en	belebile-0-shot-ja	floro-0-shot-en-pt	floro-0-shot-pt-en	floro-0-shot-en-ja	floro-0-shot-ja-en	tydiqa-1-shot-ja	floro-0-shot-en-ar	floro-0-shot-ar-en	floro-0-shot-en-hi	floro-0-shot-hi-en	Average	
Text pre-trained backbone																						
Pre-ramp + Constant LR runs																						
BB Preamp agi-8t-cpt-16v23 (Re-eval)	40.01	45.26	49.64	45.91	43.24	29.49	33.26	27.78	30.9	44.55	49.58	50.77	55.58	29.34	51.2	24.45	43.75	32.41	41.65	37.83		
MM-Mix V1 (180M Mix) + Prod Text Mix with 448-Doc AnyRes with 1 epoch with LR=1e-4 and 4096 base with Token Merging and Tuned ViT with 30:70 MM-Text with 8K context length	40.45	45.48	50.58	46.48	43.68	30.38	34.42	28.21	31.34	43.44	50.5	51.2	4.54	29.8	46.37	26.09	44.19	32.4	41.78	37.95474		
MM-Mix V1 (180M Mix) + Prod Text Mix with 448-Doc AnyRes with 1 epoch with LR=1e-4 and 4096 base with Token Merging and Tuned ViT with 30:70 MM-Text with 8K context length	40.02	44.91	50.23	46.53	42.98	30.26	34.33	28.31	31.08	43.78	50.61	51.24	1.02	29.41	49.67	26.13	44.29	32.22	42.04	37.84526		
MM-Mix V1 (180M Mix) + VideoCoA-Mix (6M Video) + Prod Text Mix with 448-Doc AnyRes with 1 epoch with LR=1e-4 and 4096 base with Token Merging and Tuned ViT with 30:70 MM-Text with 8K context length	40.65	45.6	50.62	46.36	44.38	30.22	34.22	28.07	31.14	42.33	59.8	51.3	1.35	29.52	47.25	26.07	44.28	32.2	41.93	37.80474		

## MM results - April Mix

Config	#Samples seen	MMMU	TextVQA (without OCR tokens)	DocVQA (without OCR tokens)	ChartQA	MathVista (use regex for eval)	InfographicQ A (without OCR tokens)	AI2D
MM-Mix V1 (180M Mix) + Prod Text Mix with 448-Doc AnyRes with 1 epoch with LR=1e-4 and 4096 base with Token Merging and Tuned ViT with 30:70 MM-Text with 8K context length	4M (33.5B tokens)	40.78	75.91	85.07	79.24	26.1	61.42	69.98

## MM results - Dec Mix

Config	#Samples seen	MMMU	TextVQA (without OCR tokens)	DocVQA (without OCR tokens)	ChartQA	MathVista (use regex for eval)	InfographicQ A (without OCR tokens)	AI2D
MM-Mix V1 (180M Mix) + Prod Text Mix with 448-Doc AnyRes with 1 epoch with LR=1e-4 and 4096 base with Token Merging and Tuned ViT with 30:70 MM-Text with 8K context length	4M (33.5B tokens)	40	71.37	72.93	49.92	24.2	46.69	62.99