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

Brandon McKinzie°, Zhe Gan°, Jean-Philippe Fauconnier⋆,
Sam Dodge⋆, Bowen Zhang⋆, Philipp Dufter⋆, Dhruti Shah⋆, Xianzhi Du⋆,
Futang Peng, Anton Belyi, Haotian Zhang, Karanjeet Singh, Doug Kang,
Hongyu Hè, Max Schwarzer, Tom Gunter, Xiang Kong, Aonan Zhang,
Jianyu Wang, Chong Wang, Nan Du, Tao Lei, Sam Wiseman,
Mark Lee, Zirui Wang, Ruoming Pang, Peter Grasch⋆,
Alexander Toshev†, and Yinfei Yang†

Apple
{bmckinzie,zhe.gan,toshev,yinfeiy}@apple.com
°First authors; ⋆Core authors; †Senior authors

A Dataset Details

A.1 Interleaved Image-Text Data

Following a process similar to OBELICS [14], we construct a dataset of 500M interleaved image-text documents, containing 1B images and 500B text tokens. These 500M documents are built from a collection of 3B HTML files described in Sec. A.2. From each of the HTML files, we extract the text body layer and all the `<img>` tags. We remove documents that have no images or more than 30 images. We then download the images and insert them at their original positions in the text. Finally, we perform **image filtering** and **image de-duplication** to remove low-quality and repetitive images.

During image filtering, we remove images that have corrupted bytes and/or header, aspect ratio less than 1/2 or greater than 2, are too small (less than 100px) or too large (larger than 10,000px), or if their URL contains *logo*, *button*, *icon*, *plugin* or *widget*. During image de-duplication, we remove images whose URL or MD5 hash have appeared more than 10 times in the dataset. Additionally, when an image appears multiple times on a single page, we only retain its first appearance.

A.2 Text-Only Data

From an initial Web corpus of 150B English HTML files, we perform boilerplate removal to arrive at the HTML representing the main content. We then follow similar processes as GPT-3 [4] and CCNet [35] to filter out documents that are too short, contain profanity, or are otherwise considered low-quality documents. We de-duplicate the data using exact-hash matching and LSH-based near-duplicate detection. Using these methods, we arrive at 3B HTML files.

Datasets	Size	Prompting Strategy
Text-only SFT	13k	–
LLaVA-Conv [20]	57k	–
LLaVA-Complex [20]	77k	
ShareGPT-4V [5]	102k	
VQAv2 [9]	83k	"Answer the question using a single word or phrase."
GQA [10]	72k	
OKVQA [24]	9k	
OCRVQA [28]	80k	
DVQA [11]	200k	
ChartQA [25]	18k	
AI2D [12]	3k	
DocVQA [27]	39k	
InfoVQA [26]	24k	
A-OKVQA [29]	66k	"Answer with the option's letter from the given choices directly."
COCO Captions [6]	83k	Sample from a pre-generated prompt list, *e.g.*, "Provide a brief description of the given image."
TextCaps [31]	22k	
SynthDog-EN [13]	500k	Sample from a pre-generated prompt list, *e.g.*, "Please transcribe all the text in the picture."
Total	**1.45M**	–

Table 1: List of datasets used for supervised fine-tuning.

A.3 Visual Instruction Tuning Data

Our final SFT data mixture contains a variety of datasets, mostly follow LLaVA-1.5 [18] and LLaVA-NeXT [19]. Specifically,

- To encourage the model to provide long-form detailed responses and perform conversations, we follow previous work, use the existing GPT-4 generated data (LLaVA-Conv and LLaVA-Complex [20]) and the existing GPT-4V generated data (ShareGPT-4V [5]) for model training. We also experimented with LAION-GPT4V, but did not observe further performance improvement, thus not included in the final mixture.
- To enhance the model with better vision-language (VL) understanding capability, we use a variety of academic task oriented VL datasets. These datasets are either in the form of image captioning, or in the form of VQA with short answers. Specifically,
  - For natural images: VQAv2 [9], GQA [10], OKVQA [24], A-OKVQA [29], and COCO Captions [6];
  - For text-rich images: OCRVQA [28], and TextCaps [31];
  - For document and chart understanding: DVQA [11], ChartQA [25], AI2D [12], DocVQA [27], InfoVQA [26], and SynthDog-En [13];
- To enhance the model's text-only instruction following capability, we also blend in a small amount of text-only SFT data.

The academic task oriented image captioning and VQA datasets are formatted into the instruction-following format, following LLaVA-1.5 [18], with detailed prompts summarized in Table 1.

B Training Details

B.1 Pre-training

Batch Size and Composition. For simplicity, all MM1 models are pre-trained with the same batch size of 512 and maximum decoder sequence length of 4096. We allow up to 16 images per input sequence, with each image resulting in 144 tokens as input to the decoder. Note that this results in roughly 1M text tokens and 1M image tokens per batch. Each input sequence is sampled from one of three types of input sources: (1) interleaved, (2) packed image-text pairs, or (3) text-only data, with sampling probability 45%, 45%, and 10%, respectively. When packing image-text pairs or interleaved documents along the sequence dimension, we modify the self-attention masks to prevent tokens from attention across example boundaries. For image-text pairs in particular, this was critical for maintaining strong few-shot performance.

Note that our sampling/mixing procedure is performed once offline and stored as a fixed *deterministic* snapshot of our pre-training mixture. This means, with the exception of our ablations on the pre-training mixture itself, all models in this paper are trained on the same examples in the same order. We found this was critical to ensure internal reproducibility of our results, as initial experiments showed that different random seeds in the input pipeline could have non-negligible impact on resulting models.

N	Pred. η	Pred. λ
1.2B	8.6e-5	5.0e-6
2.9B	5.9e-5	3.5e-6
6.4B	4.2e-5	2.5e-6
30B	2.2e-5	1.3e-6

Table 3: Predicted optimal peak learning rate η and weight decay λ for MM1 model sizes.

Learning Rate Schedule. For multimodal pre-training, MM1 employs a standard cosine learning rate decay schedule with an initial linear warmup of 2000 steps. The learning rate is then decayed to 10% of its peak value over the course of $2e5$ training steps. We perform gradient clipping with max norm 1 and use the AdamW optimizer with an implementation that decouples the learning rate and weight decay. For MM1-30B, we also add a z-loss term with scale 1e-4, as we observed this improves training stability, similar to [36].

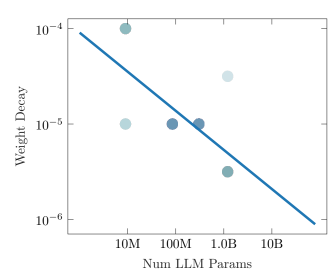
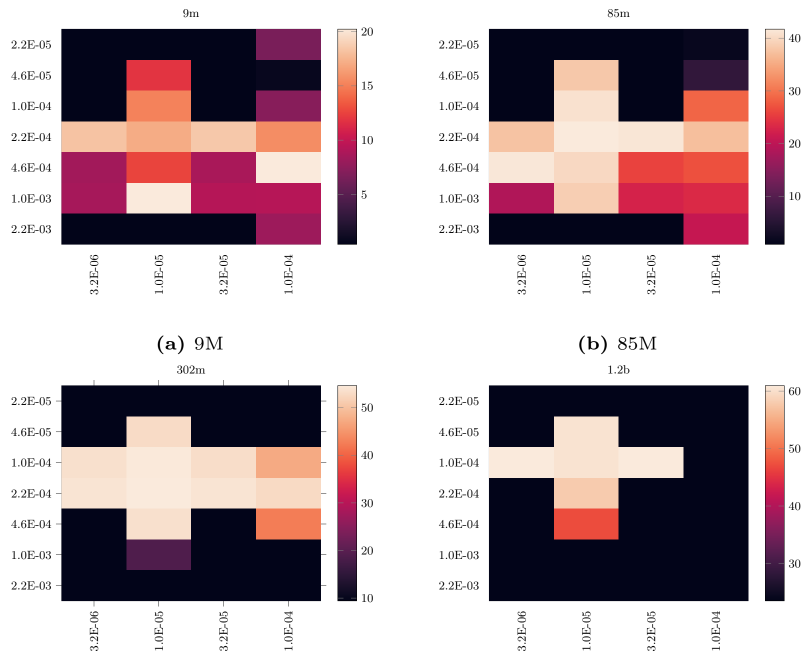


Fig. 1: Optimal weight decay as a function of model size for the grid searches described in Sec. B.1. The x-axis is the number of (non-embedding) LLM parameters and the y-axis is weight decay.

The predicted optimal (peak) learning rates for each of the main LLM sizes studied in this work are shown in Table 3. For simplicity, for the actual MM1 3B,

Model	Shot	Captioning			Visual Question Answering			
		COCO	NoCaps	TextCaps	VQAv2	TextVQA	VizWiz	OKVQA
MM1-3B Model Comparisons								
Flamingo-3B [1]	$0^\dagger$	73.0	–	–	49.2	30.1	28.9	41.2
	4	85.0	–	–	53.2	32.7	34.0	43.3
	8	90.6	–	–	55.4	32.4	38.4	44.6
	16	95.3	–	–	56.7	31.8	43.3	45.6
MM1-3B	0	73.5	55.6	63.3	46.2	29.4	15.6	26.1
	4	112.3	99.7	84.1	57.9	45.3	38.0	48.6
	8	114.6	104.7	88.8	63.6	44.6	46.4	48.4
	16	116.8	107.6	91.6	60.9	46.1	53.8	50.5
MM1-7B Model Comparisons								
IDEFICS-9B [14]	$0^\dagger$	46.0*	36.8	25.4	50.9	25.9	35.5	38.4
	4	93.0*	81.3	60.0	55.4	27.6	36.9	45.4
	8	97.0*	86.8	63.2	56.4	27.5	40.4	47.7
	16	99.7*	89.4	67.4	57.0	27.9	42.6	48.4
Flamingo-9B [1]	$0^\dagger$	79.4	–	–	51.8	31.8	28.8	44.7
	4	93.1	–	–	56.3	33.6	34.9	49.3
	8	99.0	–	–	58.0	33.6	39.4	50.0
	16	102.2	–	–	59.4	33.5	43.0	50.8
Emu2-14B [33]	$0^\dagger$	–	–	–	52.9	–	34.4	42.8
	4	–	–	–	58.4	–	41.3	–
	8	–	–	–	59.0	–	43.9	–
MM1-7B	0	76.3	61.0	64.2	47.8	28.8	15.6	22.6
	4	109.8	96.2	84.5	60.6	44.4	37.4	46.6
	8	116.3	106.6	88.2	63.6	46.3	45.3	51.4
	16	118.6	111.1	93.1	65.2	46.9	53.2	52.9
MM1-30B Model Comparisons								
IDEFICS-80B [14]	$0^\dagger$	91.8*	65.0	56.8	60.0	30.9	36.0	45.2
	4	110.3*	99.6	72.7	63.6	34.4	40.4	52.4
	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 [1]	$0^\dagger$	84.3	–	–	56.3	35.0	31.6	50.6
	4	103.2	–	–	63.1	36.5	39.6	57.4
	8	108.8	–	–	65.6	37.3	44.8	57.5
	16	110.5	–	–	66.8	37.6	48.4	57.8
Emu2-37B [33]	0	–	–	–	33.3	26.2	40.4	26.7
	4	–	–	–	67.0	48.2	54.6	53.2
	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
	4	117.9	103.8	87.5	68.8	48.1	41.7	54.9
	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

Table 2: Complete MM1 pre-training few-shot evaluation results. (*) IDEFICS includes PMD in its training data (includes COCO). (†) These models included two text-only demonstrations in their "0" prompt, whereas MM1 does not.



(a) 9M

(b) 85M

(c) 302M

(d) 1.2B

Fig. 2: 8-shot average for grid searches over peak learning rate (y-axis) and weight decay (x-axis) for different LLM sizes. Black cells correspond to settings we did not run a corresponding experiment for.

7B, and 30B models, we used η equal to 6e-5, 4e-5, and 2e-5, respectively. Finally, we fix the peak LR of the randomly initialized vision-language connector of MM1 to η =8e-5 for all model sizes. For future versions of MM1, we plan on incorporating techniques similar to [37] to avoid the need to conduct costly hyperparameter searches.

Learning Rate and Weight Decay Grid Searches. The individual grid search results corresponding to the final curve fit in Figure **??** are shown in Figure 2. We train grid search models for $5e^4$ steps, as [36] found this does not alter the conclusions. We can apply the same procedure that was used for predicting optimal learning rate to predict weight decay values, as shown in Figure 1. The blue circles correspond to actual data points from the grid search with sampling probability (and darkness of color) proportional to their 8-shot average performance. The corresponding predictions for each of the main model sizes in this work are shown in Table 3.

B.2 Supervised Fine-tuning (SFT)

The model is fine-tuned for 10k steps with batch size 256 and sequence length 2048. We employ the AdaFactor optimizer with peak learning rate 1e-5 and cosine decay to 0. We experimented different learning rates; empirically, the value of 1e-5 is optimal. During SFT, we keep both the image encoder and the LLM *unfrozen*, as empirically, we observe that finetuning the whole model achieves better performance.

C Evaluation Details

C.1 Pre-training Evaluation

Few-shot prompts are randomly sampled per-dataset from the training set if available, otherwise the validation set (ensuring the query example does not appear in any of the shots). Outputs are generated with greedy decoding until the model emits the EOS token or any additional stop tokens that can be specified on a per-task basis. The additional stop token for captioning tasks is just the newline character, and for VQA tasks we also include ".", ",", and "Question" as valid stop tokens. For postprocessing VQA predictions, we use the same logic as OpenFlamingo[1] [2]. For captioning tasks, we report CIDEr score [34] using the nlg-eval package [30]. All of our multimodal pre-training evaluations are implemented in an internal fork of EleutherAI's lm-evaluation-harness [8].

Dataset	Evaluation Split
COCO	Karpathy test
NoCaps	val
TextCaps	val
VQAv2	testdev
TextVQA	val
VizWiz	testdev
OKVQA	val

Table 4: Splits used for pre-training evaluation. Note that, unlike the main pre-training results, all pre-training ablations use the validation splits for VQAv2 and VizWiz.

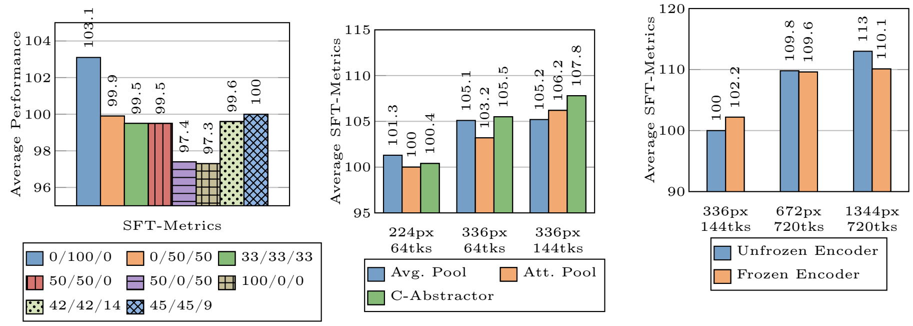
C.2 SFT Evaluation Benchmarks

We evaluate our SFT models on a collection of both traditional academic VL benchmarks and recent benchmarks specifically designed for MLLMs. For academic VL benchmarks, we include VQAv2 [9], TextVQA [32], and the image subset of ScienceQA [23]. For recent MLLM benchmarks, we include POPE [16], MME [7], MMBench [21], SEED-Bench [15], LLaVA-Bench-in-the-Wild [20], MM-Vet [39], MathVista [22], and the recent popular MMMU [40]. For all the benchmarks, we use greedy decoding to generate the responses. For MM-Vet and LLaVA-Bench-in-the-Wild, which use GPT-4 for evaluation, we run the evaluation 3 times, and report the average.

C.3 SFT Evaluation Meta-Average

In the process of SFT ablation, we synthesize all benchmark results into a single meta-average number to simplify comparisons. Because the evaluation metrics of different datasets may have different ranges, we normalize with respect to a baseline configuration. This is achieved by initially standardizing the results for

[1] Specifcally, the implementation of VQAMetric (commit 60a5fd6).



(a) Interleaved/Captions/Text data mixture. **(b)** Vision-language connector. **(c)** (Un-)freezing image encoder during SFT.

Fig. 3: SFT ablations. (a) The impact of pre-training data mixture on SFT results. Here, $x/y/z$ means that x% of the data is interleaved, y% is captions, and z% is pure text. tks: the number of image tokens. (b) The impact of different vision-language connectors on SFT results. For both (a) and (b), we first pre-train MM1-3B with the ablated setting, and then perform SFT on the pre-trained models. (c) Freezing or unfreezing the image encoder during SFT.

each task; that is, we adjust every metric by dividing it by its respective baseline, followed by averaging across all metrics. To elaborate, we establish our baseline using the performance metrics of a compact MM1 model, which is trained on 224×224 image resolution and employs attention pooling with 64 image queries.

C.4 Additional SFT Ablations

In this section, we perform SFT ablations. This section is analogous to Section 3 of the main text; here, we perform SFT on the same checkpoints and evaluate if similar lessons hold true on SFT evaluations, instead of pre-training evaluations. Furthermore, we also study whether to keep the image encoder frozen or not during SFT. For all of these ablations, we train MM1-3B-Chat.

Pre-training data mixture ablations. In Figure 3a, we compare the SFT performance with different weights for pre-training data. Pre-training with caption-only data gives the best performance across the SFT evaluation metrics. This corroborates **Data lesson 1**: caption data still lifts zero-shot performance for SFT evaluations. However, the SFT metrics do not measure few-shot performance, so the impact of the interleaved data is not noticeable in this table.

Visual-language connector ablations. In Figure 3b, we evaluate different visual-language connector configurations. As can be seen, if a low number of image tokens is used, average pooling gives similar results as C-Abstractor. When the number of image tokens is increased, the C-Abstractor configuration gives the best results. Overall, the impact of the choice of visual-language connector appears to not have a very significant impact on final test performance. Our final models use the C-Abstractor architecture.

Image encoder ablations. In Figure 3c, we study whether to keep the image encoder frozen or not during SFT. The results show that at lower image resolutions, a frozen image encoder results in better performance than an unfrozen image encoder (+2.2 points). However, at higher resolutions (*i.e.*, 1344px), it is beneficial to unfreeze the image encoder (+2.9 points). This is likely because the pre-training is performed at the base resolution without any interpolation or image sub-divisions.

C.5 Implementation Details for Few-shot MM1-30B-Chat

Our fine-tuned model can utilize in-context examples to achieve even stronger performance. Interestingly, the performance goes up when increasing the number of examples. We demonstrate this with MM1-30B-Chat.

One challenge for few-shot inputs arises due to the use of sub-image decomposition. While this strategy lifts zero-shot performance, it significantly increases the effective number of tokens consumed per image. Using 5 sub-images per input image as MM1-30B-Chat does, processing a 4-shot example where every example contains just one source image already yields 20 effective images. Representing every image with 144 tokens therefore requires 2,880 tokens for images alone, quickly exhausting limited language model context. To mitigate this limitation, we propose a new *mixed-resolution* approach. Specifically, for K in-context examples, we only encode the last N images at a high resolution via sub-image decomposition, the remaining $K-N$ in-context examples are processed at lower resolution. This makes in-context examples much less expensive to encode and allows to increase the number of in-context examples within a fixed token budget, further increasing performance. In our implementation, we set $N = 3$.

We demonstrate the effectiveness of our proposed strategy using the MathVista benchmark. Using in-context learning with chain-of-thought, the performance of MM1-30B-Chat improves from 39.4 (0-shot) to 41.9 (4-shot, with all in-context examples using full sub-image decomposition). Applying our mixed-resolution approach allows to encode additional in-context examples, enabling up to 8-shot chain of thought, further improving the performance to 44.4. This illustrates that our MM1-Chat model retains in-context learning capabilities inherited from its pre-training regimen, and that our strategy of mixed decomposition is effective at further increasing few-shot performance within a restricted context length.

D Qualitative Examples

In this section, we share qualitative examples of MM1 predictions.

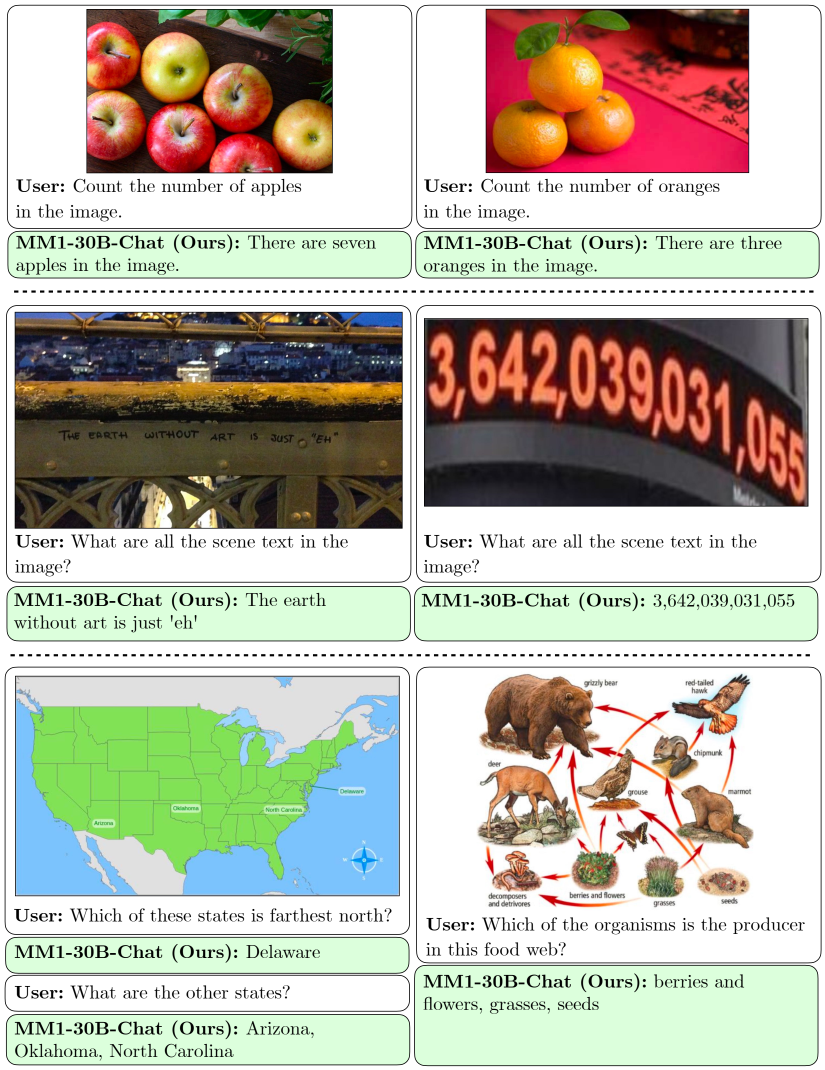
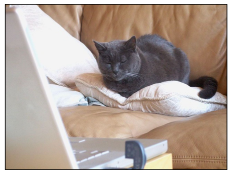


Fig. 4: Examples testing MM1 counting, OCR and scientific knowledge capabilities. Images and prompts are from COCO 2014 validation set [17] and [38].



User: Can you find the red ball near the gray cat in the image?

MM1-30B-Chat (Ours): No, there is no red ball near the gray cat in the image. The cat is sitting on a pillow on a couch, and there are no other objects like a red ball visible in the scene.

User: Can you see the bright blue feathers on the bird's wing?

MM1-30B-Chat (Ours): No, the bird in the image does not have bright blue feathers on its wing. The bird is a large bird of prey, such as a hawk or an eagle, with typical bird of prey feathers.

User: From scale 1 to 10, decide how similar the image is to the text prompt "a parrot driving a car". Explain your reasoning.

MM1-30B-Chat (Ours): I would rate the image a 1 on the scale of 1 to 10 in relation to the text prompt "a parrot driving a car." The image shows a dolphin jumping out of the water, and the text prompt is referring to a parrot driving a car. There is no similarity between the image and the text prompt, other than the fact that both contain animals. The dolphin is in the water, while the parrot is associated with a car. The actions of the animals are also different, as the dolphin is jumping out of the water, while the parrot is implied to be driving a car. The environments are completely different, with the dolphin being in a natural aquatic environment and the parrot being in a human-made, terrestrial environment.

User: From scale 1 to 10, decide how similar the image is to the text prompt "a parrot driving a car". Explain your reasoning.

MM1-30B-Chat (Ours): The image portrays a scene that is quite similar to the text prompt "a parrot driving a car". In the image, there are two parrots situated inside a blue car. The car is parked on a green field with a backdrop of palm trees and a yellow sky.

One of the parrots is perched on the steering wheel, giving the impression that it is driving the car. The other parrot is comfortably seated on the roof of the car, seemingly enjoying the view. The colors in the image are vibrant, with the blue of the car contrasting beautifully with the green field and the yellow sky. The palm trees add a tropical feel to the scene.

The image is a close representation of the text prompt, with the exception of the number of parrots and their actions. While the prompt mentions only one parrot driving a car, the image shows two parrots, one seemingly driving and the other on the roof of the car. This adds an element of surprise and humor to the image. Overall, the image scores a 9 out of 10 for its close representation of the text prompt and its creative interpretation.

Fig. 5: Examples testing MM1 against adversarial prompts and image-prompt alignment. Images and prompts are from COCO 2014 validation set [17], [38] and [3].

User: After seeing the below image, how people might emotionally feel and react?

MM1-30B-Chat (Ours): Upon witnessing this image, people might feel a mixture of emotions such as awe, fascination, and tranquility. The breathtaking view of the airplane window showing the wing of the plane and a vast, snowy mountain range beneath it offers a sense of serenity and the beauty of nature. The magnificent, orderly pattern of the mountains and their snowy peaks contrasted with the clear blue sky evokes a sense of wonder and appreciation for the natural world. This view also offers a unique perspective of the world that one might not experience daily, thus sparking curiosity and excitement. Additionally, the image can bring about feelings of calmness and peace, given the untouched and pristine nature of the landscape depicted.

User: After seeing the below image, how people might emotionally feel and react?

MM1-30B-Chat (Ours): The image shows a dark stairwell with a creepy atmosphere, which might evoke feelings of unease, fear, or curiosity in people. Some individuals may react by avoiding the stairwell altogether, while others might be intrigued by the mysterious ambiance and decide to explore it further. The darkness and the shadows in the stairwell can play with people's perceptions and trigger their imagination, possibly making them feel as if there is someone or something lurking in the shadows. Overall, the image can provoke a range of emotional responses, from apprehension to fascination, depending on each person's disposition and mindset.

User: <img1> <img2> For each of the two images comment on whether the water is salty.

MM1-30B-Chat (Ours): In the first image, there are several boats floating on a body of water. Based on the presence of boats, it's likely this is a harbor, and the water is likely salty as most harbors are typically found near the ocean. However, without more information, it's not possible to confirm the salinity of the water.

In the second image, various African animals, including elephants and giraffes, are gathered around a watering hole in a dry grass field. This body of water is most likely freshwater as animals need fresh water to drink. The presence of numerous animals drinking from the watering hole confirms that this water is essential for their survival, and it's highly unlikely they would be able to drink saltwater.

User: <img1> <img2> Which image shows the healthier meal?

MM1-30B-Chat (Ours): The healthier meal among the given images is the one that features broccoli and a bowl of vegetables, along with a plate of meat. Incorporating vegetables, like broccoli and those in the bowl, provides essential nutrients, vitamins, and fiber, which contribute to a balanced and healthier diet. The meal with various toppings on a pizza and beer would be considered less healthy as it may contain more calories, fat, and processed ingredients compared to the other meal that emphasizes vegetables and meat.

Fig. 6: Examples testing MM1 ability to perceive image aesthetics and compare multiple images. Images and prompts are from COCO 2014 validation set [17] and [38].

User: First, read the operating menus below for coffee machine buttons.

Menu 1:

Menu 2:

Menu 3:

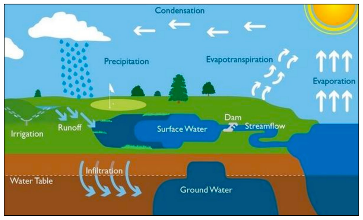
Now, let's look at the operating panel of the coffee machine in the image below, which button should I use for a 6 OZ coffee: left, right, top or bottom button?

MM1-30B-Chat (Ours): To brew a 6 OZ coffee using the coffee machine in the image, you should use the left button. This is because the buttons on the machine correspond to the menu icons, and the left button corresponds to the 6 OZ menu icon. The other buttons correspond to different cup sizes, but the left button is the one you need to press to brew a 6 OZ coffee.

User: Describe the image

MM1-30B-Chat (Ours): The image captures a moment from the perspective of a car's dashboard camera. The car is traveling on a two-lane road, divided by a yellow line, suggesting a rural or less populated area. The road is flanked by trees on both sides, their green leaves contrasting with the gray, cloudy sky above. A speed limit sign displaying "35" is visible on the right side of the road, indicating the maximum speed allowed for vehicles. The overall scene suggests a calm and serene drive through a lightly wooded area.

Fig. 7: Following [38], we tested MM1 on task-oriented scenarios such as operating machines and navigating. Images and prompts are from [38].

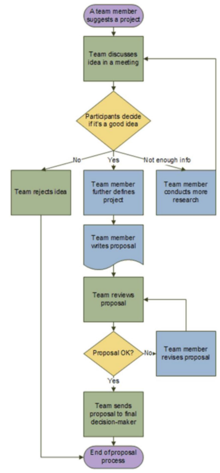


User: Suppose you are a teacher, please use the figure to explain the distinction between evaporation and evapotranspiration.

MM1-30B-Chat (Ours): Evaporation and evapotranspiration are two important processes in the water cycle that are often confused. The image provides a clear distinction between the two. Evaporation is the process where water changes from a liquid state to a gaseous state. This is represented in the image by the sun's heat causing water in the lake and ocean to turn into water vapor and rise into the atmosphere.

On the other hand, evapotranspiration is the combination of evaporation and plant transpiration. Transpiration is the process where water absorbed by plants is released into the atmosphere through the plant's leaves. In the image, evapotranspiration is represented by the water absorbed by the trees and grass being released back into the atmosphere.

In summary, evaporation is the process of water turning into vapor and rising into the atmosphere, while evapotranspiration is the combination of evaporation and plant transpiration, representing the water absorbed by plants and released back into the atmosphere.



User: Explain the figure.

MM1-30B-Chat (Ours): The image presents a flowchart that outlines the process of how a team member suggests a project. The flowchart begins with a green rectangle that contains the text "A team member suggests a project". From this starting point, the process flows through a series of diamond-shaped decision points and rectangular action points, each containing specific text that describes the next step in the process.

The process continues with the team discussing the idea in a meeting, followed by a decision point where participants decide if it's a good idea. If it's a good idea, the process moves forward. If not, the process loops back to the point where the team member conducts more research.

The flowchart is color-coded with green rectangles representing actions and yellow diamonds representing decisions. The lines connecting these shapes are black, indicating the sequence of the process. The flowchart is set against a white background, which makes the colored shapes stand out clearly.

The flowchart provides a clear and structured representation of the process, making it easy to understand the sequence of steps involved in the project suggestion and decision-making process. It's a useful tool for visualizing complex processes and ensuring that all steps are considered and followed in the correct order.

Fig. 8: Examples testing MM1 ability at extracting information from graphics. The right part shows an example of confusion, highlighted in red. Images and prompts are from [38].

User: I will show you several pictures and your task is to answer my question.

A: B: C:

Estimate the temperature in Celsius for each image. Also provide reasons.

MM1-30B-Chat (Ours):

- The image with the umbrella on the beach: 25-30 degrees Celsius. The umbrella is up, and people are at the beach, suggesting a warm, sunny day.
- The image with the plane flying over the snow: -20 to -30 degrees Celsius. The plane is flying over a snow-covered area, and there are vehicles on the ground, indicating a cold winter day.
- The image with the animals grazing in the field: 15-20 degrees Celsius. The animals are grazing, which usually occurs in milder weather conditions. The grass is green, further suggesting a temperate climate.

Fig. 9: Examples testing MM1 ability at reasoning across images and texts. Images are from COCO 2014 validation set [17] and MMMU dev set [40].

User:

Image A:

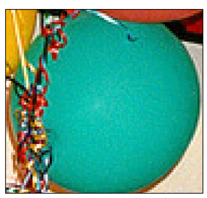
Image B:

For each image state what's the number shown on the cake. Then return the sum.

MM1-30B-Chat (Ours): The first image shows a cake with the number 3 on it, and the second image shows a cake with the number 4 on it. The sum is 7.

User: What is the color of the balloon?

The color is 🔴 R_E_D 🔴✨



The color is 🟢 G_R_E_E_N 🟢✨

MM1-30B-Chat (Ours):
The color is 🔵 B_L_U_E 🔵✨

User: I am showing you an image and you need to predict the letter or number shown when rotating the image by 180 degrees.

W

MM1-30B-Chat (Ours): 9

Fig. 10: Examples testing MM1 ability to follow instructions across multiple images (top). Examples testing MM1 at following a style or a task across few shots (bottom). Images are from COCO 2014 validation set [17] and from the authors.

References

1. Alayrac, J.B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., Ring, R., Rutherford, E., Cabi, S., Han, T., Gong, Z., Samangooei, S., Monteiro, M., Menick, J., Borgeaud, S., Brock, A., Nematzadeh, A., Sharifzadeh, S., Binkowski, M., Barreira, R., Vinyals, O., Zisserman, A., Simonyan, K.: Flamingo: a visual language model for few-shot learning (2022)
2. Awadalla, A., Gao, I., Gardner, J., Hessel, J., Hanafy, Y., Zhu, W., Marathe, K., Bitton, Y., Gadre, S., Sagawa, S., Jitsev, J., Kornblith, S., Koh, P.W., Ilharco, G., Wortsman, M., Schmidt, L.: Openflamingo: An open-source framework for training large autoregressive vision-language models. arXiv preprint arXiv:2308.01390 (2023)
3. Black, K., Janner, M., Du, Y., Kostrikov, I., Levine, S.: Training diffusion models with reinforcement learning. arXiv preprint arXiv:2305.13301 (2023)
4. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. NeurIPS (2020)
5. Chen, L., Li, J., Dong, X., Zhang, P., He, C., Wang, J., Zhao, F., Lin, D.: Sharegpt4v: Improving large multi-modal models with better captions. arXiv preprint arXiv:2311.12793 (2023)
6. Chen, X., Fang, H., Lin, T.Y., Vedantam, R., Gupta, S., Dollár, P., Zitnick, C.L.: Microsoft coco captions: Data collection and evaluation server. arXiv preprint arXiv:1504.00325 (2015)
7. Fu, C., Chen, P., Shen, Y., Qin, Y., Zhang, M., Lin, X., Yang, J., Zheng, X., Li, K., Sun, X., et al.: Mme: A comprehensive evaluation benchmark for multimodal large language models. arXiv preprint arXiv:2306.13394 (2023)
8. Gao, L., Tow, J., Abbasi, B., Biderman, S., Black, S., DiPofi, A., Foster, C., Golding, L., Hsu, J., Le Noac'h, A., Li, H., McDonell, K., Muennighoff, N., Ociepa, C., Phang, J., Reynolds, L., Schoelkopf, H., Skowron, A., Sutawika, L., Tang, E., Thite, A., Wang, B., Wang, K., Zou, A.: A framework for few-shot language model evaluation (12 2023). `https://doi.org/10.5281/zenodo.10256836`, `https://zenodo.org/records/10256836`
9. Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., Parikh, D.: Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In: CVPR (2017)
10. Hudson, D.A., Manning, C.D.: Gqa: A new dataset for real-world visual reasoning and compositional question answering. In: CVPR (2019)
11. Kafle, K., Price, B., Cohen, S., Kanan, C.: Dvqa: Understanding data visualizations via question answering. In: CVPR (2018)
12. Kembhavi, A., Salvato, M., Kolve, E., Seo, M., Hajishirzi, H., Farhadi, A.: A diagram is worth a dozen images. In: ECCV (2016)
13. Kim, G., Hong, T., Yim, M., Nam, J., Park, J., Yim, J., Hwang, W., Yun, S., Han, D., Park, S.: Ocr-free document understanding transformer. In: ECCV (2022)
14. Laurençon, H., Saulnier, L., Tronchon, L., Bekman, S., Singh, A., Lozhkov, A., Wang, T., Karamcheti, S., Rush, A.M., Kiela, D., Cord, M., Sanh, V.: Obelics: An open web-scale filtered dataset of interleaved image-text documents (2023)
15. Li, B., Wang, R., Wang, G., Ge, Y., Ge, Y., Shan, Y.: Seed-bench: Benchmarking multimodal llms with generative comprehension. arXiv preprint arXiv:2307.16125 (2023)

16. Li, Y., Du, Y., Zhou, K., Wang, J., Zhao, W.X., Wen, J.R.: Evaluating object hallucination in large vision-language models. arXiv preprint arXiv:2305.10355 (2023)
17. Lin, T., Maire, M., Belongie, S.J., Bourdev, L.D., Girshick, R.B., Hays, J., Perona, P., Ramanan, D., Doll'a r, P., Zitnick, C.L.: Microsoft COCO: common objects in context. arXiv preprint arXiv:1405.0312 (2014)
18. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning. arXiv preprint arXiv:2310.03744 (2023)
19. Liu, H., Li, C., Li, Y., Li, B., Zhang, Y., Shen, S., Lee, Y.J.: Llava-next: Improved reasoning, ocr, and world knowledge (January 2024), `https://llava-vl.github.io/blog/2024-01-30-llava-next/`
20. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning (2023)
21. Liu, Y., Duan, H., Zhang, Y., Li, B., Zhang, S., Zhao, W., Yuan, Y., Wang, J., He, C., Liu, Z., et al.: Mmbench: Is your multi-modal model an all-around player? arXiv preprint arXiv:2307.06281 (2023)
22. Lu, P., Bansal, H., Xia, T., Liu, J., Li, C., Hajishirzi, H., Cheng, H., Chang, K.W., Galley, M., Gao, J.: Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. arXiv preprint arXiv:2310.02255 (2023)
23. Lu, P., Mishra, S., Xia, T., Qiu, L., Chang, K.W., Zhu, S.C., Tafjord, O., Clark, P., Kalyan, A.: Learn to explain: Multimodal reasoning via thought chains for science question answering. NeurIPS (2022)
24. Marino, K., Rastegari, M., Farhadi, A., Mottaghi, R.: Ok-vqa: A visual question answering benchmark requiring external knowledge. In: CVPR (2019)
25. Masry, A., Long, D.X., Tan, J.Q., Joty, S., Hoque, E.: Chartqa: A benchmark for question answering about charts with visual and logical reasoning. arXiv preprint arXiv:2203.10244 (2022)
26. Mathew, M., Bagal, V., Tito, R., Karatzas, D., Valveny, E., Jawahar, C.: Infographicvqa. In: WACV (2022)
27. Mathew, M., Karatzas, D., Jawahar, C.: Docvqa: A dataset for vqa on document images. In: WACV (2021)
28. Mishra, A., Shekhar, S., Singh, A.K., Chakraborty, A.: Ocr-vqa: Visual question answering by reading text in images. In: ICDAR (2019)
29. Schwenk, D., Khandelwal, A., Clark, C., Marino, K., Mottaghi, R.: A-okvqa: A benchmark for visual question answering using world knowledge. In: ECCV (2022)
30. Sharma, S., El Asri, L., Schulz, H., Zumer, J.: Relevance of unsupervised metrics in task-oriented dialogue for evaluating natural language generation. arXiv preprint arXiv:1706.09799 (2017)
31. Sidorov, O., Hu, R., Rohrbach, M., Singh, A.: Textcaps: a dataset for image captioning with reading comprehension. In: ECCV (2020)
32. Singh, A., Natarjan, V., Shah, M., Jiang, Y., Chen, X., Parikh, D., Rohrbach, M.: Towards vqa models that can read. In: CVPR (2019)
33. Sun, Q., Cui, Y., Zhang, X., Zhang, F., Yu, Q., Luo, Z., Wang, Y., Rao, Y., Liu, J., Huang, T., et al.: Generative multimodal models are in-context learners. arXiv preprint arXiv:2312.13286 (2023)
34. Vedantam, R., Zitnick, C.L., Parikh, D.: Cider: Consensus-based image description evaluation. arXiv preprint arXiv:1411.5726 (2014)
35. Wenzek, G., Lachaux, M.A., Conneau, A., Chaudhary, V., Guzmán, F., Joulin, A., Grave, E.: Ccnet: Extracting high quality monolingual datasets from web crawl data. arXiv preprint arXiv:1911.00359 (2019)
36. Wortsman, M., Liu, P.J., Xiao, L., Everett, K., Alemi, A., Adlam, B., Co-Reyes, J.D., Gur, I., Kumar, A., Novak, R., Pennington, J., Sohl-dickstein, J., Xu, K.,

Lee, J., Gilmer, J., Kornblith, S.: Small-scale proxies for large-scale transformer training instabilities (2023)
37. Yang, G., Hu, E.J., Babuschkin, I., Sidor, S., Liu, X., Farhi, D., Ryder, N., Pachocki, J., Chen, W., Gao, J.: Tensor programs v: Tuning large neural networks via zero-shot hyperparameter transfer (2022)
38. Yang, Z., Li, L., Lin, K., Wang, J., Lin, C.C., Liu, Z., Wang, L.: The dawn of lmms: Preliminary explorations with gpt-4v (ision). arXiv preprint arXiv:2309.17421 (2023)
39. Yu, W., Yang, Z., Li, L., Wang, J., Lin, K., Liu, Z., Wang, X., Wang, L.: Mm-vet: Evaluating large multimodal models for integrated capabilities. arXiv preprint arXiv:2308.02490 (2023)
40. Yue, X., Ni, Y., Zhang, K., Zheng, T., Liu, R., Zhang, G., Stevens, S., Jiang, D., Ren, W., Sun, Y., et al.: Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. arXiv preprint arXiv:2311.16502 (2023)