The Scalability of Simplicity: Empirical Analysis of Vision-Language Learning with a Single Transformer

Supplementary Material

In this supplementary material, we provide additional experimental details and results for SAIL.

6. Additional Experimental Details

Training Configurations. In this section, we provide the corresponding setups for our experiment series in the main paper, including the default setting, the data scaling series, the model scaling series, and ablation experiment settings. The detailed configurations are shown in Table 17.

Evaluation Configurations. In the main paper, we measure the model performance on several benchmarks: VQAv2 [27], GQA [35], ScienceQA-IMG [51], TextVQA [64], POPE [43], MME [45], MMBench [49], and SEEDBench-IMG [39]. We normalized the performance to a full score of 100 and averaged the performance across these benchmarks to plot the curves shown in Figure 1(A) and Figure 3. The detailed experimental results are shown in Table 16.

7. Additional Experimental Results

Additional Experiments on Data Scaling. In Section 4.3, we compared SAIL with its modular MLLM counterpart in terms of data scaling capability. For the modular MLLM baseline, we employed SigLIP-SO [84] as the vision encoder while maintaining identical architecture and initialization parameters for the language model component as SAIL. Both models underwent Pretraining stage-1 with identical settings, where SAIL processed 32M, 128M, and 512M image-text pairs respectively, with all parameters being trainable during pretraining. Here, our purpose was to test whether SAIL can match a strong modular MLLM by scaling up pretraining data. To this end, we used a strong pretrained ViT in the modular MLLM and ensured both models saw the same amount of data. This strict setup allows us to justify SAIL 's effectiveness if it achieves comparable performance.

Additionally, we conducted complementary experiments under a modified configuration. In this setup, both SAIL and the modular MLLM's *ViT* were pretrained on identical image-text pairs from the same data sources. The modular MLLM subsequently underwent an alignment phase using LLaVA-558K [48] to train its projector. Both models then received identical supervised fine-tuning (SFT) on the same downstream data. As in Figure 5, SAIL outperforms the modular MLLM at 128M pretraining data and demonstrates better data scaling.

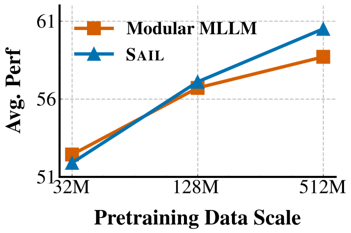


Figure 5. **Data scaling curve.** For the experiment setting, both SAIL and the ViT of the modular MLLM were trained from scratch with the same amount and source of image-text pair data.

Additional Experiments on ARO. We add several VLMs for ARO in Table 8: all achieve near-perfect performance, showing that current VLMs are effective at recognizing object attributes and relationships. Combined with our image classification and segmentation results, this further highlights SAIL 's strong visual capabilities compared to both leading vision models and VLM baselines.

Model	Rel	Attr	C.Ord	F.Ord
LLaVA-ov-7B	98.8	99.2	97.8	100.0
Qwen2VL-7B	98.4	99.2	100.0	100.0
EVEv2	94.4	98.8	99.9	100.0
SOLO	96.5	98.6	95.7	96.7
SAIL	100.0	99.5	100.0	100.0

Table 8. Performance on ARO (%).

Measure SAIL's Language Capabilities. As demonstrated in Table 9, multimodal pretraining causes some drop in language performance compared to the original LLM. However, as more text data in pretraining is included, SAIL 's language ability improved.

Model	ARC-e	HellaSwag	PIQA	naturalQA-5	triviaQA-5
Mistral-7B	80.0	81.3	83.0	29.1	69.9
SAIL-128M	68.8	76.7	79.4	23.8	59.3
SAIL-512M	71.2	77.9	79.8	25.6	63.8

Table 9. Language capability performance (%).

Vision Task Performance at Different Stages. We evaluated SAIL after Pretraining Stage 2 and SFT in Table 10. SAIL-S2 shows slight improvement over SAIL-S1, likely due to more seen data. SAIL-SFT maintains strong classification performance but shows some drop in segmentation, likely due to adaptation to downstream tasks during SFT, but it still outperforms other VLMs in Table 3.

SAIL-S2					SAIL-SFT				
cls@1	cls@5	mIoU	mAcc	aAcc	cls@1	cls@5	mIoU	mAcc	aAcc
86.28	97.92	56.28	67.83	85.12	83.77	97.38	46.18	57.76	80.92

Table 10. Vision task performance across different stages (%).

A Comparison of SAIL and LLaVA-1.5. In this section, we conduct an experiment to compare SAIL with LLaVA-1.5 [46]. In this experiment, our SAIL is trained on 512M image-text pairs in Pretraining Stage 1, followed by fine-tuning on the LLaVA-mix-665K dataset. To fairly compare the performance of the two models, we do not use the anyres strategy during SFT. Instead, we adopt the same image processing approach as LLaVA-1.5, ensuring that the aspect ratio and number of image tokens are consistent across both models.

The experimental results are presented in Table 11. Despite our model being trained on only 512M image-text pairs, which is significantly smaller than the CLIP pretraining data used in the LLaVA-1.5 model, the results show that our model achieves comparable performance to LLaVA-1.5 across various benchmarks. Remarkably, our model even outperforms LLaVA-1.5 on specific benchmarks such as DocVQA and ChartVQA.

These findings highlight the strong potential of Single Transformer models in terms of data scaling. Specifically, they suggest that even with a relatively smaller pretraining dataset, Single Transformer models can perform on par with, or even exceed, more extensively trained modular MLLMs like LLaVA-1.5 when similar preprocessing and controlled variables are applied.

Compare SAIL and LLaVA on MMVP. We compare SAIL and LLaVA-1.5 [46] on MMVP [69] to dissect the behavior of the two models. The results are shown in Figure 6. From examples (A) and (B), we observe that SAIL performs better in perceiving minor regions and objects. Examples (C) and (D) illustrate that SAIL can more accurately distinguish the states of objects.

Additional Experiments on Information Flow Pattern Analysis. In the main paper, we analyzed the distribution patterns of image attention scores for different Single Transformer-based MLLMs and modular MLLMs. The results showed that Single Transformer-based MLLMs allocate more attention weights to image tokens. However, this could be due to different models processing varying numbers of image tokens, where more image tokens lead to higher aggregated attention scores.

To analyze this in a more controlled manner, we designed an additional experiment. Using the data scaling setup at 512M, we pretrained SAIL and its modular MLLM counterpart. After pretraining, we fine-tuned both models using the LLaVA-mix-665K dataset, fixing the resolution size to 224x224 during SFT, instead of using any resolution.

The results, shown in Figure 7, reveal that SAIL allocates higher attention scores to image tokens across all transformer layers compared to the modular MLLM, particularly in medium layers (+43.5% in layer 14) and deep layers (+41.2% in layer 31).

From this, we can conclude that Single Transformer-based MLLMs tend to allocate a significant portion of attention to previous image tokens during prediction. In contrast, modular MLLMs allocate a smaller portion of their attention directly to image tokens, indicating a less image-centric approach in their prediction mechanism.

Attention Map Visualization. In the main paper, we found that Single Transformer-based MLLMs allocate a large portion of attention weights to image tokens during inference, indicating a more vision-centric model. Here, we visualize the attention distribution of SAIL across different regions of the image when predicting tokens.

The results in Figure 8 illustrate the attention maps for specific tokens to the image portion when SAIL generates predictions for multimodal queries. The visualizations show that in the early transformer layers, the predicted tokens primarily focus on the salient regions of the image. As the model progresses to deeper layers, the attention shifts to areas more relevant to the predicted tokens. This behavior demonstrates that SAIL has the potential to function as a grounding model, effectively correlating text tokens with their corresponding image regions.

In other words, during inference, the model incrementally concentrates attention weights on relevant regions, aiding in decision-making. This progressive focusing of attention signifies the model's capability to ground text tokens in the corresponding visual context, enhancing its performance in vision-language tasks.

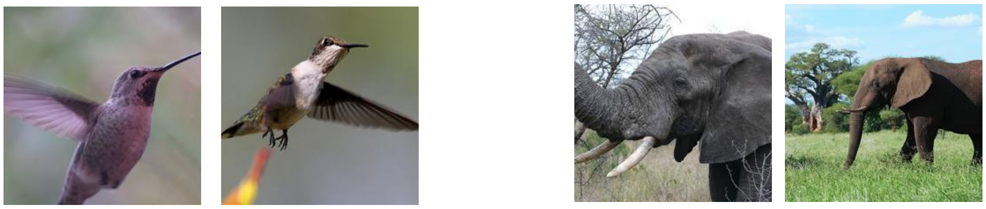
Visual Understanding Demonstration. We investigate several vision perception and reasoning capabilities of our SAIL. These include its ability to understand rich OCR information (Table 12), interpret real-world scenes (Table 13), comprehend scientific charts (Table 14), and analyze poster contents (Table 15).

Method	Pretrain	SFT	VQAv2	GQA	SciQA-IMG	TextVQA	POPE	MMBench	SEEDBench	DocVQA	ChartQA	AI2D	MMStar	avg
LLaVA-1.5-336px [46]	*12.8B*+558K	665K	78.5	62.0	66.8	58.2	85.9	64.3	66.1	28.1	18.2	54.8	32.4	58.3
SAIL	512M	665K	77.8	61.6	68.0	56.4	86.6	61.3	69.8	29.3	21.5	58.7	37.1	59.1

Table 11. **Comparison of SAIL and LLaVA1.5.** We evaluate the models on VQAv2 [27], GQA [35], ScienceQA [51], TextVQA [64], POPE [43], MMBench [49], SEEDBench [39], DocVQA [55], ChartQA [54], AI2D [38] and MMStar [9].

(A) Are there patterns on the easter eggs?
GT: Yes; No
SAIL: Yes; No
LLaVA1.5: Yes; Yes

(B) Are there any words displayed on the vehicle's lightbar?
GT: Yes; No
SAIL: Yes; No
LLaVA1.5: Yes; Yes



(C) Are the birds flapping upward or downward?
GT: Upward; Downward
SAIL: Upward; Downward
LLaVA1.5: Upward; Upward

(D) Is the elephant's trunk raised or lowered?
GT: Raised; Lowered
SAIL: Raised; Lowered
LLaVA1.5: Lowered; Lowered

Figure 6. Comparison of SAIL and LLaVA-1.5 on MMVP examples. SAIL demonstrates better performance in perceiving minor regions and objects, as well as more accurately distinguishing object states.

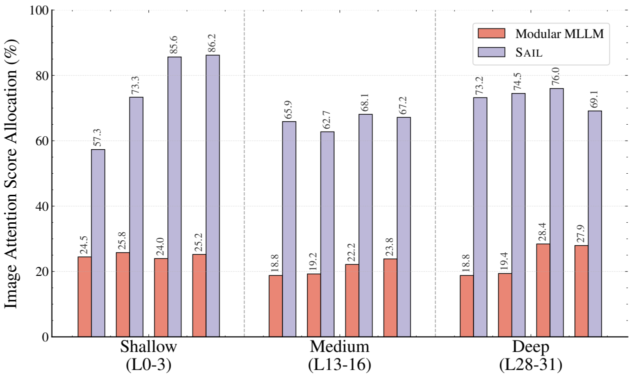
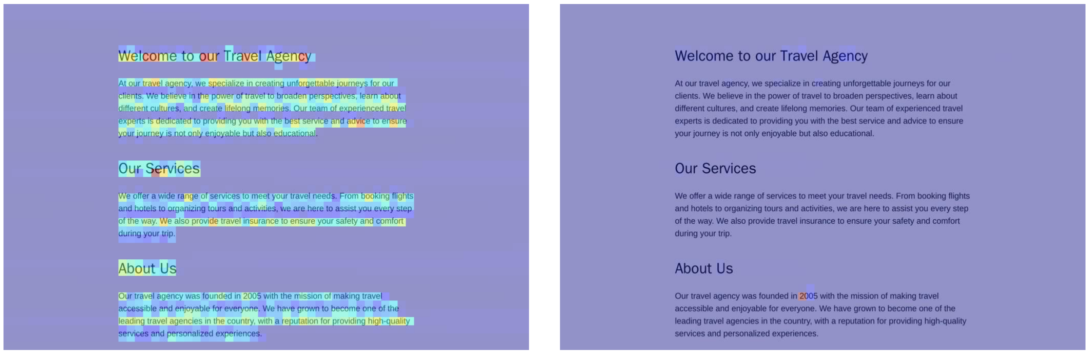


Figure 7. **Image attention score allocation for SAIL and its modular MLLM counterpart.** We compared the attention score allocation distribution for shallow layers, medium layers, and deep layers between these two models. The Single Transformer-based MLLM model significantly allocates a higher proportion of attention score to image tokens during prediction than the modular MLLM.

Query: When was the travel agency founded?



Welcome to our Travel Agency

At our travel agency, we specialize in creating unforgettable journeys for our clients. We believe in the power of travel to broaden perspectives, learn about different cultures, and create lifelong memories. Our team of experienced travel experts is dedicated to providing you with the best service and advice to ensure your journey is not only enjoyable but also educational.

Our Services

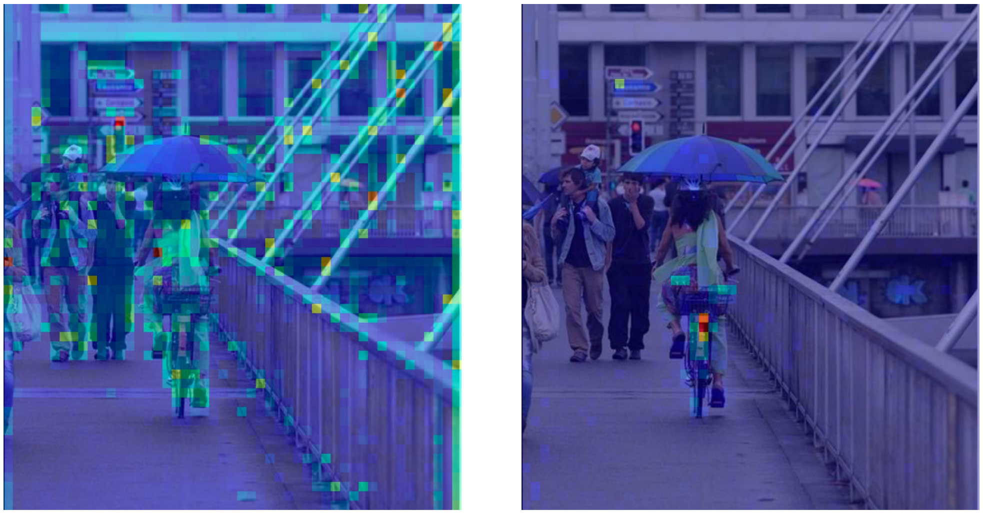
We offer a wide range of services to meet your travel needs. From booking flights and hotels to organizing tours and activities, we are here to assist you every step of the way. We also provide travel insurance to ensure your safety and comfort during your trip.

About Us

Our travel agency was founded in 2005 with the mission of making travel accessible and enjoyable for everyone. We have grown to become one of the leading travel agencies in the country, with a reputation for providing high-quality services and personalized experiences.

Output token: “2”, Layer 2 Output token: “2”, Layer 25

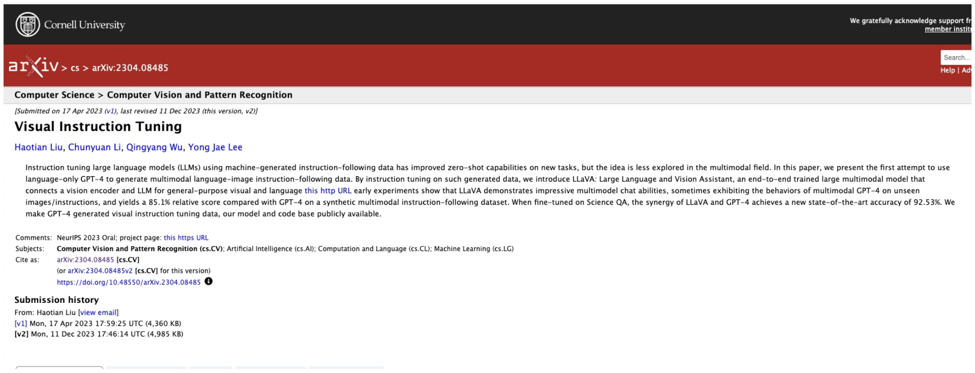
Query: What color is the Bicycle?



Output token: “red”, Layer 0 Output token: “red”, Layer 31

Figure 8. **Visualization of SAIL’s attention distribution across image regions during token prediction.** In early transformer layers, attention primarily focuses on the salient regions of the image. As the model progresses to deeper layers, attention shifts to areas more relevant to the predicted tokens.

Example 1: Understanding OCR Information in Webpage.



Cornell University

We gratefully acknowledge support fr
member instit

arXiv > cs > arXiv:2304.08485

Search...
Help | Ad

Computer Science > Computer Vision and Pattern Recognition

[Submitted on 17 Apr 2023 (v1), last revised 11 Dec 2023 (this version, v2)]

Visual Instruction Tuning

Haotian Liu, Chunyuan Li, Qingyang Wu, Yong Jae Lee

Instruction tuning large language models (LLMs) using machine-generated instruction-following data has improved zero-shot capabilities on new tasks, but the idea is less explored in the multimodal field. In this paper, we present the first attempt to use language-only GPT-4 to generate multimodal language-image instruction-following data. By instruction tuning on such generated data, we introduce LLaVA: Large Language and Vision Assistant, an end-to-end trained large multimodal model that connects a vision encoder and LLM for general-purpose visual and language this http URL early experiments show that LLaVA demonstrates impressive multimodel chat abilities, sometimes exhibiting the behaviors of multimodal GPT-4 on unseen images/instructions, and yields a 85.1% relative score compared with GPT-4 on a synthetic multimodal instruction-following dataset. When fine-tuned on Science QA, the synergy of LLaVA and GPT-4 achieves a new state-of-the-art accuracy of 92.53%. We make GPT-4 generated visual instruction tuning data, our model and code base publicly available.

Comments:	NeurIPS 2023 Oral; project page: this https URL
Subjects:	**Computer Vision and Pattern Recognition (cs.CV)**; Artificial Intelligence (cs.AI); Computation and Language (cs.CL); Machine Learning (cs.LG)
Cite as:	arXiv:2304.08485 **[cs.CV]** (or arXiv:2304.08485v2 **[cs.CV]** for this version) https://doi.org/10.48550/arXiv.2304.08485

Submission history

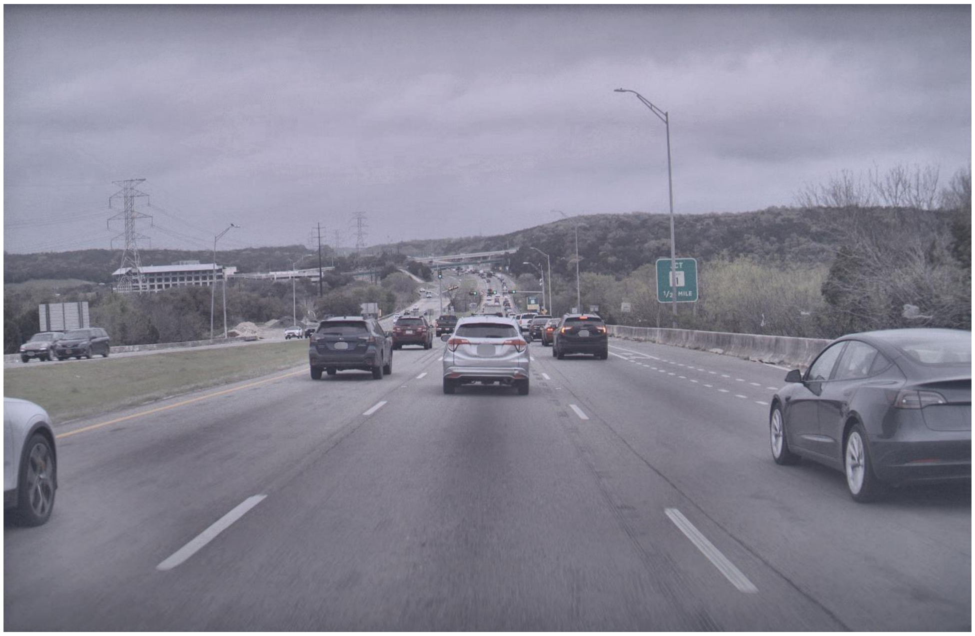
From: Haotian Liu [view email]
[v1] Mon, 17 Apr 2023 17:59:25 UTC (4,360 KB)
[v2] Mon, 11 Dec 2023 17:46:14 UTC (4,985 KB)

User Please output all the text information.

SAIL Cornell University arXiv cs arXiv:2304.08485 Computer Science Computer Vision and Pattern Recognition
Visual Instruction Tuning HaoTian Liu, Chunyuan Li, Qingyang Wu, Yong Jae Lee Instruction tuning large language models (LLMs) using machine-generated instruction-following data has improved zero-shot capabilities on new tasks, but the idea is less explored in the multimodal field. In this paper, we present the first attempt to use language-only GPT-4 to generate multimodal language-image instruction-following data. By instruction tuning on such generated data, we introduce LLM for general-purpose visual and language visual and language this http URL URL Early Experiments: LLMs demonstrate impressive multimodel chat abilities, sometimes exhibiting the behaviors of multimodal GPT-4 on a synthetic multimodal instruction-following dataset. When fine-tuned on Science QA, the synergy of LLAVA and GPT-4 achieves a new state-of-the-art accuracy of 92.53%. We make GPT-4 generated visual instruction tuning data, our model and code base publicly available. Comments: NeurIPS 2023 Oral; project page: this https URL Subjects: Computer Vision and Pattern Recognition (cs.CV); Artificial Intelligence (cs.AI); Computation and Language (cs.CL); Machine Learning (cs.LG) Cite as: arXiv:2304.08485 [cs.CV] (or arXiv:2304.08485v2 [cs.CV] for this version) https://doi.org/10.48550/arXiv.2304.08485 Submission history From: Haotian Liu [view email] [v1] Mon, 17 Apr 2023 17:59:25 UTC (4,360 KB) [v2] Mon, 11 Dec 2023 17:46:14 UTC (4,985 KB)

Table 12. SAIL demonstrates a strong capability to extract information in OCR-rich scenarios.

Example 2: Understanding Real-World Scene.

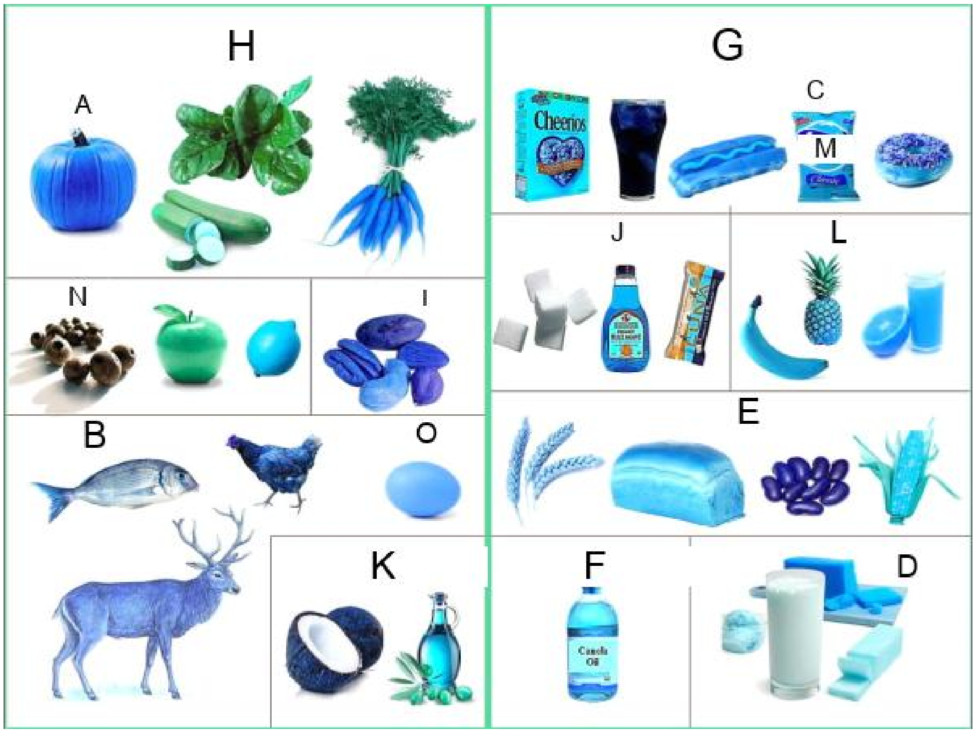


User Please describe the image in detail.

SAIL There are several notable elements in the image:
1. The highway is multi-lane, with multiple lanes visible. 2. The traffic is dense, with vehicles in various lanes. 3. The weather seems to be overcast, as indicated by the grayish tones of the image. 4. There are several traffic signs and signs visible, including a green highway sign indicating “1/2 mile” to the right, and a green sign with the number “16” and the word “MILE” on the left side of the image. 5. The background includes a hill or mountain, with trees and greenery visible. 6. There are also several power lines and electrical pylons visible along the sides of the highway. The image’s focus is on the traffic and the surrounding environment. The vehicles are primarily in the center and right lanes of the highway, with some vehicles in the left lanes as well. The highway is flanked by lush greenery, including trees and grassy areas, suggesting that this is a rural or possibly a suburban area. The traffic appears to be moving at a moderate pace. The weather conditions suggest that this is a daytime scene, possibly during the morning or late afternoon. The image is taken from a perspective that suggests the viewer is in a vehicle on the highway, observing the traffic ahead. In summary, the image captures a typical highway scene with moderate traffic, surrounded by a mix of natural and built environments, under an overcast sky. The road is in good condition, and the vehicles are traveling smoothly in both directions.

Table 13. SAIL demonstrates a strong capability about understanding the real-world scene.

Example 3: Understanding the Scientific Charts.

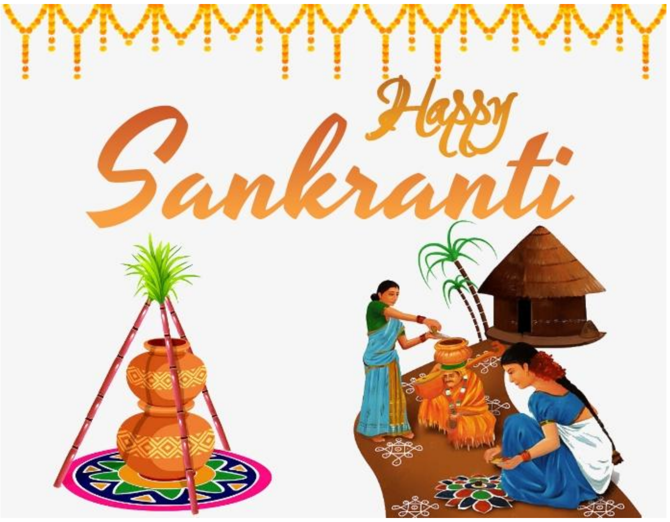


User Please describe the image in detail.

SAIL This image is a collage of various objects, foods, and animals, all of which are predominantly blue or have blue elements. The image is divided into sections labeled with letters from A to O. Here is a detailed description of each section: 1. A - Pumpkin: A blue pumpkin is depicted, which is a common symbol associated with autumn and Halloween. 2. B - Fish: A blue fish is shown, which could be a representation of a common aquatic animal. 3. C - M&M: A blue M&M candy is present.4. D - Milk and Soap: A glass of milk, a bar of soap, and a blue sponge.5. E - Bread and Jelly Beans: Blue wheat, a blue loaf of bread, and blue jelly beans.6. F - Castor Oil: A bottle of castor oil.7. G - Cheerios and Snacks: A box of Cheerios, a glass of a dark blue drink, a blue popsicle, a blue donut, and a blue muffin. 8. H - Vegetables: A collection of vegetables including a blue cucumber, green leafy vegetables, and a bunch of blue carrots.9. I - Apple and Stones: A green apple, a blue candy, and blue stones.10. J - Syrup and Candy: Sugar cubes, a bottle of blue syrup, and a blue candy bar.11. K - Coconut and Aloe: A blue coconut, a bottle of blue liquid, and a blue aloe vera plant.12. L - Pineapple and Juice: A blue pineapple, a blue banana, and a glass of blue juice.13. M - Candy Bar: A blue candy bar.14. N - Blueberries: Blueberries.15. O - Oval Object: A blue oval-shaped object.

Table 14. SAIL demonstrates a strong capability about understanding the scientific charts.

Example 4: Understanding Poster Content.



User Please describe the image in detail.

SAIL At the top of the image, there is a decorative border with a pattern of hanging golden ornaments, which adds a festive and celebratory feel to the overall design. The background is plain white, which helps the colorful elements stand out.

In the center of the image, the text "Happy Sankranti" is prominently displayed in a large, elegant, and stylized font. The text is in a warm orange color, which is traditionally associated with the festival.

Below the text, there is a depiction of a woman dressed in traditional Indian attire, likely a saree, in shades of blue and green. She is engaged in the act of preparing a traditional Sankranti sweet, which is a common practice during the festival. The sweet is shaped like a pyramid and is decorated with intricate patterns, reflecting the artistic and cultural heritage of the region.

To the left of the woman, there is a three-tiered pot, known as a "patala", which is a traditional Indian cooking vessel. The pot is adorned with a colorful design and is placed on a colorful circular mat, which is also a traditional element in Indian households.

On the right side of the woman, there is a small hut with a thatched roof, which is a common architectural style in rural Indian villages. The hut is depicted in a simple and rustic manner, emphasizing the rural and traditional aspect of the scene.

Overall, the image captures the essence of the Sankranti festival, highlighting the cultural and religious aspects of the celebration. The use of traditional clothing, the preparation of traditional sweets, and the depiction of a rural village scene all contribute to a sense of cultural authenticity and celebration.

Table 15. SAIL demonstrates a strong capability about understanding the poster content.

Model	VQAv2	GQA	SciQA-IMG	TextVQA	POPE	MME	MMBench	SEEDBench-IMG	Norm.AVG
Figure 1, modular MLLM, 32M	76.96	58.7	68.48	58.68	88.17	1599	69.44	70.31	61.41
Figure 1, modular MLLM, 128M	78.47	59.78	70.05	59.82	86.78	1638	68.57	68.11	61.52
Figure 1, modular MLLM, 512M	80.06	62.38	70.34	57.85	83.14	1379	70.82	69.83	61.86
Figure 1 & 5, SAIL, 32M	70.51	57.95	63.32	31.67	81.77	1421	48.22	61.51	51.93
Figure 1 & 5, SAIL, 128M	76.36	60.93	62.61	56.86	85.5	1458	53.94	66.60	57.91
Figure 1 & 5, SAIL, 512M	78.51	62.06	67.48	63.94	86.04	1530	56.71	68.83	60.51
Figure 3, SAIL-3B	67.3	53.2	63.8	30.9	66.9	820.8	44.6	55.4	47.80
Figure 3, SAIL-0.5B	59.1	46.9	59.6	20.1	59.8	761.45	38.5	35.1	39.92
Figure 5, modular MLLM, 32M	70.60	58.03	65.83	29.63	82.68	1486	50.02	62.12	52.43
Figure 5, modular MLLM, 128M	74.52	61.22	66.24	48.27	85.20	1489	52.76	64.98	56.72
Figure 5, modular MLLM, 512M	76.20	63.04	67.23	56.87	85.50	1513	54.01	66.20	58.70

Table 16. Detailed experimental results in the paper.

Exp	Model	LLM	Stage 1		Stage 2		SFT	
			Data	LR	Data	LR	Data	LR
Figure 1(A)	SAIL, point 32M	Mistral-7B-v0.1	Standard Stage 1 Data (32M image-text pairs)	(5e-5, 5e-6)	-	-	LLaVA-mix-665K	(1e-5,0)
Figure 1(A)	SAIL, point 128M	Mistral-7B-v0.1	Standard Stage 1 Data (128M image-text pairs)	(5e-5, 5e-6)	-	-	LLaVA-mix-665K	(1e-5,0)
Figure 1(A), Table 6	SAIL, point512M	Mistral-7B-v0.1	Standard Stage 1 Data (512M image-text pairs)	(5e-5, 5e-6)	-	-	LLaVA-mix-665K	(1e-5,0)
Figure 1(B), Table 2	SAIL	Mistral-7B-v0.1	Standard Stage 1 Data (512M image-text pairs)	(5e-5, 5e-6)	Standard Stage 2 Data	(1e-5, 5e-6)	Standard SFT Data	(1e-5, 0)
Table 3, 4, 5	SAIL	Mistral-7B-v0.1	Standard Stage 1 Data (512M image-text pairs)	(5e-5, 5e-6)	-	-	-	-
Figure 3, Table 7	SAIL-0.5B	Qwen2.5-0.5B	Standard Stage 1 Data (128M image-text pairs)	(5e-4, 5e-6)	-	-	LLaVA-mix-665K	(1e-5, 0)
Figure 3	SAIL-3B	Qwen2.5-3B	Standard Stage 1 Data (128M image-text pairs)	(1e-4, 5e-6)	-	-	LLaVA-mix-665K	(1e-5, 0)
Figure 3	SAIL-7B	Mistral-7B-v0.1	Standard Stage 1 Data (128M image-text pairs)	(5e-5, 5e-6)	-	-	LLaVA-mix-665K	(1e-5,0)

Table 17. **Experimental Configurations for Various Setups.** The table lists the models used, the specific LLM variants, the datasets, and learning rates (LR) applied during each training stage (Pretraining Stage 1, Pretraining Stage 2, and SFT). “Standard Stage 1 Data”, “Standard Stage 2 Data” and “Standard SFT Data” are listed in Table 1. Specific points and tables/figures referred to in the text are also indicated.