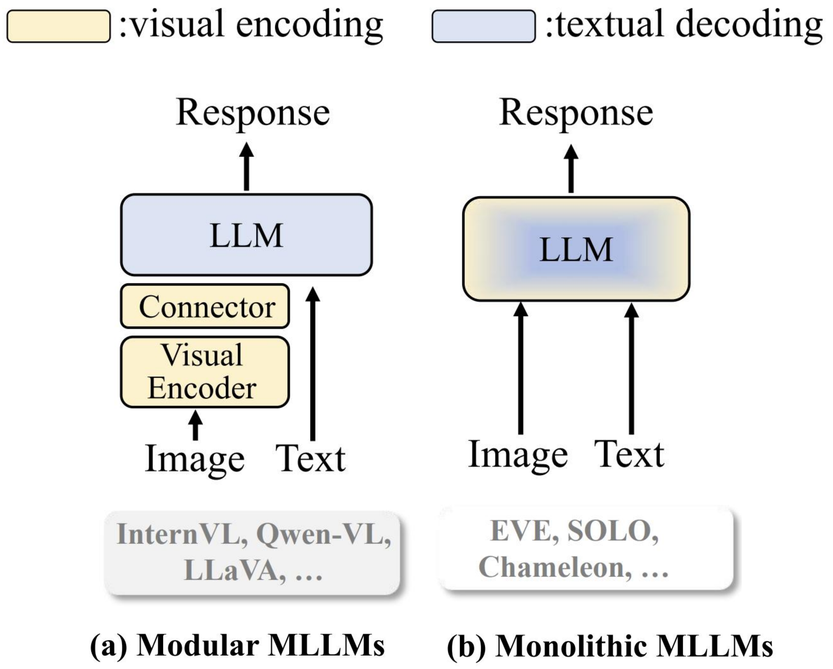
Mono-InternVL: Pushing the Boundaries of Monolithic Multimodal Large Language Models with Endogenous Visual Pre-training

Gen Luo[1]*, Xue Yang[1]*, Wenhan Dou[2]*, **Zhaokai Wang**[3,1]*,
Jiawen Liu[4], Jifeng Dai[2,1], Yu Qiao[1], Xizhou Zhu[2,1]

[1]OpenGVLab, Shanghai AI Laboratory [2]Tsinghua University
[3]Shanghai Jiao Tong University [4]Johns Hopkins University

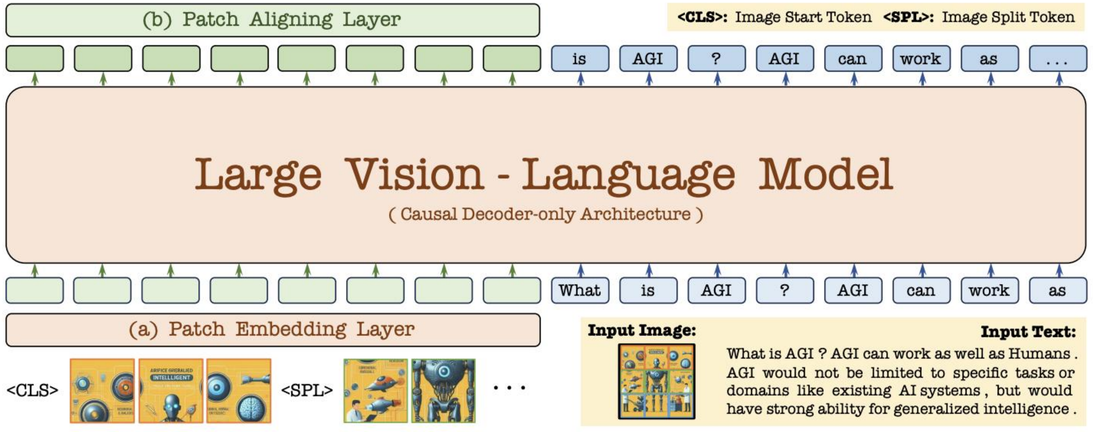
Motivation

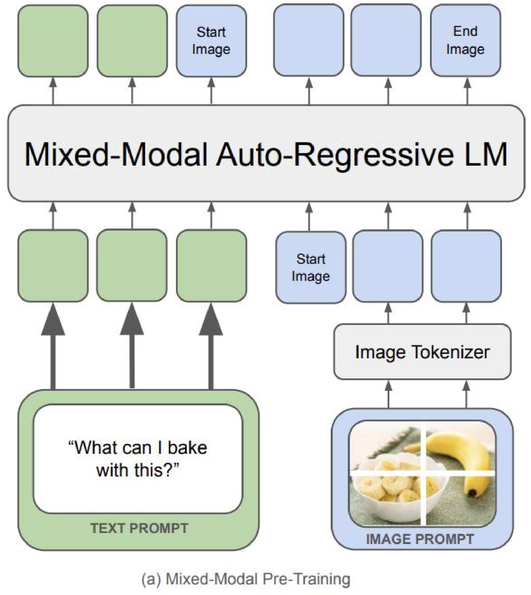
- Monolithic (i.e., **w/o vision encoder**) MLLMs show potential in design simplicity and deployment efficiency
- Weaker visual perception capabilities compared with modular MLLMs



Motivation

- Native pretraining (e.g. Chameleon): Train from scratch
  - High training costs; Unstable optimization
- Continuous pretraining (e.g. EVE): Extend pretrained LLMs
  - Catastrophic forgetting of pretrained language knowledge
- Our analysis: due to **shared parameter architecture** for vision and text
  - Optimization for vision can negatively impact language capabilities





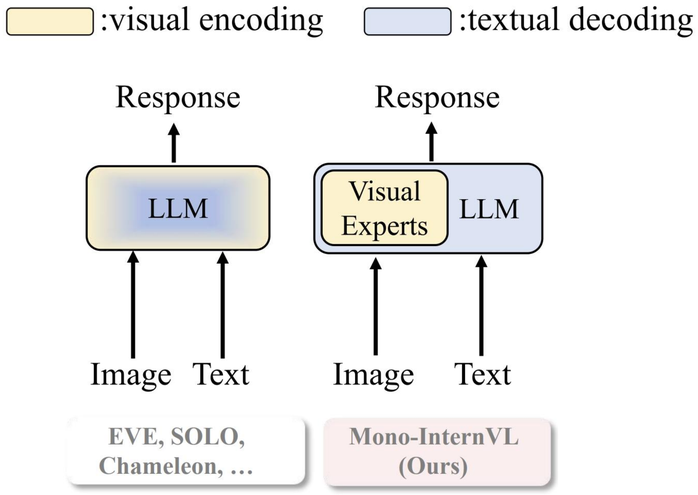
(a) Mixed-Modal Pre-Training

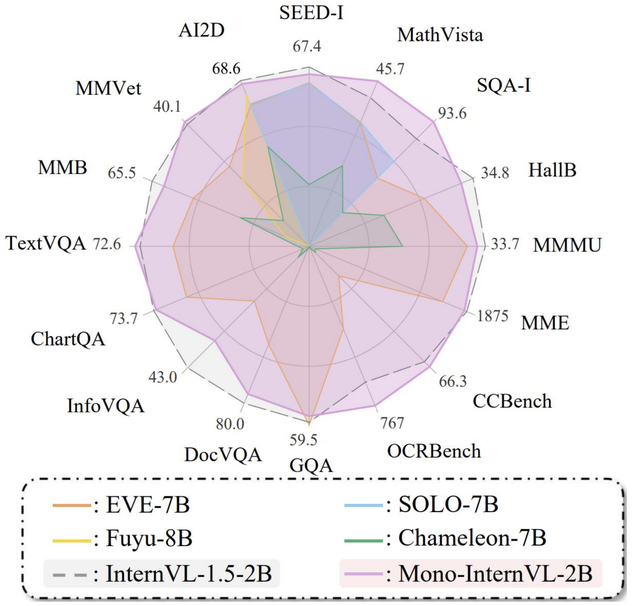
[1] Chameleon: Mixed-Modal Early-Fusion Foundation Models

[2] Unveiling Encoder-Free Vision-Language Models

Method - Overview

- **Mono-InternVL:** retains language knowledge of pretrained LLM by introducing **visual experts** through a MoE design, and applies **delta tuning** (training partial parameters)
- Endogenous visual pretraining (**EViP**) strategy for progressive coarse-to-fine pretraining
- Surpasses other monolithic MLLMs (e.g. +2.5% over Emu3-8B); Comparable with leading modular MLLMs with only **2B** parameters

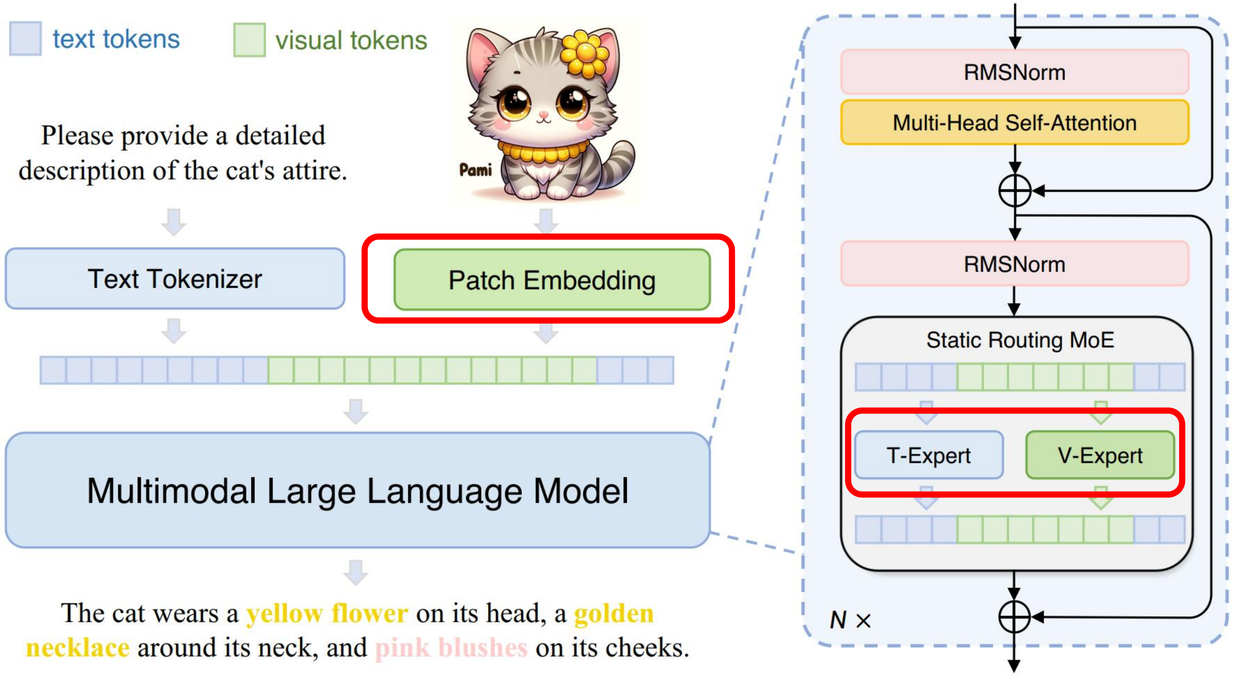




Method - Monolithic Architecture

(1) **Visual and textual embeddings:** Directly patchifies images using a lightweight module

(2) **Multimodal mixture-of-experts (MoE) structure:** Embed visual experts into a pre-trained LLM to utilize pretrained language knowledge and mitigate the catastrophic forgetting issue

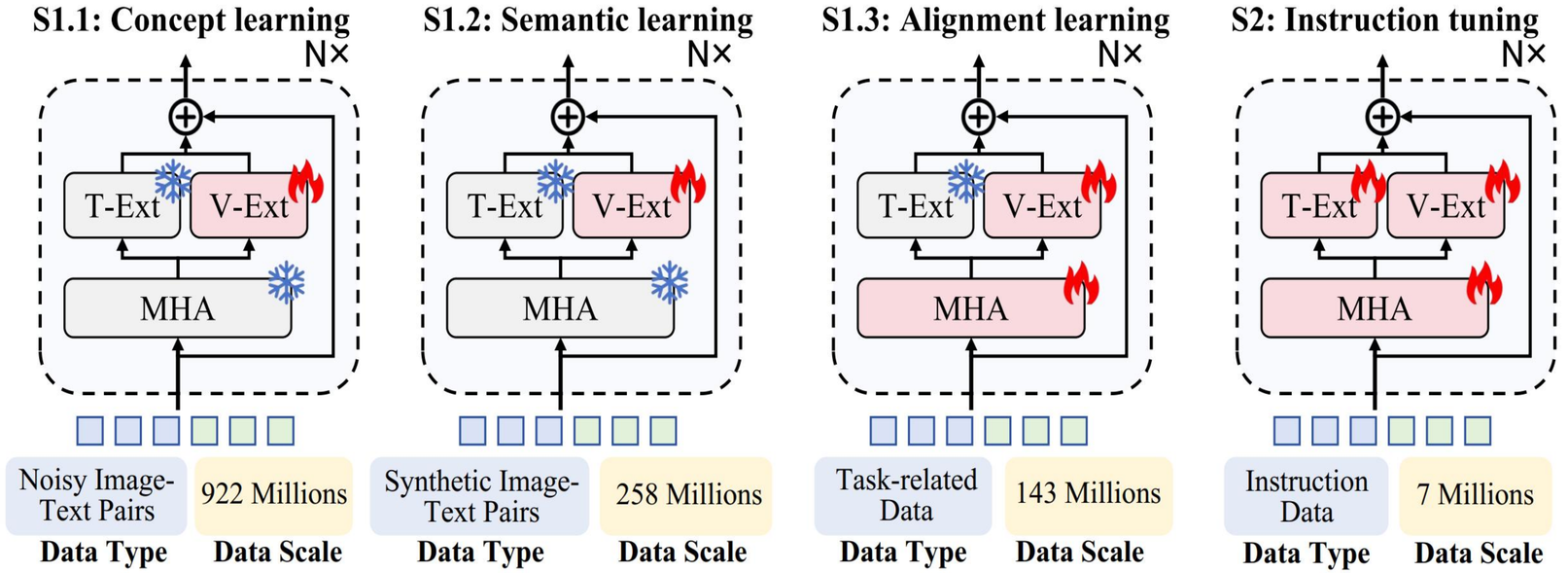


Method - Endogenous Visual Pretraining (EViP)

S1.1: **Concept learning** to grasp basic visual concepts.

S1.2: **Semantic learning** to capture high-level semantics.

S1.3: **Alignment learning** to align knowledge with downstream tasks.



Experiments

Model	#A-Param	MMB	MMVet	MMMU	MME	MathVista	SEED-I	OCRBench	HallB	CCB	Avg_{MM}	TextVQA	SQA-I	GQA	DocVQA	AI2D	ChartQA	InfoVQA	Avg_{QA}
▼ *Modular MLLMs:*																			
MobileVLM-V2-3B [20]	3.0B	63.2	–	–	–	–	–	–	–	–	–	57.5	70.0	66.1	–	–	–	–	–
Mini-Gemini-2B [52]	3.5B	59.8	31.1	31.7	1653	29.4	–	–	–	–	–	56.2	–	–	34.2	–	–	–	–
MM1-3B-MoE-Chat [74]	3.5B	70.8	42.2	38.6	1772	32.6	69.4	–	–	–	–	72.9	76.1	–	–	–	–	–	–
DeepSeek-VL-1.3B [65]	2.0B	64.6	34.8	32.2	1532	31.1	66.7	409	27.6	37.6	43.4	57.8	–	–	–	51.5	–	–	–
PaliGemma-3B [7]	2.9B	71.0	33.1	34.9	1686	28.7	69.6	614	32.2	29.6	46.7	68.1	–	–	–	68.3	–	–	–
MiniCPM-V-2 [107]	2.8B	69.1	41.0	38.2	1809	38.7	67.1	605	36.1	45.3	51.2	74.1	–	–	71.9	62.9	–	–	–
†InternVL-1.5-2B [18]	2.2B	70.9	39.3	34.6	1902	41.1	69.8	654	37.5	63.5	54.4	70.5	84.9	61.6	85.0	69.8	74.8	55.4	71.7
Qwen2VL-2B [101]	2.1B	74.9	49.5	41.1	1872	43.0	-	809	41.7	-	-	79.7	-	-	90.1	74.7	73.5	65.5	-
▼ *Monolithic MLLMs:*																			
Fuyu-8B (HD) [6]	8B	10.7	21.4	–	–	–	–	–	–	–	–	–	–	–	–	64.5	–	–	–
SOLO [16]	7B	–	–	–	1260	34.4	64.4	–	–	–	–	–	73.3	–	–	61.4	–	–	–
Chameleon-7B[1] [12]	7B	31.1	8.3	25.4	170	22.3	30.6	7	17.1	3.5	16.1	4.8	47.2	–	1.5	46.0	2.9	5.0	17.9
EVE-7B [26]	7B	49.5	25.6	32.3	1483	25.2	61.3	327	21.1	12.4	34.8	51.9	63.0	60.8	22.0	48.5	19.5	20.0	40.8
EVE-7B (HD) [26]	7B	52.3	25.7	32.6	1628	34.2	64.6	398	26.4	16.3	38.9	56.8	64.9	62.6	53.0	61.0	59.1	25.0	54.6
Emu3 [104]	8B	58.5	37.2	31.6	–	–	**68.2**	687	–	–	–	64.7	89.2	60.3	76.3	**70.0**	68.6	**43.8**	67.6
Mono-InternVL-2B	1.8B	**65.5**	**40.1**	**33.7**	**1875**	**45.7**	67.4	**767**	**34.8**	**66.3**	**55.2**	**72.6**	**93.6**	59.5	**80.0**	68.6	**73.7**	43.0	**70.1**

Table 2: **Comparison with existing MLLMs on general MLLM benchmarks and visual question answering benchmarks.** "#A-Param" denotes the number of activated parameters. For MME, we sum the perception and cognition scores. Avg_{MM} and Avg_{QA} denote the normalized average performance of MLLM benchmarks and VQA benchmarks, respectively. † InternVL-1.5-2B adopts the same LLM and high-quality data with Mono-InternVL-2B, so we mark it as the modular counterpart. **Bold** indicates the highest among monolithic MLLMs.

Experiments

Model	#A-Param	Data	Shots	COCO Caps	Flickr30k	NoCaps	VQAv2
Flamingo [1]	3B	>2.1B	0	73.0	–	–	49.2
MM1 [74]	3.5B	>2.3B	0	73.5	–	55.6	46.2
Chameleon [12]	34B	>1.4B	2	120.2	74.7	–	66.0
Mono-InternVL-S1.2	1.8B	0.9B	0	87.3	72.7	54.1	–
Mono-InternVL-S1.3	1.8B	1.1B	0	135.6	77.3	116.5	71.1

Table 3: **Zero-shot pre-training performance of Mono-InternVL and existing MLLMs.** "S1.2" and "S1.3" denote pre-training stages of semantic learning and alignment learning, respectively. Images of COCO have been seen in S1.3, so we mark its performance in gray.

Models	#A-Param	MMLU	CMMLU	AGIEval	MATH
InternLM2-Chat [10]	1.8B	47.1	46.1	38.8	13.9
EVE [26]	7B	43.9	33.4	22.6	0.7
Chameleon [12]	7B	52.1	-	-	11.5
Mono-InternVL	2B	45.1	44.0	40.9	12.3

Table 4: **Comparison of Mono-InternVL and existing monolithic MLLMs on four common NLP tasks.** Except for Chameleon, models are evaluated using OpenCompass toolkit [22].

Experiments

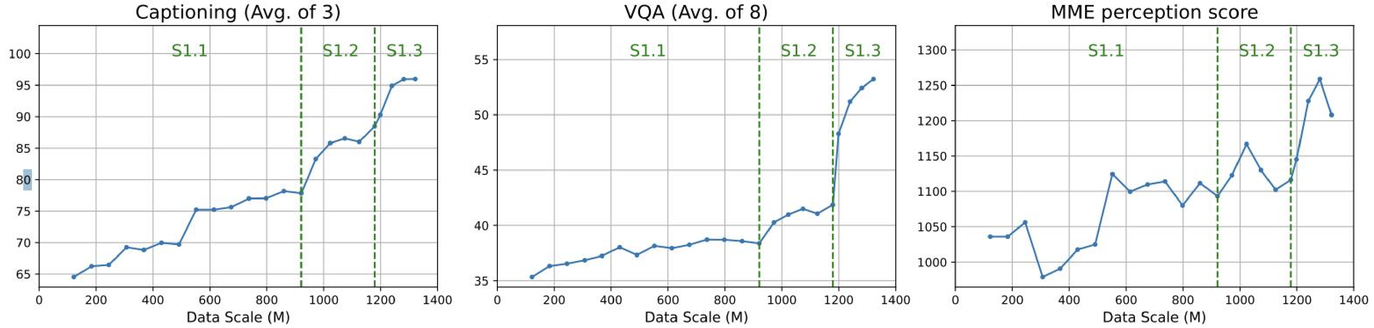


Figure 4: **Ablation studies of EViP with the increase of pre-training data size across three sub-stages: (S1.1) Concept learning; (S1.2) Semantic learning; (S1.3) Alignment learning.** For each data point, we fine-tune the corresponding pre-trained model on the instruction data of LLaVA-665*k* and obtain the downstream performance. Results of captioning and VQA are averaged from 3 and 8 tasks, respectively. See Appendix §A.3 for complete results.

Model	#T-Param	Strategy	MME-P	DocVQA	InfoVQA	SQA-I	GQA	ChartQA	AI2D
InternLM2	1.8B	Full	753	16.1	11.6	36.7	51.4	10.8	27.7
+ V-Expert	3.0B	Full	948	18.6	11.9	37.7	53.0	11.1	26.6
+ V-Expert	1.2B	Delta	995	18.9	14.6	56.5	53.4	13.5	42.7

Table 5: **Ablation of different strategies for visual pre-training.** All models are pre-trained on 61 million image-text pairs from Laion-2B [81] and fine-tuned on instruction data from LLaVA-665*k*. [59]. "Full" and "Delta" denote full tuning and delta tuning, respectively. "T-Param" refers to trainable parameters.

Experiments

Model	#Image Tokens	#Text Tokens	#Total Input Tokens	TTFT	TPS
InternVL-1.5-2B	768	256	1024	0.24	382
Mono-InternVL-2B	768	256	1024	0.09 (-63%)	436 (+14%)
InternVL-1.5-2B	1792	256	2048	0.45	183
Mono-InternVL-2B	1792	256	2048	0.15 (-67%)	232 (+27%)
InternVL-1.5-2B	3840	256	4096	1.93	52
Mono-InternVL-2B	3840	256	4096	0.79 (-59%)	68 (+31%)

Table 6: **Inference speed comparison of Mono-InternVL and InternVL-1.5.** Models are deployed on an NVIDIA A100 using LMDeploy with Pytorch backend [64], with a concurrency of 16 and the number of output tokens fixed as 120. "TTFT" and "TPS" denotes the time to first token in seconds and throughput in tokens per second, respectively.

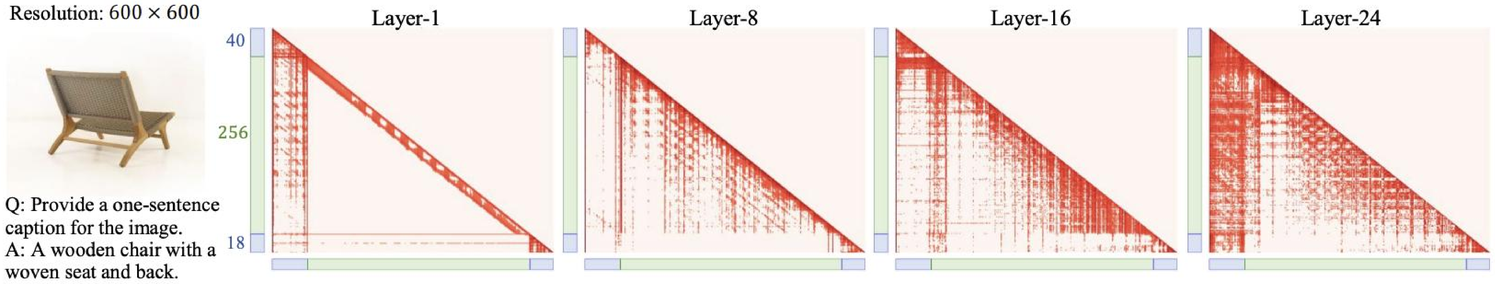


Figure 5: **Visualization of attention maps in Mono-InternVL.** The first blue segment, green segment and the second green segment in the axes represent the system prompt tokens (text), image tokens (visual) and user prompt tokens (text), respectively. The numbers on the left side of attention maps indicate the number of tokens.

Visualization

Prompt: Detect the <ref>the American Shorthair</ref> in the image with its bounding box.

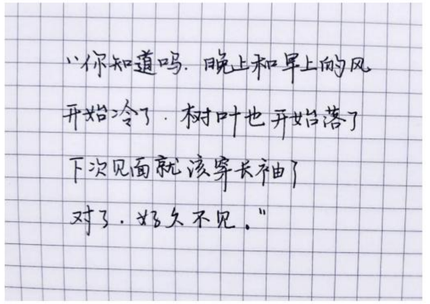
Response: <ref> the American Shorthair </ref> <box> [[229, 125, 660, 932]] </box>

Prompt: Please describe the detected object.

Response: The object in the image is a cat. The cat has gray fur with black stripes and is wearing a pink sweater. It appears to be sitting on a windowsill, looking up attentively towards the camera.

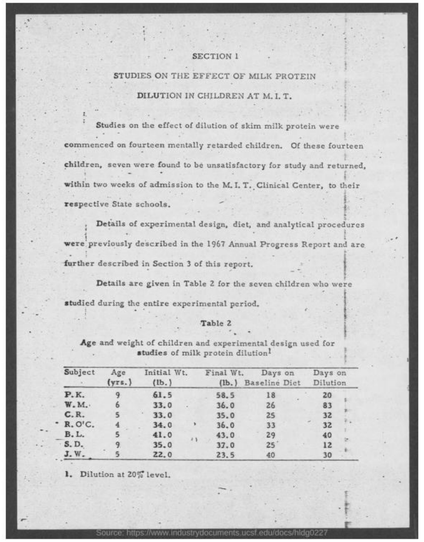
Prompt: Describe this image.

Response: This image depicts a bustling urban scene at night in an animation style, with a large electronic billboard prominently displaying the words "YUNIKA VISION" in bright white letters against a black background. The billboard is surrounded by several other tall buildings, including those of well-known companies such as HITACHI, TOSHIBA, SHARP, SONY, PANASONIC, and NEC.
In the foreground, there are multiple train tracks that intersect near what appears to be an elevated railway or subway system. Several trains are visible, indicating active transportation amidst the busy cityscape.
The skyline is dotted with various other commercial and residential structures, contributing to the sense of a thriving metropolis. Trees can also be seen around the area, providing some greenery amidst the urban environment. Overall, the image captures the dynamic energy of a modern city at night, with its blend of technology, commerce, and human activity.



Prompt: 识别图中文字。

Response: 图中文字如下：
“你知道吗。晚上和早上的风开始冷了，树叶也开始落了下次见面就该穿长袖了对了。好久不见。”



SECTION 1

STUDIES ON THE EFFECT OF MILK PROTEIN DILUTION IN CHILDREN AT M. I. T.

Studies on the effect of dilution of skim milk protein were commenced on fourteen mentally retarded children. Of these fourteen children, seven were found to be unsatisfactory for study and returned, within two weeks of admission to the M. I. T. Clinical Center, to their respective State schools.

Details of experimental design, diet, and analytical procedures were previously described in the 1967 Annual Progress Report and are further described in Section 3 of this report.

Details are given in Table 2 for the seven children who were studied during the entire experimental period.

Table 2

Age and weight of children and experimental design used for studies of milk protein dilution[1]

Subject	Age (yrs.)	Initial Wt. (lb.)	Final Wt. (lb.)	Days on Baseline Diet	Days on Dilution
P.K.	9	61.5	58.5	18	20
W.M.	6	33.0	36.0	26	83
C.R.	5	33.0	35.0	25	32
R.O'C.	4	34.0	36.0	33	32
B.L.	5	41.0	43.0	29	40
S.D.	9	35.0	37.0	25	12
J.W.	5	22.0	23.5	40	30

1. Dilution at 20% level.

Source: https://www.industrydocuments.ucsf.edu/docs/hldg0227

Prompt: What is the age of subject B.L.?

Response: B.L. is 5 years old at the time of study on the effect of milk protein dilution.

Prompt: Whose initial weight is the largest? What is it?

Response: The largest initial weight in the table is for P.K. with 61.5 pounds (lb.).

Prompt: How many children participated in the experiment in total?

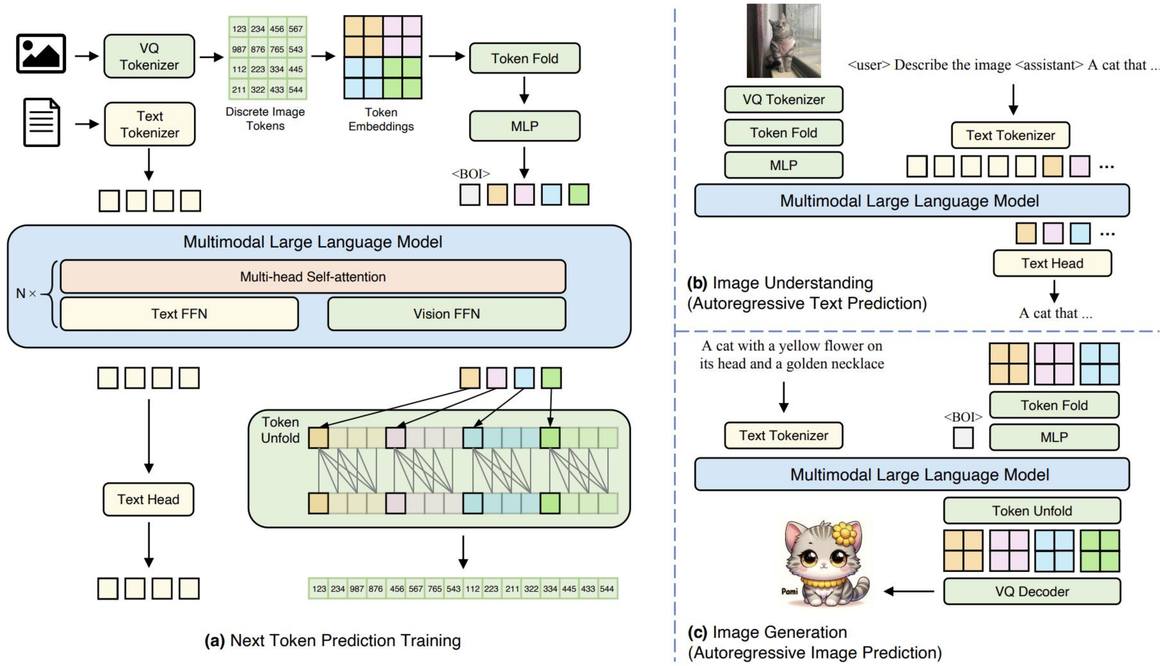
Response: There are a total of 14 children who participated in the experiment.

Takeaways

- Monolithic MLLMs are weak due to **shared parameters** for vision and text
- Introduces **vision experts** as a MoE structure and an **EViP** pretraining strategy
- Mono-InternVL surpasses existing monolithic MLLMs and close the gap with modular MLLMs
- Provides new directions for designing future MLLMs

Another Work - SynerGen-VL

- Another work of ours in CVPR 2025
- Extend the monolithic structure to **unified multimodal understanding and generation**
- Fully autoregressive, discrete tokenizer, visual experts, token folding&unfolding
- Comparable to Emu3-8B but with only 2B parameters



[1] SynerGen-VL: Towards Synergistic Image Understanding and Generation with Vision Experts and Token Folding

Thanks for Listening !

arXiv

Project Page

Code

Email: wangzhaokai@sjtu.edu.cn