IEEE Hong Kong 6G Wireless Summit
11-12 September 2024
Hong Kong, China

IEEE ComSoc IEEE Communications Society | IEEE

Department of Computer Science and Engineering, CUHK

Empowering Mobile Devices with Multimodal Large Language Models in 6G Era

Chengyao Wang, from Prof. Jiaya Jia’s Group

Motivation

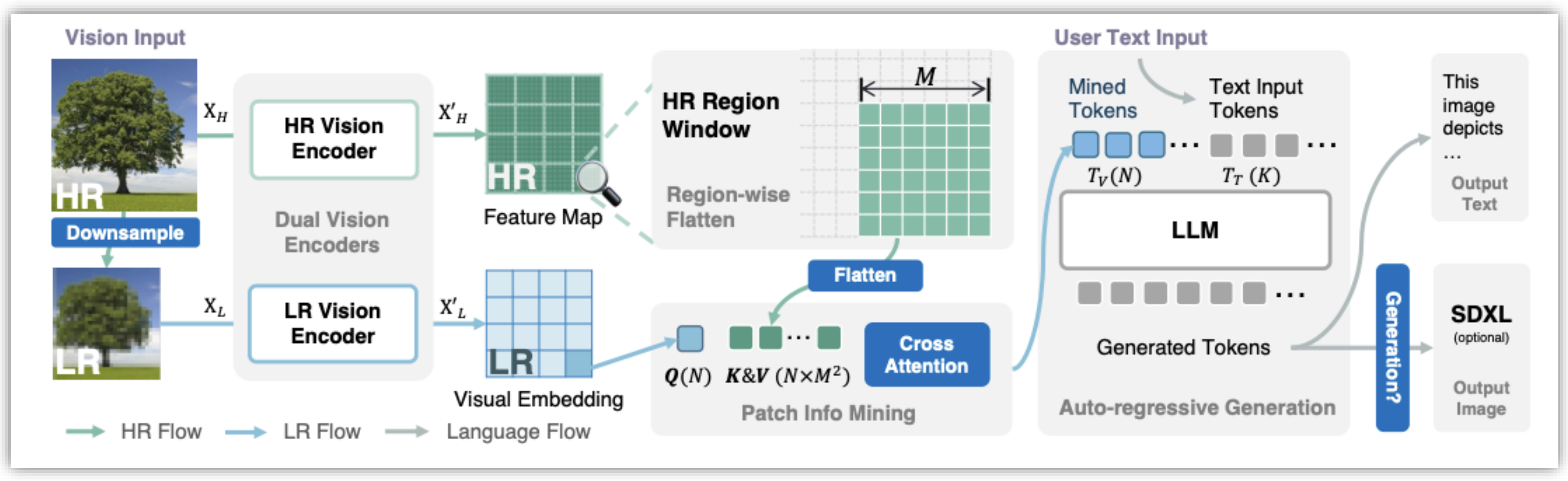
The rapid advancements in multimodal large language models offer significant potential for 6G wireless communication systems, where AI-driven applications will require real-time data processing, visual understanding, and reasoning capabilities. However, the computational demands of these models, especially for cloud-based inference, remain a challenge. To address this, we propose Mini-Gemini, an initiative focused on improving open-source models and optimizing them for efficient local inference on mobile devices and personal computers, reducing reliance on closed-source systems. This approach aligns with 6G’s vision of low-latency, distributed AI, enabling seamless, accessible, and efficient AI-powered applications across a wide range of devices in the 6G era.

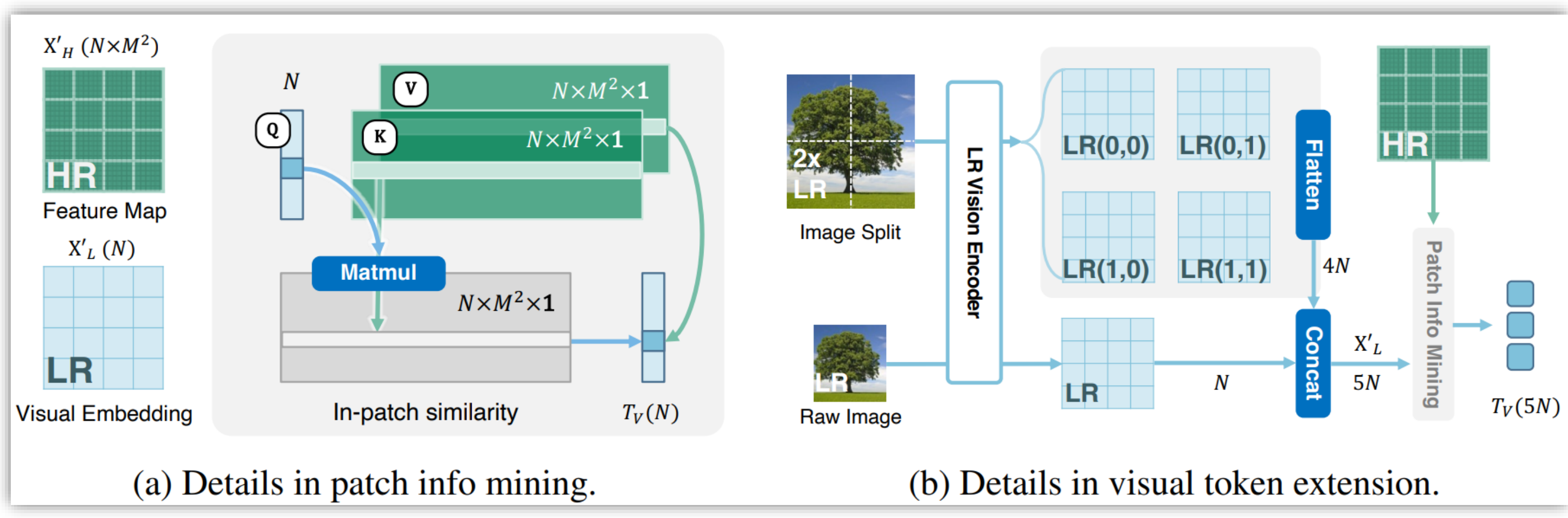
Challenge

1. There is a huge performance gap between close source and open source VLMs.
2. Large language model is hard to deployed on edge devices, due to huge inference cost.

Contribution

1. With data iteration and better utilization of the image data resolution, we propose token simplification to reduce computation.
2. Our approach attains leading performance in various settings and even surpasses many private models.
3. With instruction driven visual information mining approach, our model can efficiently deployed on laptop and mobile phone.





(a) Details in patch info mining.

(b) Details in visual token extension.

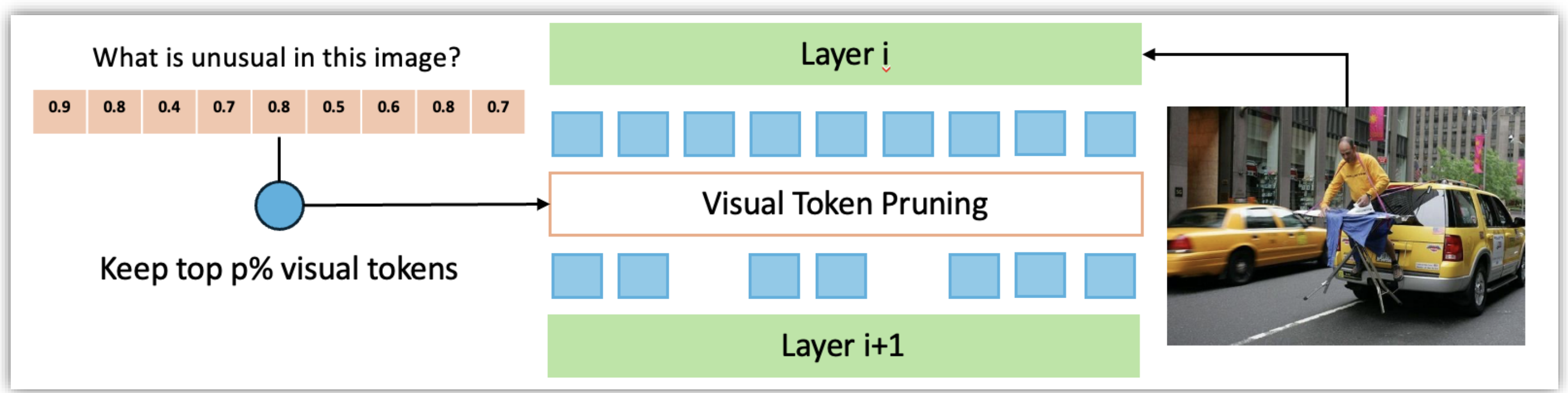
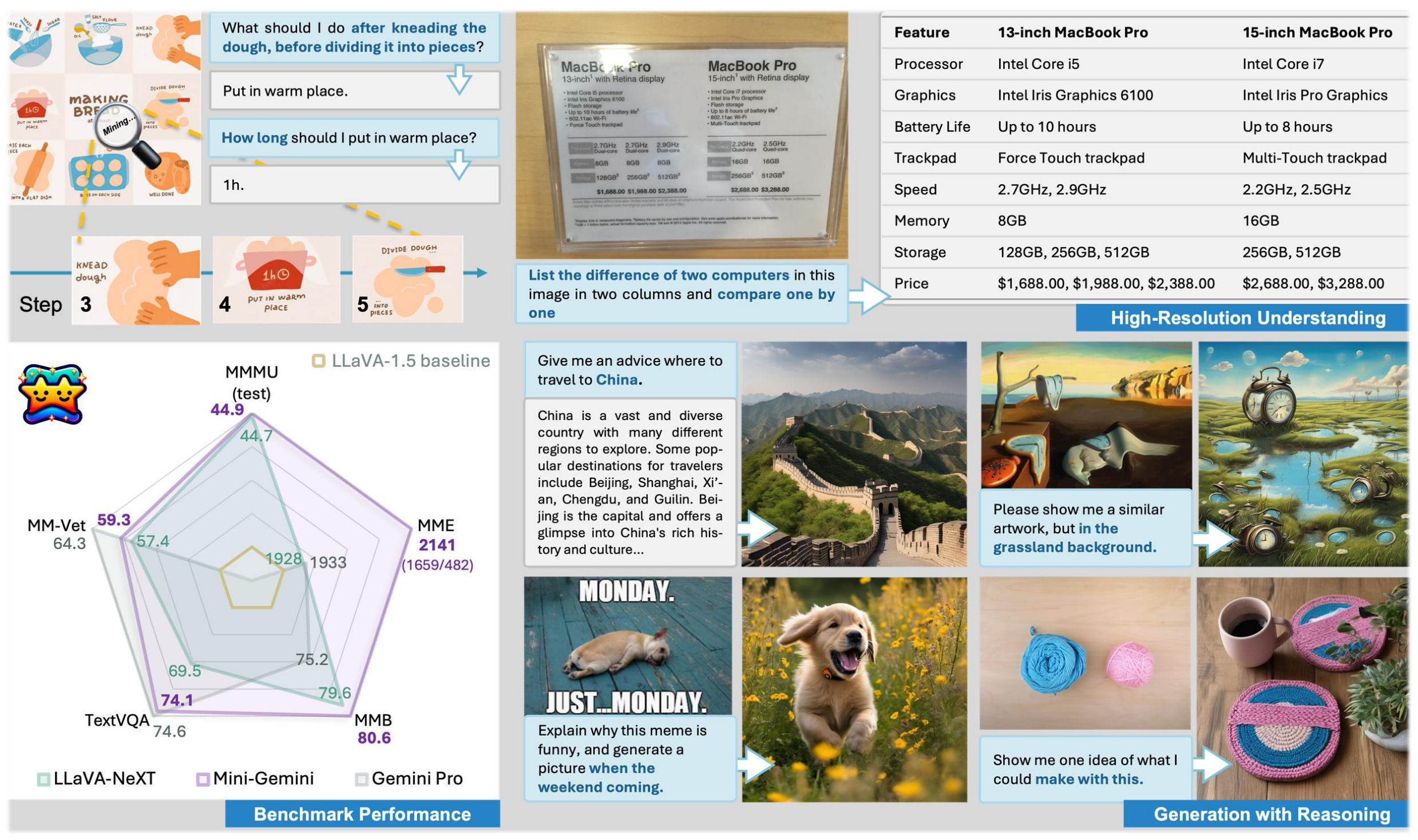


Image Understanding and Generation



What should I do after kneading the dough, before dividing it into pieces?

Put in warm place.

How long should I put in warm place?

1h.

List the difference of two computers in this image in two columns and compare one by one

Feature	13-inch MacBook Pro	15-inch MacBook Pro
Processor	Intel Core i5	Intel Core i7
Graphics	Intel Iris Graphics 6100	Intel Iris Pro Graphics
Battery Life	Up to 10 hours	Up to 8 hours
Trackpad	Force Touch trackpad	Multi-Touch trackpad
Speed	2.7GHz, 2.9GHz	2.2GHz, 2.5GHz
Memory	8GB	16GB
Storage	128GB, 256GB, 512GB	256GB, 512GB
Price	$1,688.00, $1,988.00, $2,388.00	$2,688.00, $3,288.00

High-Resolution Understanding

Benchmark Performance

Give me an advice where to travel to China.

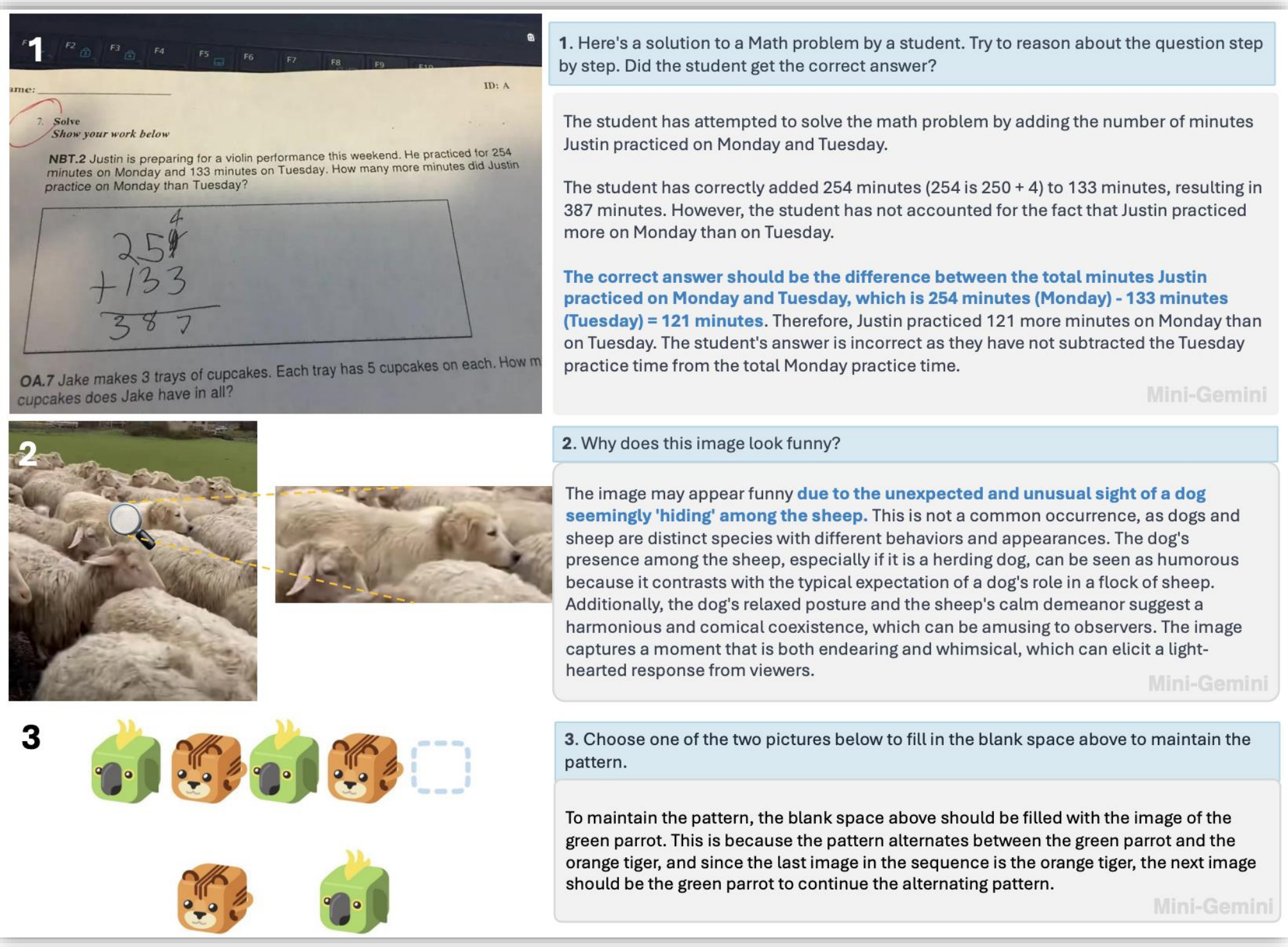
China is a vast and diverse country with many different regions to explore. Some popular destinations for travelers include Beijing, Shanghai, Xi’an, Chengdu, and Guilin. Beijing is the capital and offers a glimpse into China's rich history and culture...

Please show me a similar artwork, but in the grassland background.

Explain why this meme is funny, and generate a picture when the weekend coming.

Show me one idea of what I could make with this.

Generation with Reasoning



1. Here's a solution to a Math problem by a student. Try to reason about the question step by step. Did the student get the correct answer?

The student has attempted to solve the math problem by adding the number of minutes Justin practiced on Monday and Tuesday.

The student has correctly added 254 minutes (254 is 250 + 4) to 133 minutes, resulting in 387 minutes. However, the student has not accounted for the fact that Justin practiced more on Monday than on Tuesday.

The correct answer should be the difference between the total minutes Justin practiced on Monday and Tuesday, which is 254 minutes (Monday) - 133 minutes (Tuesday) = 121 minutes. Therefore, Justin practiced 121 more minutes on Monday than on Tuesday. The student's answer is incorrect as they have not subtracted the Tuesday practice time from the total Monday practice time.

Mini-Gemini

2. Why does this image look funny?

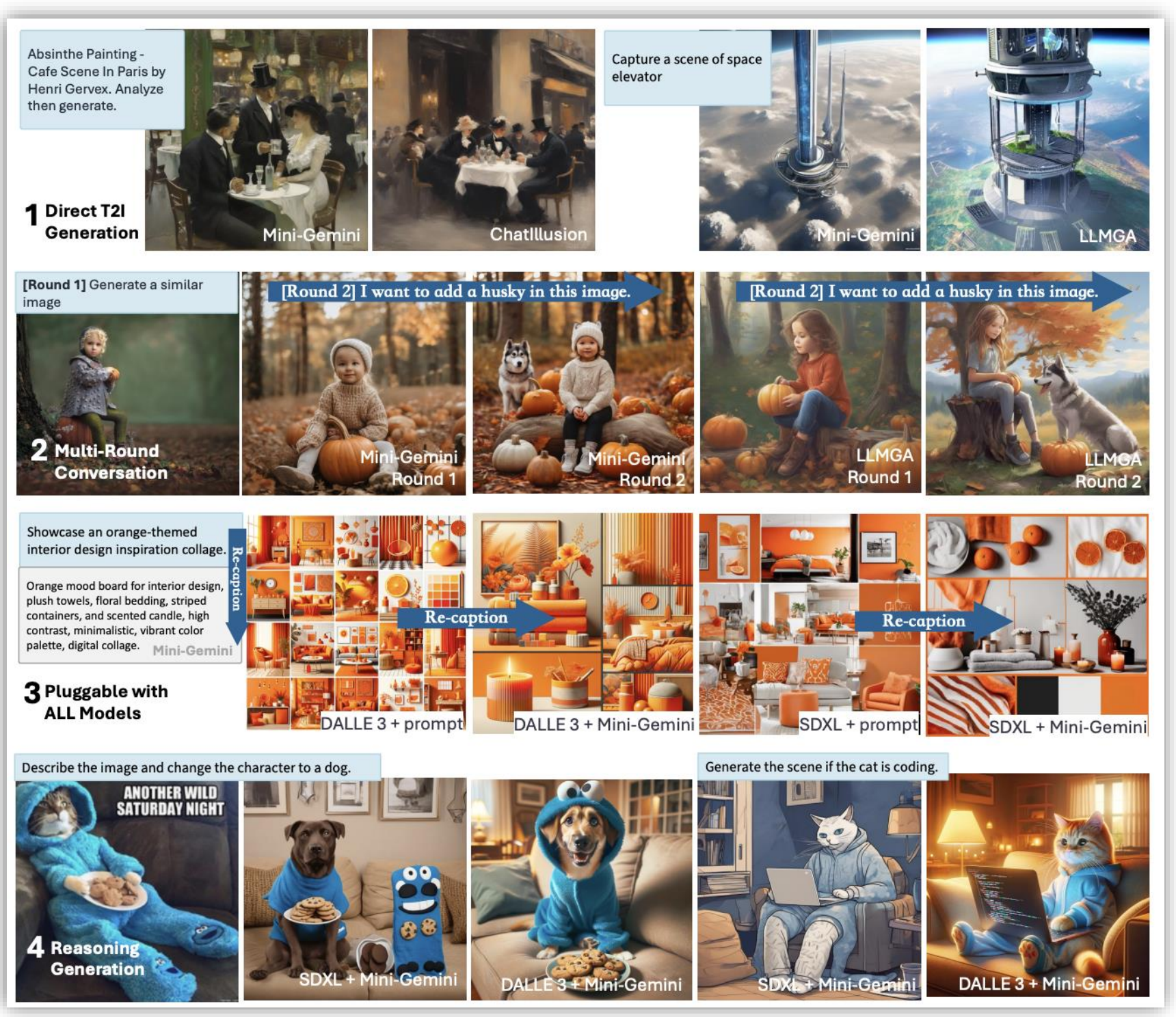
The image may appear funny due to the unexpected and unusual sight of a dog seemingly 'hiding' among the sheep. This is not a common occurrence, as dogs and sheep are distinct species with different behaviors and appearances. The dog's presence among the sheep, especially if it is a herding dog, can be seen as humorous because it contrasts with the typical expectation of a dog's role in a flock of sheep. Additionally, the dog's relaxed posture and the sheep's calm demeanor suggest a harmonious and comical coexistence, which can be amusing to observers. The image captures a moment that is both endearing and whimsical, which can elicit a light-hearted response from viewers.

Mini-Gemini

3. Choose one of the two pictures below to fill in the blank space above to maintain the pattern.

To maintain the pattern, the blank space above should be filled with the image of the green parrot. This is because the pattern alternates between the green parrot and the orange tiger, and since the last image in the sequence is the orange tiger, the next image should be the green parrot to continue the alternating pattern.

Mini-Gemini



Absinthe Painting - Cafe Scene In Paris by Henri Gervex. Analyze then generate.

Capture a scene of space elevator

[Round 1] Generate a similar image

[Round 2] I want to add a husky in this image.

[Round 2] I want to add a husky in this image.

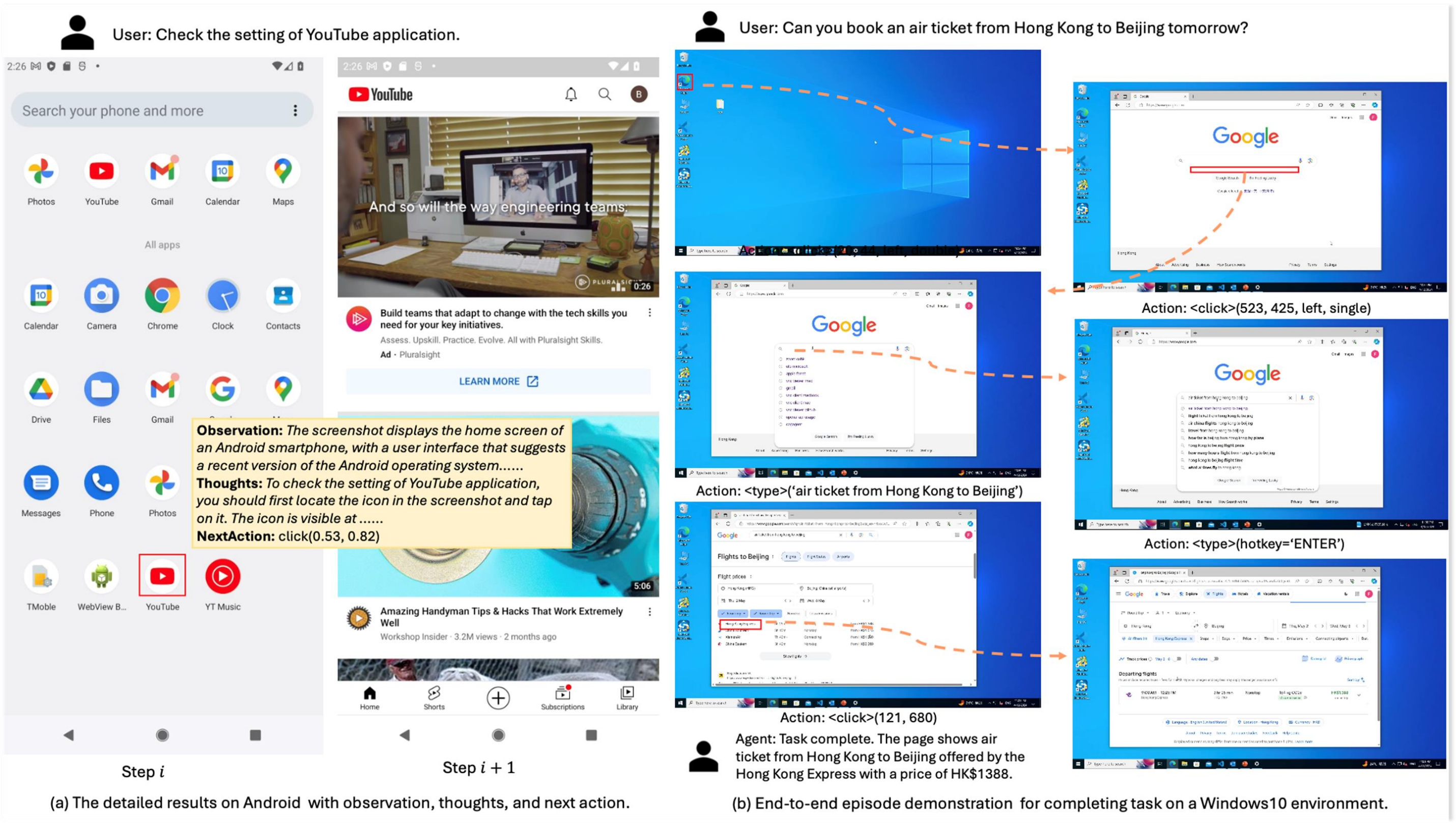
Showcase an orange-themed interior design inspiration collage.

Orange mood board for interior design, plush towels, floral bedding, striped containers, and scented candle, high contrast, minimalistic, vibrant color palette, digital collage. Mini-Gemini

Describe the image and change the character to a dog.

Generate the scene if the cat is coding.

Multimodal Agent



User: Check the setting of YouTube application.

Build teams that adapt to change with the tech skills you need for your key initiatives.
Assess. Upskill. Practice. Evolve. All with Pluralsight Skills.
Ad · Pluralsight

Amazing Handyman Tips & Hacks That Work Extremely Well
Workshop Insider · 3.2M views · 2 months ago

Observation: *The screenshot displays the home screen of an Android smartphone, with a user interface that suggests a recent version of the Android operating system......*
Thoughts: *To check the setting of YouTube application, you should first locate the icon in the screenshot and tap on it. The icon is visible at ......*
NextAction: click(0.53, 0.82)

Step i

Step $i+1$

(a) The detailed results on Android with observation, thoughts, and next action.

User: Can you book an air ticket from Hong Kong to Beijing tomorrow?

Action: <click>(523, 425, left, single)

Action: <type>(‘air ticket from Hong Kong to Beijing’)

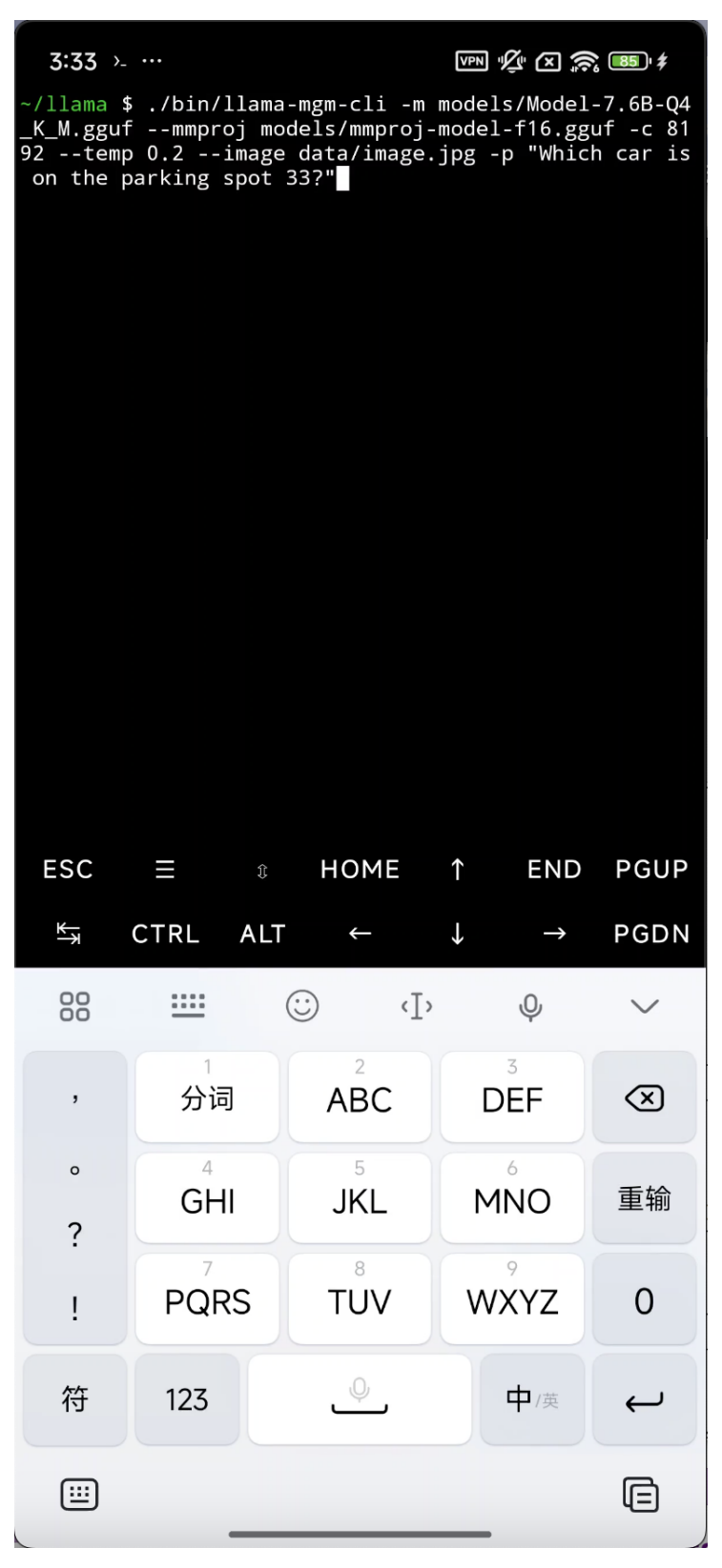
Action: <type>(hotkey=‘ENTER’)

Action: <click>(121, 680)

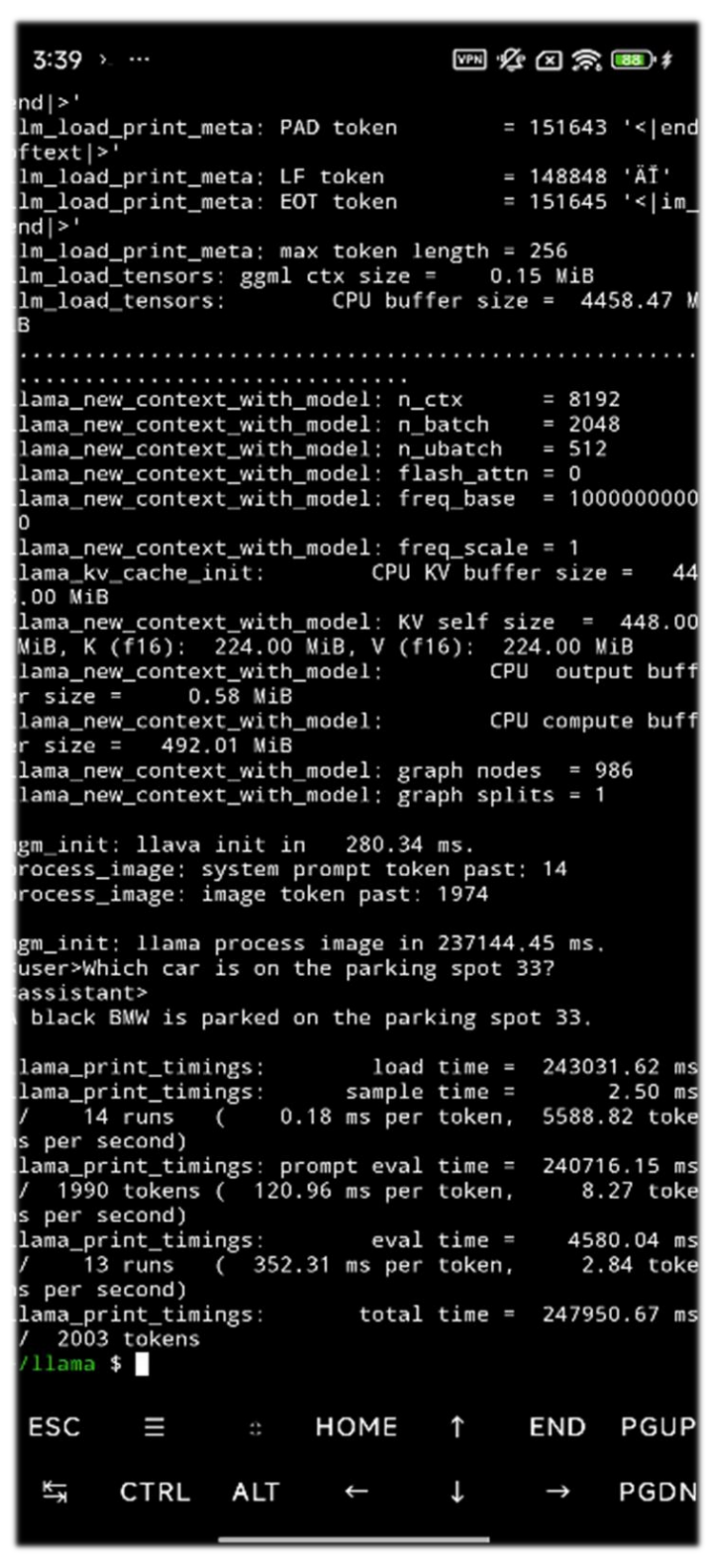
Agent: Task complete. The page shows air ticket from Hong Kong to Beijing offered by the Hong Kong Express with a price of HK$1388.

(b) End-to-end episode demonstration for completing task on a Windows10 environment.

Mobile Deployment



```
3:33
~/llama $ ./bin/llama-mgm-cli -m models/Model-7.6B-Q4
_K_M.gguf --mmproj models/mmproj-model-f16.gguf -c 81
92 --temp 0.2 --image data/image.jpg -p "Which car is
 on the parking spot 33?"
```



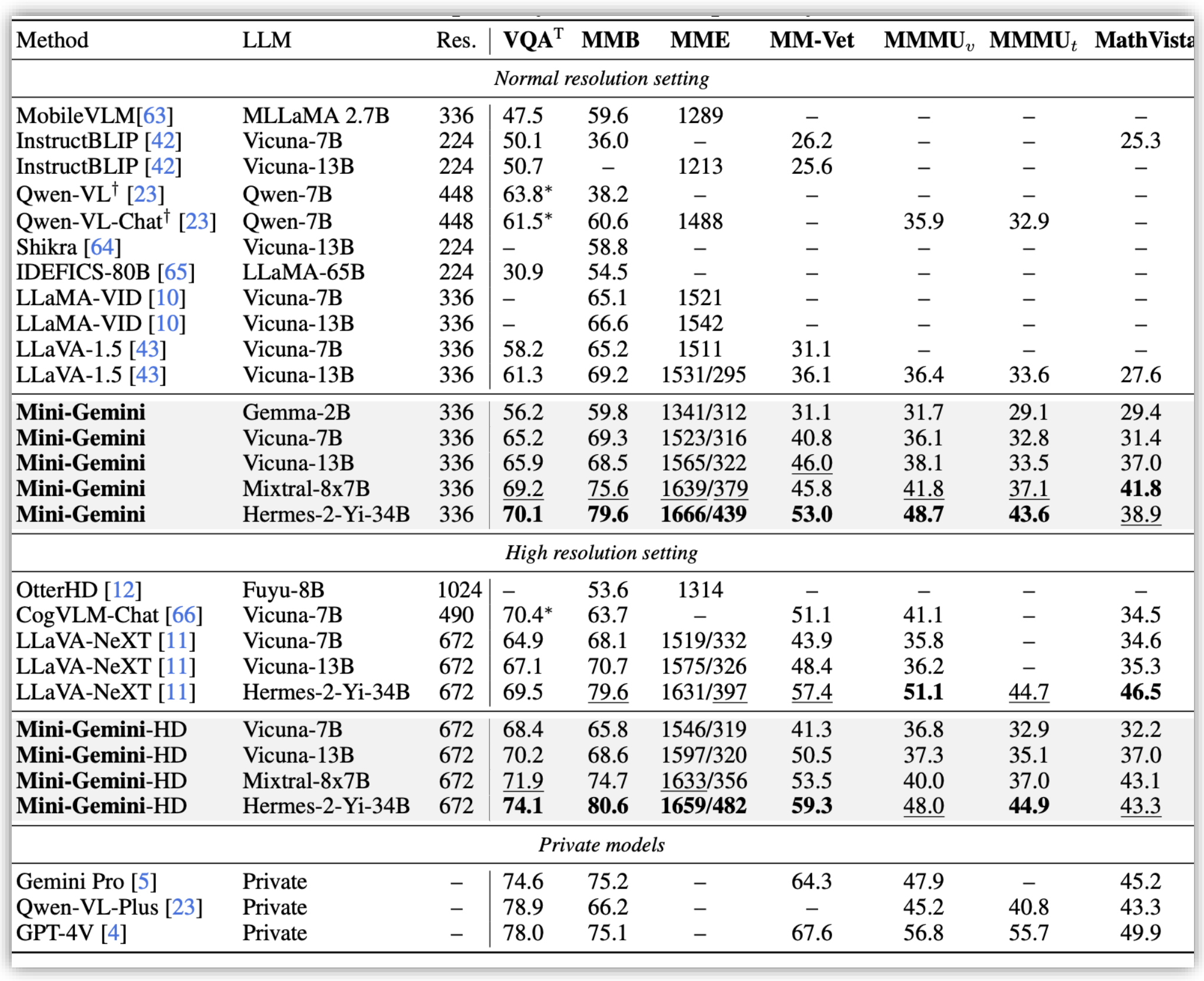
```
3:39
nd|>'
lm_load_print_meta: PAD token          = 151643 '<|end
ftext|>'
lm_load_print_meta: LF token           = 148848 'ÄĪ'
lm_load_print_meta: EOT token          = 151645 '<|im_
nd|>'
lm_load_print_meta: max token length = 256
lm_load_tensors: ggml ctx size =    0.15 MiB
lm_load_tensors:        CPU buffer size =  4458.47 M
B
.................................................
...........................
lama_new_context_with_model: n_ctx      = 8192
lama_new_context_with_model: n_batch    = 2048
lama_new_context_with_model: n_ubatch   = 512
lama_new_context_with_model: flash_attn = 0
lama_new_context_with_model: freq_base  = 1000000000
0
lama_new_context_with_model: freq_scale = 1
lama_kv_cache_init:        CPU KV buffer size =   44
.00 MiB
lama_new_context_with_model: KV self size  =  448.00
MiB, K (f16):  224.00 MiB, V (f16):  224.00 MiB
lama_new_context_with_model:        CPU  output buff
r size =     0.58 MiB
lama_new_context_with_model:        CPU compute buff
r size =   492.01 MiB
lama_new_context_with_model: graph nodes  = 986
lama_new_context_with_model: graph splits = 1

gm_init: llava init in    280.34 ms.
rocess_image: system prompt token past: 14
rocess_image: image token past: 1974

gm_init: llama process image in 237144.45 ms.
user>Which car is on the parking spot 33?
assistant>
 black BMW is parked on the parking spot 33.

lama_print_timings:        load time =  243031.62 ms
lama_print_timings:      sample time =      2.50 ms
/    14 runs   (    0.18 ms per token,  5588.82 toke
s per second)
lama_print_timings: prompt eval time =  240716.15 ms
/  1990 tokens (  120.96 ms per token,     8.27 toke
s per second)
lama_print_timings:        eval time =    4580.04 ms
/    13 runs   (  352.31 ms per token,     2.84 toke
s per second)
lama_print_timings:       total time =  247950.67 ms
/  2003 tokens
/llama $
```

Quantitative Results



Method	LLM	Res.	VQA^T	MMB	MME	MM-Vet	MMMU$_v$	MMMU$_t$	MathVista
Normal resolution setting									
MobileVLM[63]	MLLaMA 2.7B	336	47.5	59.6	1289	–	–	–	–
InstructBLIP [42]	Vicuna-7B	224	50.1	36.0	–	26.2	–	–	25.3
InstructBLIP [42]	Vicuna-13B	224	50.7	–	1213	25.6	–	–	–
Qwen-VL† [23]	Qwen-7B	448	63.8*	38.2	–	–	–	–	–
Qwen-VL-Chat† [23]	Qwen-7B	448	61.5*	60.6	1488	–	35.9	32.9	–
Shikra [64]	Vicuna-13B	224	–	58.8	–	–	–	–	–
IDEFICS-80B [65]	LLaMA-65B	224	30.9	54.5	–	–	–	–	–
LLaMA-VID [10]	Vicuna-7B	336	–	65.1	1521	–	–	–	–
LLaMA-VID [10]	Vicuna-13B	336	–	66.6	1542	–	–	–	–
LLaVA-1.5 [43]	Vicuna-7B	336	58.2	65.2	1511	31.1	–	–	–
LLaVA-1.5 [43]	Vicuna-13B	336	61.3	69.2	1531/295	36.1	36.4	33.6	27.6
Mini-Gemini	Gemma-2B	336	56.2	59.8	1341/312	31.1	31.7	29.1	29.4
Mini-Gemini	Vicuna-7B	336	65.2	69.3	1523/316	40.8	36.1	32.8	31.4
Mini-Gemini	Vicuna-13B	336	65.9	68.5	1565/322	46.0	38.1	33.5	37.0
Mini-Gemini	Mixtral-8x7B	336	69.2	75.6	1639/379	45.8	41.8	37.1	**41.8**
Mini-Gemini	Hermes-2-Yi-34B	336	**70.1**	**79.6**	**1666/439**	**53.0**	**48.7**	**43.6**	38.9
High resolution setting									
OtterHD [12]	Fuyu-8B	1024	–	53.6	1314	–	–	–	–
CogVLM-Chat [66]	Vicuna-7B	490	70.4*	63.7	–	51.1	41.1	–	34.5
LLaVA-NeXT [11]	Vicuna-7B	672	64.9	68.1	1519/332	43.9	35.8	–	34.6
LLaVA-NeXT [11]	Vicuna-13B	672	67.1	70.7	1575/326	48.4	36.2	–	35.3
LLaVA-NeXT [11]	Hermes-2-Yi-34B	672	69.5	79.6	1631/397	57.4	**51.1**	44.7	**46.5**
Mini-Gemini-HD	Vicuna-7B	672	68.4	65.8	1546/319	41.3	36.8	32.9	32.2
Mini-Gemini-HD	Vicuna-13B	672	70.2	68.6	1597/320	50.5	37.3	35.1	37.0
Mini-Gemini-HD	Mixtral-8x7B	672	71.9	74.7	1633/356	53.5	40.0	37.0	43.1
Mini-Gemini-HD	Hermes-2-Yi-34B	672	**74.1**	**80.6**	**1659/482**	**59.3**	48.0	**44.9**	43.3
Private models									
Gemini Pro [5]	Private	–	74.6	75.2	–	64.3	47.9	–	45.2
Qwen-VL-Plus [23]	Private	–	78.9	66.2	–	–	45.2	40.8	43.3
GPT-4V [4]	Private	–	78.0	75.1	–	67.6	56.8	55.7	49.9

Our Related Publications

1. Li, Yanwei*, Chengyao Wang*, and Jiaya Jia. "Llama-vid: An image is worth 2 tokens in large language models." In *European Conference on Computer Vision*, pp. 323-340. Springer, Cham, 2025.
2. Li, Yanwei*, Yuechen Zhang*, Chengyao Wang*, Zhisheng Zhong, Yixin Chen, Ruihang Chu, Shaoteng Liu, and Jiaya Jia. "Mini-gemini: Mining the potential of multi-modality vision language models." *arXiv preprint arXiv:2403.18814* (2024).

Acknowledgment

This work was supported in part by the Hong Kong Research Grants Council under the Areas of Excellence Scheme Grant AoE/E-601/22-R