香港中文大學
The Chinese University of Hong Kong

Mini-Gemini:

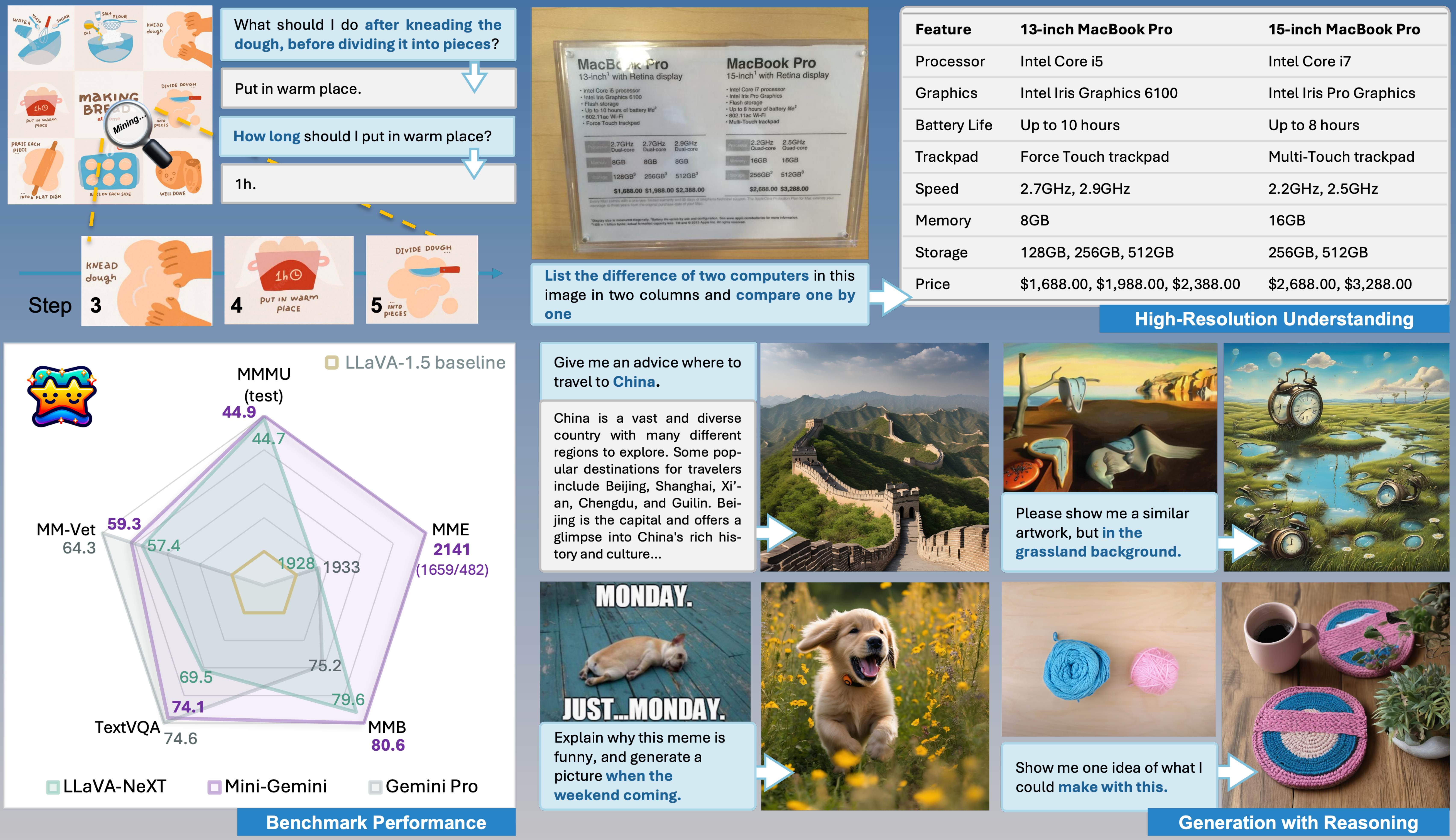
Mining the Potential of Multi-modality Vision Language Models

Yanwei Li

Introduction

How to further improve current VLMs?

3 key parts in Mini-Gemini: High-resolution, better data, and extended application



Introduction

How to further improve current VLMs?

Current Vision Language Model (VLM):

Representation

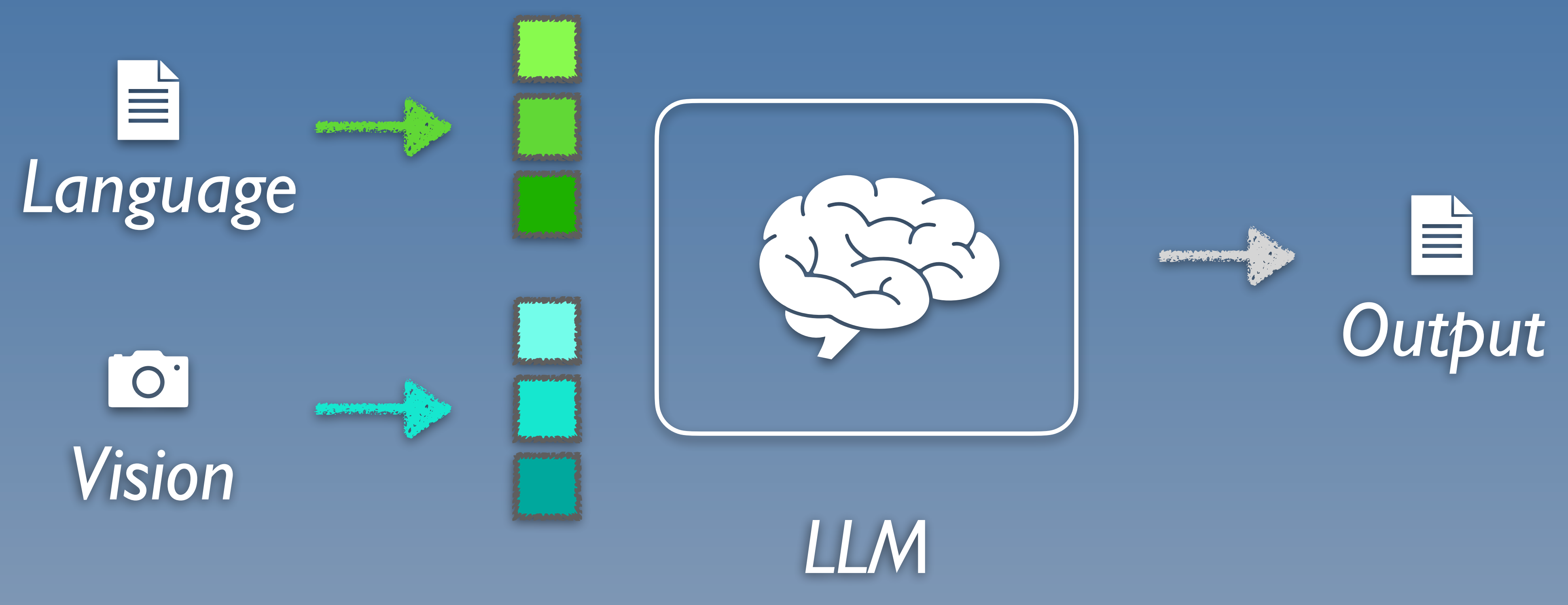
- *Language —> Tokenizer —> Text Token*
- *Vision —> Transformer —> Image Token*

Processing

- *Process tokens from different modalities in LLM*

Prediction

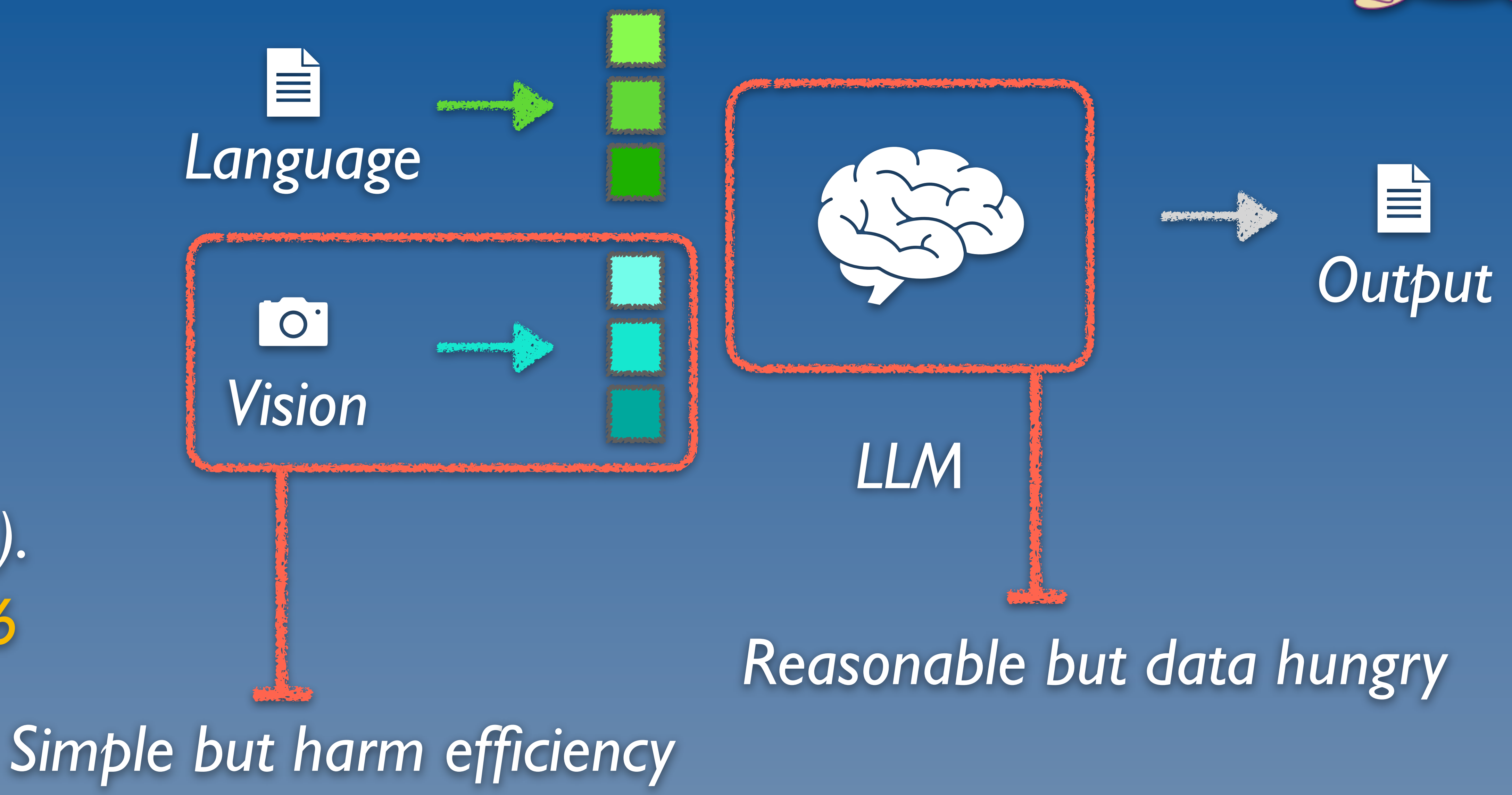
- *Predict text or images from the generated token*



General pipeline of current VLM

VLM for High-Resolution

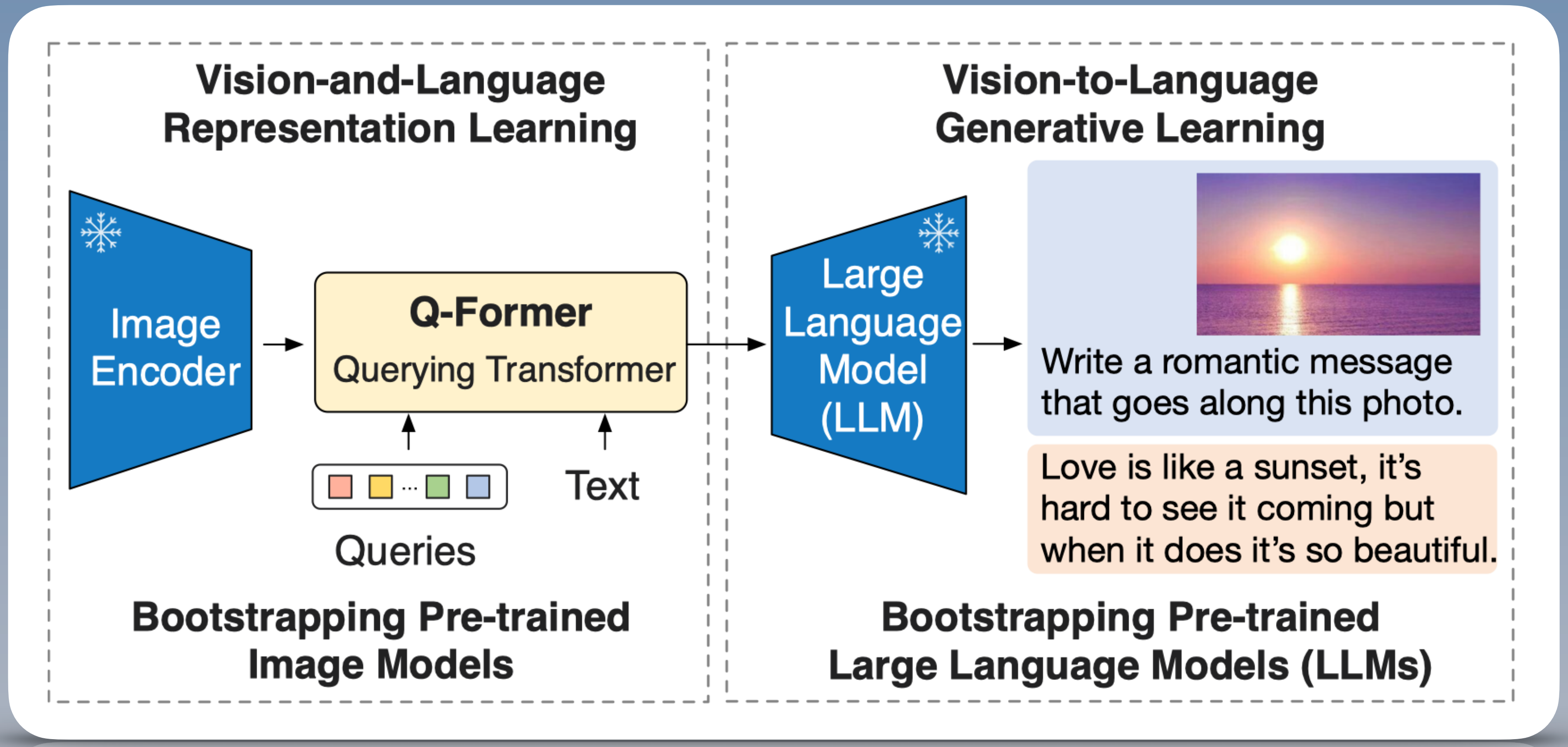
What if we want to process high-resolution image/video?



Token Cost

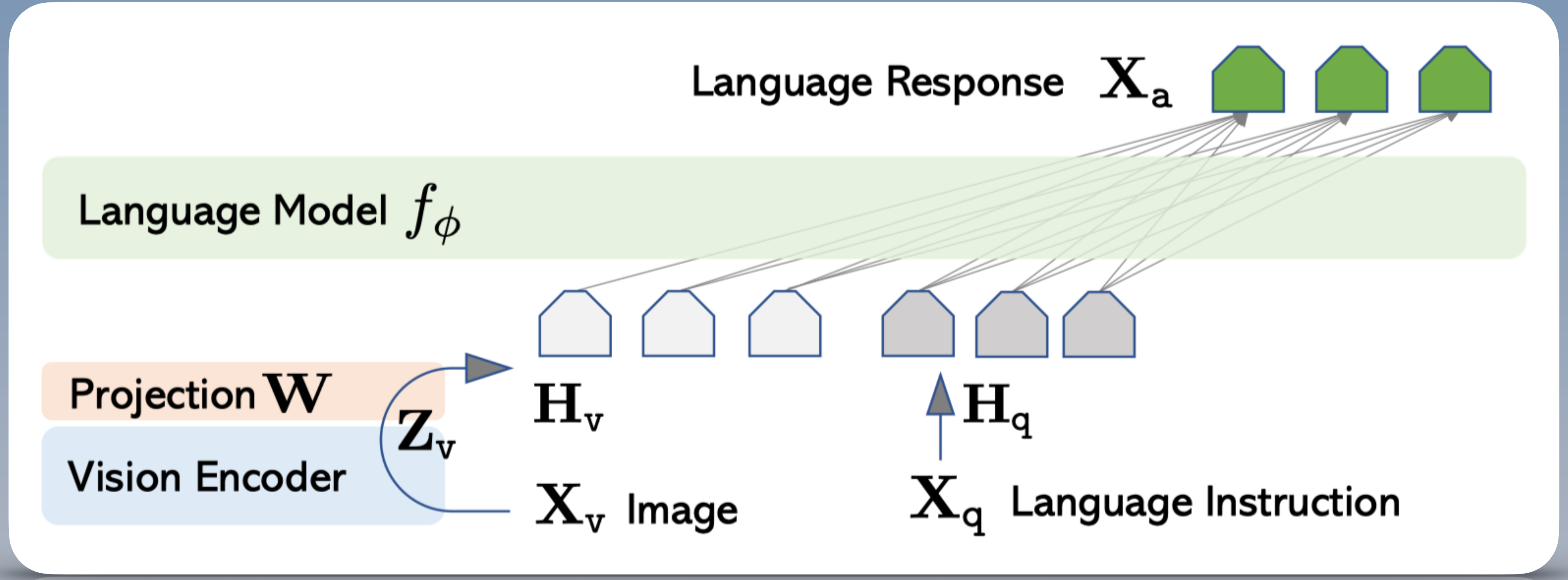
- *Query-based:* efficient but loss detail, Flamingo[1] (64), BLIP2[2] (32).
- *Projector-based:* token number increase with resolution, 224->256 Token, 336->576 Token, and 672->2304 Token.

Too many token harm the efficiency, especially for training and multi-image.



Query-based Vision token generation in BLIP2[2].

32 queries for each image.



Projector-based Vision token generation in LLaVA[3].

256 queries for each image with 224 size.

VLM for High-Resolution

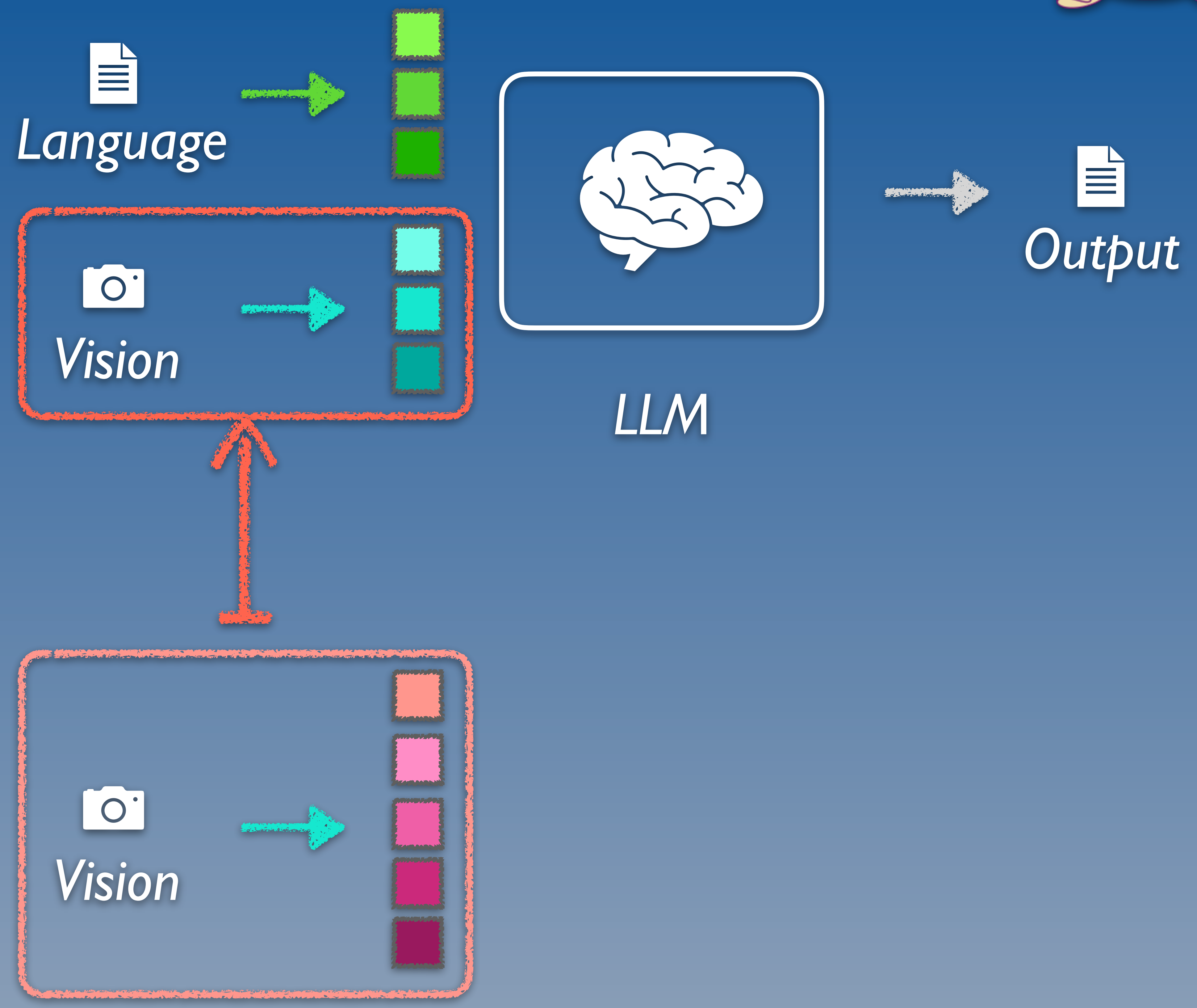
What if we want to process high-resolution image/video?

Token Cost

- *Query-based:* efficient but loss detail, Flamingo[1] (64), BLIP2[2] (32).
- *Projector-based:* token number increase with resolution, 224->256 Token, 336->576 Token, and 672->2304 Token.

Too many token harm the efficiency, especially for training and multi-image.

Can we utilize another high-resolution encoder to provide detailed spatial info?



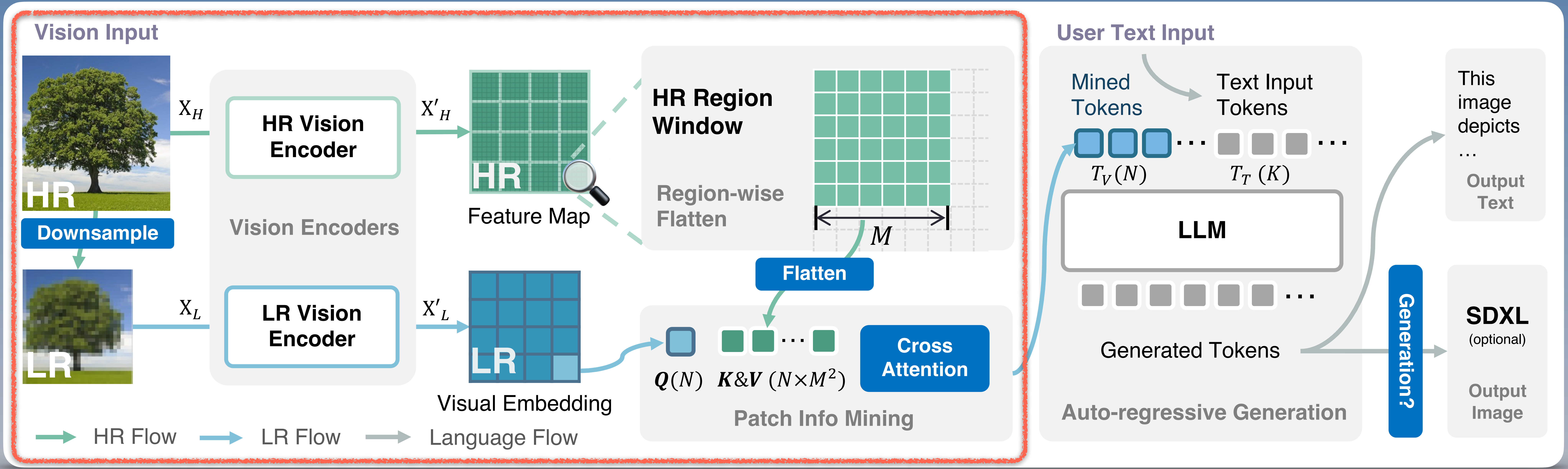
A High-Resolution but Efficient Net

[1] Jean-Baptiste Alayrac, et.al., Flamingo: a Visual Language Model for Few-Shot Learning, In NeurIPS, 2022.
[2] Junnan Li, et.al., BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models, arXiv, 2023.
[3] Haotian Liu, et.al., Visual Instruction Tuning, In NeurIPS, 2023.

VLM for High-Resolution

HR Solution

- *ViT-based LR Encoder: CLIP-pretrained ViT model with low resolution image for query.*
- *CNN-based HR Encoder: CLIP-pretrained CNN model with adaptive high resolution input for key and value.*

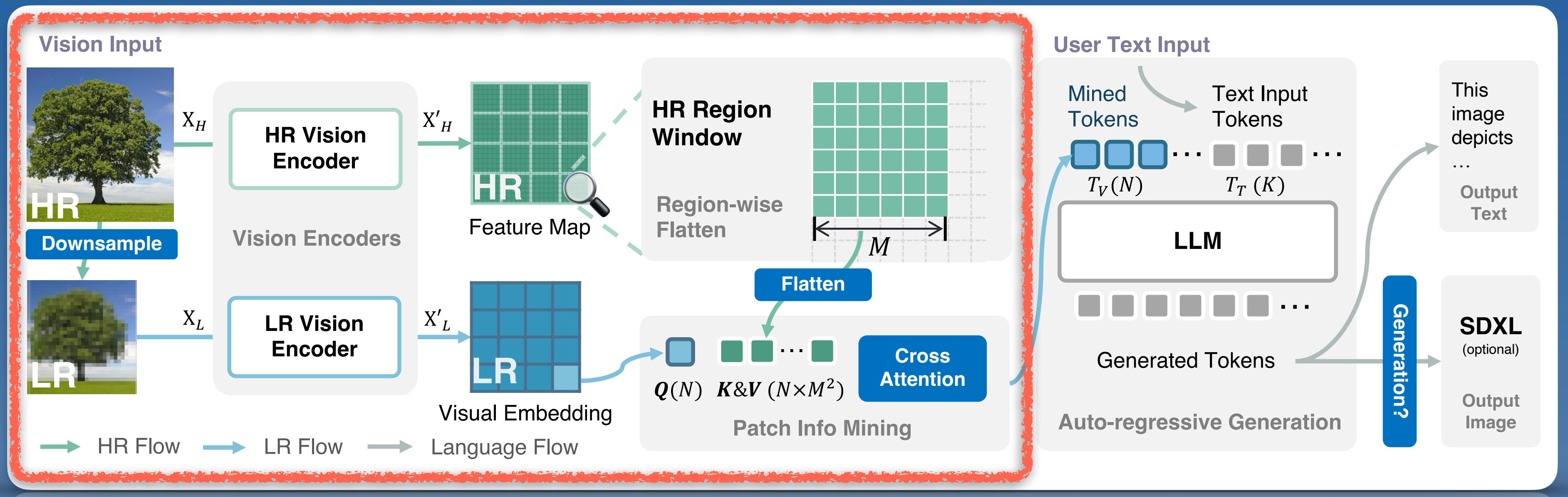


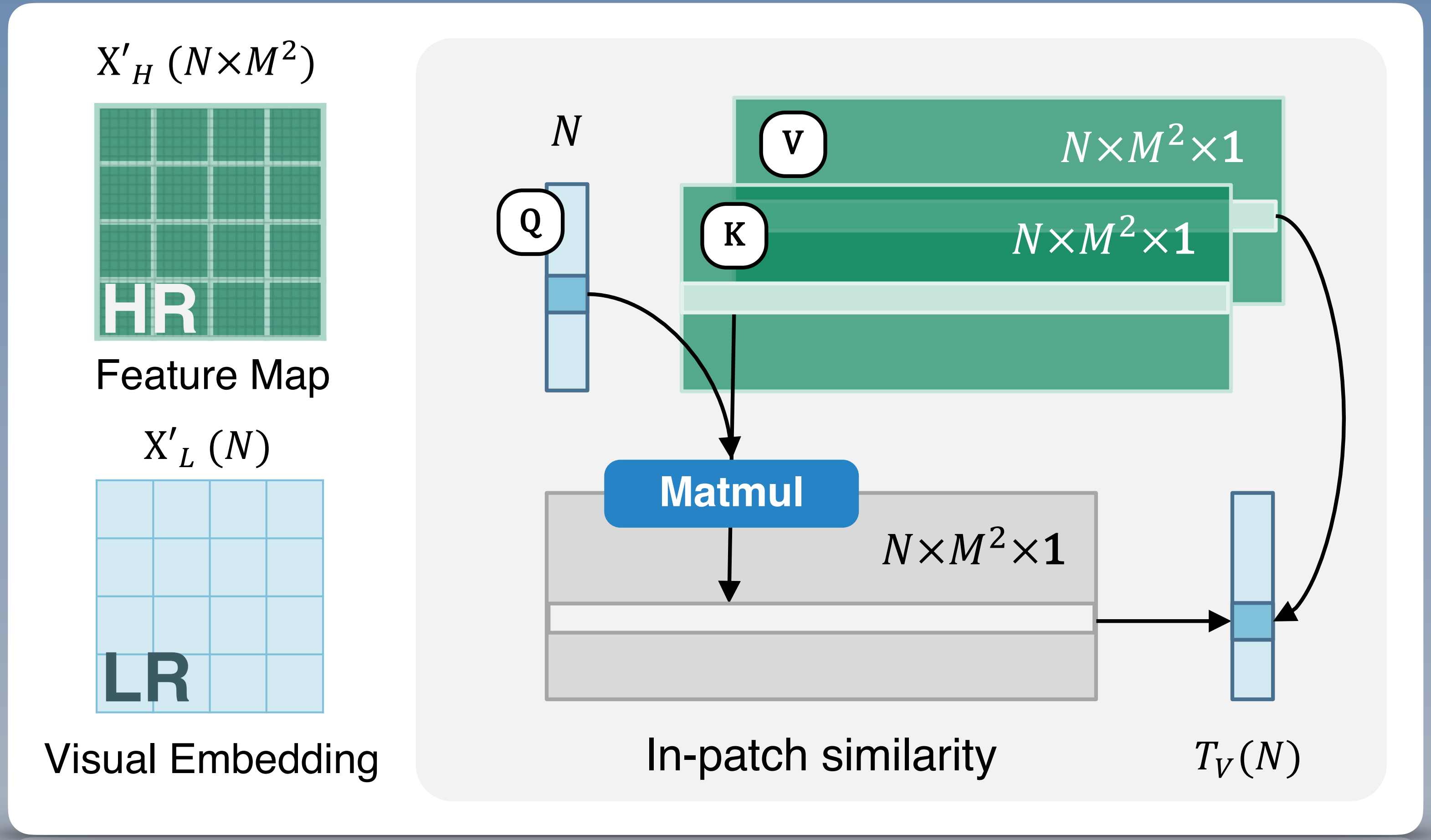
The framework of Mini-Gemini with any-to-any workflow.

VLM for High-Resolution

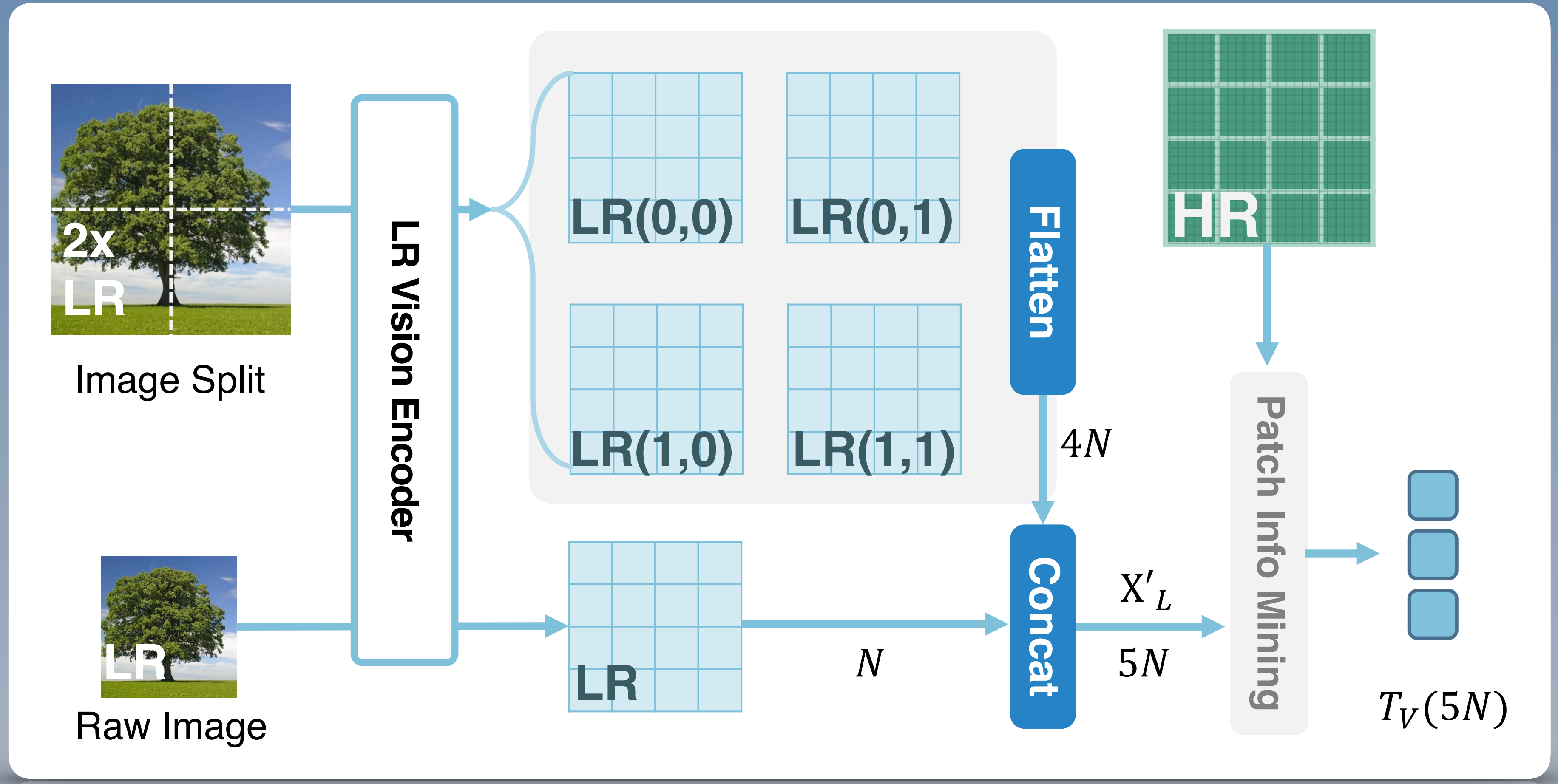
HR Solution

- ViT-based LR Encoder: CLIP-pretrained ViT model with low resolution image for query.
- CNN-based HR Encoder: CLIP-pretrained CNN model with adaptive high resolution input for key and value.
- Visual Token Extension: Extend LR image to support higher resolution and more visual tokens.





Details in Patch Info Mining.

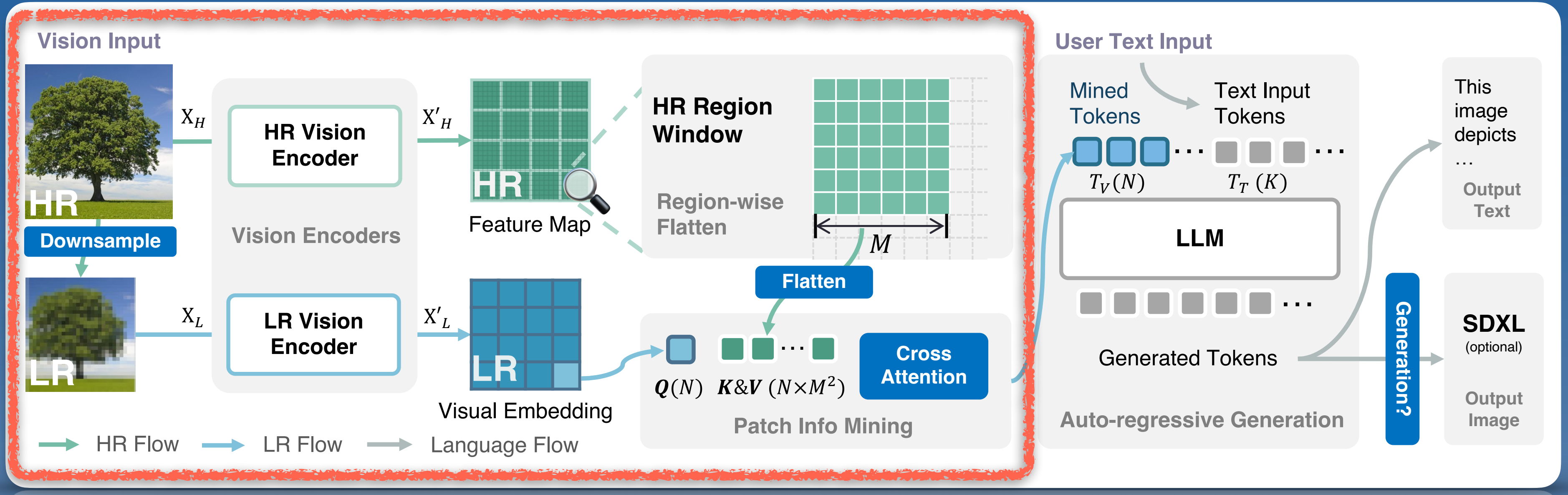


Details in Visual Token Extension.

VLM for High-Resolution

HR Solution

- *Patch Info Mining:* Significant improvement in detail-related evaluation, like TextVQA.
- *Vision Encoder:* Improve the HR feature quality is important, but it will converge with large encoder.
- *Image Resolution:* Both LR and HR resolution contributes a lot to the final results, while LR query improves more.



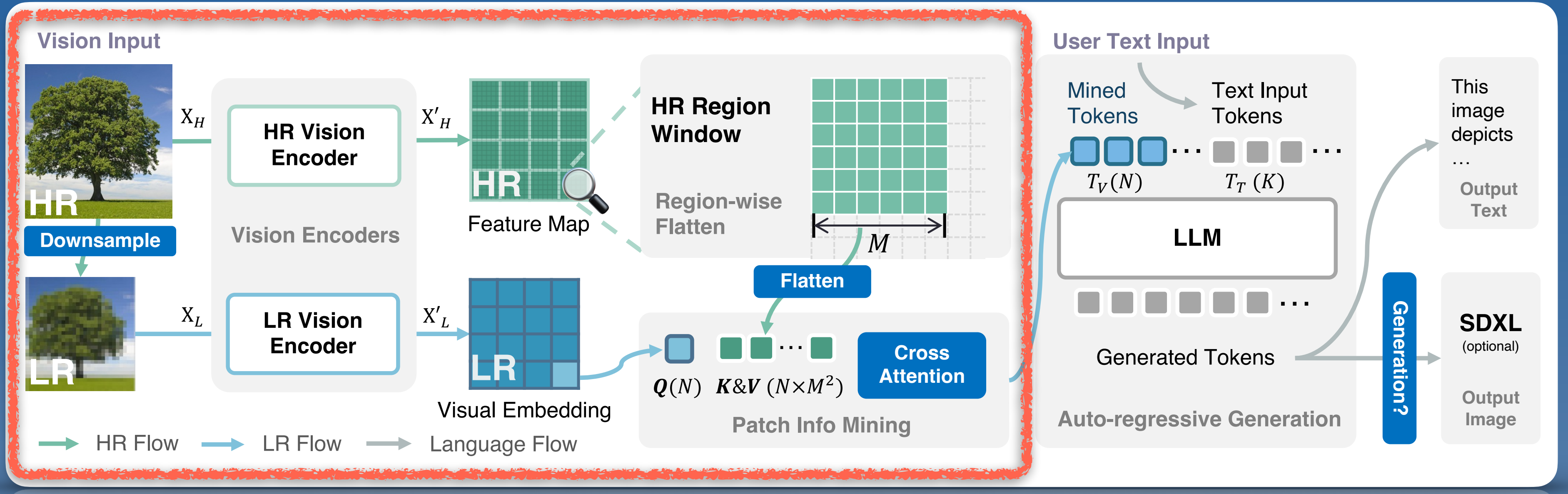
Method	VE-HR	LR	HR	Token Num.	VQA^T	MME	MM-Vet
Baseline	–	224	–	256	54.1*	1467.1	30.7
+ Info mining	ConvX-L	224	512	256	58.1* +4.0	**1485.2** +18.1	31.3 +0.6
+ Higher res.	ConvX-L	224	768	256	**59.8*** +1.7	1478.3 -6.9	**31.9** +0.6
Baseline	–	336	–	576	58.2*	1510.7	31.1
+ Info mining	ConvX-B	336	768	576	58.4* +0.2	1451.7 -59.0	33.8 +2.7
+ Larger VE-HR	ConvX-L	336	768	576	61.5* +3.1	**1517.0** +65.3	**34.6** +0.8
+ Larger VE-HR	ConvX-XXL	336	768	576	**62.0*** +0.5	1505.7 -11.3	33.8 -0.8

Comparison with different info mining settings.

VLM for High-Resolution

HR Solution

- *Patch Info Mining:* Significant improvement in detail-related evaluation, like TextVQA.
- *Vision Encoder:* Improve the HR feature quality is important, but it will converge with large encoder.
- *Image Resolution:* Both LR and HR resolution contributes a lot to the final results, while LR query improves more.
- *Visual Token Num:* Token count, especially global image contributes a lot, but converge over 8K.



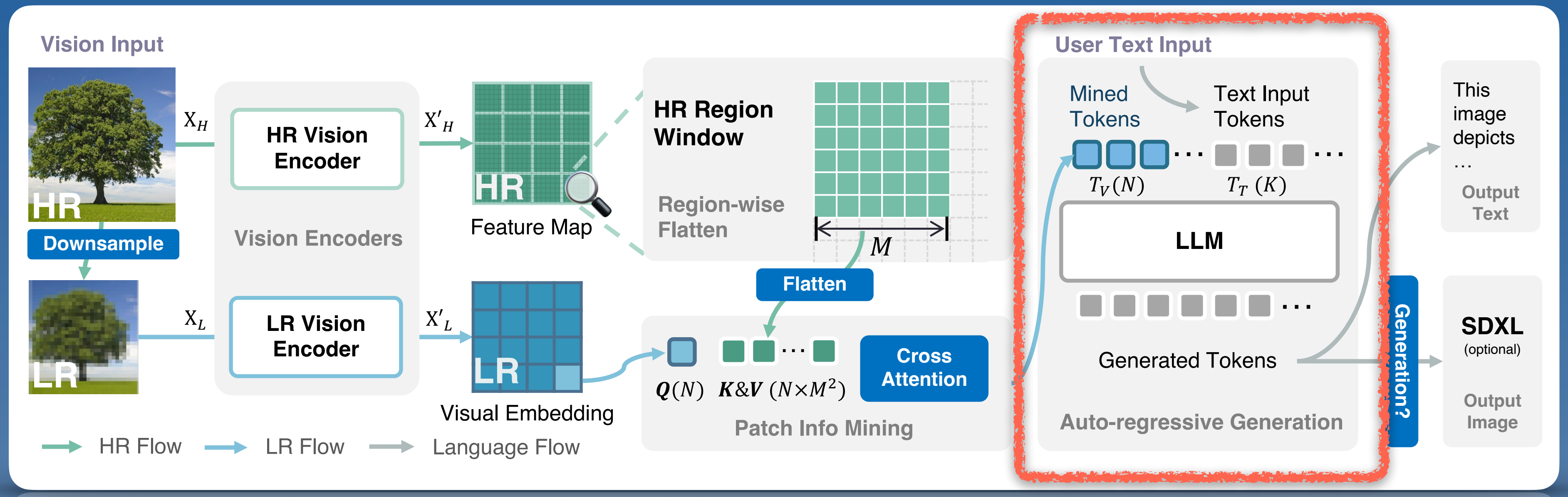
Method	LR	HR	Token Num.	VQA^T		MME		MM-Vet	
Baseline	336	–	576	58.2*		1510.7		31.1	
+ *Info mining*	336	768	576	61.5*	+3.3	1517.0	+6.3	34.6	+3.5
+ *ShareGPT4V*	336	768	576	63.2*	+1.7	1527.6	+10.6	34.2	-0.4
– *TextCaps*	336	768	576	59.0	-4.2	1465.2	-62.4	35.0	+0.8
+ *LAION-GPT-4V*	336	768	576	58.7	-0.3	1521.8	+56.6	33.4	-1.6
+ *OCR-related*	336	768	576	61.6	+2.9	1523.5	+1.7	33.7	+0.3
+ *Gen-related*	336	768	576	62.2	+0.6	1521.2	-2.3	37.0	+3.3
+ *ALLaVA*	336	768	576	65.2	+3.0	1523.3	+2.1	40.8	+3.8
+ *Token extension*	672	1536	2880	**68.4**	+3.2	**1546.2**	+22.9	**41.3**	+0.5

Comparison with different info mining settings.

VLM with Better Data

Better Data

- Detailed Visual Instruction: High quality data is quite important, especially in SFT stage.
- Text-related Data: High quality OCR data plays a vital role in text-related QA.
- Text-only Data: Balanced text-only data is essential, too less or too much greatly harms the performance. We try drop all and add extra 100K text-only data but get bad results.



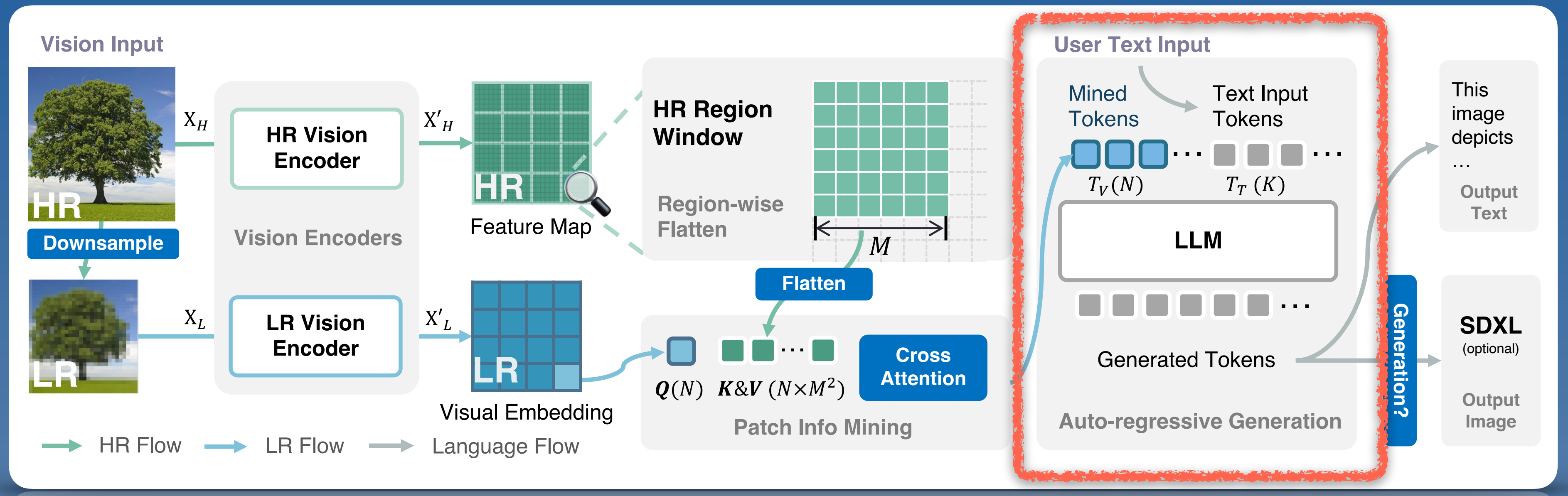
Method	LR	HR	Token Num.	VQA^T		MME		MM-Vet	
Baseline	336	–	576	58.2*		1510.7		31.1	
+ *Info mining*	336	768	576	61.5*	+3.3	1517.0	+6.3	34.6	+3.5
+ *ShareGPT4V*	336	768	576	63.2*	+1.7	1527.6	+10.6	34.2	-0.4
– *TextCaps*	336	768	576	59.0	-4.2	1465.2	-62.4	35.0	+0.8
+ *LAION-GPT-4V*	336	768	576	58.7	-0.3	1521.8	+56.6	33.4	-1.6
+ *OCR-related*	336	768	576	61.6	+2.9	1523.5	+1.7	33.7	+0.3
+ *Gen-related*	336	768	576	62.2	+0.6	1521.2	-2.3	37.0	+3.3
+ *ALLaVA*	336	768	576	65.2	+3.0	1523.3	+2.1	40.8	+3.8
+ *Token extension*	672	1536	2880	**68.4**	+3.2	**1546.2**	+22.9	**41.3**	+0.5

Comparison with different info mining settings.

VLM with Better Data

Better Data

- Detailed Visual Instruction: High quality data is quite important, especially in SFT stage.
- Text-related Data: High quality OCR data plays a vital role in text-related QA.
- Text-only Data: Balanced text-only data is essential, too less or too much greatly harms the performance. We try drop all and add extra 100K text-only data but get bad results.



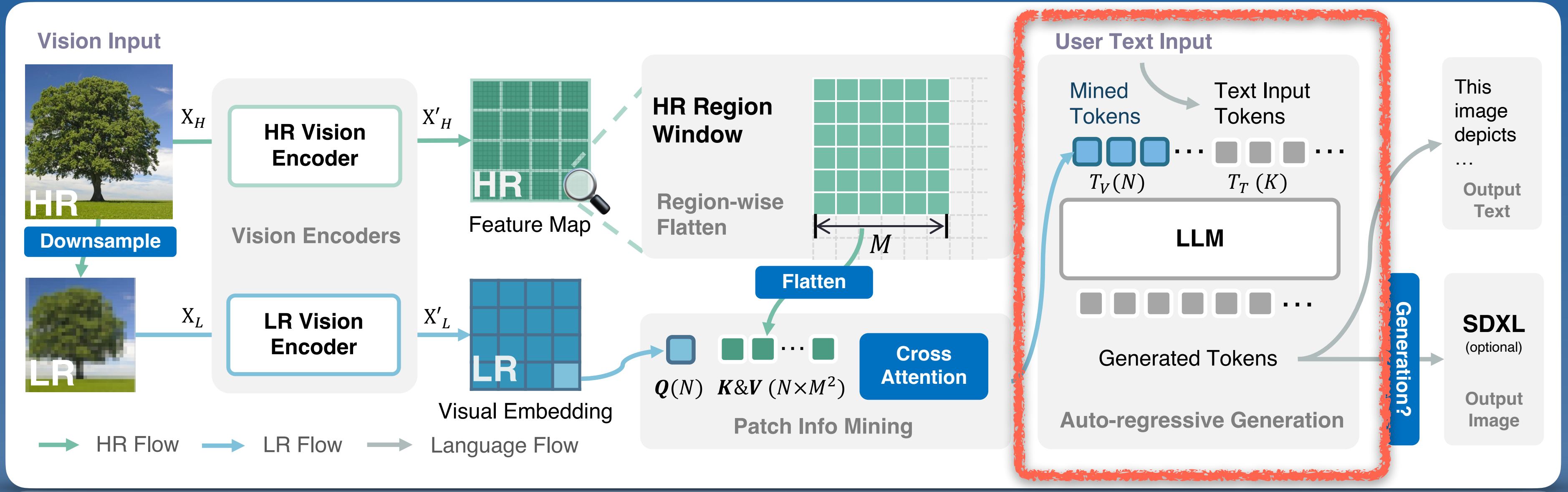
Method	LR	HR	Token Num.	VQA^T		MME		MM-Vet	
Baseline	336	–	576	58.2*		1510.7		31.1	
+ *Info mining*	336	768	576	61.5*	+3.3	1517.0	+6.3	34.6	+3.5
+ *ShareGPT4V*	336	768	576	63.2*	+1.7	1527.6	+10.6	34.2	-0.4
– *TextCaps*	336	768	576	59.0	-4.2	1465.2	-62.4	35.0	+0.8
+ *LAION-GPT-4V*	336	768	576	58.7	-0.3	1521.8	+56.6	33.4	-1.6
+ *OCR-related*	336	768	576	61.6	+2.9	1523.5	+1.7	33.7	+0.3
+ *Gen-related*	336	768	576	62.2	+0.6	1521.2	-2.3	37.0	+3.3
+ *ALLaVA*	336	768	576	65.2	+3.0	1523.3	+2.1	40.8	+3.8
+ *Token extension*	672	1536	2880	**68.4**	+3.2	**1546.2**	+22.9	**41.3**	+0.5

Comparison with different info mining settings.

VLM with Better Data

Better Data

- Detailed Visual Instruction: High quality data is quite important, especially in SFT stage.
- Text-related Data: High quality OCR data plays a vital role in text-related QA.
- Text-only Data: Balanced text-only data is essential, too less or too much greatly harms the performance. We try drop all and add extra 100K text-only data but get bad results.



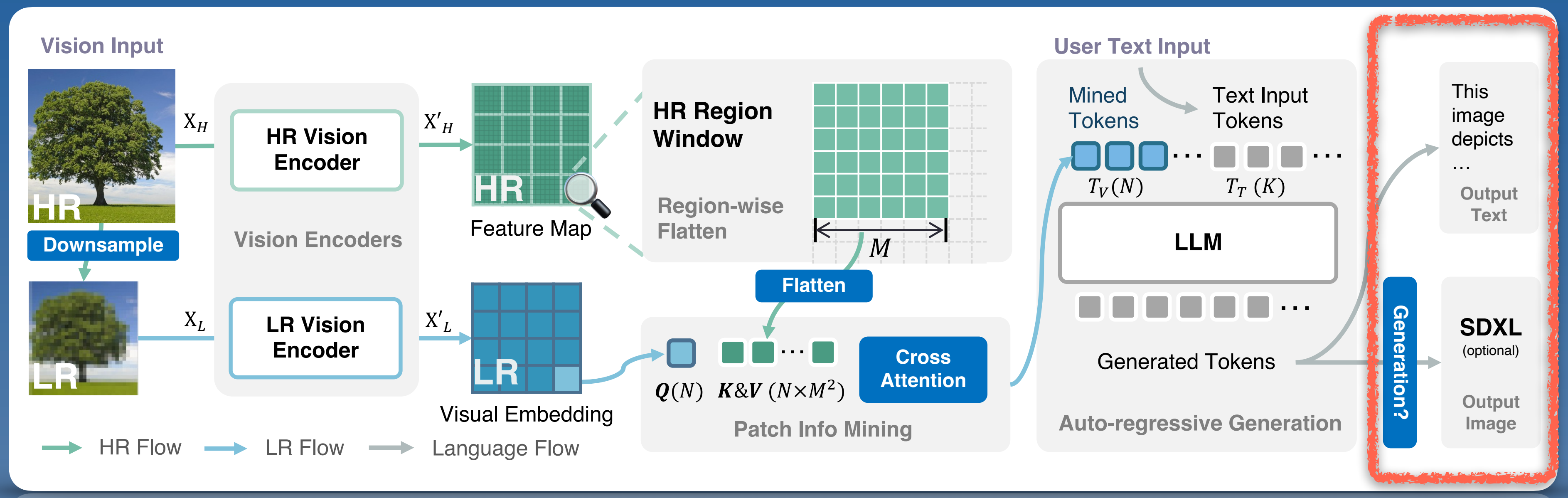
Method	LR	HR	Token Num.	VQA^T		MME		MM-Vet	
Baseline	336	–	576	58.2*		1510.7		31.1	
+ *Info mining*	336	768	576	61.5*	+3.3	1517.0	+6.3	34.6	+3.5
+ *ShareGPT4V*	336	768	576	63.2*	+1.7	1527.6	+10.6	34.2	-0.4
– *TextCaps*	336	768	576	59.0	-4.2	1465.2	-62.4	35.0	+0.8
+ *LAION-GPT-4V*	336	768	576	58.7	-0.3	1521.8	+56.6	33.4	-1.6
+ *OCR-related*	336	768	576	61.6	+2.9	1523.5	+1.7	33.7	+0.3
+ *Gen-related*	336	768	576	62.2	+0.6	1521.2	-2.3	37.0	+3.3
+ *ALLaVA*	336	768	576	65.2	+3.0	1523.3	+2.1	40.8	+3.8
+ *Token extension*	672	1536	2880	**68.4**	+3.2	**1546.2**	+22.9	**41.3**	+0.5

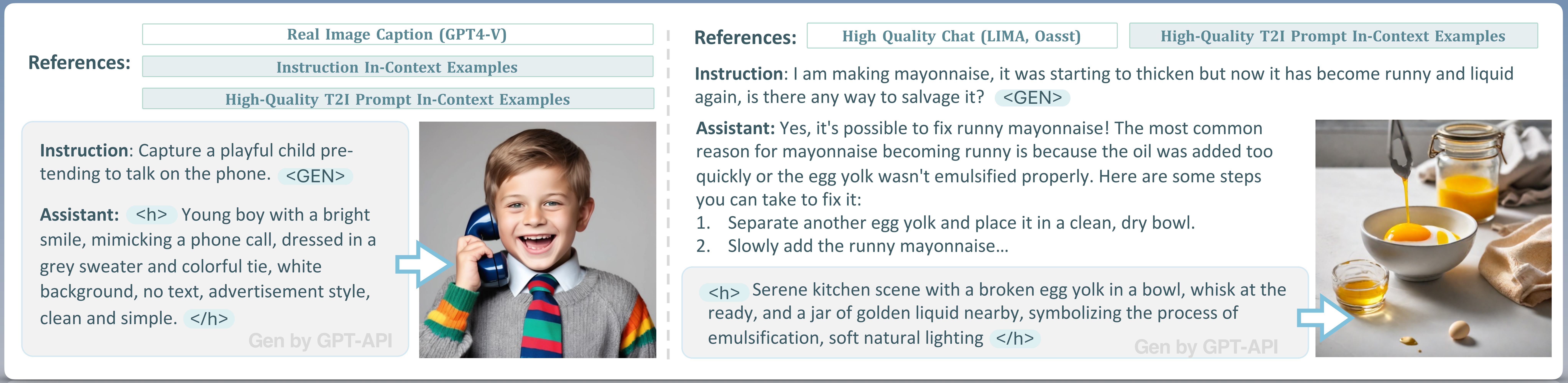
Comparison with different info mining settings.

VLM for Generation

Generation Data

- *Target:* *Use text to link reasoning results in VLM and generation model. We tried use predicted embedding but get worse or comparable results.*
- *Re-caption Data:* *Given high quality caption, generate potential user prompt and short caption for SD.*
- *In-context Prompt:* *Enable the model to well generate the generation captions in multi-round conversations.*





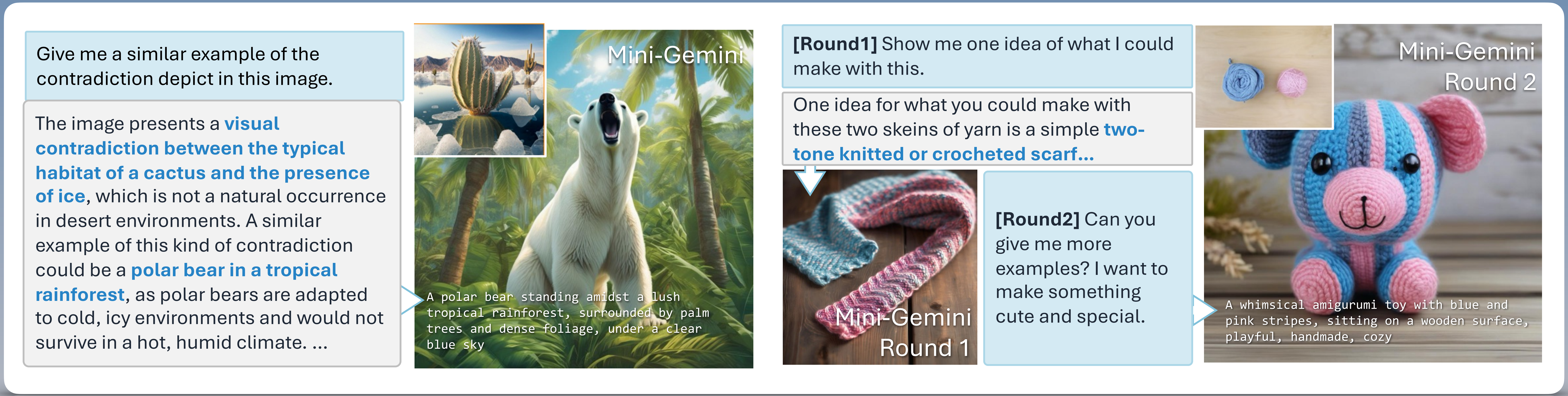
Simple instruction re-caption and In-context prompt generation.

VLM for Generation

Some interesting points

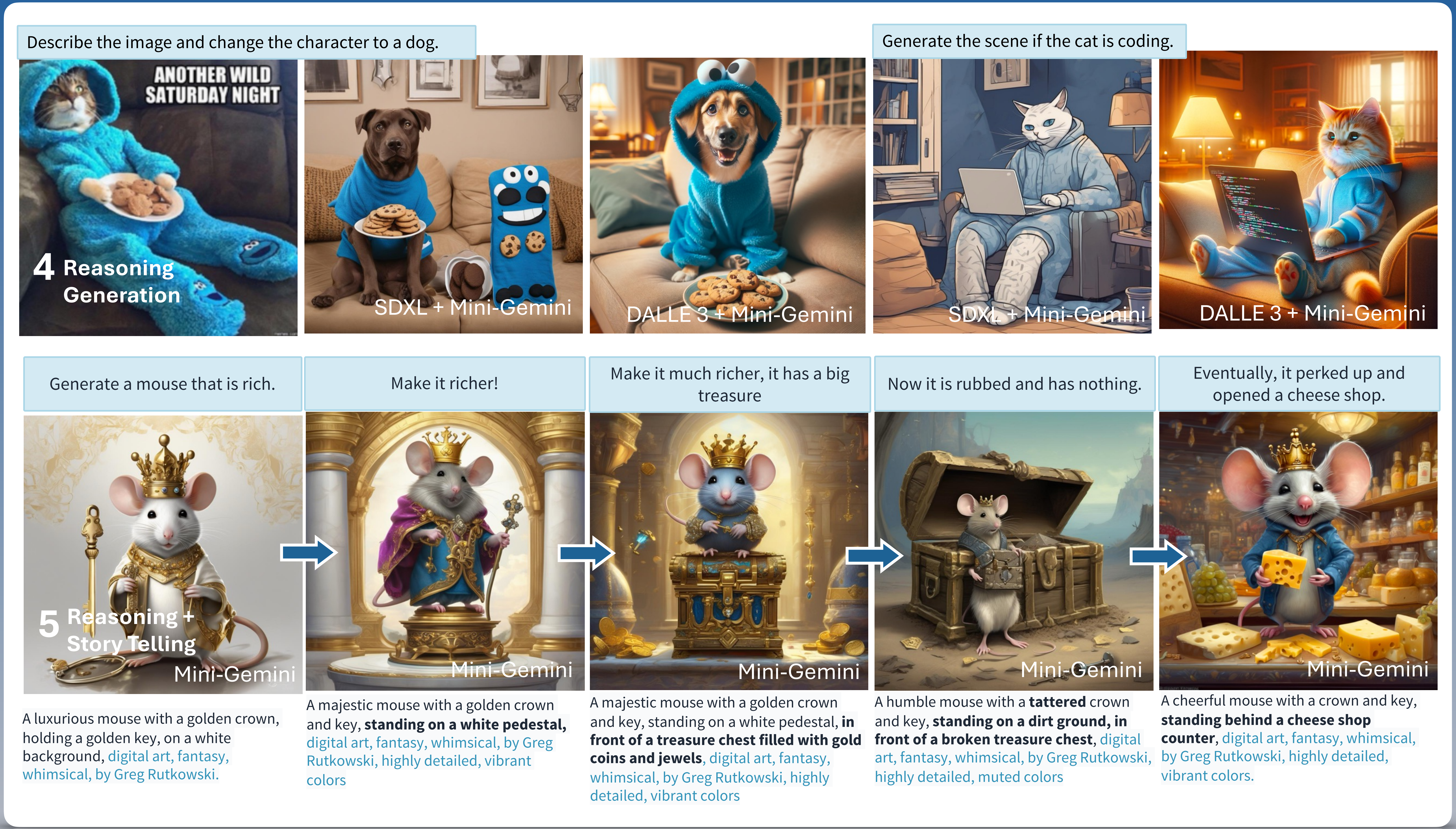
- *High-quality Text-only Data:* *The high-quality text-only data for generation further improve the understanding.*
- *Text-only input-> Image input:* *The text-only pipeline can be generalized to image and text input without modification.*
- *Generation with Reasoning:* *With good image-text alignment, it can combine understanding, reasoning, generation.*

Input with pure text.



Input with text and image.

VLM for Generation



Different generative models and reasoning with story-telling.

VLM for Generation

Generation with different VLMs


Source Image


Try 1


Try 2

Mini-Gemini-34B-HD

Prompt: Give me a similar example of the contradiction depict in this image.

Response: The image presents a visual contradiction by juxtaposing a cactus, which is typically associated with hot, arid environments, with a frozen landscape. A similar example of this type of contradiction could be a polar bear in a desert setting, as polar bears are adapted to cold climates and are commonly found in the Arctic, where ice and snow are prevalent. This scenario would be equally unexpected and visually striking, as it contrasts the natural habitat and adaptations of the polar bear with the extreme heat and lack of water in a desert.
<h> A polar bear standing in a desert, surrounded by sand dunes and a clear blue sky, digital art, surreal, by Greg Rutkowski </h>

VLM for Generation

Generation with different VLMs

Source Image

Try 1

Try 2

Mini-Gemini-13B-HD

Prompt: Give me a similar example of the contradiction depict in this image.

Response: The image presents a visual contradiction of a cactus in a frozen environment, which is not a natural occurrence. Cacti are plants that are typically found in arid, hot environments and are adapted to conserve water. The presence of ice and snow is incongruous with the cactus's natural habitat, which is usually characterized by extreme heat and dry conditions. This juxtaposition creates a surreal and thought-provoking image that challenges the viewer's expectations of the natural world.

A similar example of a contradiction could be found in a scenario where a tropical bird species is depicted in a snowy landscape, or a desert plant is shown thriving in a rainforest setting. These examples also defy the natural order and create a visual paradox that can provoke curiosity and reflection on the adaptability of life and the changing conditions of our planet.

<h> A tropical bird perched on a snowy branch, surrounded by a serene, snow-covered forest, digital art, surreal, vibrant colors, peaceful, detailed, artstation </h>

VLM for Generation

Generation with different VLMs

Source Image

Try 1

Try 2

Mini-Gemini-7B-HD

Prompt: Give me a similar example of the contradiction depict in this image.

Response: A similar example of the contradiction depicted in this image could be a scene of a tropical beach during winter. In this scenario, the beach would be covered in snow, and the palm trees, which are typically associated with warm climates, would be surrounded by snow-covered sand. This would be a
contradiction because palm trees are not native to cold climates and are typically found in tropical regions where they thrive in warm, sunny environments.
<h> A serene tropical beach blanketed in snow, with palm trees standing tall amidst the frosty landscape, a stark contrast to their usual warm habitat. </h>

VLM for Generation

Generation with different VLMs

Source Image

Try 1

Try 2

Mini-Gemini-2B

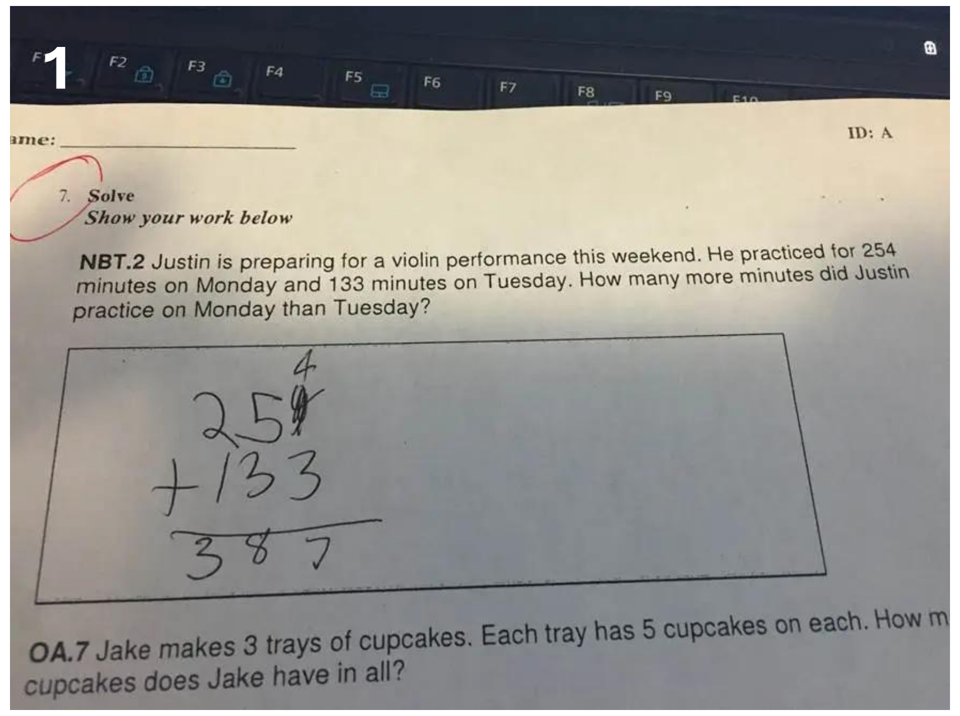
Prompt: Give me a similar example of the contradiction depict in this image.

Response: The image presents a striking contrast between the harsh, icy conditions of a desert and the life-sustaining water that is essential for life. A similar example of this contradiction could be the contrast between the arid, dry conditions of a desert and the lush, green vegetation that thrives in a rainforest. This juxtaposition highlights the resilience of life in extreme environments and the importance of water in sustaining ecosystems.

More Results

More interesting demos.

1



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?

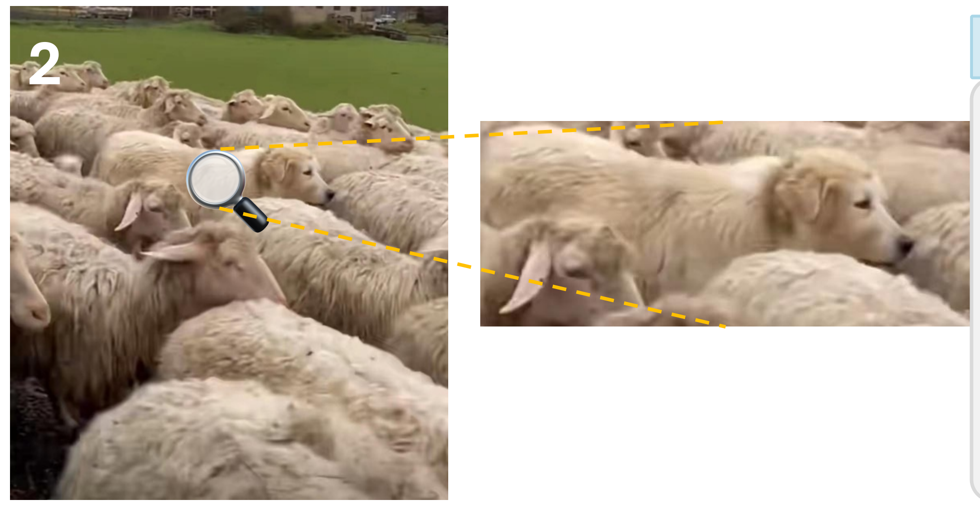
The student has attempted to solve the math problem by adding the number of minutes Justin practiced on Monday and Tuesday.

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



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

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

Comparisons with leading models.

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

Thanks!

[Project] [Code] [Demo] [Paper] [Model]