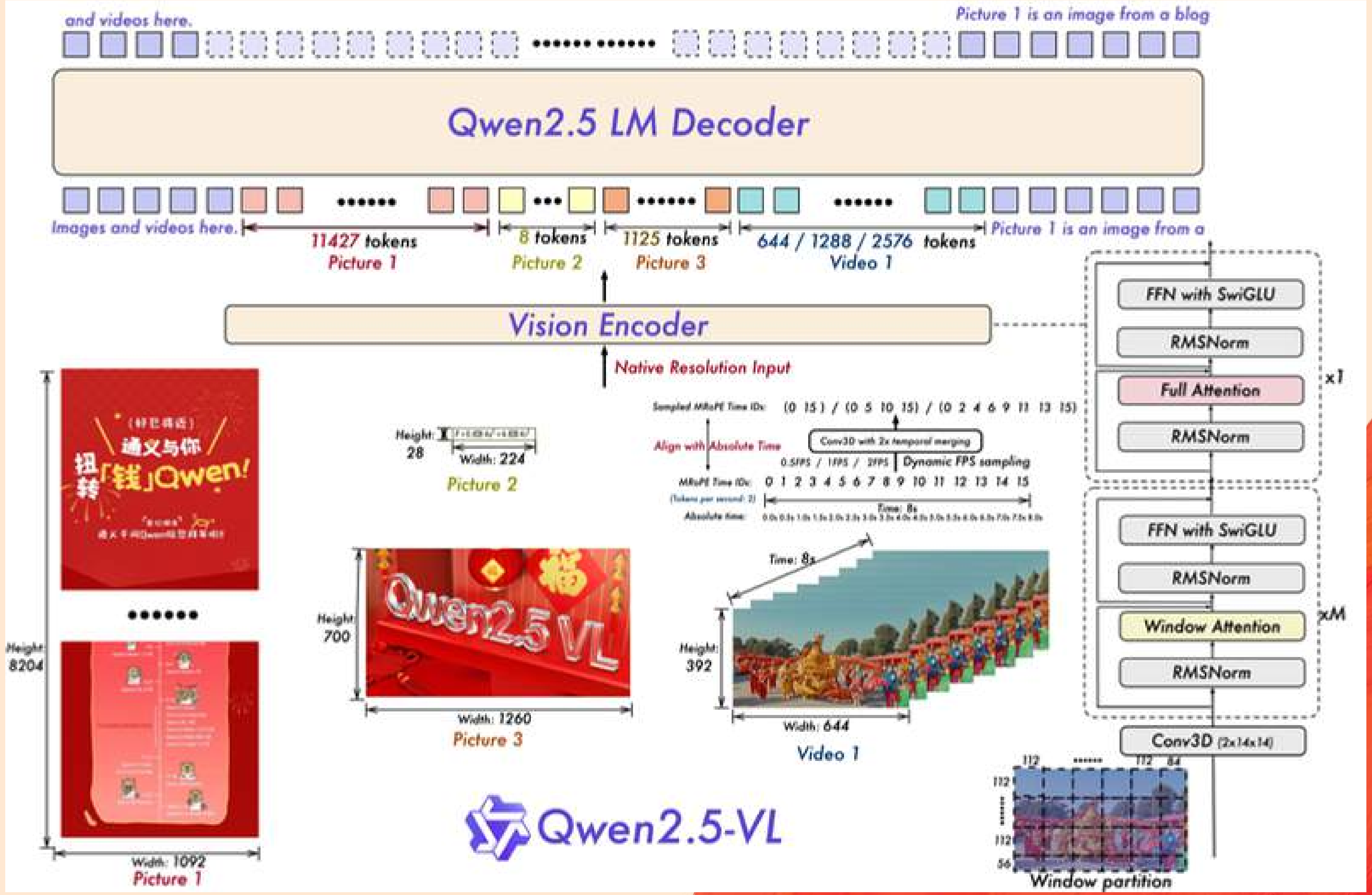
Qwen2.5 -VL

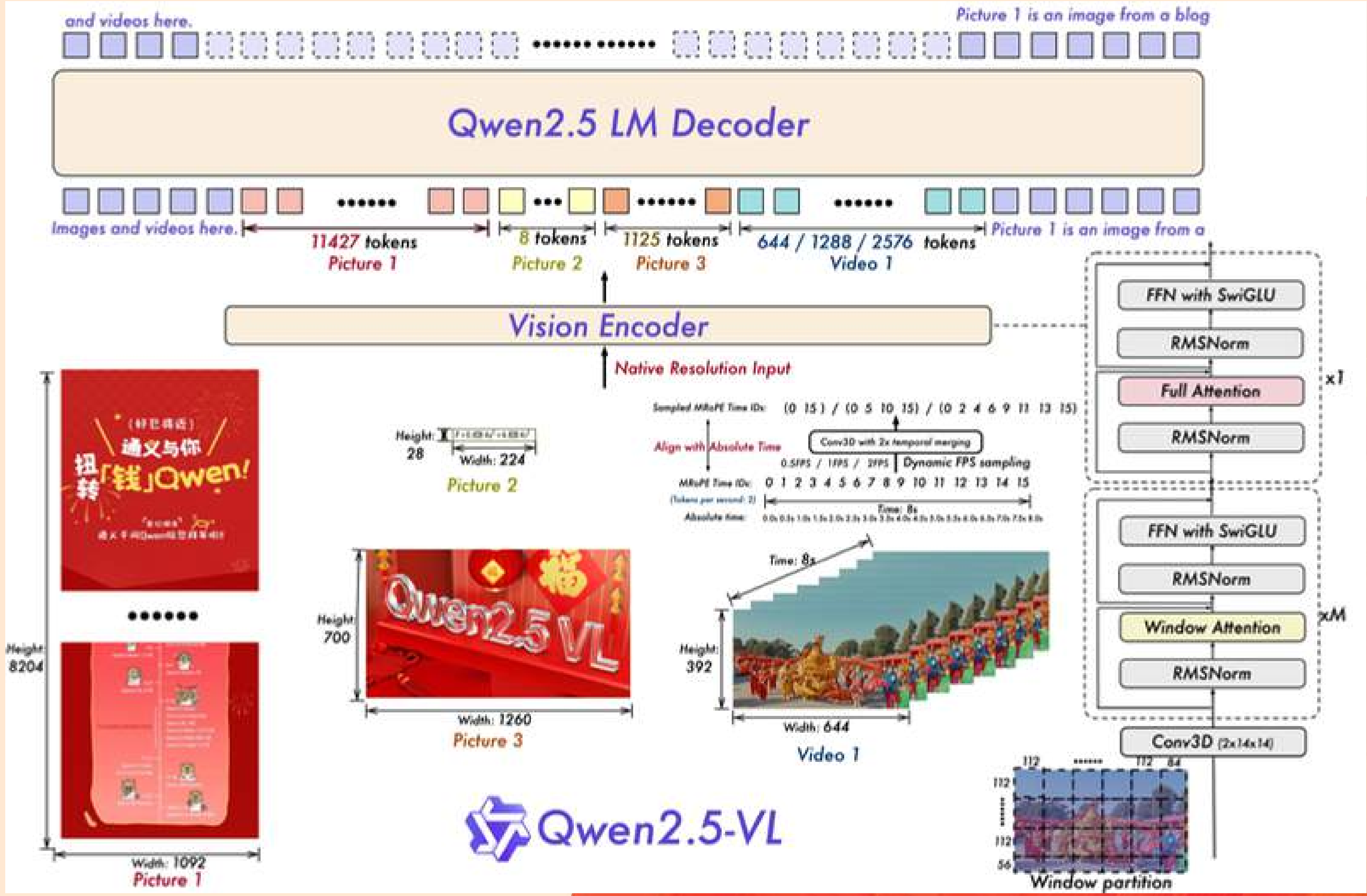
Technical Report

What is it being done?

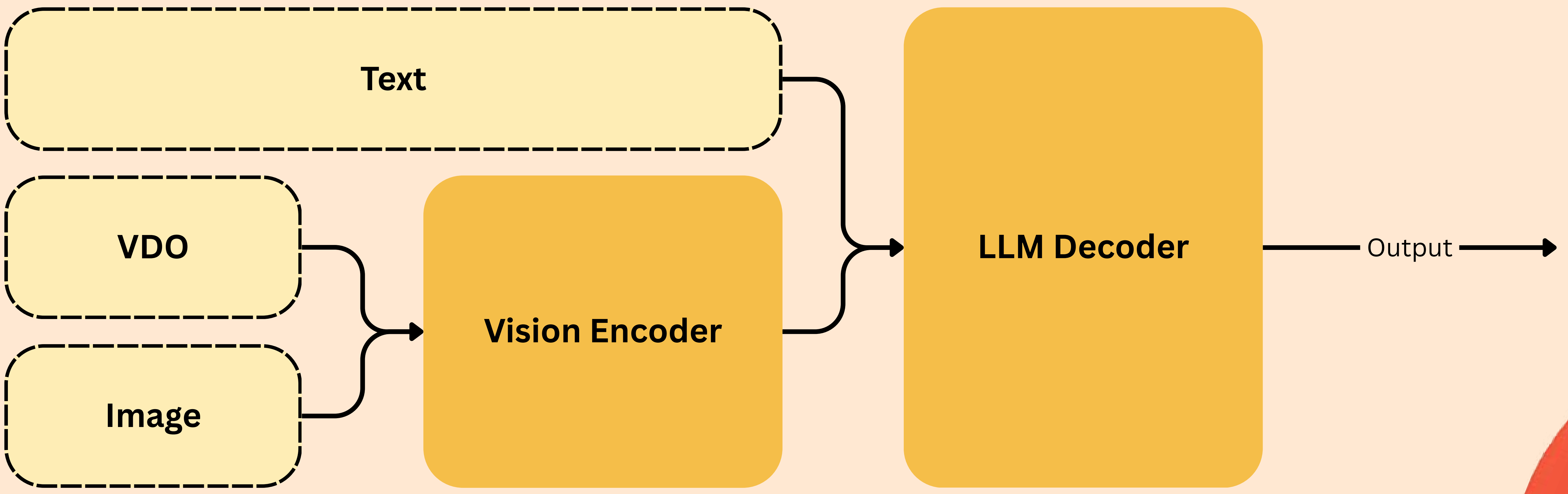
Qwen2.5-VL is a new Large vision-language models (LVLMs) that significantly advances visual understanding and interaction capabilities.

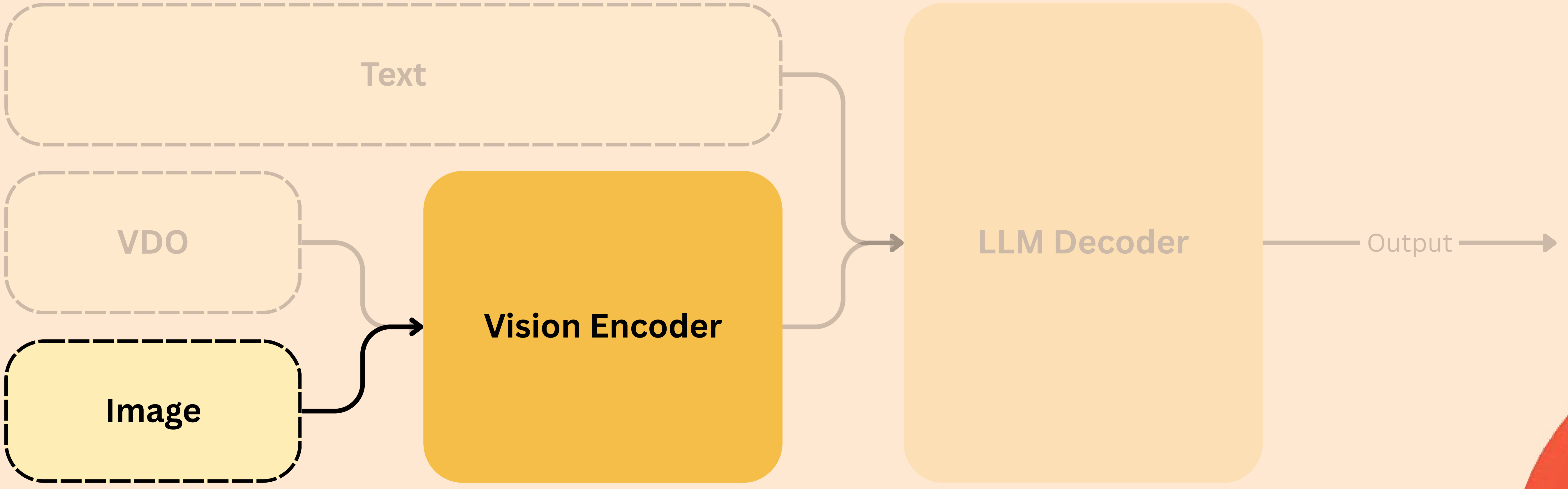


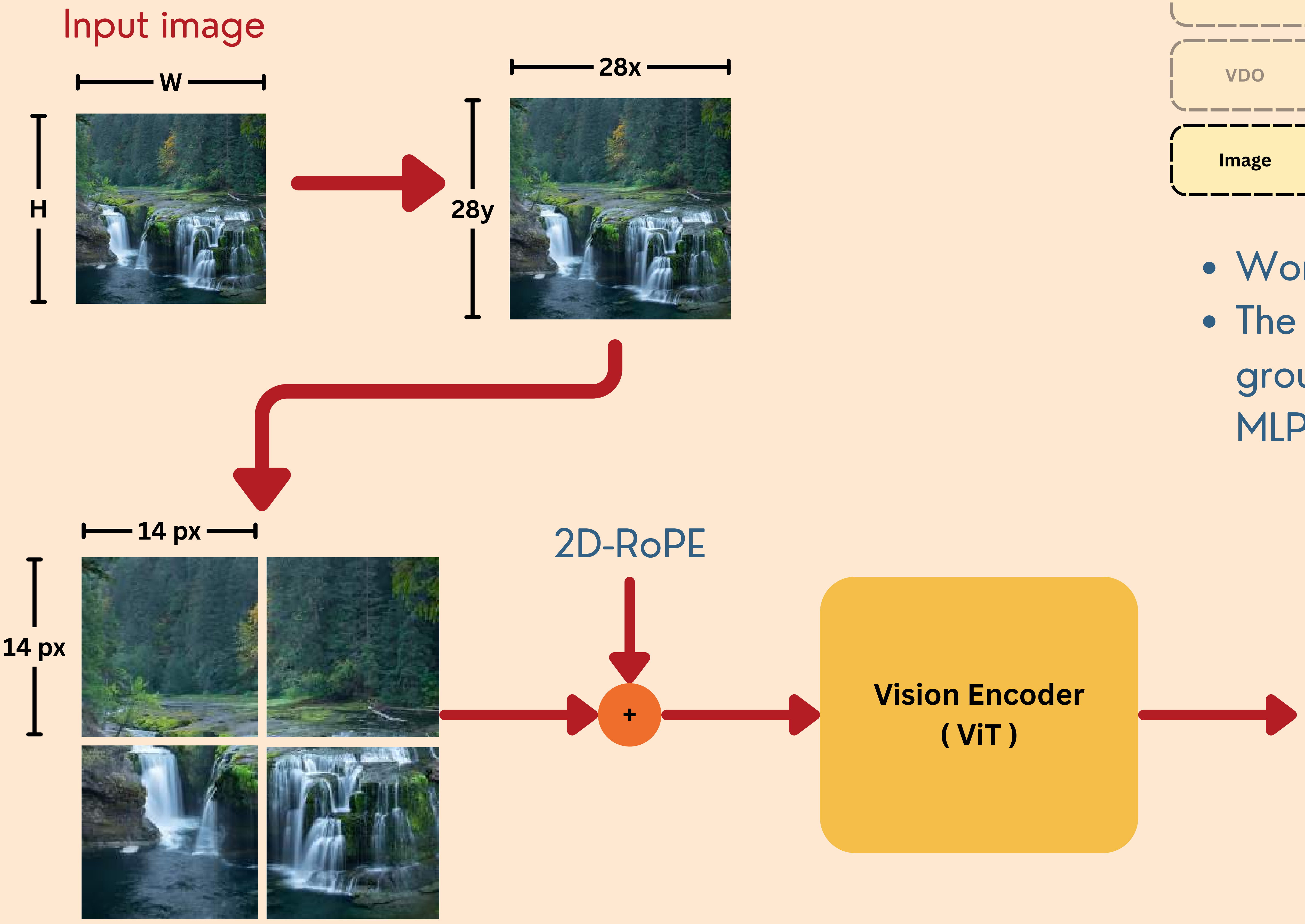
- Major advancements include enhanced visual recognition, precise object localization, robust document parsing, and long-video comprehension.
- Native dynamic-resolution processing and absolute time encoding for improved spatial and temporal understanding.

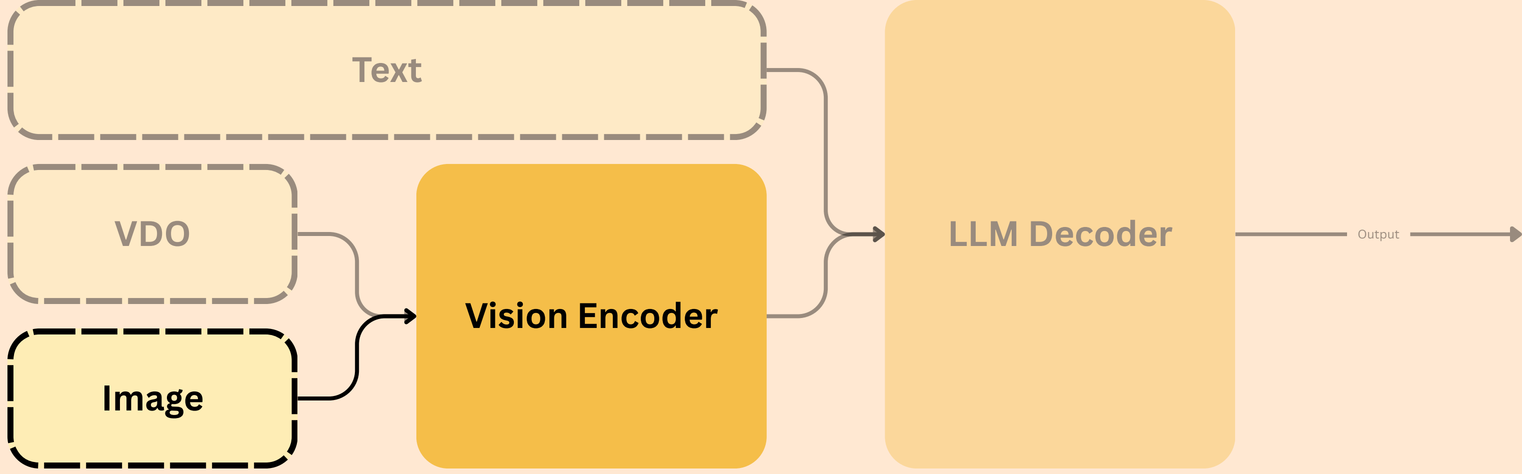


How is it being done?

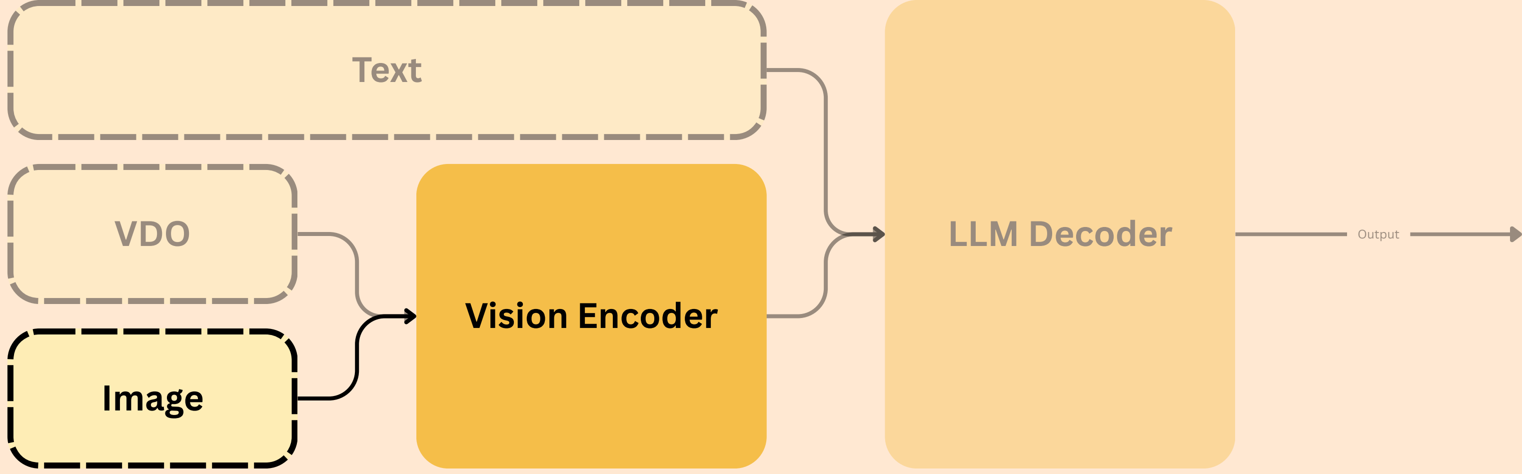

Text
VDO
Image
Vision Encoder
LLM Decoder
Output


Text
VDO
Image
Vision Encoder
LLM Decoder
Output

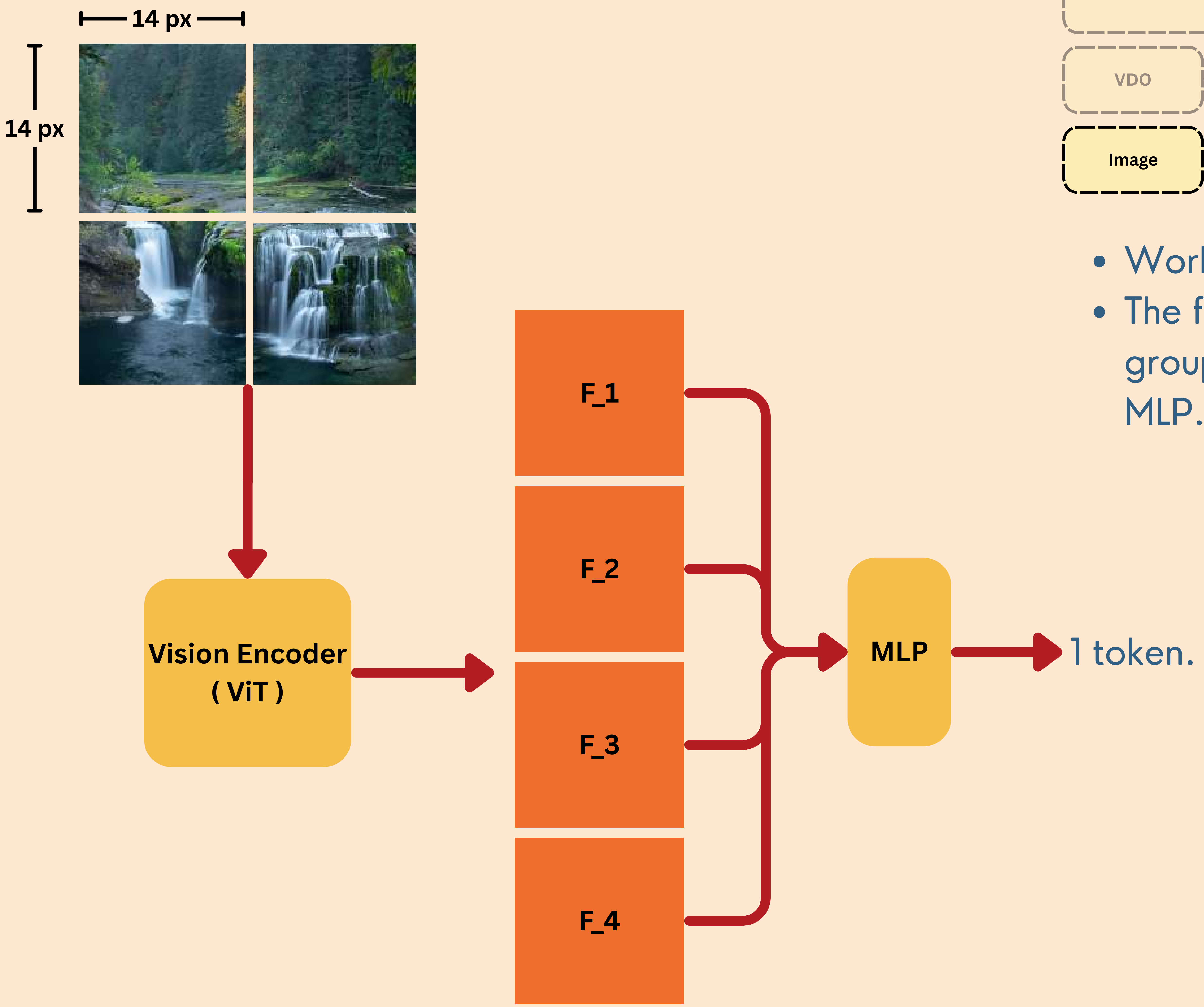


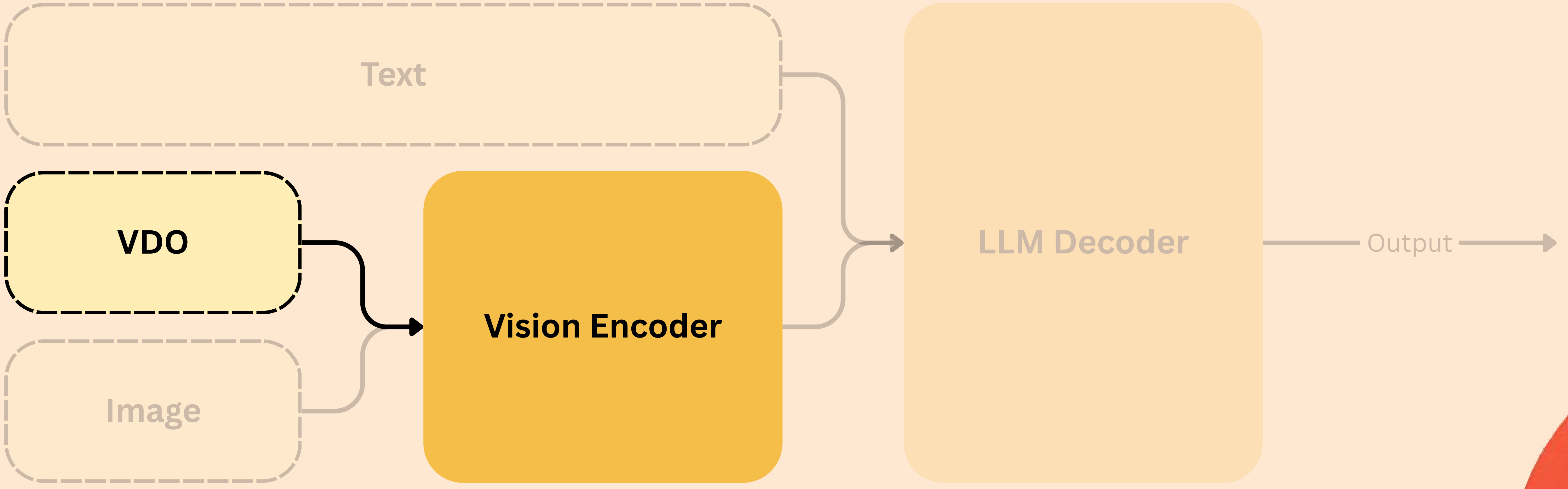


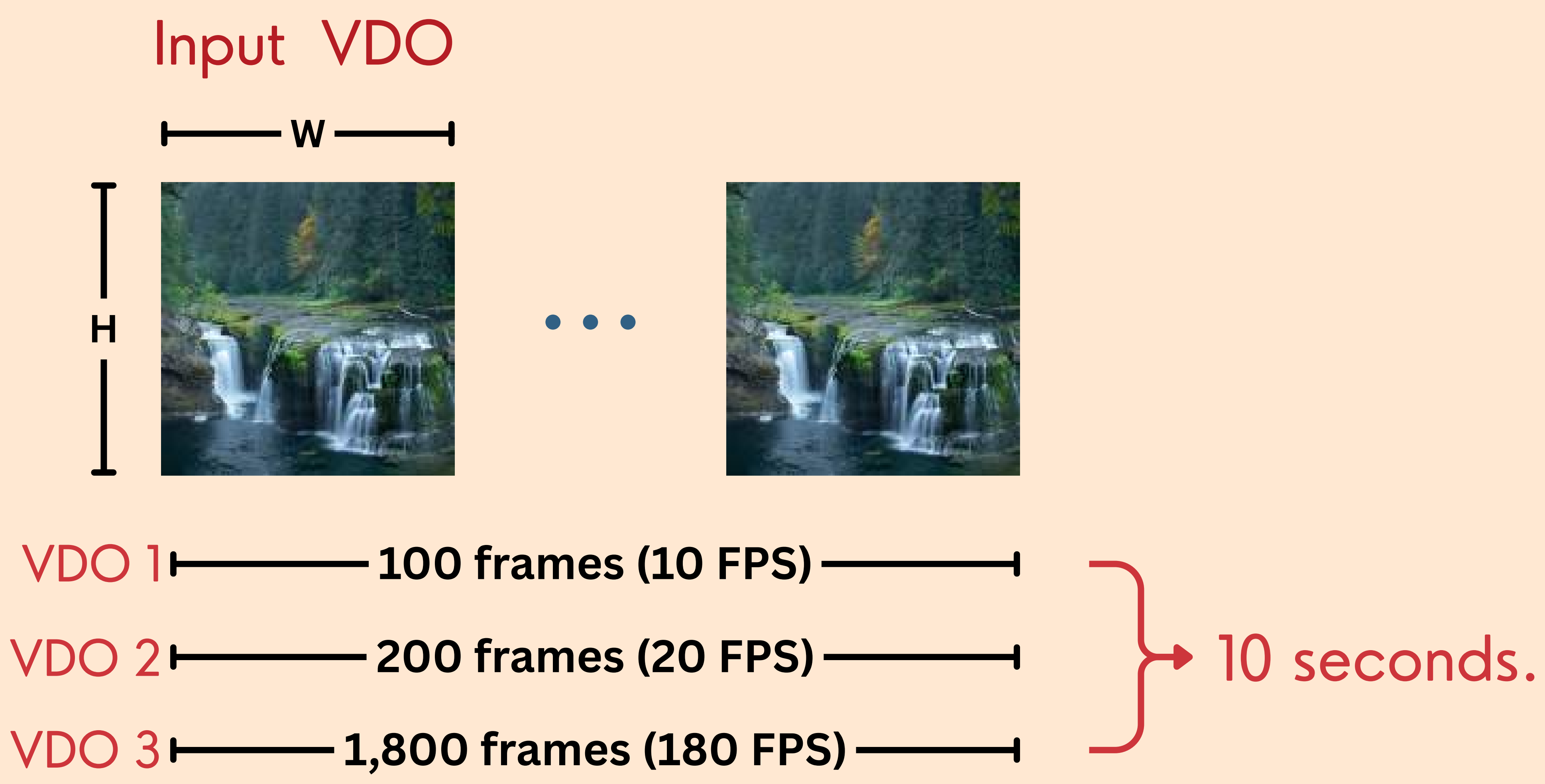
- Work with native resolution.
- The features of 4 adjacent patches grouped together then projects with MLP.

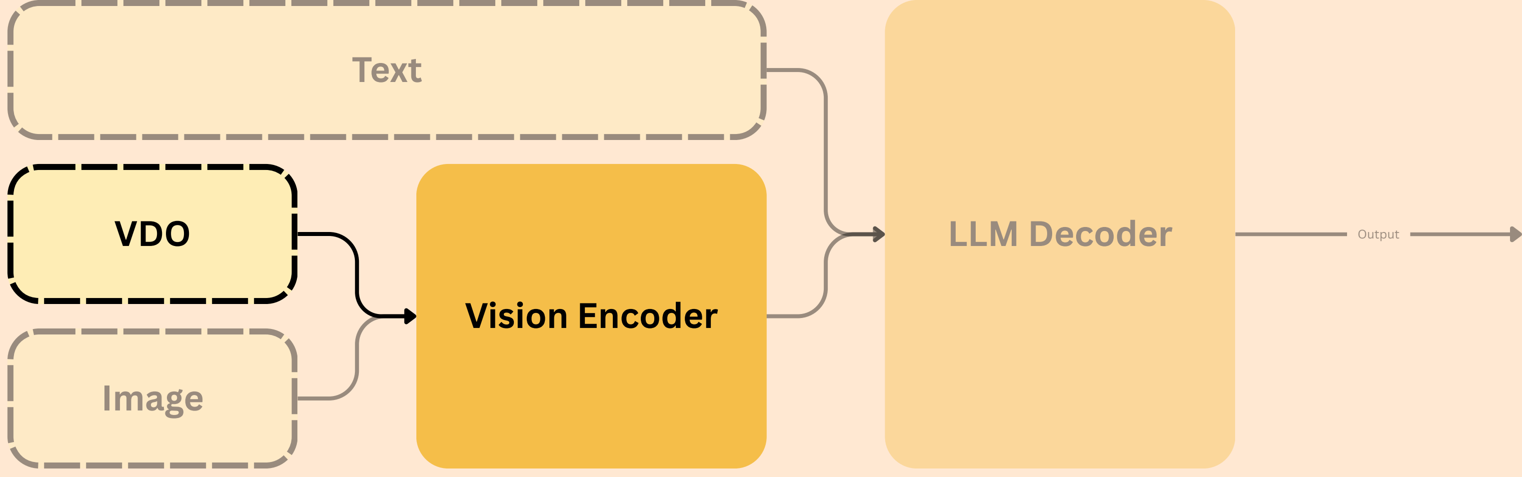


- Work with native resolution.
- The features of 4 adjacent patches grouped together then projects with MLP.



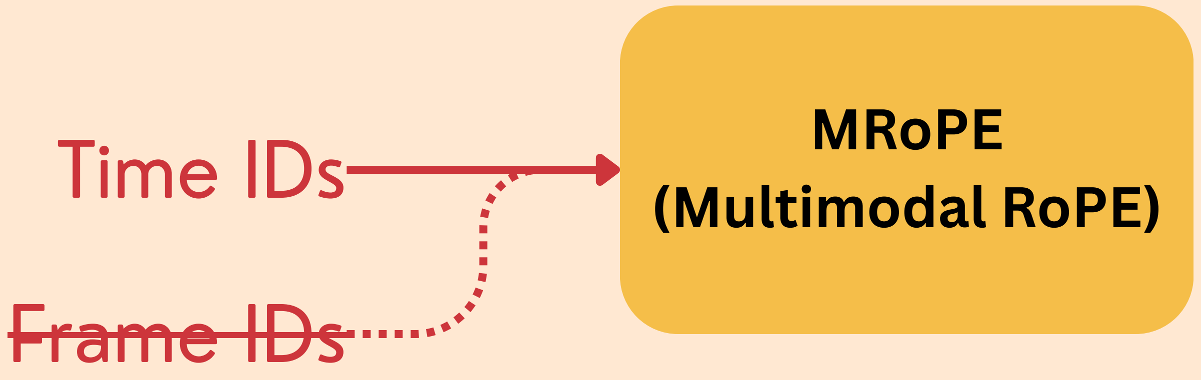

Text
VDO
Image
Vision Encoder
LLM Decoder
Output





- Work with native resolution.
- Support dynamic FPS sampling.
- Positional Encoding with absolute time.
- The features of 2 consecutive frames are group together.

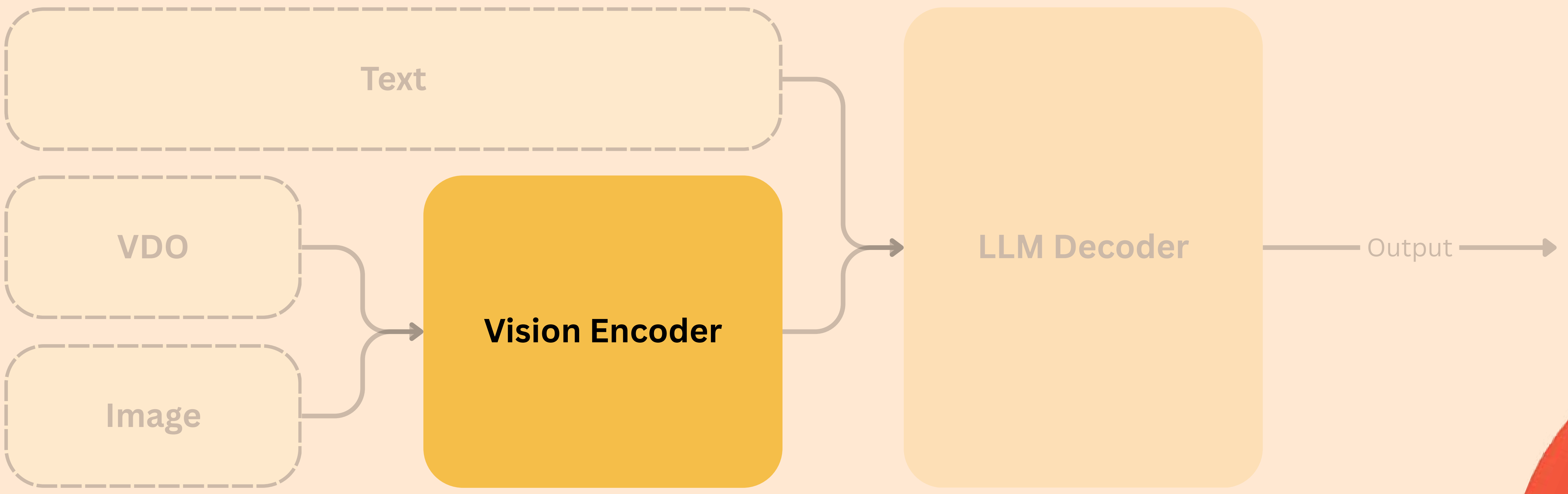
!!! So, we use **Time IDs** instead of **Frame IDs** !!!

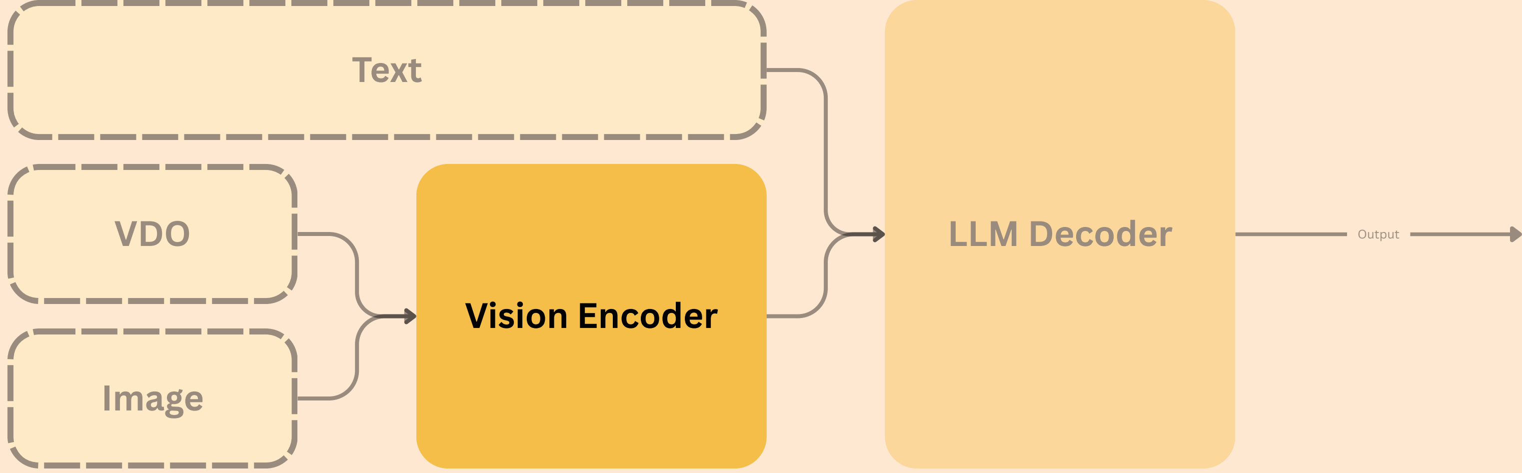


Explaination

MRoPE (Multimodal RoPE)

	Width	Height	Temporal
Text	Position IDs	Position IDs	Position IDs
Image	Image width	Image height	Constants
Video	Video width	Video height	Time IDs


Text
VDO
Image
Vision Encoder
LLM Decoder
Output



- Apply LLM design into the Vision Transformer
  - Activation function → **SwiGLU**
  - Normalization → **RMSNorm (Root Mean Square Normalization)**

- window-based attention to reduce computational efforts
  - window size = 112x112 (**8x8 patches**), no padding if smaller

HOW IS IT DIFFERENT FROM PREVIOUS WORK

หัวข้อ	**Transformer ปกติ**	**Qwen2.5-VL**
Input	Text หรือ Image (fixed)	Text + Image + Video (dynamic)
Position Encoding	1D RoPE	3D MRoPE (temporal + spatial)
Attention	Full Attention	Windowed Attention + Selective full attention (4X)
ViT	Standard ViT	Custom ViT + RMSNorm + SwiGLU
Multi-modal Fusion	Concatenation/linear	MLP Merger (compressed patches)
Video Support	จำกัด ความยาววิดีโอ	รองรับ long video, absolute time encoding
Image Resolution	Resize required	รองรับ native resolution

Dataset

PRE-TRAINING DATA

Stages	Visual Pre-Training	Multimodal Pre-Training	Long-Context Pre-Training
Data	Image Caption Knowledge OCR	+ Pure text Interleaved Data VQA, Video Grounding, Agent	+ Long Video Long Agent Long Document
Tokens	1.5T	2T	0.6T
Sequence length	8192	8192	32768
Training	ViT	ViT & LLM	ViT & LLM

- Approximately 4 trillion tokens.
- It includes cleaned web data, synthetic data, and multimodal sources such as image captions, interleaved image-text pairs, OCR data, visual knowledge datasets, multimodal academic questions, localization and grounding datasets, document parsing data, video descriptions, and agent interaction data.

INTERLEAVED IMAGE-TEXT DATA

Purpose:

- Enables multimodal learning by offering both visual and textual cues simultaneously.
- Ensures that the model maintains robust text-only capabilities when images are absent.
- Provides a wide range of general information, even though raw interleaved data can be noisy.

Example not real dataset.

INTERLEAVED IMAGE-TEXT DATA

Preprocessing

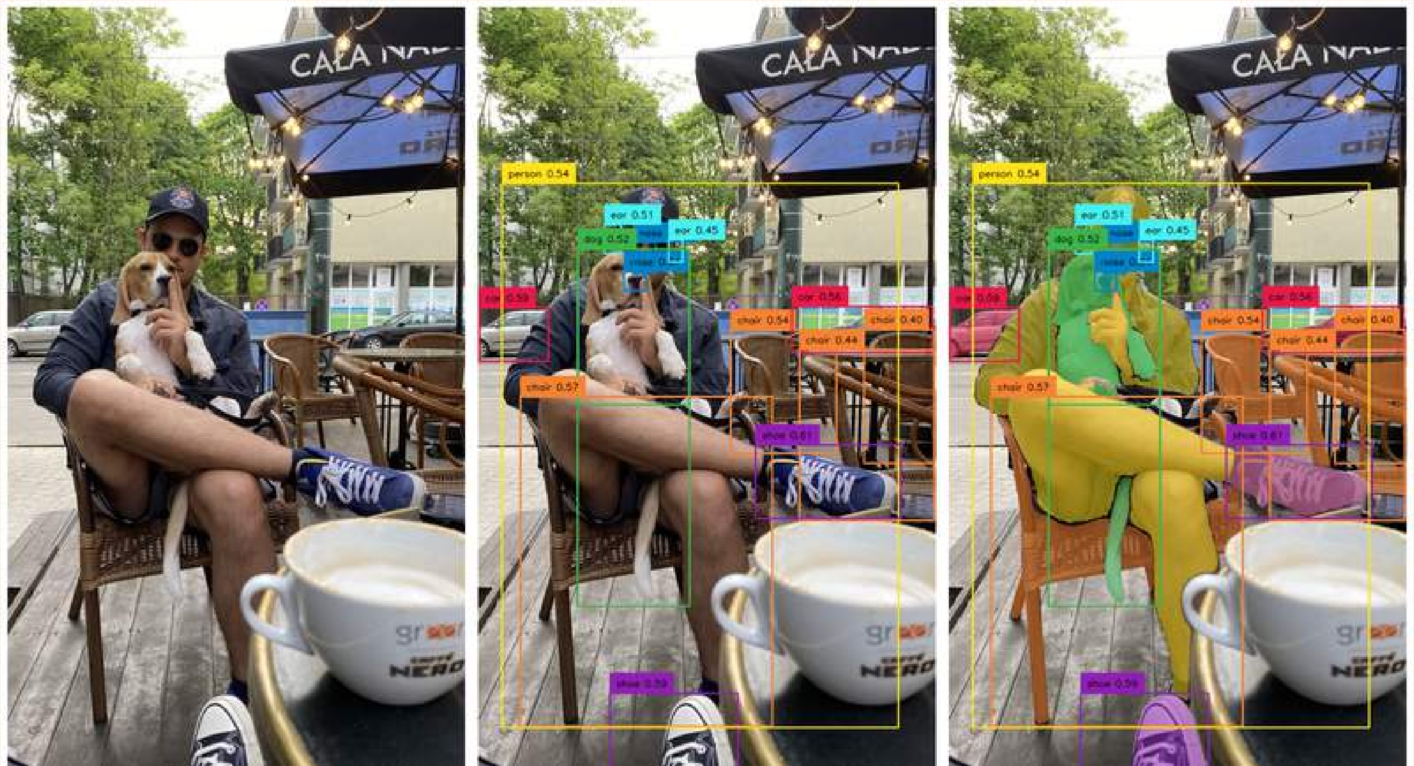
- Scoring System:
  - A four-stage evaluation using an internal model based on:
    1. Text-Only Quality
    2. Image-Text Relevance – Ensuring that the image adds meaningful context.
    3. Information Complementarity – Both modalities provide unique, complementary information.
    4. Information Density Balance – Balancing the amount of information from both the image and the text.
- This scoring helps in filtering and selecting only high-quality, useful image-text pairs.

GROUNDING DATA WITH ABSOLUTE POSITION COORDINATES

Purpose:

- To accurately capture the size and location of objects within images.
- Supports tasks such as object detection and localization by preserving the real-world scale.

-Includes over 10,000 object categories.
-Synthesizes non-existent object categories and constructs images with multiple instances per object for robustness.



Example not real dataset.

DOCUMENT OMNI-PARSING DATA

Purpose:

- To enable the model to parse and understand complex document layouts and elements, including text, tables, charts, formulas, images, and more.

- Uniformly formatted in HTML.

QwenVL HTML Format

```
<html><body>
# paragraph
<p data-bbox="x1 y1 x2 y2"> content </p>
# table
<style>table{id} style</style><table data-bbox="x1 y1 x2 y2" class="table{id}"> table content
</table>
# chart
<div class="chart" data-bbox="x1 y1 x2 y2"> <img data-bbox="x1 y1 x2 y2" /><table> chart content
</table></div>
# formula
<div class="formula" data-bbox="x1 y1 x2 y2"> <img data-bbox="x1 y1 x2 y2" /> <div> formula
content </div></div>
# image caption
<div class="image caption" data-bbox="x1 y1 x2 y2"> <img data-bbox="x1 y1 x2 y2" /><p> image
caption </p></div>
# image ocr
<div class="image ocr" data-bbox="x1 y1 x2 y2"> <img data-bbox="x1 y1 x2 y2" /><p> image ocr
</p></div>
# music sheet
<div class="music sheet" format="abc notation" data-bbox="x1 y1 x2 y2"> <img data-bbox="x1 y1
x2 y2" /> <div> music sheet content </div></div>
# chemical formula content
<div class="chemical formula" format="smile" data-bbox="x1 y1 x2 y2"> <img data-bbox="x1 y1
x2 y2" /> <div> chemical formula content </div></div>
</html></body>
```

OCR DATA

Purpose:

- To improve the model's ability to recognize and process textual content in images.
- Supports multilingual OCR capabilities.

-Includes languages such as French, German, Italian, Spanish, Portuguese, Arabic, Russian, Japanese, Korean, and Vietnamese.
-1 million samples are synthesized, 6 million real-world samples

Example not real dataset.

VIDEO DATA

Purpose:

- To ensure the model can robustly understand video content across varying frame rates (FPS) and long-duration videos.

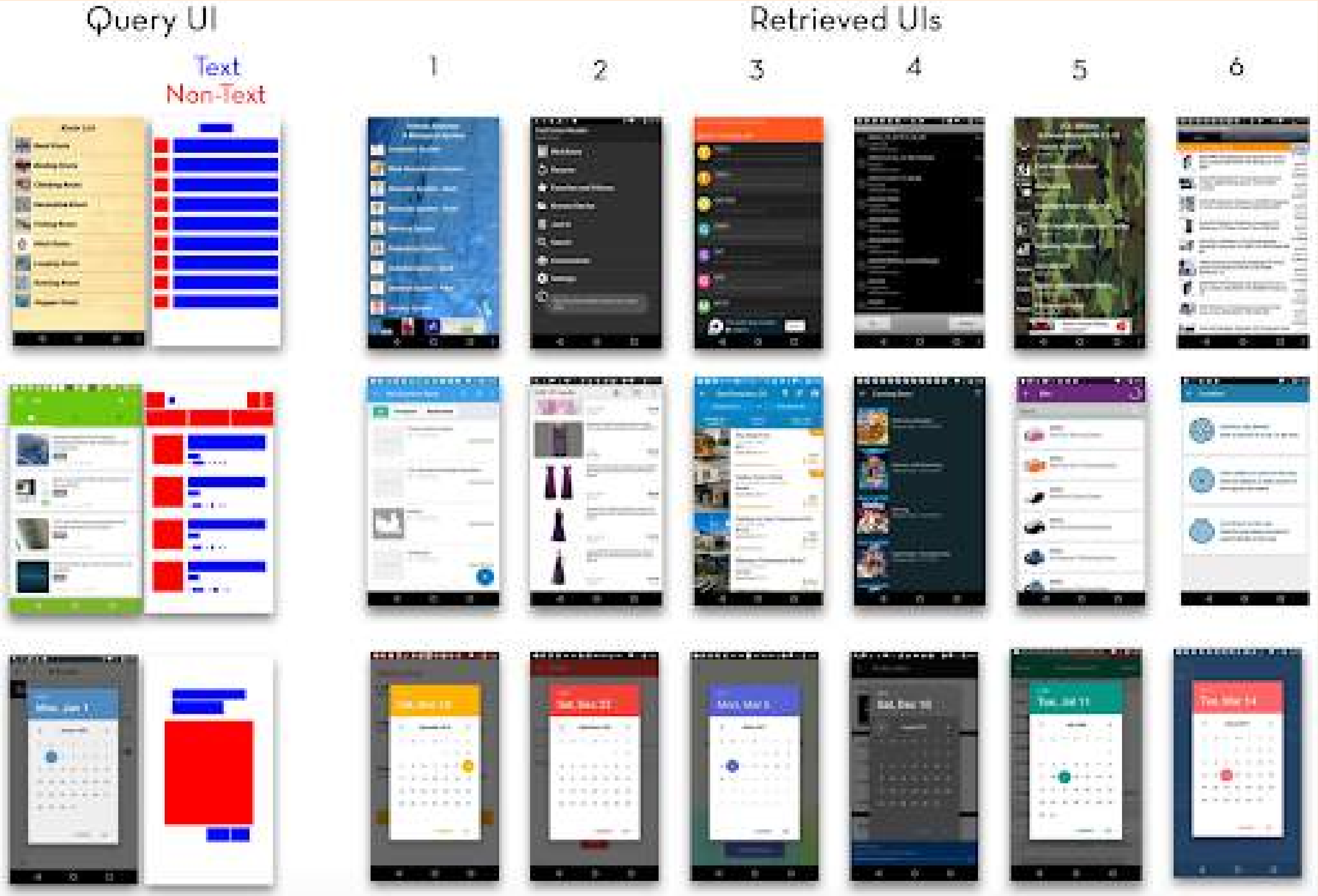
Example not real dataset.

AGENT DATA

Purpose:

- Enhances the model's perception and decision-making abilities in user interface (UI) contexts across various platforms (mobile, web, desktop).

-Screenshots from different platforms.



Example not real dataset.

INSTRUCTION DATA

- Used in the Supervised Fine-Tuning (SFT) phase.
- Approximately 2 million entries.
- Pure text data (50%) and multimodal data (50%).
- Primary Languages is Chinese and English and has additional multilingual entries.
- Data is structured to support a wide range of queries, ensuring broad coverage.

INSTRUCTION DATA

Dialogue Complexity:

- Single-Turn Interactions: Simpler, one-shot inputs where the query is addressed in a single interaction.
- Multi-Turn Interactions: More complex, where context is maintained over several conversational turns.
- Visual Dynamics: The data simulates realistic scenarios ranging from single-image inputs to sequences of multiple images, adding layers of context and complexity.

-

INSTRUCTION DATA

include:

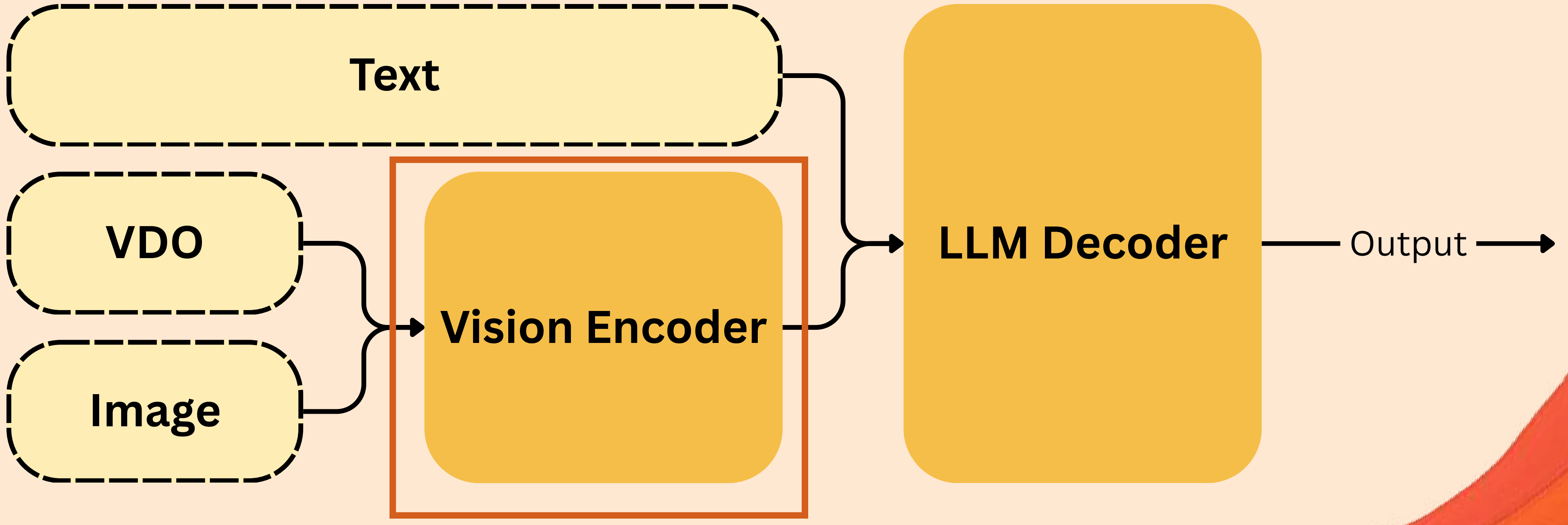
- General Visual Question Answering (VQA) and Image Captioning
- Mathematical Problem-Solving and Coding Tasks
- Security-Related Queries
- Document and OCR Tasks
- Grounding and Video Analysis
- Agent Interactions (focused on UI and operational decision-making)

-

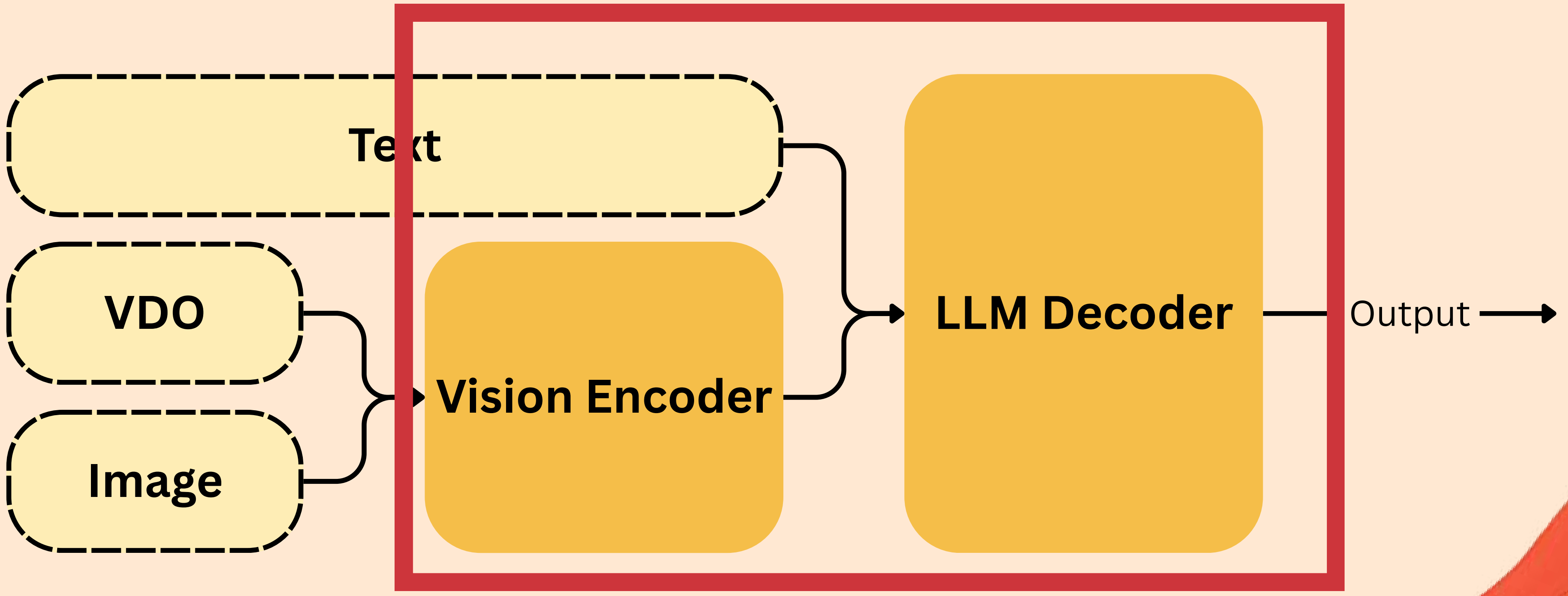
Training and fine-tuning

PHASE 1: VISION ENCODER TRAINING

- Only Vision Encoder training is trained in this phase
- Datasets such as image captions, visual knowledge, and OCR data are used to train in this phase



PHASE 2,3: ENTIRE MODEL TRAINING



- All model parameters are unfreeze.
- Datasets such as interleaved data, multi-task learning, visual question answering, multimodal mathematics, agent-based task, video understanding and pure text are introduce to train in the 2nd phase.
- The 3rd phase introduce data that has longer token lengths compared to 2nd phase. (8192 tokens compares to 32768)

FINE-TUNING

- Instruction-response based data are used to fine tune the model.
- Both single-turn and multi-turn interactions are introduced.
- Domain-specific categorization, Domain tailored filtering, Rule-Based filtering, and Model-Based filtering are applied to ensure dataset quality before the fine-tuning process


```
{"instruction": "What is Agentic AI?", "response": "Agentic AI is a type of artificial intelligence ch
{"instruction":"What is the core characteristic of Agentic AI that sets it apart from other types of A
{"instruction":"How does Agentic AI gather information about its environment?","context":"","response"
{"instruction":"What is the purpose of the reasoning stage in Agentic AI?","context":"","response":"To
{"instruction":"Can you describe a scenario where Agentic AI would be useful?","context":"","response"
{"instruction":"How does Agentic AI improve over time?","context":"","response":"Through the learning
{"instruction":"What are the four key stages of Agentic AI, and how do they relate to each other?","co
{"instruction":"How does Agentic AI differ from traditional AI systems that require constant human inp
{"instruction":"Can Agentic AI learn from its mistakes?","context":"","response":"Yes, through the lea
{"instruction":"What are some potential applications of Agentic AI in industries such as healthcare or
{"instruction": "How does Agentic AI maintain balance between autonomous decision-making and safety co
{"instruction": "How do the four stages interact with each other?", "context": "Agentic AI Operational
{"instruction": "What distinguishes Agentic AI's autonomy from traditional AI systems?", "context": "A
{"instruction": "Which stage is most critical for overall system performance?", "context": "Agentic AI
{"instruction": "How does autonomous learning contribute to improved decision-making?", "context": "Ag
{"instruction": "How does the perception stage filter relevant information?", "context": "Agentic AI O
{"instruction": "What role does initial programming play in autonomous behavior?", "context": "Agentic
{"instruction": "What role does reasoning play in decision optimization?", "context": "Agentic AI Oper
{"instruction": "How is autonomy measured or quantified in Agentic AI systems?", "context": "Agentic A
{"instruction": "How does the action stage implement decisions?", "context": "Agentic AI Operational S
{"instruction": "What are the ethical implications of AI autonomy?", "context": "Agentic AI Autonomy",
{"instruction": "What feedback mechanisms exist between stages?", "context": "Agentic AI Operational S
```

Evaluation

Overview

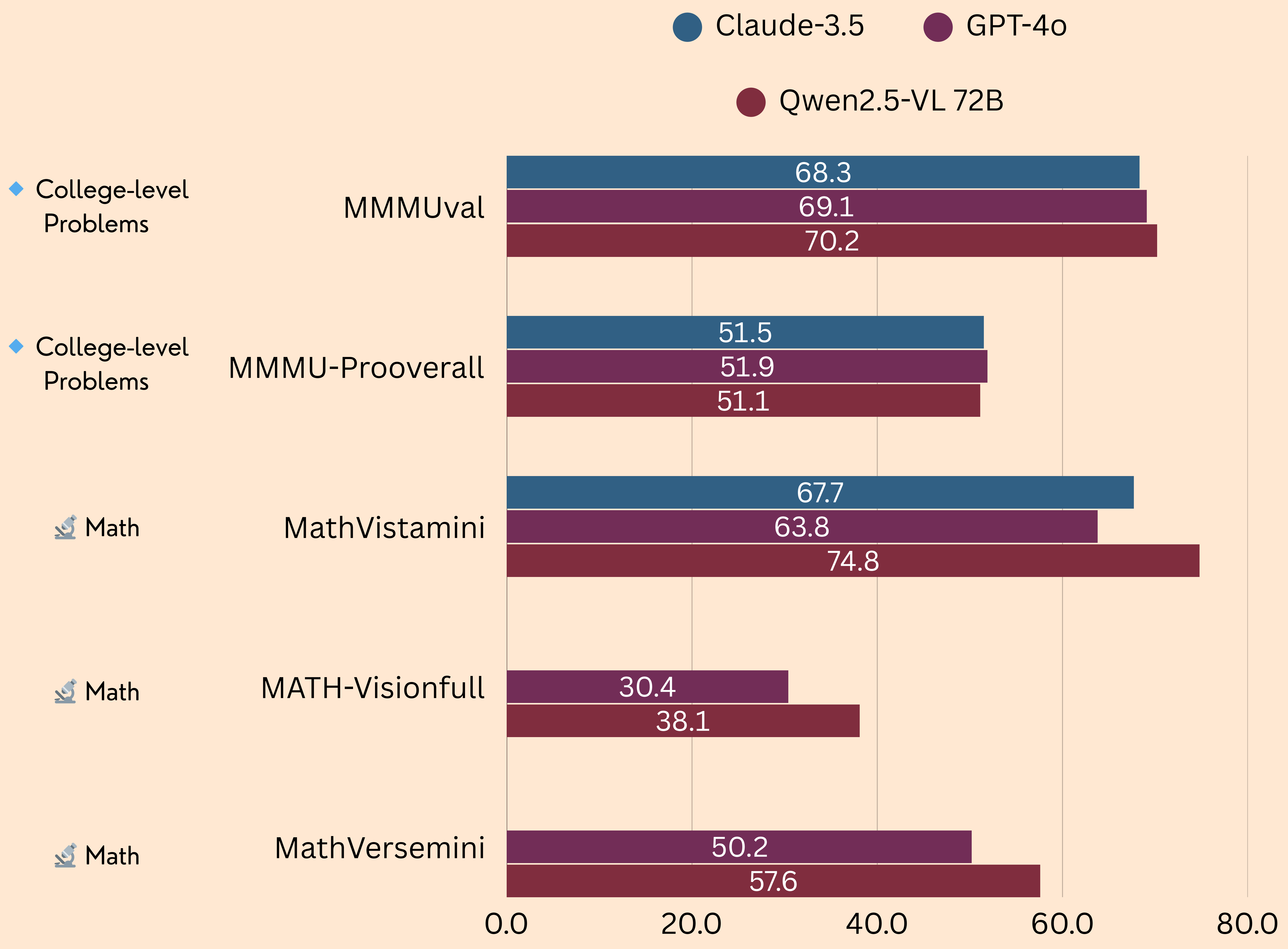
On Evaluation Task

1. General Language Understanding
2. Multimodal Tasks

Qwen2.5-VL-72B

VS

LLaMA 3.1 405B
GPT-4
Claude 3.5
Gemini 1.5
และ SOTA อื่น ๆ



General Language Understanding

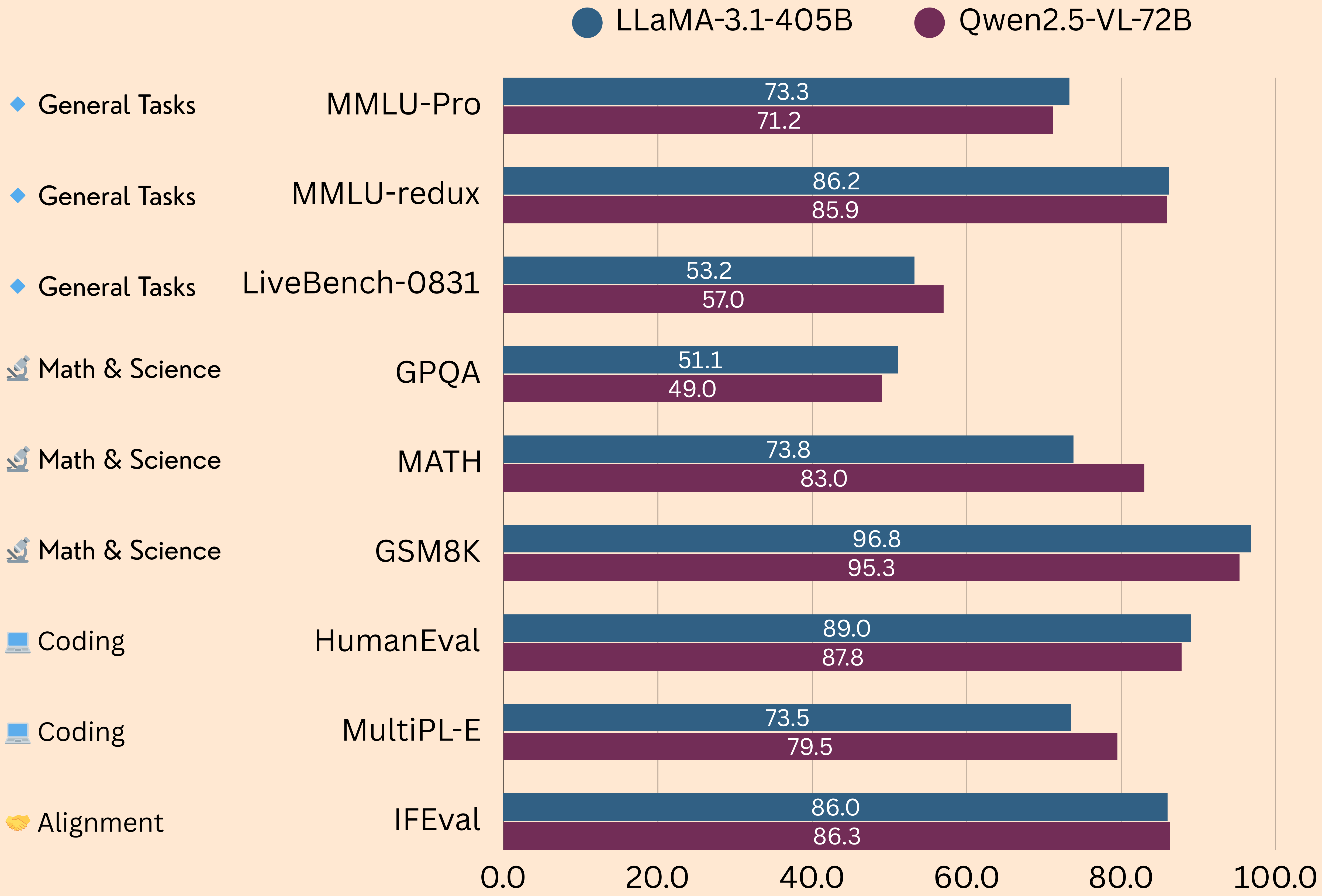
GPO-4o

- Evenly matched with GPO-4o in College-level Problems
- Qwen outperform in Math Problem

Claude-3.5

- Qwen outperform Claude-3.5 in all dataset

General Language Understanding



Qwen2.5-VL-72B vs LLaMA-3.1-405B

- Most of the dataset, The performance is Evenly matched
- Better performance in MATH dataset

Multimodel Task

OCR-related Document Understading

- Best performace in 6 out of 12 dataset

Table 5: **Performance of Qwen2.5-VL and other models on OCR, chart, and document understanding benchmarks.**

Datasets	Claude-3.5 Sonnet	Gemini 1.5 Pro	GPT 4o	InternVL2.5 78B	Qwen2.5-VL 72B	Qwen2.5-VL 7B	Qwen2.5-VL 3B
OCR-related Parsing Tasks							
CC-OCR	62.5	73.0	66.9	64.7	**79.8**	77.8	74.5
$\text{OmniDocBench}_{\text{edit en/zh}}\downarrow$	0.330/0.381	0.230/**0.281**	0.265/0.435	0.275/0.324	**0.226**/0.324	0.308/0.398	0.409/0.543
OCR-related Understanding Tasks							
$\text{AI2D}_{\text{w. M.}}$	81.2	88.4	84.6	**89.1**	88.7	83.9	81.6
$\text{TextVQA}_{\text{val}}$	76.5	78.8	77.4	83.4	83.5	**84.9**	79.3
$\text{DocVQA}_{\text{test}}$	95.2	93.1	91.1	95.1	**96.4**	95.7	93.9
$\text{InfoVQA}_{\text{test}}$	74.3	81.0	80.7	84.1	**87.3**	82.6	77.1
$\text{ChartQA}_{\text{test Avg.}}$	**90.8**	87.2	86.7	88.3	89.5	87.3	84.0
$\text{CharXiv}_{\text{RQ/DQ}}$	**60.2**/84.3	43.3/72.0	47.1/84.5	42.4/82.3	49.7/**87.4**	42.5/73.9	31.3/58.6
SEED-Bench-2-Plus	71.7	70.8	72.0	71.3	**73.0**	70.4	67.6
OCRBench	788	754	736	854	**885**	864	797
$\text{VCR}_{\text{En-Hard-EM}}$	41.7	28.1	73.2	-	79.8	**80.5**	37.5
OCR-related Comprehensive Tasks							
$\text{OCRBench\_v2}_{\text{en/zh}}$	45.2/39.6	51.9/43.1	46.5/32.2	49.8/52.1	**61.5/63.7**	56.3/57.2	54.3/52.1

12 Datasets

+

“What is the total cost of this bill”

→ **“400 Bath”**

Multimodel Task

Counting Task

- Better than Gemini 1.5 Pro, Claude, GPT, the best performance

Table 7: **Performance of Qwen2.5-VL and other models on counting.**

Datasets	Gemini 1.5-Pro	GPT-4o	Claude-3.5 Sonnet	Molmo-72b	InternVL2.5-78B	Qwen2.5-VL-72B
CountBench	85.5	87.9	89.7	91.2	72.1	93.6

1 Dataset

“How many cows in the picture”

“4 Cows”

Multimodel Task

Grounding Task

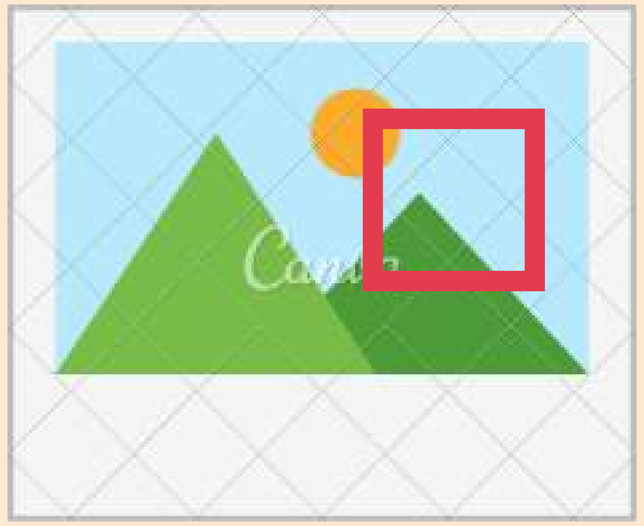
- Better than Gemini 1.5 Pro in all dataset
- But lower performance than InternVL

Table 6: **Performance of Qwen2.5-VL and other models on grounding.**

Datasets	Gemini 1.5 Pro	Grounding DINO	Molmo 72B	InternVL2.5 78B	Qwen2.5-VL 72B	Qwen2.5-VL 7B	Qwen2.5-VL 3B
Refcoco$_{val}$	73.2	90.6	-	93.7	92.7	90.0	89.1
Refcoco$_{testA}$	72.9	93.2	-	95.6	94.6	92.5	91.7
Refcoco$_{testB}$	74.6	88.2	-	92.5	89.7	85.4	84.0
Refcoco+$_{val}$	62.5	88.2	-	90.4	88.9	84.2	82.4
Refcoco+$_{testA}$	63.9	89.0	-	94.7	92.2	89.1	88.0
Refcoco+$_{testB}$	65.0	75.9	-	86.9	83.7	76.9	74.1
Refcocog$_{val}$	75.2	86.1	-	92.7	89.9	87.2	85.2
Refcocog$_{test}$	76.2	87.0	-	92.2	90.3	87.2	85.7
ODinW	36.7	55.0	-	31.7	43.1	37.3	37.5
PointGrounding	-	-	69.2	-	67.5	67.3	58.3

10 Datasets

"Where is the red car"



Multimodel Task

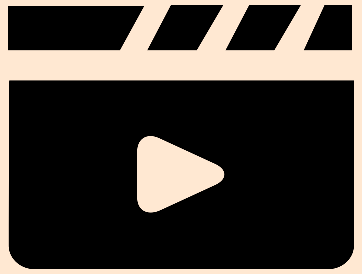
Video Task

- Better than Gemini 1.5 Pro, GPT-4o in most dataset
- Best performance in 8 out of 13 datasets

Table 8: **Performance of Qwen2.5-VL and other models on video benchmarks.**

Datasets	Gemini 1.5-Pro	GPT-4o	Qwen2.5-VL-72B	Qwen2.5-VL-7B	Qwen2.5-VL-3B
Video Understanding Tasks					
Video-MME$_{w/o\ sub.}$	**75.0**	71.9	73.3	65.1	61.5
Video-MME$_{w\ sub.}$	**81.3**	77.2	79.1	71.6	67.6
Video-MMMU	53.9	**61.2**	60.2	47.4	-
MMVU$_{val}$	65.4	**67.4**	62.9	50.1	-
MVBench	60.5	64.6	**70.4**	69.6	67.0
MMBench-Video	1.30	1.63	**2.02**	1.79	1.63
LongVideoBench$_{val}$	64.0	**66.7**	60.7	56.0	54.2
LVBench	33.1	30.8	**47.3**	45.3	43.3
EgoSchema$_{test}$	71.2	72.2	**76.2**	65.0	64.8
PerceptionTest$_{test}$	-	-	**73.2**	70.5	66.9
MLVU$_{M\text{-}Avg}$	-	64.6	**74.6**	70.2	68.2
TempCompass$_{Avg}$	67.1	73.8	**74.8**	71.7	64.4
Video Grounding Tasks					
Charades-STA$_{mIoU}$	-	35.7	**50.9**	43.6	38.8

13 Dataset



"What happen at 1:30?"

"Red car pass by"

Multimodel Task

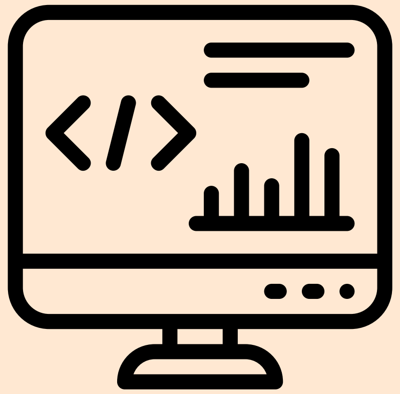
GUI Task

- Better than Gemini 2.0, GPT-4o, Claude in most dataset
- Best performance in 5 out of 7 dataset

Table 9: **Performance of Qwen2.5-VL and other models on GUI Agent benchmarks.**

Benchmarks	GPT-4o	Gemini 2.0	Claude	Aguvis-72B	Qwen2-VL-72B	Qwen2.5-VL-72B
ScreenSpot	18.1	84.0	83.0	**89.2**	-	87.1
ScreenSpot Pro	-	-	17.1	23.6	1.6	**43.6**
Android Control High$_{EM}$	20.8	28.5	12.5	66.4	59.1	**67.36**
Android Control Low$_{EM}$	19.4	60.2	19.4	84.4	59.2	**93.7**
AndroidWorld$_{SR}$	34.5% (SoM)	26% (SoM)	27.9%	26.1%	6% (SoM)	**35%**
MobileMiniWob++$_{SR}$	61%	42% (SoM)	61% (SoM)	66%	50% (SoM)	**68%**
OSWorld	5.03	4.70	**14.90**	10.26	2.42	8.83

7 Dataset



+

"Click Setting Button"

Click Setting Button

Conclusion

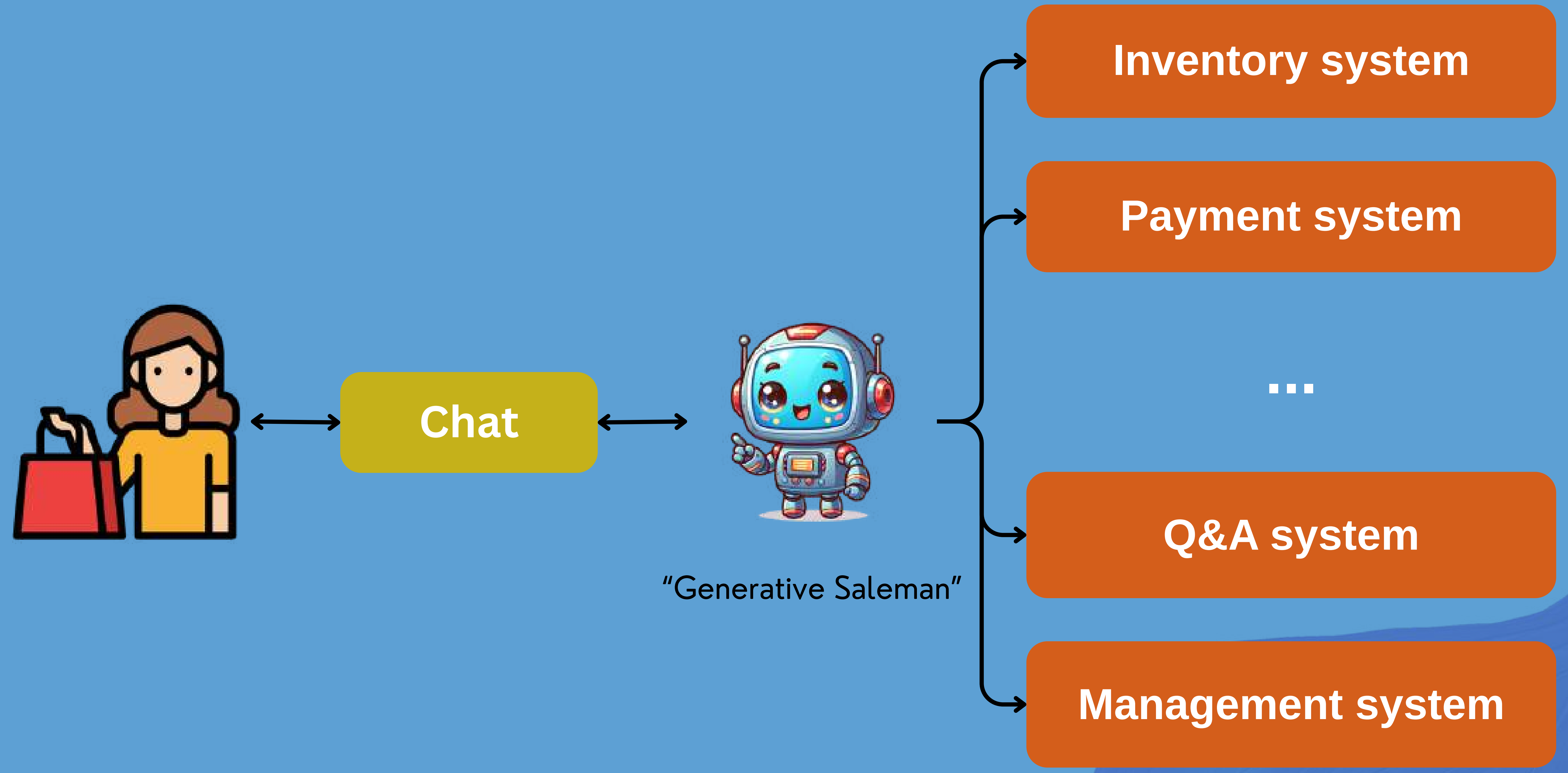
Qwen2.5-VL-72B

VS

LLaMA 3.1 405B
GPT-4
Claude 3.5
Gemini 1.5
และ SOTA อื่น ๆ

- Qwen2.5-VL-72B is clearly better than other model in Multimodel Task
- Qwen2.5-VL-72B is better perfomace in Math problem
- And have similar performance in General Language Understanding

Our Project Update



Chat
"Generative Saleman"
Inventory system
Payment system
...
Q&A system
Management system

Scope and Goal

- Thai language only
- For use exclusively in Board Game trading context only
- Not yet connected to social media chat application, initially includes only a simple and user-friendly UI

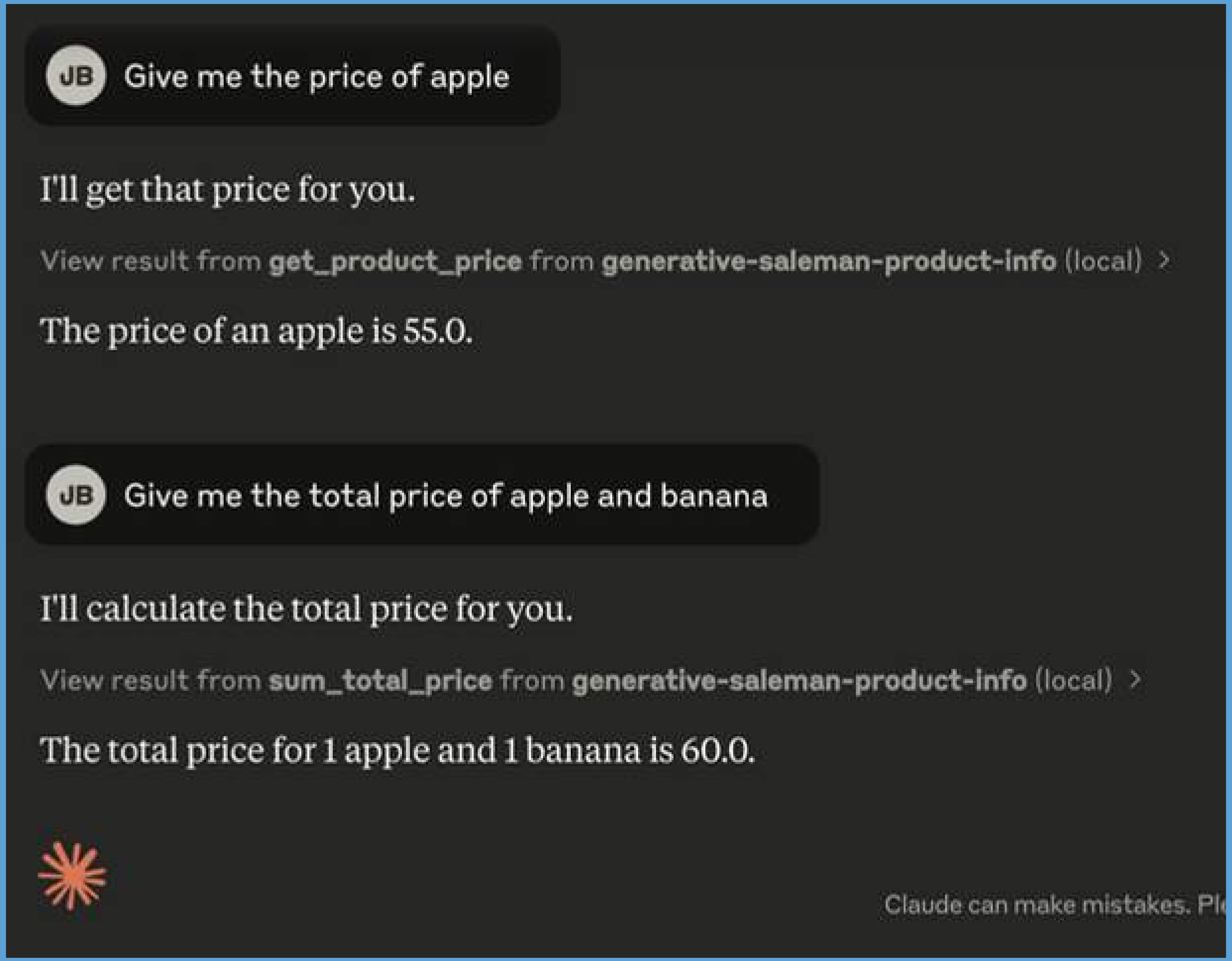
```python
# Add sum_total_price tool
@mcp.tool()
def sum_total_price(product_list: list[tuple[str, int]]) -> float:
    """
    Calculate the total price of a list of products.

    :param product_list: A list of product names which in form (product_name, quantity).
    :type product_list: list[tuple[str, int]]

    :return: The total price of the products.
    :rtype: float

    ## Example
    > sum_total_price([('apple', 1), ('banana', 1)])
    60.0
    > sum_total_price([('grape', 1), ('watermelon', 2)])
    95.0
    """
    total_price = 0
    for product, quantity in product_list:
        total_price += PRODUCT_NAME_TO_PRICE.get(product, 0) * quantity
    return total_price
```

Example code in MCP frameworks

Model Context Protocol

An open protocol that enables seamless integration between LLM applications and external data sources and tools. - Model Context Protocol

GitHub



Example usage in Claude Desktop

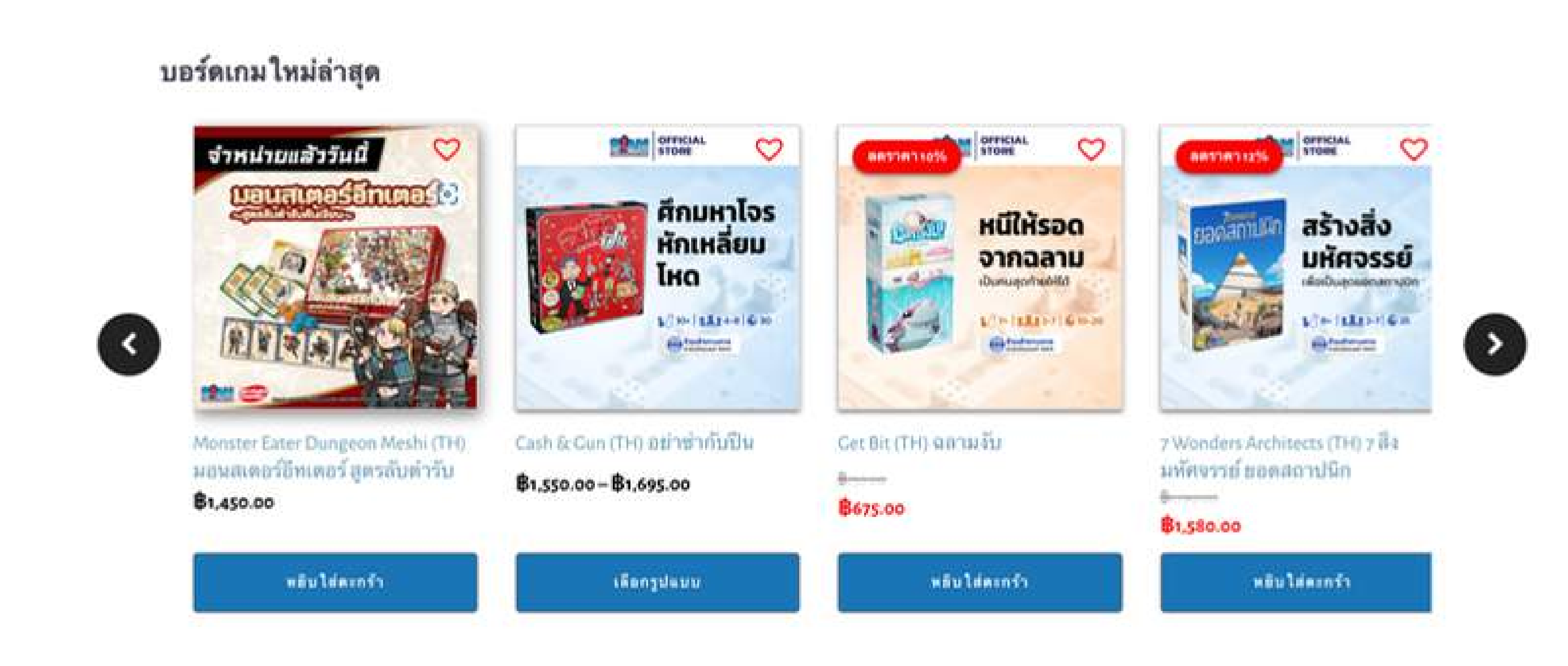
Payment Solutions

- Generate QR code from API https://promptpay.io/
- Using API from https://openslipverify.com/ to verify slip


```
{'success': True,
 'statusMessage': 'SUCCESS',
 'data': {'receivingBank': '004',
  'sendingBank': '004',
  'transRef': '0450855o5op7poh5wDDM',
  'transDate': '20250326',
  'transTime': '15:17',
  'sender': {'displayName': 'นาย ศุภฤกษ์ ค',
   'name': 'MR. SUPHAROEK K',
   'account': {'value': 'xxx-x-x8695-x'}},
  'billerID': '010753600031508',
  'billerName': 'MIXUE CU',
  'amount': 20,
  'compCode': '-',
  'ref1': 'KB000001896105',
  'ref2': 'KPS004KB000001896105',
  'ref3': '-'}}
```

Dataset

- Scaping Data from https://siamboardgames.com/
- Open Boardgame dataset https://www.kaggle.com/datasets/andrewmvd/board-games/data
- Huggingface https://huggingface.co/datasets/goendalf666/sales-conversations



ID BoardGamesGeek ID	Name Board game name	Year Published Year published	Min Players Min suggested players	Max Players Max suggested players	Play Average suggest creators
0 total values	20338 unique values	1948 total values	6875 total values	11135 total values	
174430	Gloomhaven	2017	1	4	120
161936	Pandemic Legacy: Season 1	2015	2	4	60
224517	Brass: Birmingham	2018	2	4	120
167791	Terraforming Mars	2016	1	5	120
233078	Twilight Imperium: Fourth Edition	2017	3	6	480
291457	Gloomhaven: Jaws of the Lion	2020	1	4	120
182028	Through the Ages: A New Story of Civilization	2015	2	4	120
220308	Gaia Project	2017	1	4	150
187645	Star Wars: Rebellion	2016	2	4	240