An Introduction to World Model and Beyond

World Model Study Group @ STRUCT

Presented by ZhangJiahang, KuangHaowei, GaoWenshuo

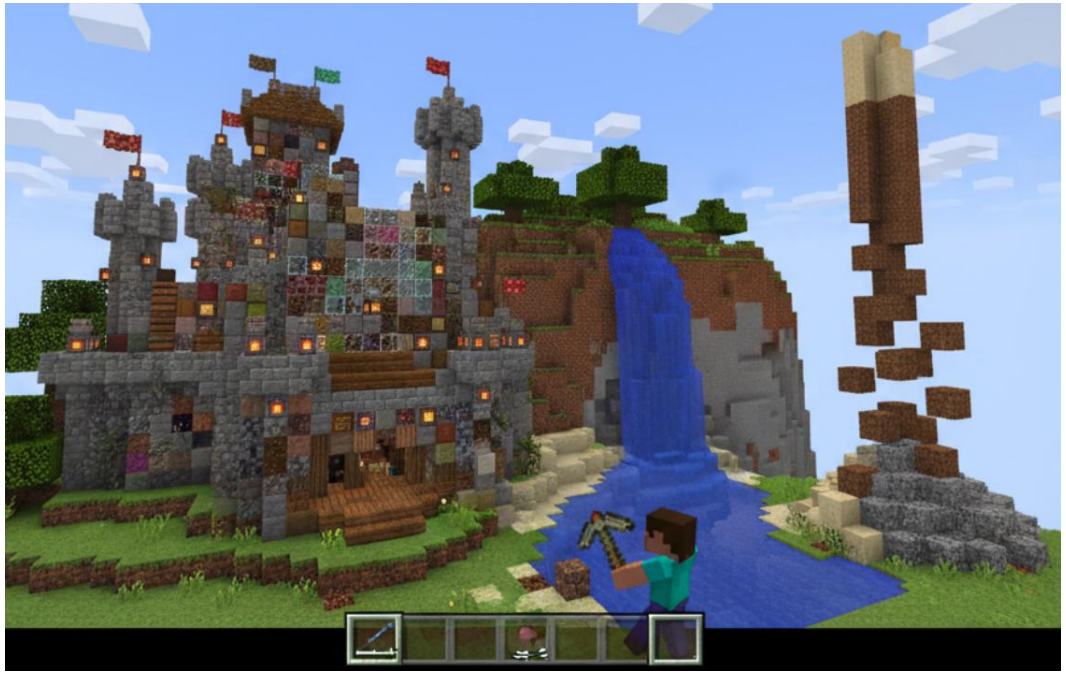
2026.1.27

Content

- World Model Taxonomy
  - Generation Focused
  - Representation Focused
- Multi-Modal Large Language Models
  - Foundation MLLMs with Qwen2.5-VL
  - Chain-of-Thought
- Unified Generation-Understanding Models

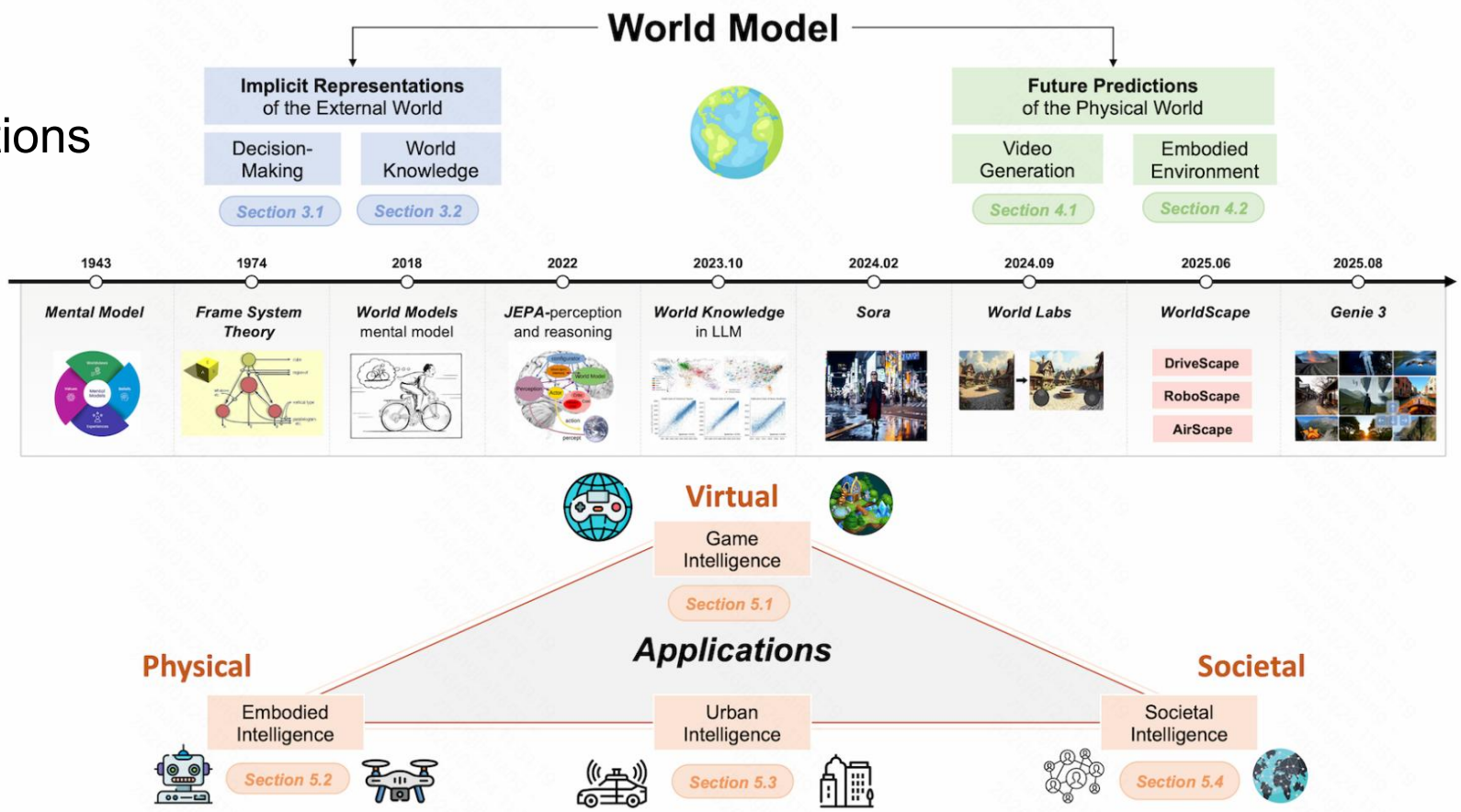
World Model Introduction

- World Model: AI Model vs. Game Engine
  - Knowledge-Driven World Predictor & Rule-based World Simulator

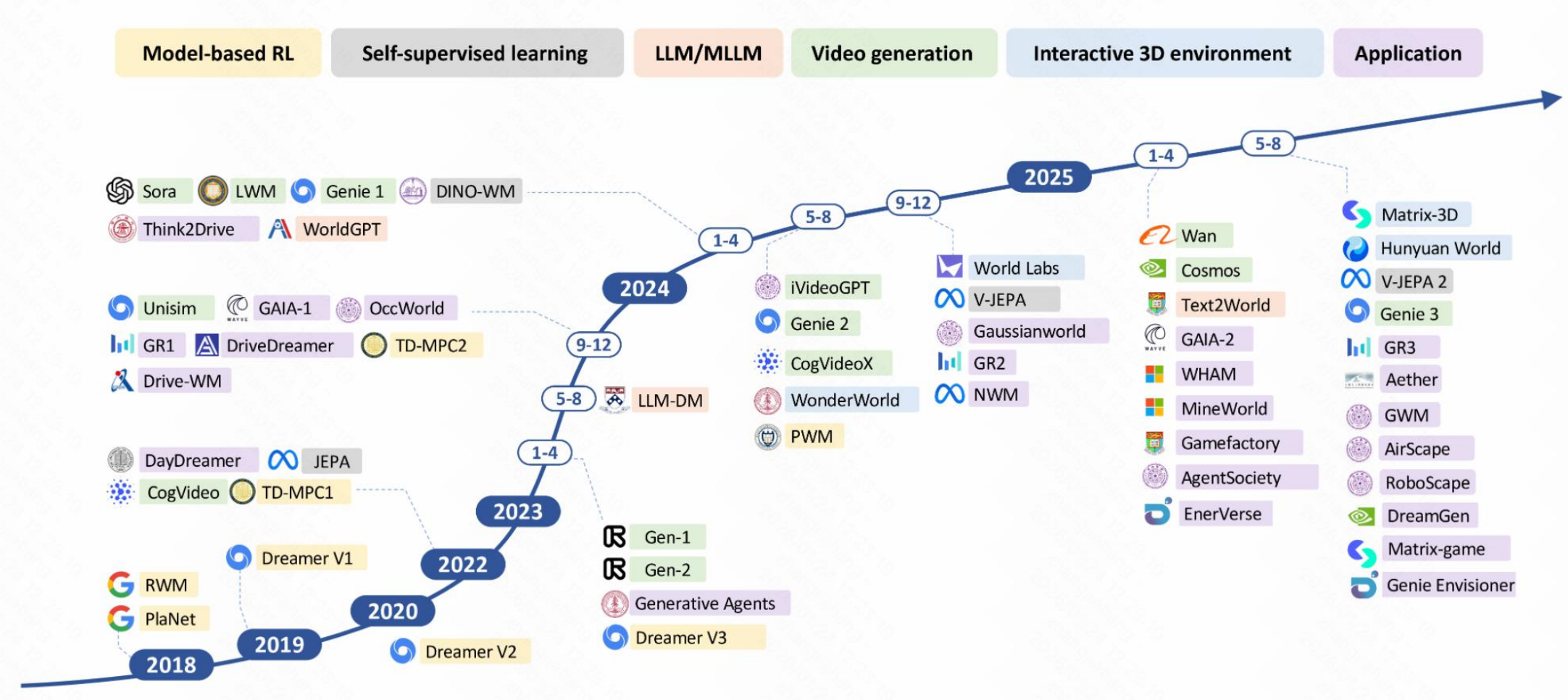


World Model Introduction

- Taxonomy
  - Implicit Representations
  - Pixel Prediction
- Key Capacity
  - Future Prediction



World Model Introduction



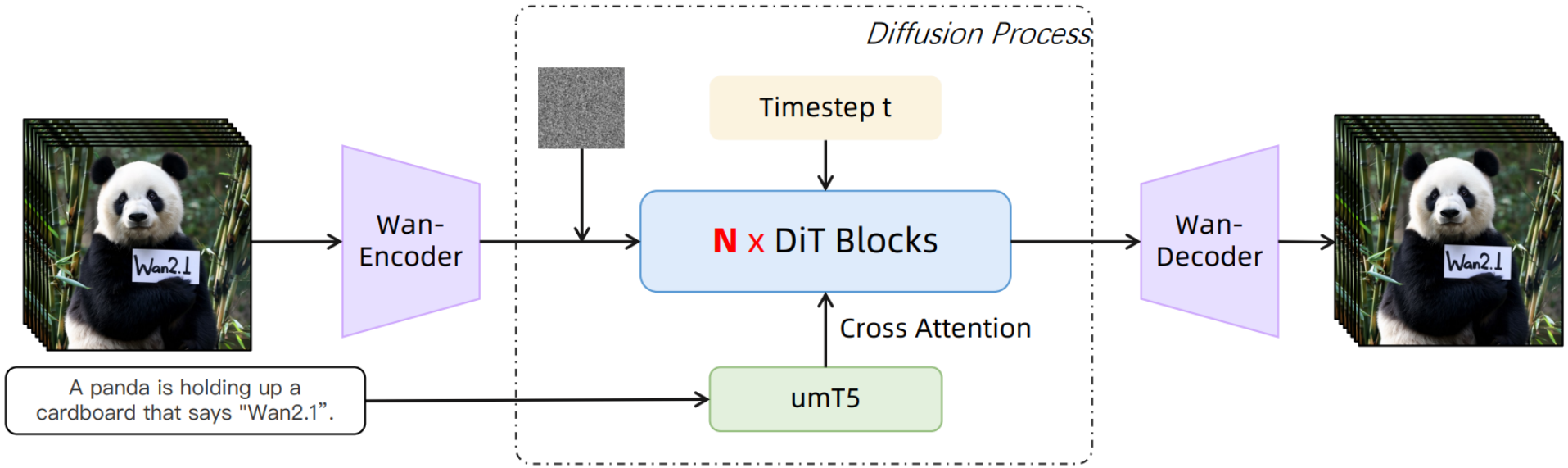
Pixel-Level Generation-Based Model

- Future Prediction Capacity:
  - Temporal Prediction
- Video Generation Models
  - Wan Model, Sora ...

Wan Team, WAN: Open and Advanced Large-scale Video Generation Models, arXiv 2503.20314

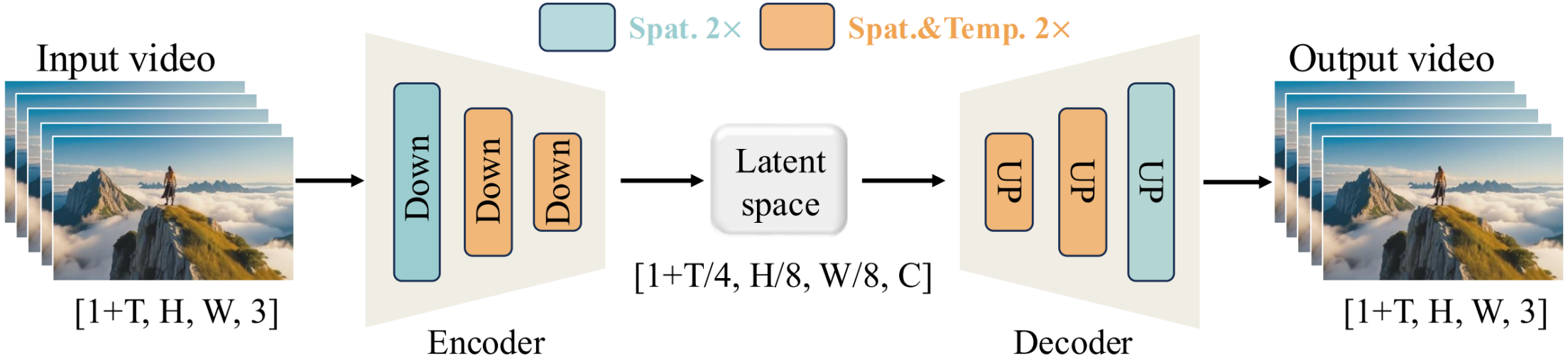
Pixel-Level Generation-Based Model

- Wan Model
  - Diffusion Transformer (DiT) Backbone



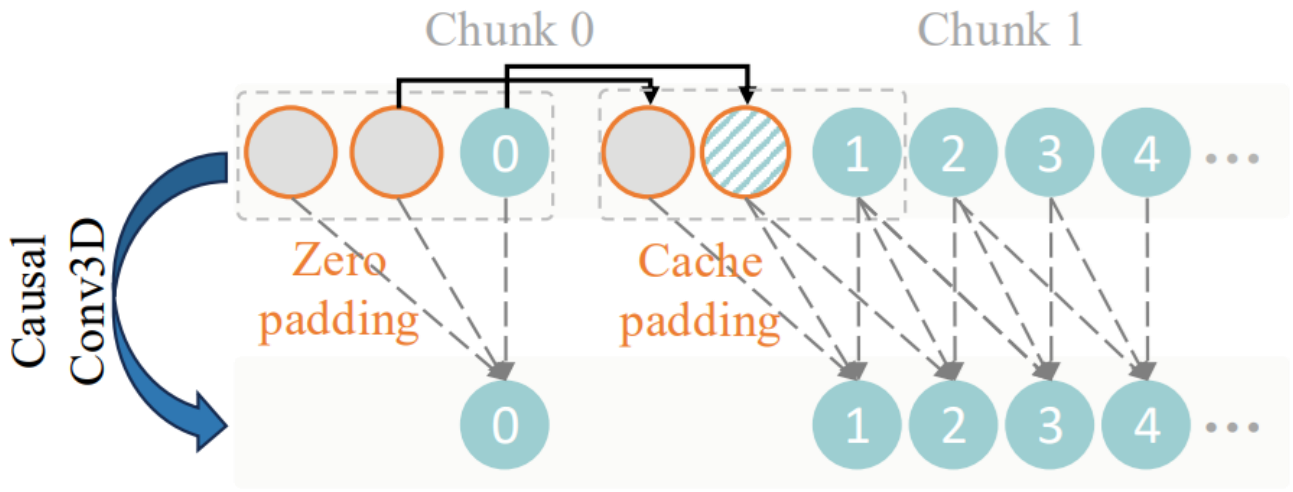
Pixel-Level Generation-Based Model

- Wan Model
  - Diffusion Transformer (DiT) Backbone
  - **Wan-Enc/Dec**



Pixel-Level Generation-Based Model

- Wan Model
  - Diffusion Transformer (DiT) Backbone
  - **Wan-Enc/Dec**



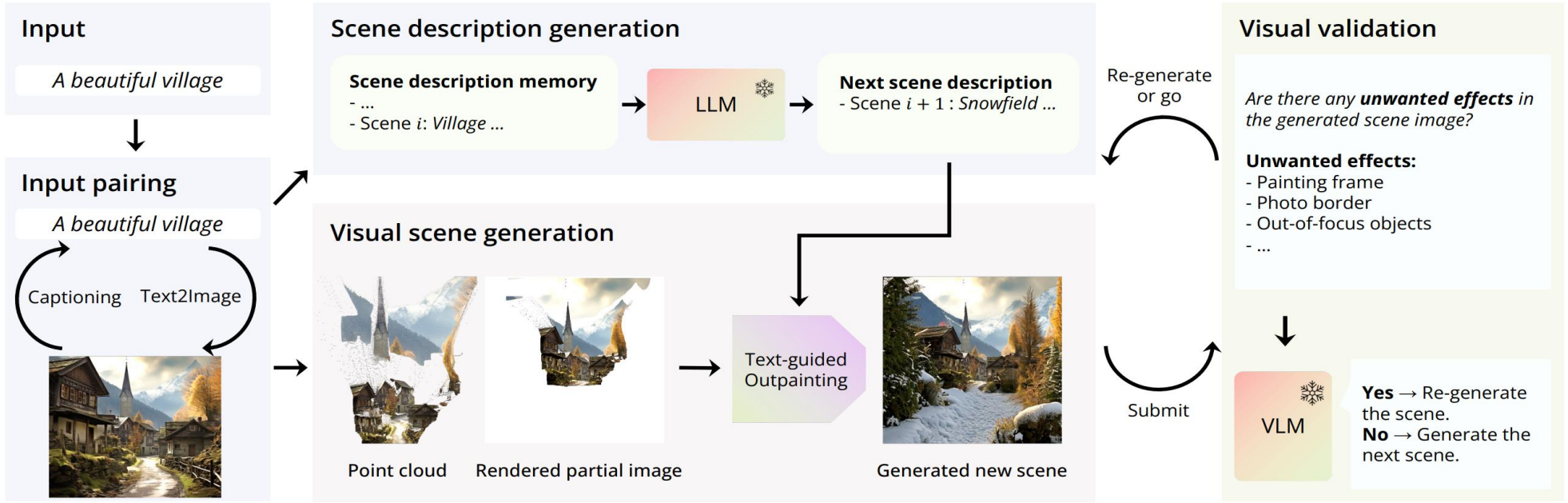
3D Generation-Based Model

- Future Prediction Capacity:
  - Spatial Prediction

Yu et al., WonderJourney: Going from Anywhere to Everywhere, CVPR, 2024

3D Generation-Based Model

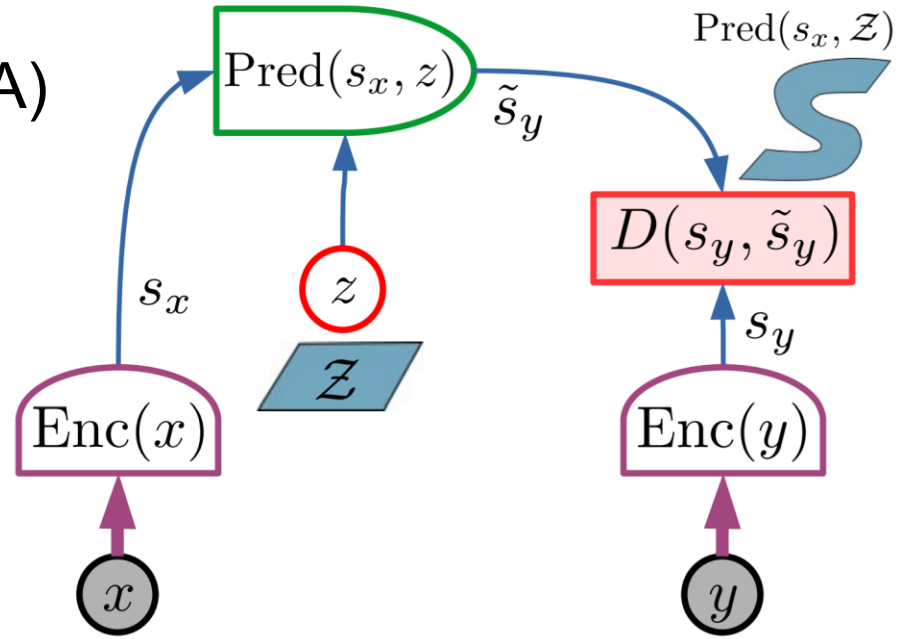
- Future Prediction Capacity:
  - Spatial Prediction



Representation Prediction Model

- Future Prediction Capacity:
  - Representation/Semantic Prediction

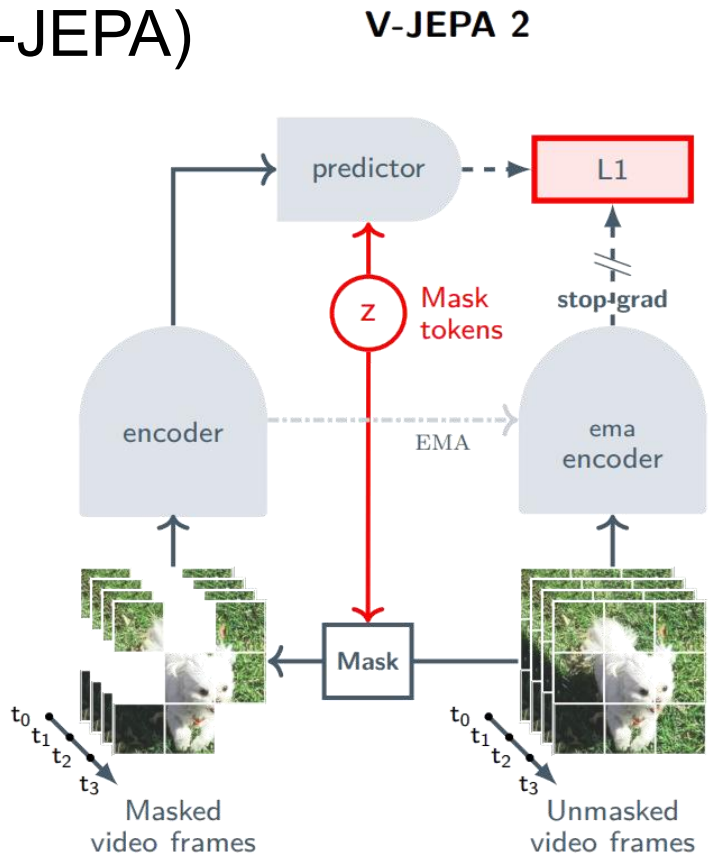
- Joint-Embedding Predictive Architecture (JEPA)



Representation Prediction Model

- Video Joint Embedding Predictive Architecture (V-JEPA)

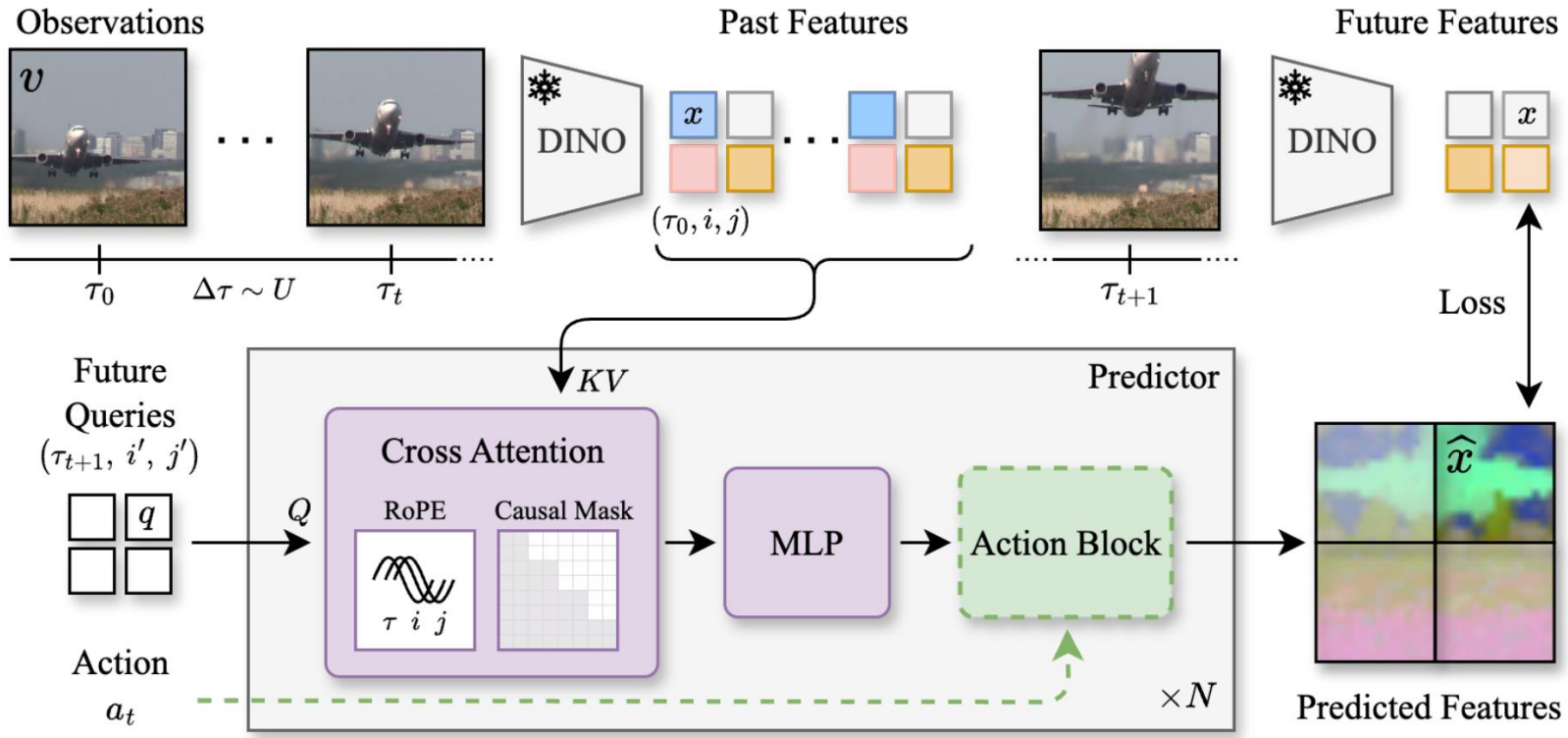
Method	Param.	*Action Anticipation*		
		Verb	Noun	**Action**
InAViT (Roy et al., 2024)	160M	51.9	52.0	25.8
Video-LLaMA (Zhang et al., 2023)	7B	52.9	52.0	26.0
PlausiVL (Mittal et al., 2024)	8B	55.6	54.2	27.6
Frozen Backbone				
V-JEPA 2 ViT-L	300M	57.8	53.8	32.7
V-JEPA 2 ViT-H	600M	59.2	54.6	36.5
V-JEPA 2 ViT-g	1B	61.2	55.7	38.0
V-JEPA 2 ViT-g$_{384}$	1B	**63.6**	**57.1**	**39.7**



Representation Prediction Model

- DINO-World

- Input
  - past frames
  - future query tokens
- Encoder
  - Frozen DINO
- Predictor
- Output
  - Future Frames



Representation Prediction Model

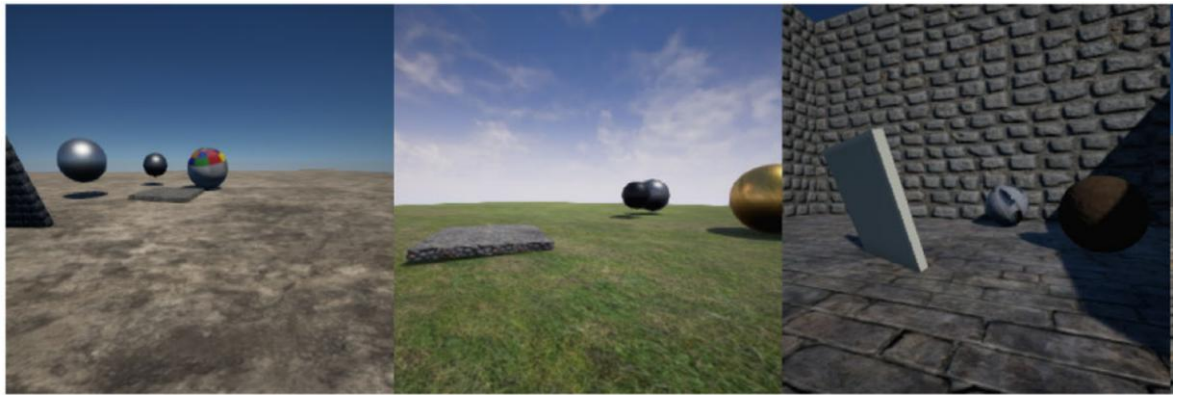
- DINO-World
- Experiments： Future semantic segmentation & depth estimation

		VSPW mIoU (↑)			Cityscapes mIoU (↑)			KITTI RMSE (↓)		
	Encoder	Present	Short	Mid	Present	Short	Mid	Present	Short	Mid
Copy Last	ViT-B	52.8	47.9	42.1	68.6	53.2	39.7	2.963	3.778	4.745
COSMOS-4B	ViT-B	52.8	46.6	40.2	68.6	55.4	46.2	2.963	4.178	4.742
COSMOS-12B	ViT-B	52.8	46.6	40.7	68.6	55.6	45.9	2.963	4.157	4.617
V-JEPA	ViT-L	29.1	8.2	7.7	48.8	15.5	14.0	3.502	7.217	7.491
V-JEPA	ViT-H	28.0	4.9	4.6	49.7	13.3	12.2	3.402	5.458	5.785
DINO-Foresight	ViT-B	50.6	44.7	37.7	66.9	64.5	**57.2**	2.882	3.562	**3.740**
DINO-world	ViT-B	52.8	**51.6**	**47.0**	68.6	**64.7**	55.1	2.963	**3.214**	4.268

Baldassarre et al., Back to the Features: DINO as a Foundation for Video World Models, arXiv 2507.19468

Representation Prediction Model

- DINO-World
- Experiments： Understanding physical law

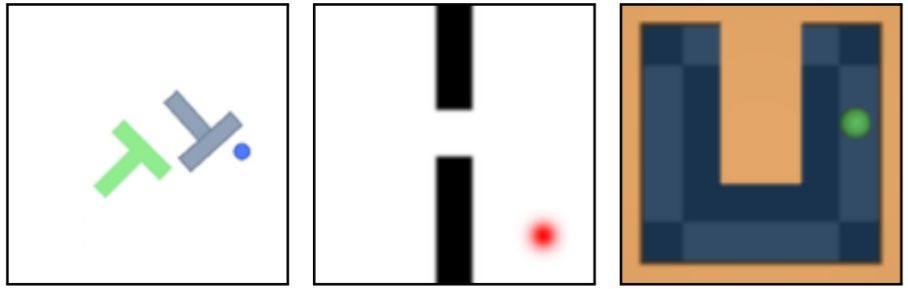


	Encoder	Predictor	IntPhys	GRASP	InfLevel
COSMOS-4B	VAE	4B	**99.5**	60.1	44.8
V-JEPA	ViT-L	22M	92.2	67.0	58.9
V-JEPA	ViT-H	22M	89.4	73.0	59.9
DINO-Foresight	ViT-B	193M	87.8	64.9	62.8
DINO-world	ViT-B	1.1B	91.3	**76.0**	**63.7**

Baldassarre et al., Back to the Features: DINO as a Foundation for Video World Models, arXiv 2507.19468

Representation Prediction Model

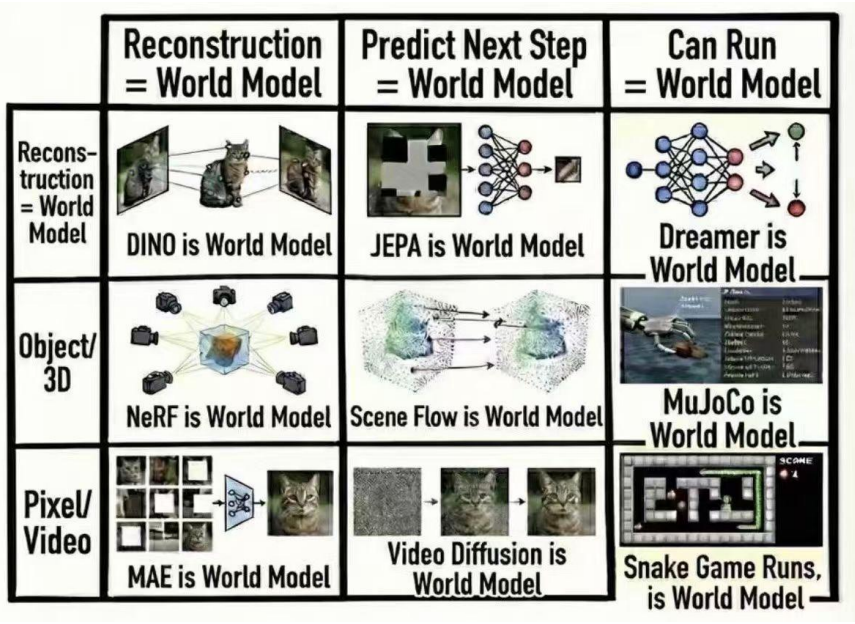
- DINO-World
- Experiments： Planning



Model	PushT	Wall	PointMaze
Scratch	46.9	87.1	59.4
Action-only	49.4	91.1	61.6
Fine-tuned	59.4	93.8	68.7

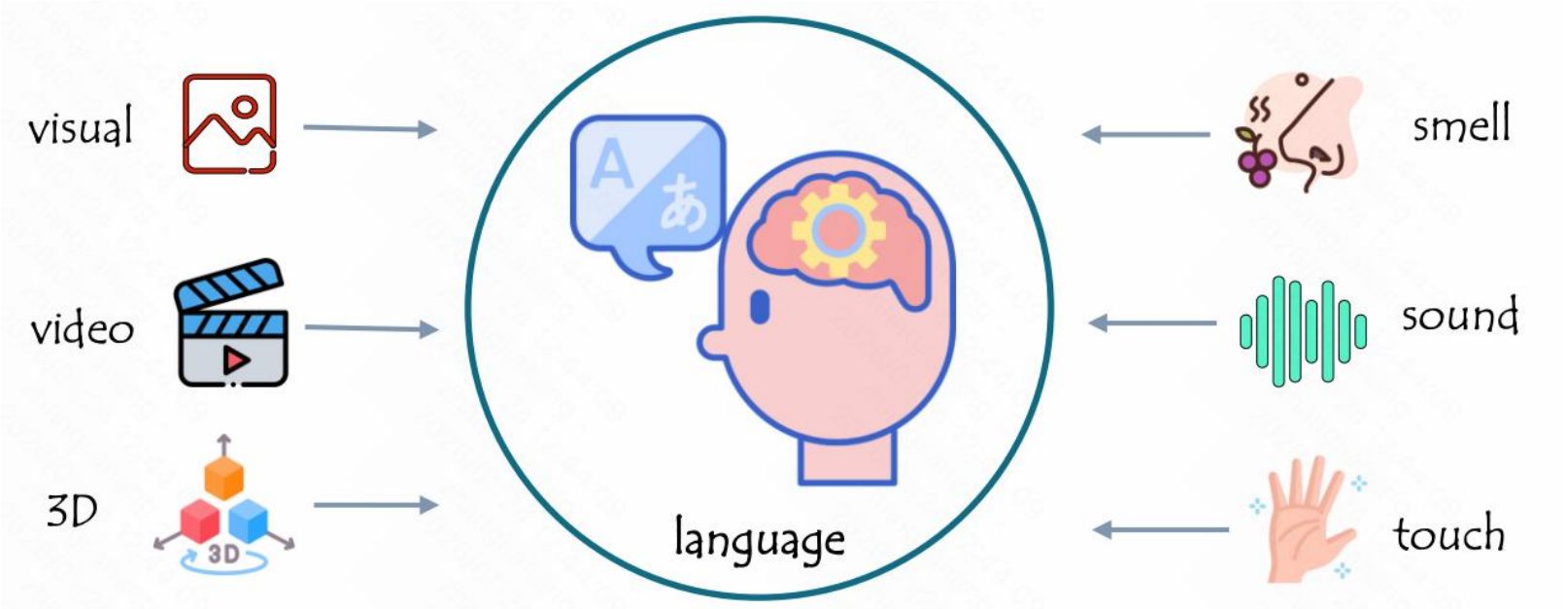
World Model Summary

- Key Capacity: **Future Prediciton**
- Two Types
  - Generation-Based: Wan, Sora (pixel-level), WonderJourney (3D model)
  - Representation-Based: JEPA



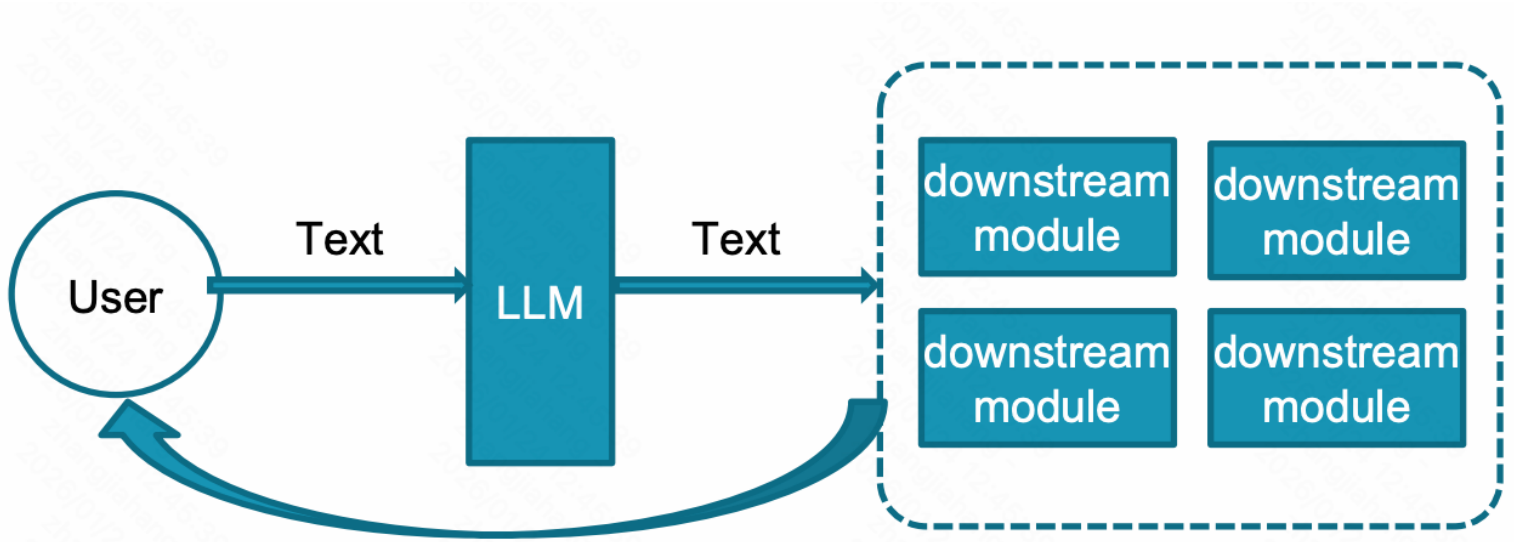
Multi-Modal Large Language Models

- MLLM: Multi-Modal Large Language Models



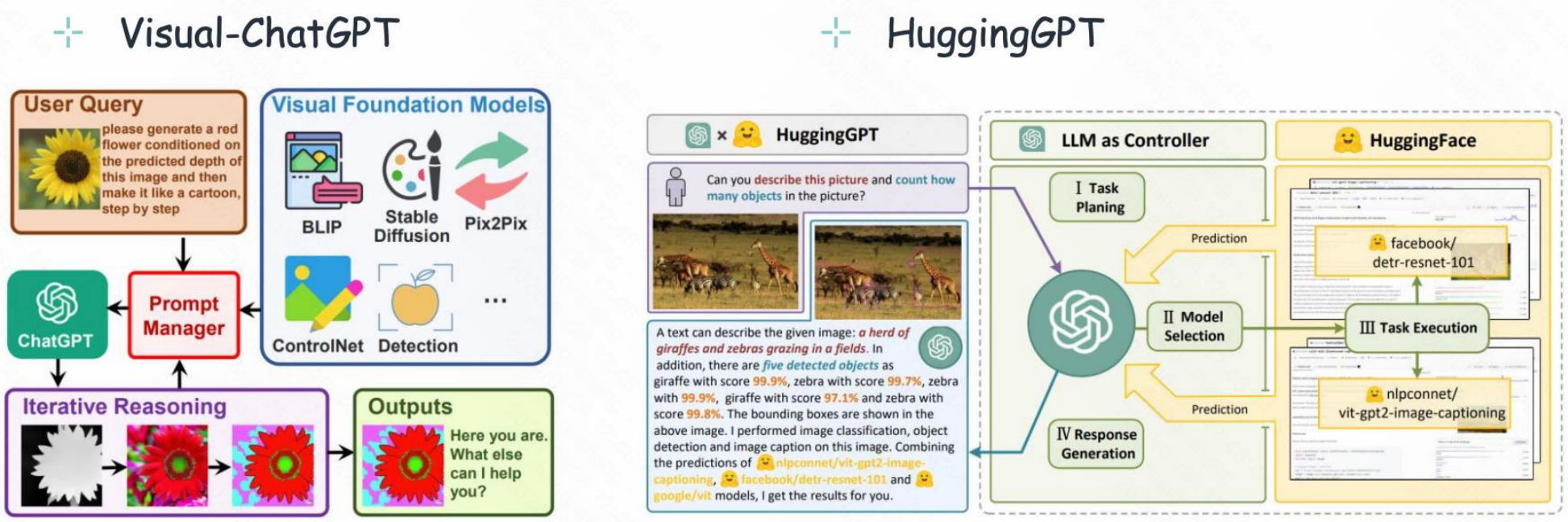
Multi-Modal Large Language Models

- Architecture-I: LLM as Discrete Scheduler/Controller
  - Quick to build (without training), flexible extension to many tool features
  - Information loss in text medium, the bottle-neck



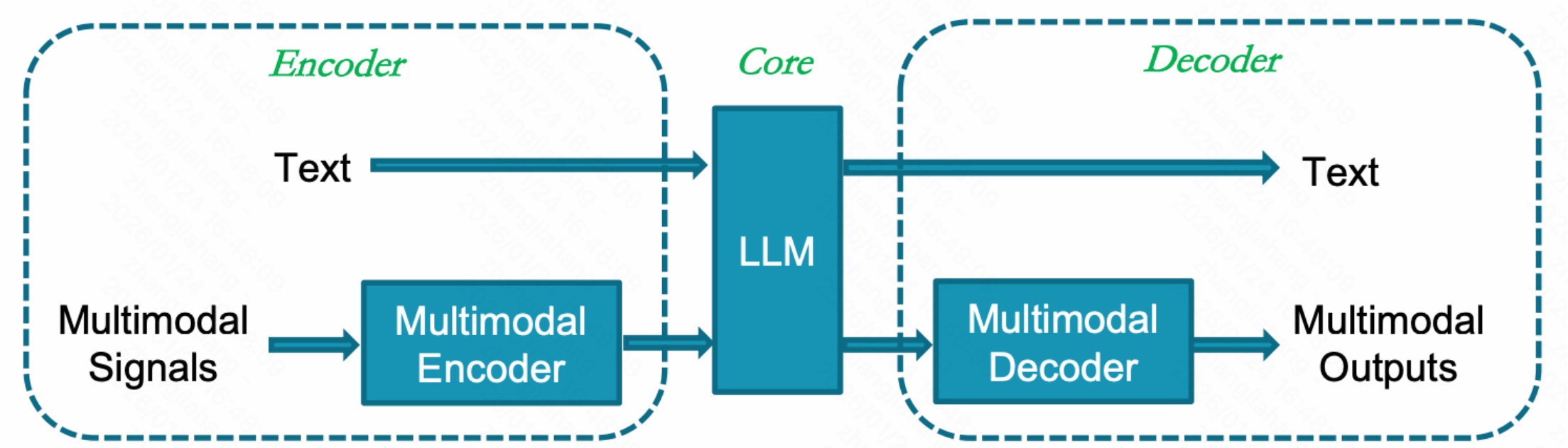
Multi-Modal Large Language Models

- Architecture-I: LLM as Discrete Scheduler/Controller
  - Quick to build (without training), flexible extension to many tool features
  - Information loss in text medium, the bottle-neck



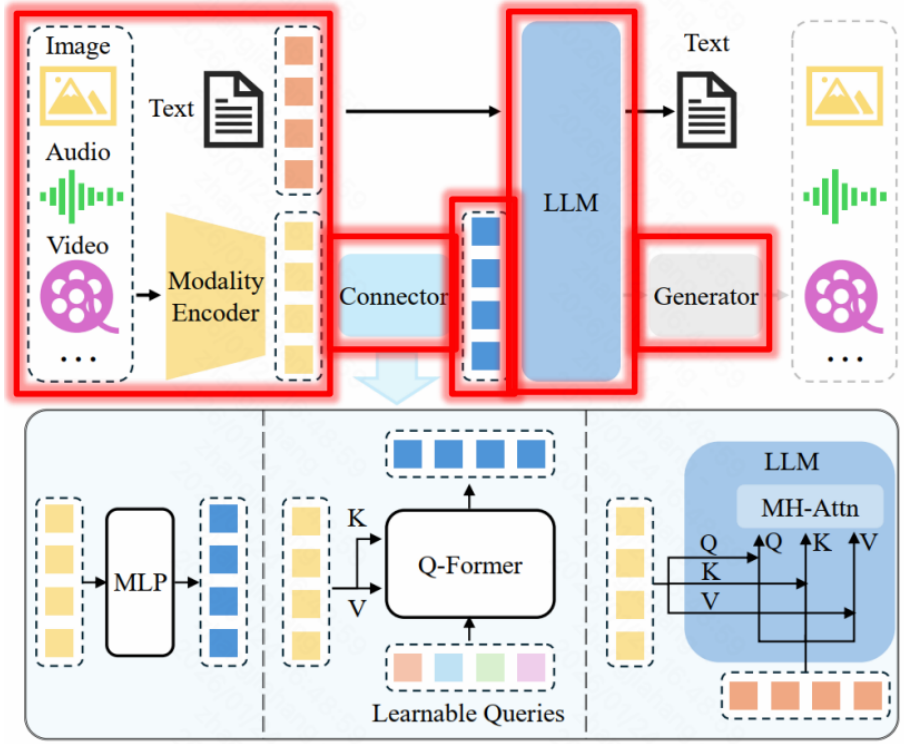
Multi-Modal Large Language Models

- Architecture-II: LLM as Joint Part of Pipeline
  - Perceive multimodal information, and react by itself,
  - Encoder-LLM-Decoder Structure



Multi-Modal Large Language Models

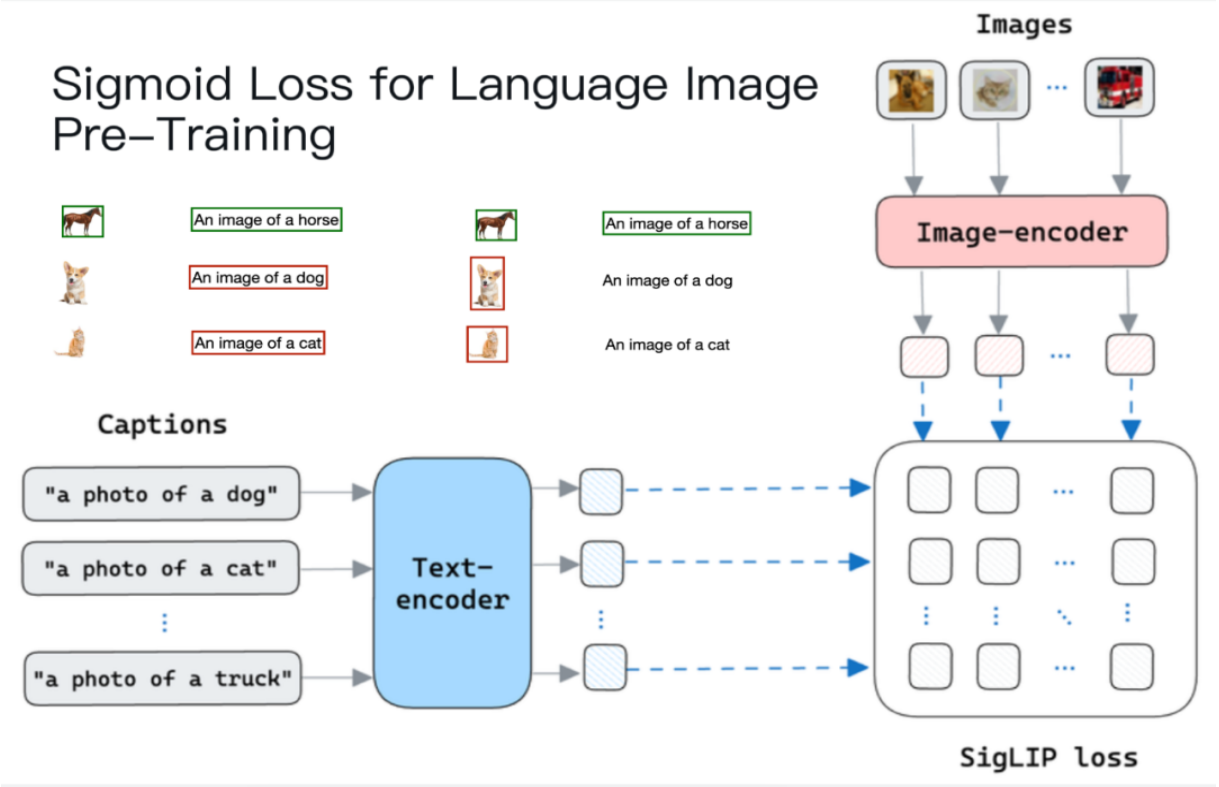
- Architecture-II: LLM as Joint Part of Pipeline
  - > 90% MLLMs belong to this category
  - Higher Performance
  - Multi-modal tokens as LLM inputs



A Survey on Multimodal Large Language Models. https://github.com/BradyFU/A wesome-Multimodal-LargeLanguage-Models, 2023.

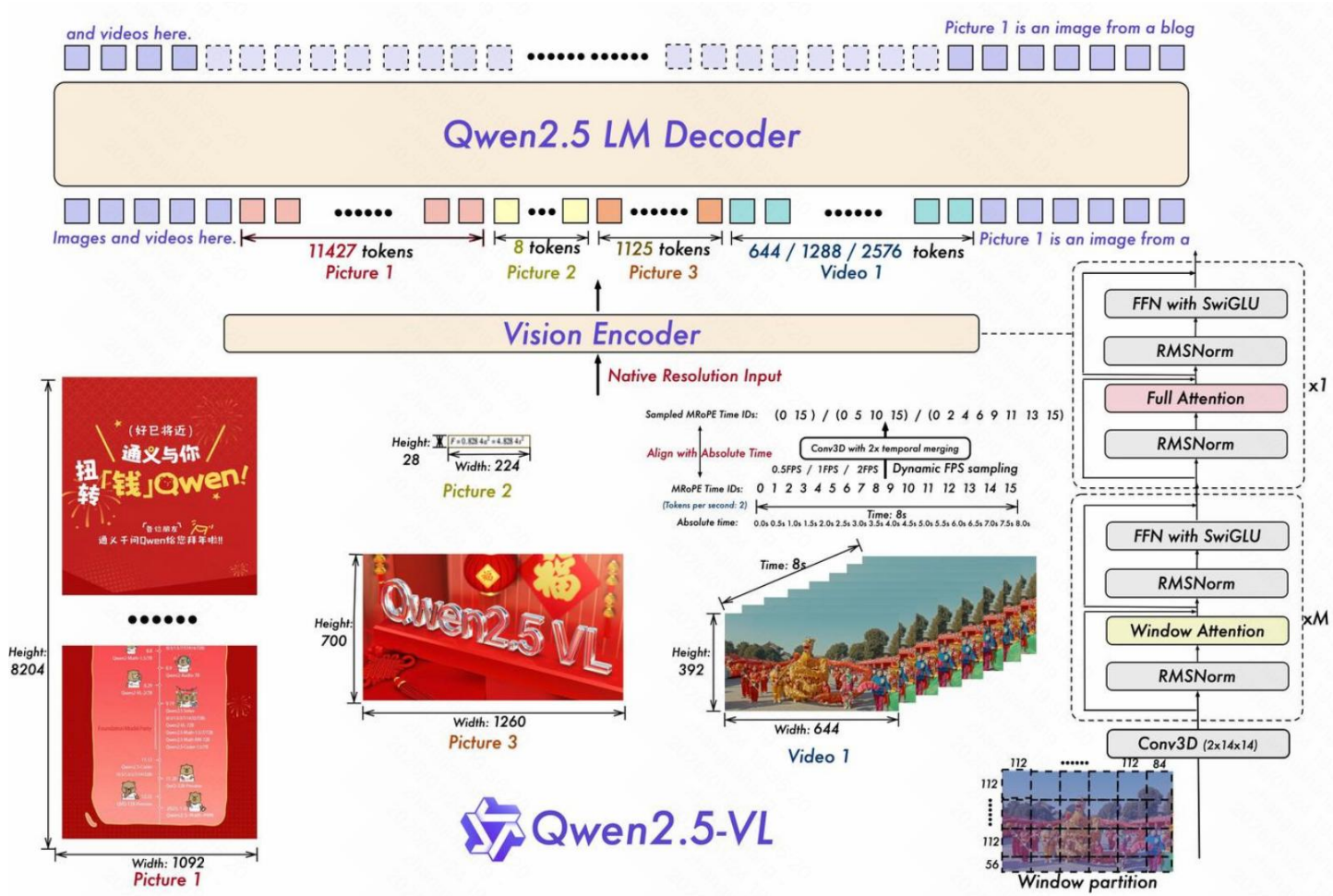
Multi-Modal Large Language Models

- Architecture-II: LLM as Joint Part of Pipeline
- Visual Encoder
  - CLIP: contrastive learning
  - SigLIP: sigmoid + binary cls.



Representative MLLM Works

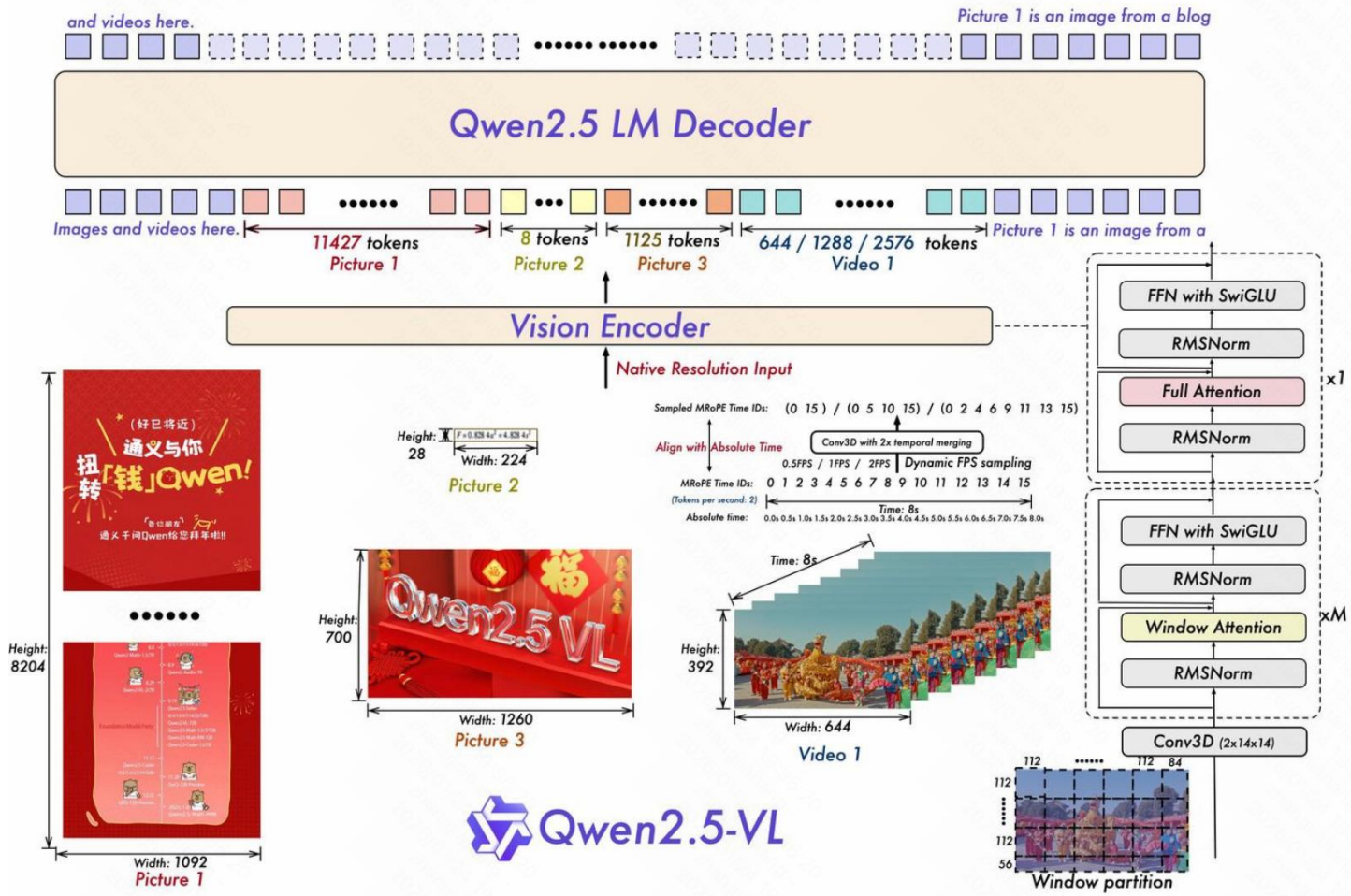
- Qwen2.5-VL
- Arch.
  - Vision Encoder
  - LLM
  - Connector (MLP)



Qwen2.5-VL Technical Report, https://arxiv.org/abs/2502.13923

Representative MLLM Works

- Qwen2.5-VL
  - **Native Resolution**
- Details:
  - Packed data training



Qwen2.5-VL Technical Report, https://arxiv.org/abs/2502.13923

Representative MLLM Works

- Qwen2.5-VL
  - Native Resolution
  - **3D MRoPE**

Qwen2.5-VL Technical Report, https://arxiv.org/abs/2502.13923

Representative MLLM Works

- RoPE (Rotary Positional Embedding)
- Most Popular PE Method for LLM/MLLM

- Compared with other PEs, e.g., Sinusoidal PE

$$E_i = X_i + P_i$$

$$(X_i + P_i)^T (X_j + P_j)$$

Other PEs : Absolute Encoding

$$\tilde{q}_m = R_m q_m, \quad \tilde{k}_n = R_n k_n$$

$$\langle \tilde{q}_m, \tilde{k}_n \rangle = (R_m q_m)^T (R_n k_n) = q_m^T (R_m^T R_n) k_n$$

RoPE

RoFormer: Enhanced Transformer with Rotary Position Embedding, Neurocomputing 2024

Representative MLLM Works

- RoPE (Rotary Positional Embedding)
- Key advantages compared with other PEs, e.g., Sinusoidal PE

$$E_i = X_i + P_i$$

$$(X_i + P_i)^T (X_j + P_j)$$

Other PEs : Absolute Encoding

$$\tilde{q}_m = R_m q_m, \quad \tilde{k}_n = R_n k_n$$

$$\langle \tilde{q}_m, \tilde{k}_n \rangle = (R_m q_m)^T (R_n k_n) = q_m^T (R_m^T R_n) k_n$$

RoPE

$$R_m^T R_n = R_{n-m}$$

RoFormer: Enhanced Transformer with Rotary Position Embedding, Neurocomputing 2024

Representative MLLM Works

- Multimodal Rotary Position Embedding (MRoPE) in QwenVL
- Motivation: Extend 1D RoPE to 2D images
  - 1: Split channels

$$x = [x_{height}, x_{width}]$$

  - 2: Apply RoPE separately

$$\text{MRoPE}(x, h, w) = [\text{RoPE}(x_{height}, h), \text{RoPE}(x_{width}, w)]$$

  - 3. Recall attention computation

$$\text{Score} = \underbrace{(R_{h_1} q_h)^T (R_{h_2} k_h)}_{\text{高度部分}} + \underbrace{(R_{w_1} q_w)^T (R_{w_2} k_w)}_{\text{宽度部分}}$$

$$\text{Score} \propto \text{RelativeDistance}(h_1 - h_2) + \text{RelativeDistance}(w_1 - w_2)$$

Qwen2.5-VL Technical Report, https://arxiv.org/abs/2502.13923

Representative MLLM Works

- 3D MRoPE in Qwen2.5-VL
- Further entend to 3D video from MRoPE
  - video patches: (t, h, w)

$$x = [x_t, x_h, x_w]$$

$$\text{3D-RoPE}(x, t, h, w) = \text{Concat}\Big(\text{RoPE}(x_t, t),\ \text{RoPE}(x_h, h),\ \text{RoPE}(x_w, w)\Big)$$

$$\text{Score} \propto \underbrace{\cos(\theta_t \cdot \Delta t)}_{\text{时间距离}} + \underbrace{\cos(\theta_h \cdot \Delta h)}_{\text{垂直距离}} + \underbrace{\cos(\theta_w \cdot \Delta w)}_{\text{水平距离}}$$

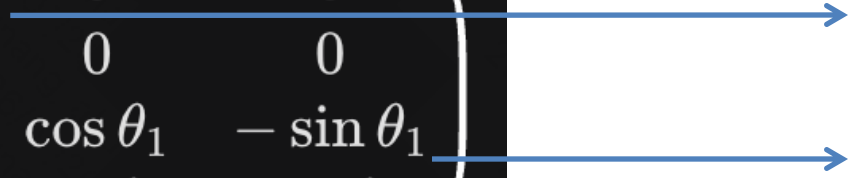
Qwen2.5-VL Technical Report, https://arxiv.org/abs/2502.13923

Representative MLLM Works

- 3D MRoPE Interleaved in Qwen3-VL
- Diving into the details of RoPE
  - a token dimension = D

D=4

$$R = \begin{pmatrix} \cos\theta_0 & -\sin\theta_0 & 0 & 0 \\ \sin\theta_0 & \cos\theta_0 & 0 & 0 \\ 0 & 0 & \cos\theta_1 & -\sin\theta_1 \\ 0 & 0 & \sin\theta_1 & \cos\theta_1 \end{pmatrix}$$



high frequency, x0,x1

low frequency, x2x3

Qwen2.5-VL Technical Report, https://arxiv.org/abs/2502.13923

Representative MLLM Works

- 3D MRoPE Interleaved in Qwen3-VL
- Diving into the details of RoPE
  - recall 3D MRoPE

$$[\underbrace{t, t, t, \ldots}_{\text{低维(高频)}}, \underbrace{h, h, h, \ldots}_{\text{中维(中频)}}, \underbrace{w, w, w, \ldots}_{\text{高维(低频)}}]$$

D=4

$$R = \begin{pmatrix} \cos\theta_0 & -\sin\theta_0 & 0 & 0 \\ \sin\theta_0 & \cos\theta_0 & 0 & 0 \\ 0 & 0 & \cos\theta_1 & -\sin\theta_1 \\ 0 & 0 & \sin\theta_1 & \cos\theta_1 \end{pmatrix}$$

high frequency

low frequency

Qwen2.5-VL Technical Report, https://arxiv.org/abs/2502.13923

Representative MLLM Works

- 3D MRoPE Interleaved in Qwen3-VL
- Diving into the details of RoPE
  - recall 3D MRoPE

$$[\underbrace{t, t, t, \ldots}_{\text{低维(高频)}}, \underbrace{h, h, h, \ldots}_{\text{中维(中频)}}, \underbrace{w, w, w, \ldots}_{\text{高维(低频)}}]$$

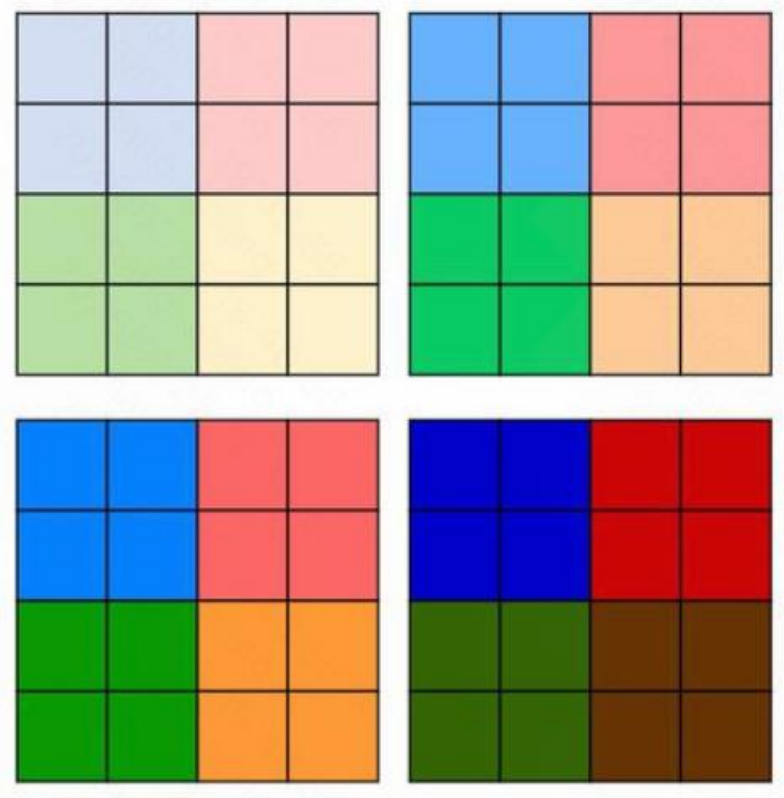
  - improve it:

$$[t, h, w,\ t, h, w,\ t, h, w, \ldots]$$

Qwen2.5-VL Technical Report, https://arxiv.org/abs/2502.13923

Representative MLLM Works

- Qwen2.5-VL
  - Native Resolution
  - 3D MRoPE
  - **Window Attention & Full Attention**



Qwen2.5-VL Technical Report, https://arxiv.org/abs/2502.13923

Representative MLLM Works

- Qwen2.5-VL
- Training Strategies
- **Stage 1**: train a ViT-Encoder from scratch
  - image captions
  - visual knowledge
  - contrastive learning

Qwen2.5-VL Technical Report, https://arxiv.org/abs/2502.13923

Representative MLLM Works

- Qwen2.5-VL
- Training Strategies
- **Stage 1**: train a ViT-Encoder from scratch
- **Stage 2:** jointly train ViT & QwenVL decoder
  - multiple diverse visual understanding datasets

Qwen2.5-VL Technical Report, https://arxiv.org/abs/2502.13923

Representative MLLM Works

- Qwen2.5-VL
- Training Strategies
- **Stage 1**: train a ViT-Encoder from scratch
- **Stage 2:** jointly train ViT & QwenVL decoder
- **Stage 3:** long-context understanding
  - long-range videos
  - complex reasoning
  - tokens extension from 8192 to 32768

Qwen2.5-VL Technical Report, https://arxiv.org/abs/2502.13923

Representative MLLM Works

- Qwen2.5-VL
- Training Strategies
- **Stage 1**: train a ViT-Encoder from scratch
- **Stage 2:** jointly train ViT & QwenVL decoder
- **Stage 3:** long-context understanding
- **Stage 4:** post-training
  - Supervised Fine-Tuning (SFT) with *Rejection Sampling*
  - Direct Preference Optimization (DPO)

Qwen2.5-VL Technical Report, https://arxiv.org/abs/2502.13923

Representative MLLM Works

- Qwen2.5-VL
- Data Preparation

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

Qwen2.5-VL Technical Report, https://arxiv.org/abs/2502.13923

Representative MLLM Works

- Qwen2.5-VL
- Experiments

Table 3: **Performance of Qwen2.5-VL and State-of-the-art.**

Datasets	Previous Open-source SoTA	Claude-3.5 Sonnet-0620	GPT-4o 0513	InternVL2.5 78B	Qwen2-VL 72B	Qwen2.5-VL 72B	Qwen2.5-VL 7B	Qwen2.5-VL 3B
College-level Problems								
$MMMU_{val}$ (Yue et al., 2023)	70.1 Chen et al. (2024d)	68.3	69.1	70.1	64.5	**70.2**	58.6	53.1
$MMMU\text{-}Pro_{overall}$ (Yue et al., 2024)	48.6 Chen et al. (2024d)	51.5	**51.9**	48.6	46.2	51.1	38.3	31.56
Math								
$MathVista_{mini}$ (Lu et al., 2024)	72.3 Chen et al. (2024d)	67.7	63.8	72.3	70.5	**74.8**	68.2	62.3
$MATH\text{-}Vision_{full}$ (Wang et al., 2024d)	32.2 Chen et al. (2024d)	-	30.4	32.2	25.9	**38.1**	25.1	21.2
$MathVerse_{mini}$ (Zhang et al., 2024c)	51.7 Chen et al. (2024d)	-	50.2	51.7	-	**57.6**	49.2	47.6
General Visual Question Answering								
MegaBench (Chen et al., 2024b)	47.4 MiniMax et al. (2025)	52.1	**54.2**	45.6	46.8	51.3	36.8	28.9
$MMBench\text{-}EN_{test}$ (Liu et al., 2023d)	88.3 Chen et al. (2024d)	82.6	83.4	88.3	86.9	**88.6**	83.5	79.1
$MMBench\text{-}CN_{test}$ (Liu et al., 2023d)	88.5 Chen et al. (2024d)	83.5	82.1	**88.5**	86.7	87.9	83.4	78.1
$MMBench\text{-}V1.1\text{-}EN_{test}$ (Liu et al., 2023d)	87.4 Chen et al. (2024d)	80.9	83.1	87.4	86.1	**88.4**	82.6	77.4
MMStar (Chen et al., 2024c)	69.5 Chen et al. (2024d)	65.1	64.7	69.5	68.3	**70.8**	63.9	55.9
MME_{sum} (Fu et al., 2023)	**2494** Chen et al. (2024d)	1920	2328	**2494**	2483	2448	2347	2157
MuirBench (Wang et al., 2024a)	63.5 Chen et al. (2024d)	-	68.0	63.5	-	**70.7**	59.6	47.7
$BLINK_{val}$ (Fu et al., 2024c)	63.8 Chen et al. (2024d)	-	**68.0**	63.8	-	64.4	56.4	47.6
$CRPE_{relation}$ (Wang et al., 2024h)	78.8 Chen et al. (2024d)	-	76.6	78.8	-	**79.2**	76.4	73.6
$HallBench_{avg}$ (Guan et al., 2023)	**58.1** Wang et al. (2024f)	55.5	55.0	57.4	**58.1**	55.2	52.9	46.3
MTVQA (Tang et al., 2024)	**31.9** Chen et al. (2024d)	25.7	27.8	31.9	30.9	31.7	29.2	24.8
$RealWorldQA_{avg}$ (X.AI, 2024)	78.7 Chen et al. (2024d)	60.1	75.4	**78.7**	77.8	75.7	68.5	65.4
$MME\text{-}RealWorld_{en}$ (Zhang et al., 2024f)	62.9 Chen et al. (2024d)	51.6	45.2	62.9	-	**63.2**	57.4	53.1
$MMVet_{turbo}$ (Yu et al., 2024)	74.0 Wang et al. (2024f)	70.1	69.1	72.3	74.0	**76.2**	67.1	61.8
MM-MT-Bench (Agrawal et al., 2024)	7.4 Agrawal et al. (2024)	7.5	**7.72**	-	6.59	7.6	6.3	5.7

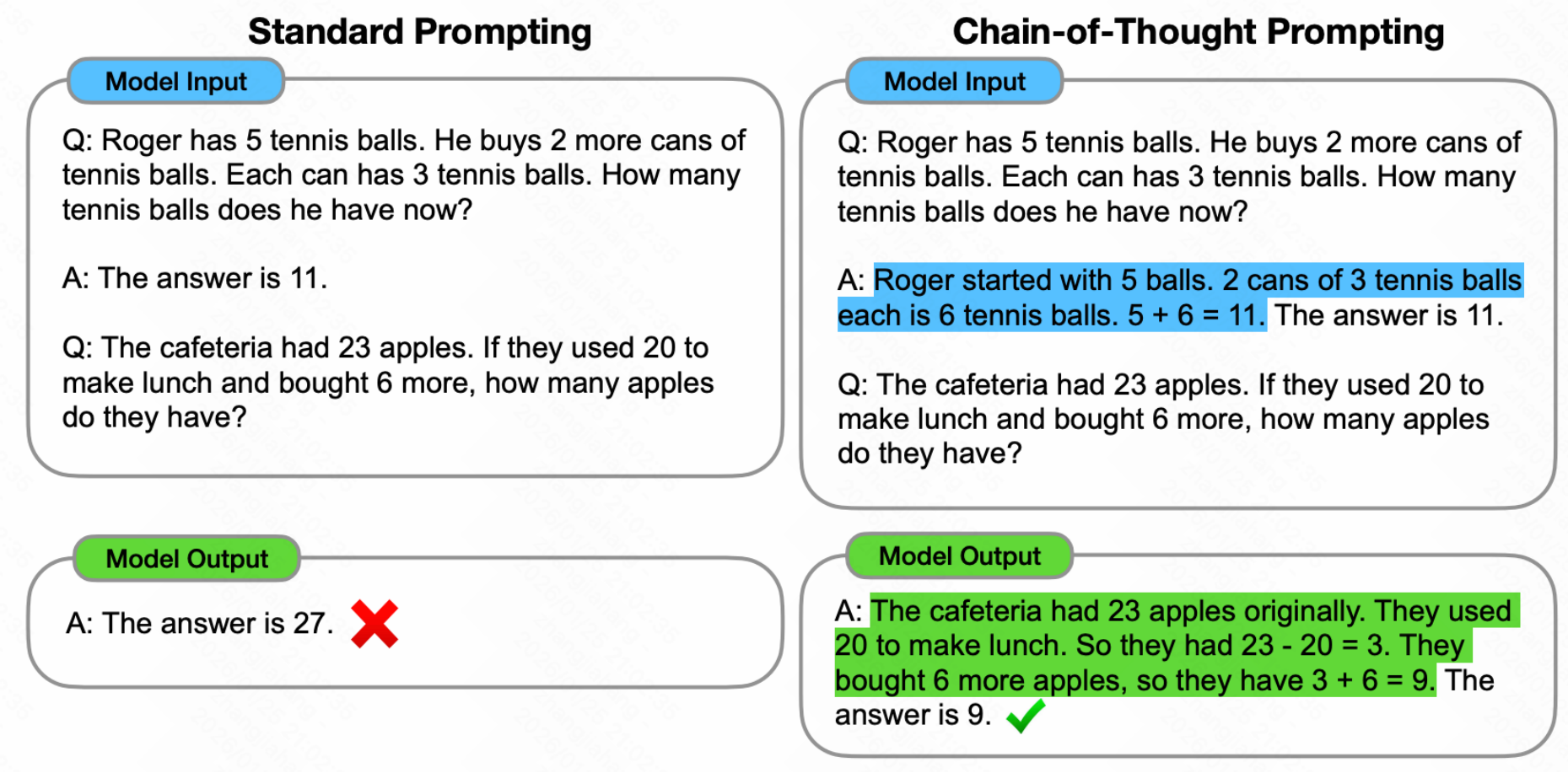
Qwen2.5-VL Technical Report, https://arxiv.org/abs/2502.13923

Rank ↑	Rank Spread ⓘ	Model ↑↓	Score ↑↓	95% CI (±) ↑↓	Votes ↑↓	Organization ↑↓	License ↑↓
1	1 ◄–► 1	gemini-3-pro	1320	±10	6,370	Google	Proprietary
2	2 ◄–► 2	gemini-3-flash	1296	±12	3,858	Google	Proprietary
3	3 ◄–► 9	gemini-3-flash (thinking-minimal)	1264	±15	1,724	Google	Proprietary
4	3 ◄–► 7	gemini-2.5-pro	1263	±7	76,223	Google	Proprietary
5	3 ◄–► 10	gemini-2.5-flash-preview-09-2025	1250	±10	5,307	Google	Proprietary
6	3 ◄–► 11	gpt-5.1-high	1249	±11	4,147	OpenAI	Proprietary
7	3 ◄–► 12	ernie-5.0-preview-1220	1249 ⓘ Preliminary	±12	3,601	Baidu	Proprietary
8	4 ◄–► 12	qwen3-vl-235b-a22b-instruct	1243	±9	7,585	Alibaba	Apache 2.0
9	5 ◄–► 12	chatgpt-4o-latest-20250326	1240	±6	20,118	OpenAI	Proprietary
10	4 ◄–► 12	gpt-5.1	1239	±11	4,412	OpenAI	Proprietary
11	6 ◄–► 12	gemini-2.5-flash	1233	±7	44,278	Google	Proprietary

https://lmarena.ai/zh/leaderboard/vision/overall-no-style-control

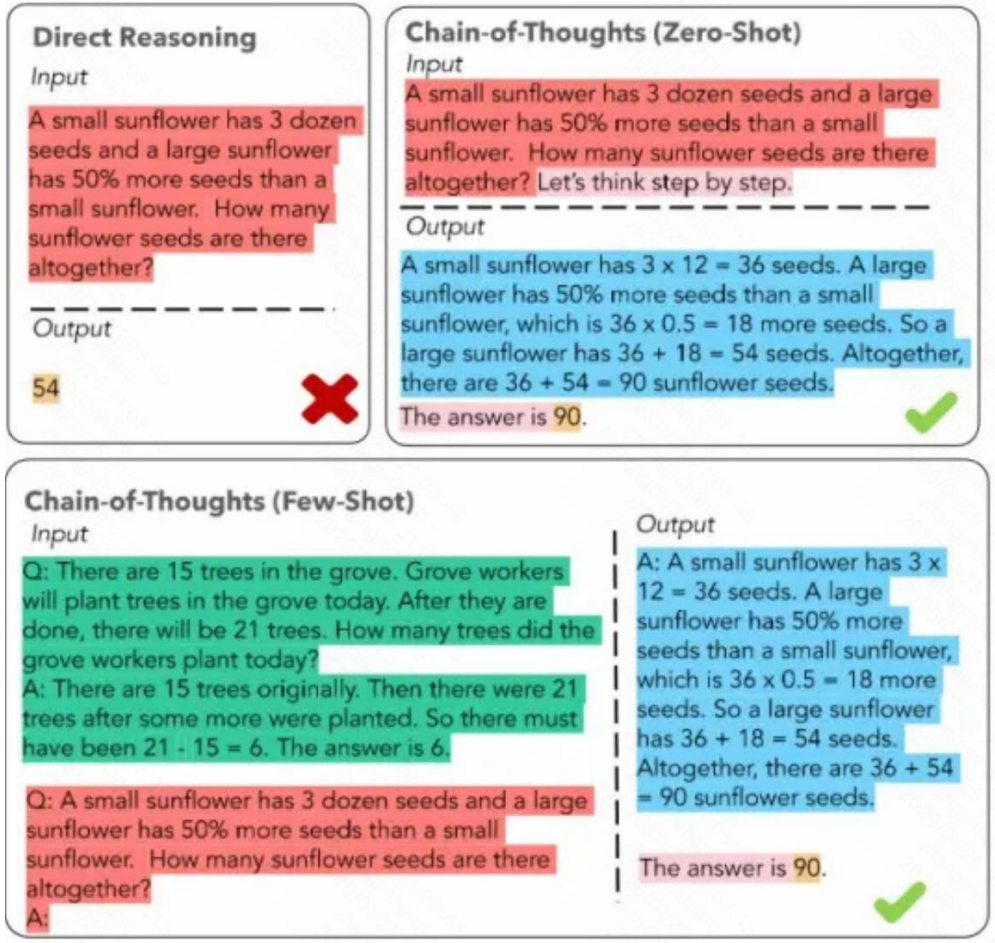
Chain-of-Thought

- CoT



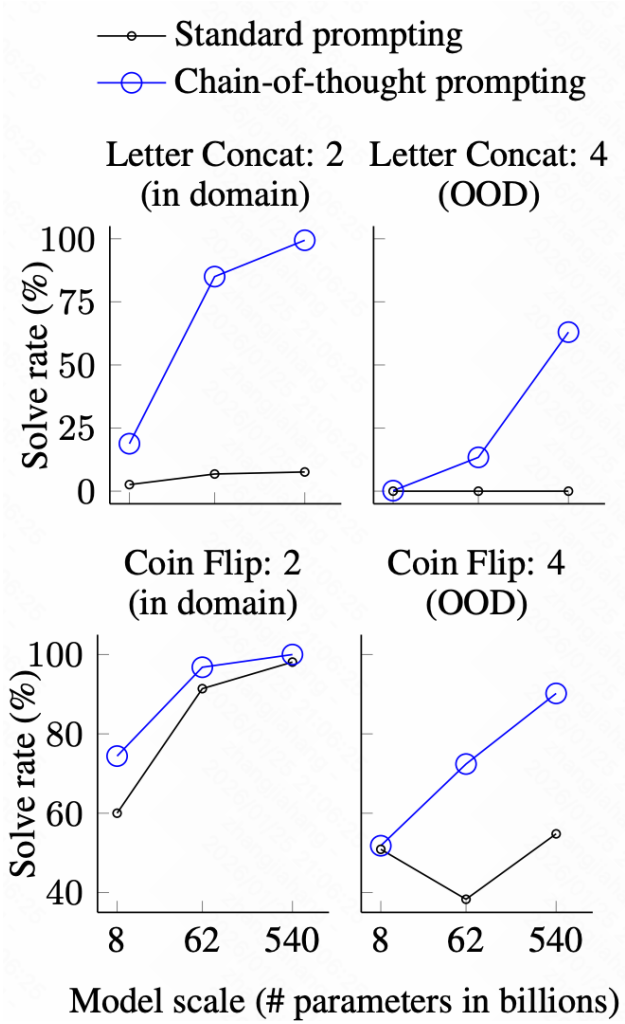
Chain-of-Thought

- CoT



Chain-of-Thought

- CoT Advantages
  - reasoning process, more computation for complex questions
  - interpretability
  - stronger performance on mathematic reasoning
  - can be training-free



Google, Chain-of-Thought Prompting Elicits Reasoning in Large Language Models, NeurIPS 2022

Chain-of-Thought

- Video CoT: Spatial-temporal thought

Question: According to the video, which of the following is the main reason why people commemorate Qu Yuan?
A. Because people love Zongzi.
B. Because he committed suicide by drowing himself in Miluo River.
C. Because he brought peace and prosperity to the state.
D. Because he was exile.

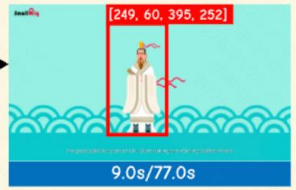
Ground Truth Answer: C

Qwen2.5-VL-7B: B

Video-R1-7B: <think>The video suggests the commemoration of Qu Yuan is associated with significant historical and cultural storytelling. Qu Yuan was noted for his contributions in poetry and governance. Despite his upright, loyal, and notably significant contributions, leading to peace and prosperity, he was maliciously accused. The vivid recall for his sorrow leads him towards a tragic, symbolic ending of diving into the Miluo River. The focus is on the ethical life symbolism of his personal loss and resilience. </think><answer>B</answer>

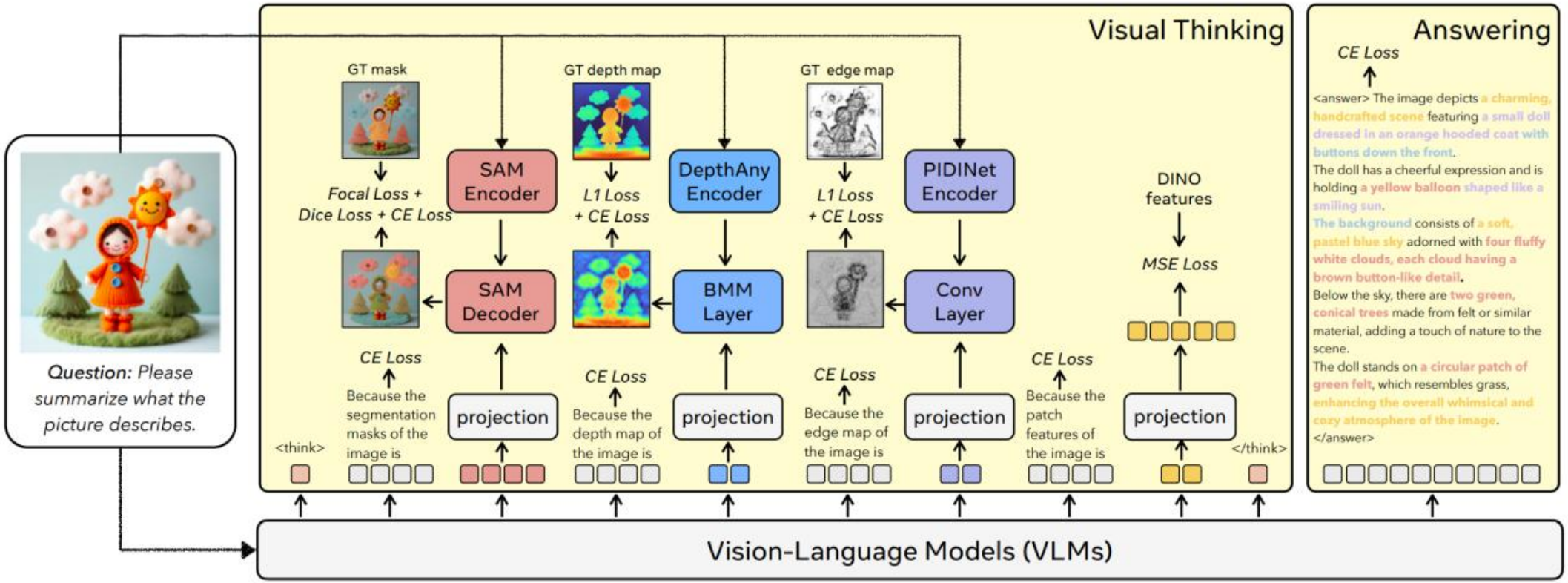
VideoRFT-7B: <think>To determine the main reason for commemorating Qu Yuan, I examined the video content. The video begins with a vibrant blue screen displaying Chinese characters and dragon motifs, which transitions into an animated sequence about Qu Yuan. The narrative shifts to his betrayal and exile, leading to his eventual suicide by drowning in the Miluo River. Later, the video mentions that people eat Zongzi and hold dragon boat races on the fifth day of May to honor Qu Yuan. Therefore, the key reason for commemorating Qu Yuan is his suicide, which is the central event depicted. Therefore, the answer is B, as the other options do not align with the main reason provided in the video. </think><answer>B</answer>

Open-o3 Video: <think>Think process: The video provides historical context about Qu Yuan. Initially, it shows a <obj>man</obj><box>[249,60,395,252]</box>at<t>9.0</t>s who is identified as the "great patriotic poet of Chu State." The narrative then describes his character, stating that he was "upright, loyal and highly esteemed," which led to the people's respect for him, brought peace and prosperity, and influenced the dragon boat festival. The video concludes by mentioning the damming of the river and Qu Yuan's suicide by diving into it. This information is visualized through various scenes, such as people bowing to <obj>the man</obj><box>[472,70,612,269]</box>at<t>24.0</t>s and the silhouette of him near mountains, which align with the narration. Therefore, the cause for celebrating him is his significant role in bringing peace and prosperity.</think> <answer>C</answer>



Chain-of-Thought

- Latent CoT

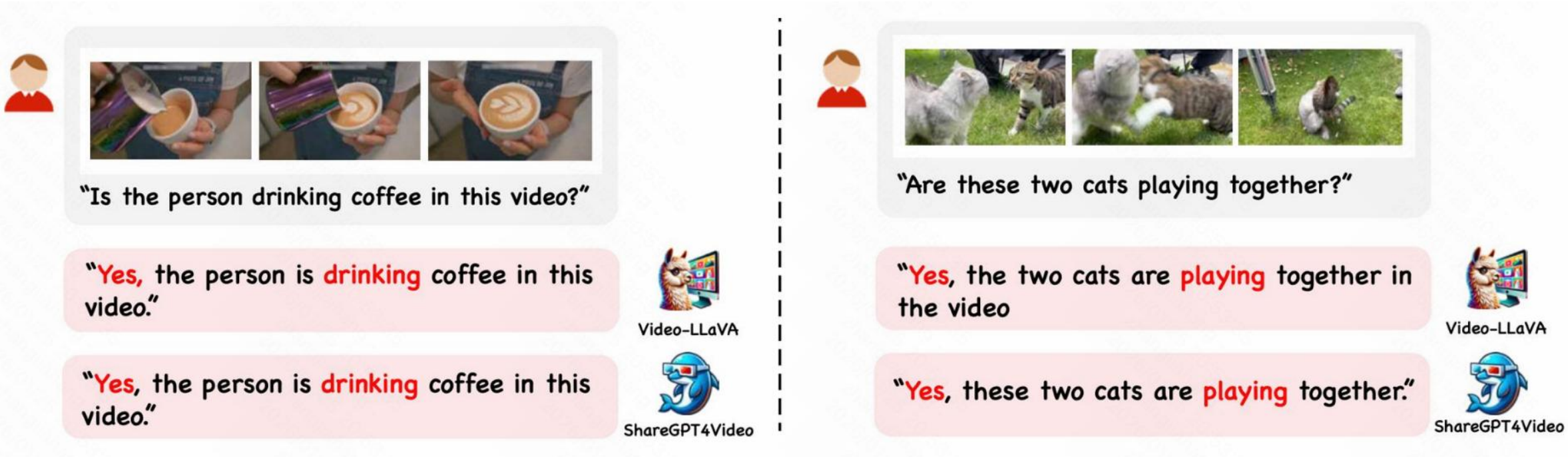


Multi-Modal Large Language Models

- Challenges
  - Spatial Reasoning
  - Fine-grained Perception: small object, category calssification, counting ...
  - Long-Term Modeling & Forgetting Problem
  - Efficiency
  - ...

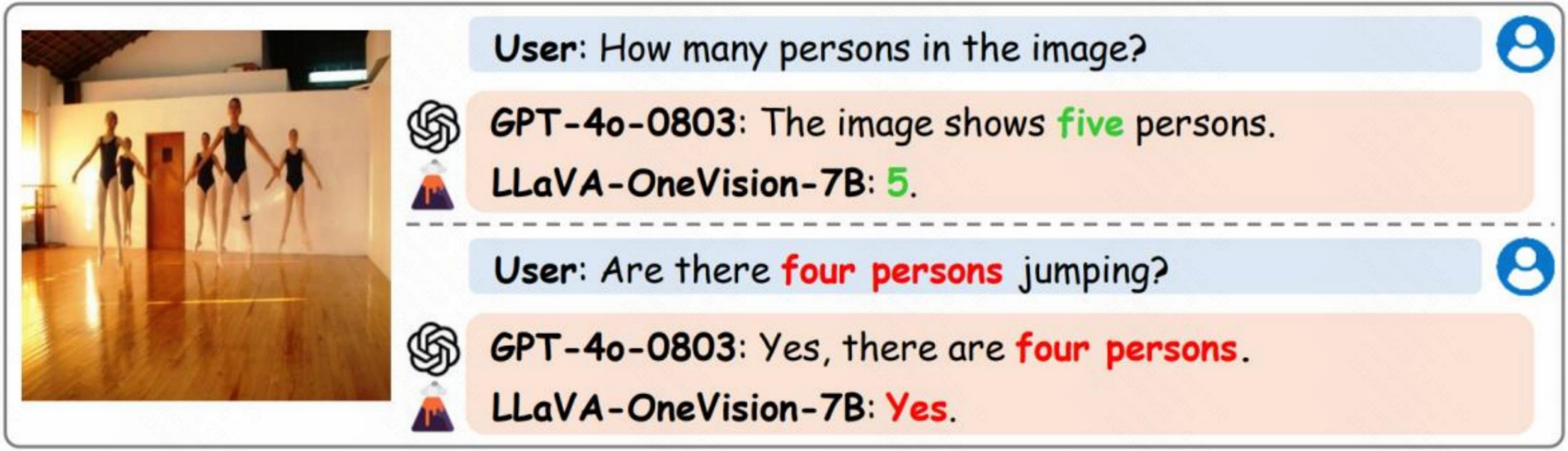
Multi-Modal Large Language Models

- Hallucinations



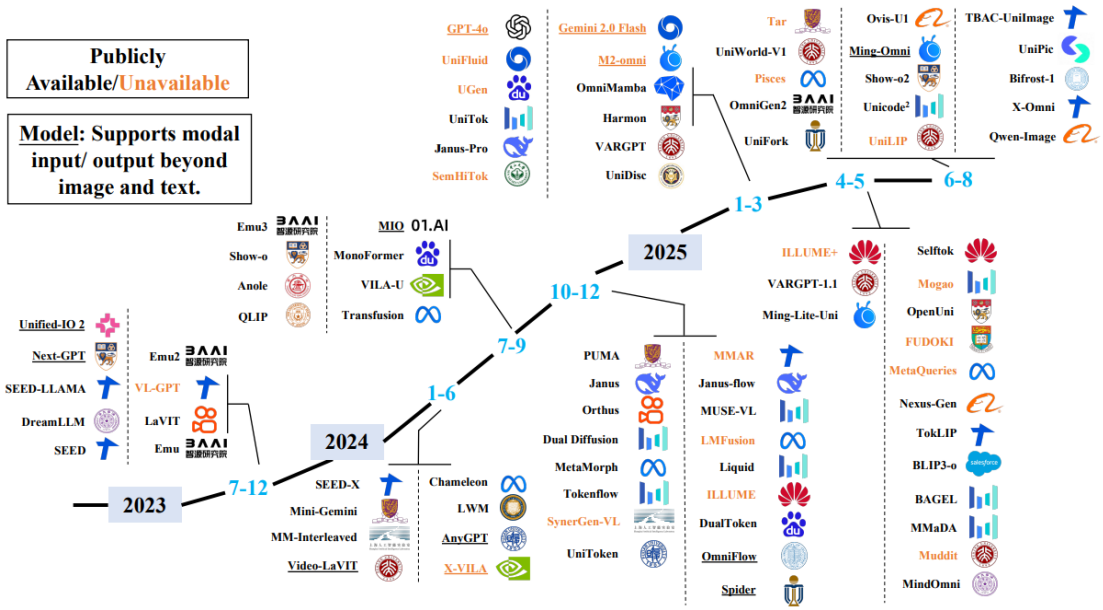
Multi-Modal Large Language Models

- Hallucinations



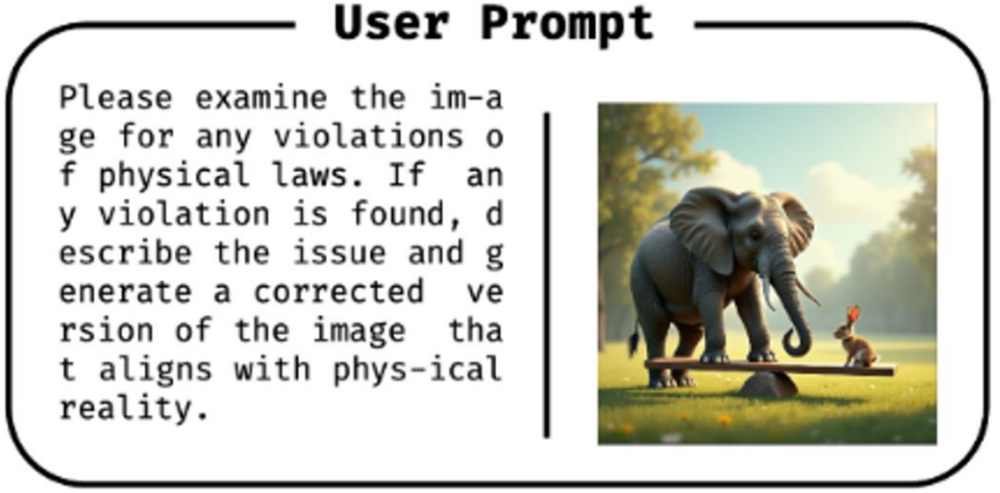
Unified Generation-Understanding Models

- What is Unified Multimodal Model?
  - Generation & understanding capability in a unified model
  - “Any-to-any” flexibility
    - image2text
    - text2image
    - interleaved multimodal sequences
    - ...
  - Shared Semantic Space
    - Align different data into a unified representation



Unified Generation-Understanding Models

- Why we need Unified Multimodal Model?
  - Reduced Complexity
    - Simplified E2E framework instead of complex pipelines
      - Pipeline
        - Understanding (in: text, img; out: text)
        - Generation (in: text; out: img)
      - UMM
        - input: text, img
        - out: text, img



Unified Generation-Understanding Models

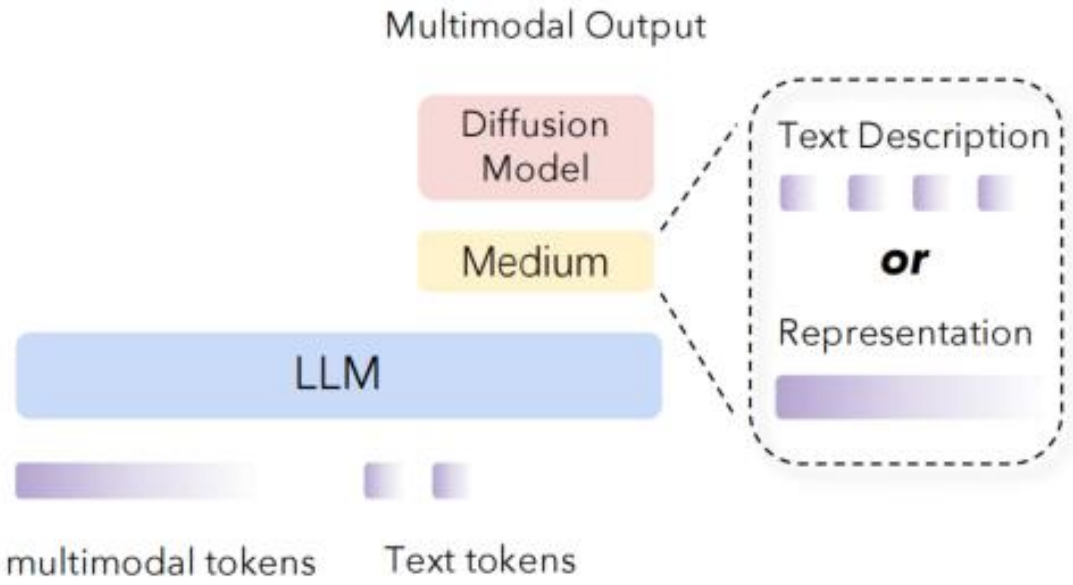
- Why we need Unified Multimodal Model?
  - Cross-Modal Synergy
    - Generation&understanding tasks are training on a unified backbone
    - Generation training improves understanding
      - Spatial awareness
      - Finer details
    - Understanding training improves generation
      - logical consistency

Unified Generation-Understanding Models

- Two Types of Unified Multimodal Model
  - Modular Joint Modeling
    - Prompt-Mediated
    - Representation-Mediated
  - End-to-End Unified Modeling
    - Autoregressive
    - Diffusion
    - Autoregressive-Diffusion Hybrid

Modular Joint Modeling

- Bridge understanding and generation model with a alignment module
- Directly utilize the pre-trained model
- Improve generation with understanding model
- Two types:
  - Prompt-Mediated
  - Representation-Mediated



End-to-End Unified Modeling

- Jointly modeling understanding and generation through end-to-end training
- Unified modeling input and target modality within the model itself
  - Reduce information loss during modality transformation
  - Enhancing both performance
- Three types:
  - Autoregressive
  - Diffusion
  - Autoregressive-Diffusion Hybrid

End-to-End Unified Modeling

- Autoregressive: Emu3

BAAI, Next-token predicition is all you need, arXiv 2024

End-to-End Unified Modeling

- Autoregressive: Emu3
  - Architecture baseline: Transformer like Llama-2
  - Architecture improvement
    - Vision Tokenizer: SBER-MoVQGAN5
    - Longer Context Length: 131072
    - Rotary Positional Embeddings (RoPE)

Configurations	Emu3
Parameters	8B
Layers	32
Hidden Size	4096
Intermediate Size	14336
Heads	32
KV Heads	8
Vocabulary Size	184622
RoPE Base	1000000
Context Length	131072

BAAI, Next-token predicition is all you need, arXiv 2024

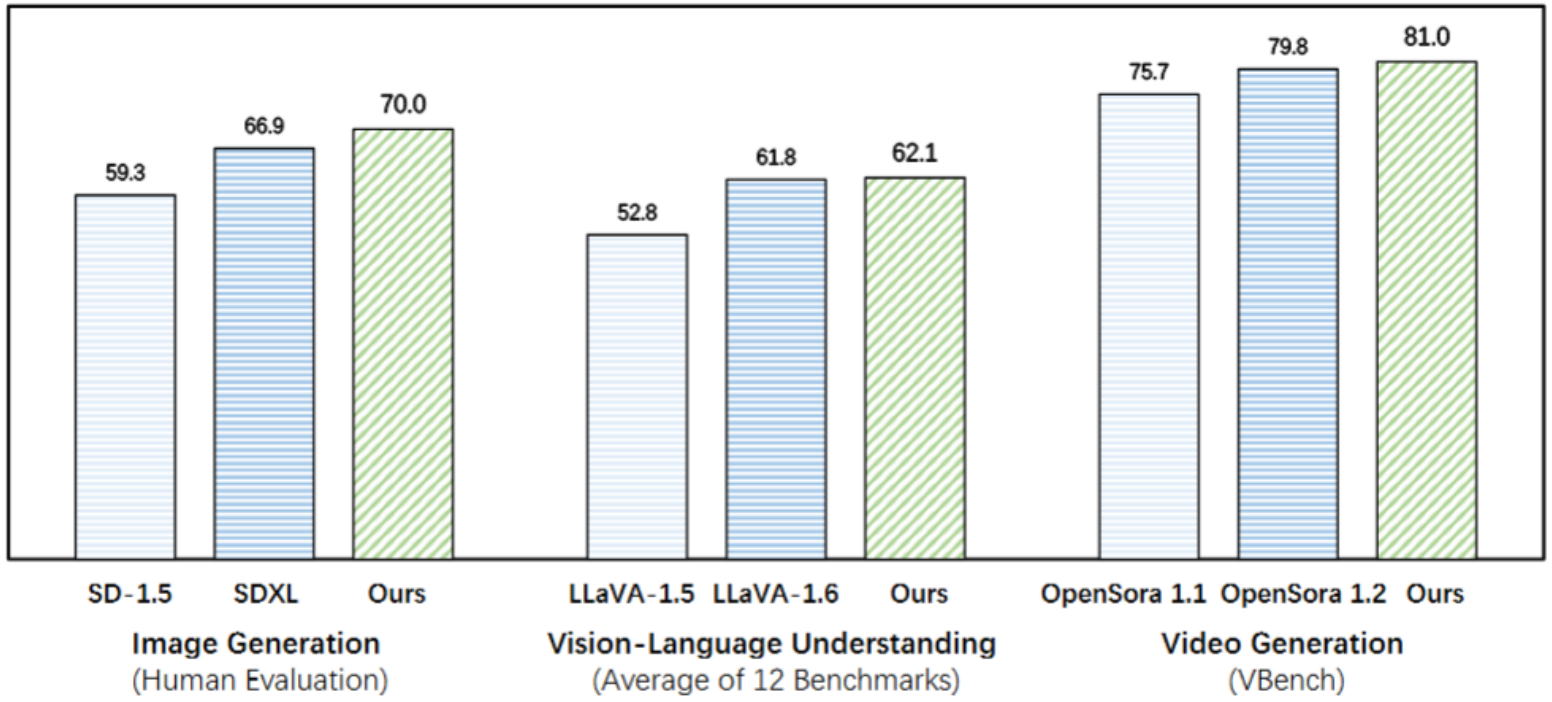
End-to-End Unified Modeling

- Autoregressive: Emu3
  - Training Mechanism
    - Data: Text/Image/Video
    - Loss: Standard cross-entropy loss for next-token prediction task
    - Pre-Training
      - Stage 1: Training on text and image, with context length=5120
      - Stage 2: Training on text, image and video, with context length=131072
    - Post-Training
      - Quality Fine-Tuning
      - Direct Preference Optimization
      - Vision-Language Understanding

BAAI, Next-token predicition is all you need, arXiv 2024

End-to-End Unified Modeling

- Autoregressive: Emu3
  - Performance



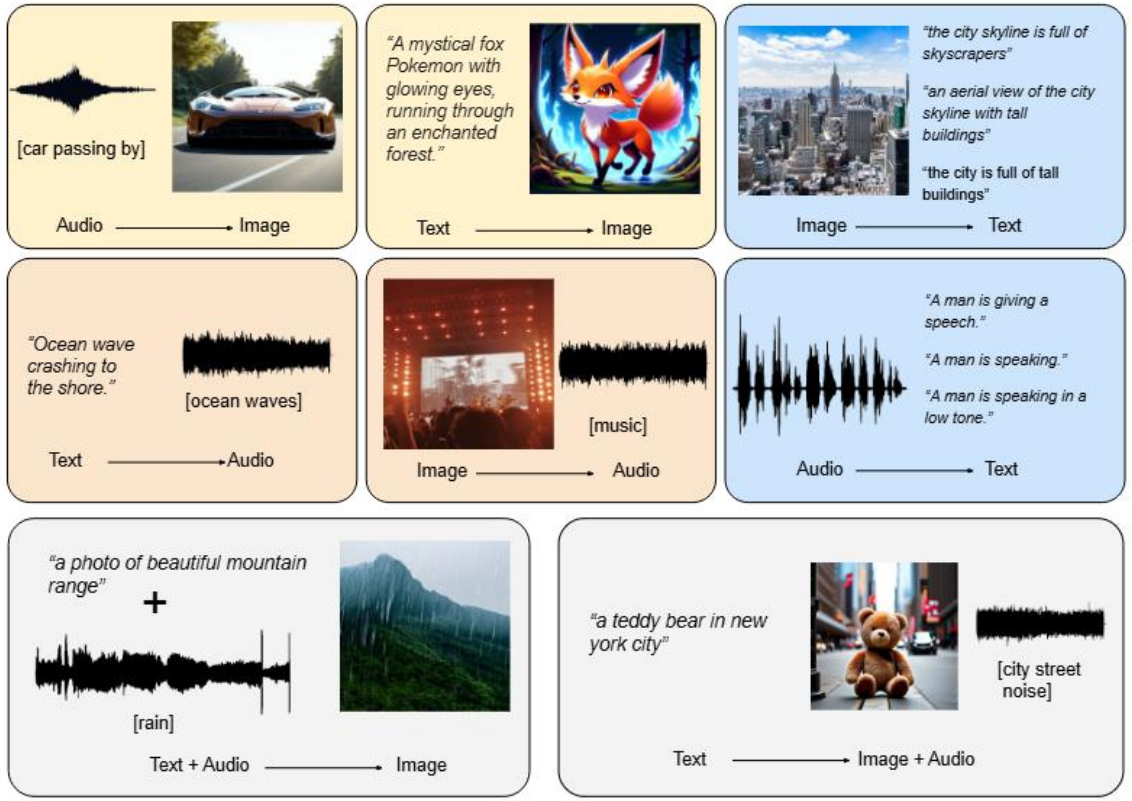
BAAI, Next-token predicition is all you need, arXiv 2024

End-to-End Unified Modeling

- Autoregressive
  - Pros
    - Aligns naturally with LLMs, simplicity
    - Jointly optimize understanding and generation within a shared semantic space
  - Cons
    - Substantial differences across modalities
    - Token-based generation introduces information bottlenecks
    - Error accumulation during autoregressive generation
    - Sequential generation process lead to low efficiency

End-to-End Unified Modeling

- Diffusion: OmniFlow



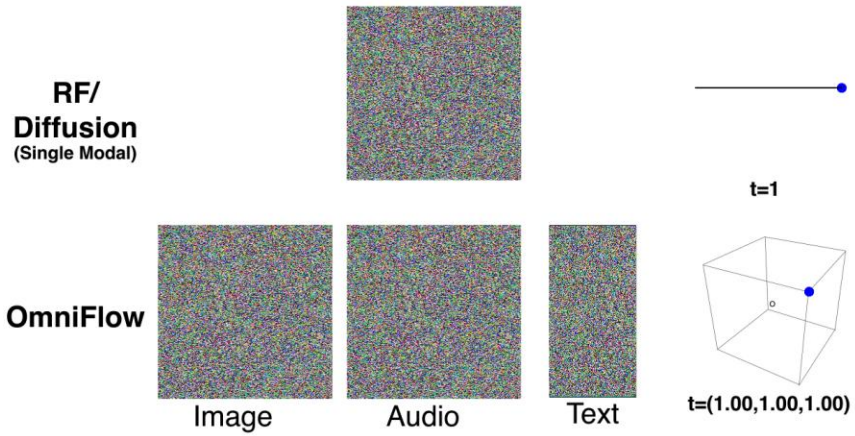
UCLA, OmniFlow: Any-to-any generation with multi-modal rectified flows, CVPR 2025

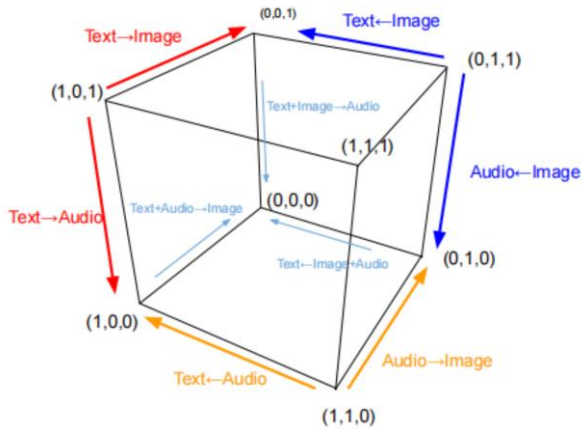
End-to-End Unified Modeling

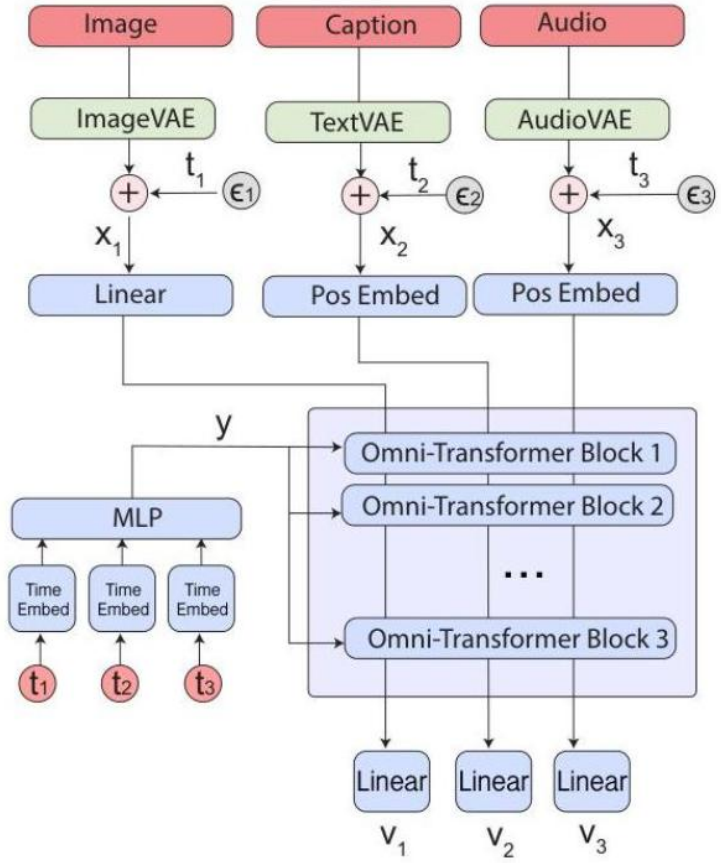
- Diffusion: OmniFlow
  - Modeling: Rectified Flow

$$x_i^t = (1 - t_i)x_i^0 + t_i x_i^1$$

  - Architecture: MMDiT







End-to-End Unified Modeling

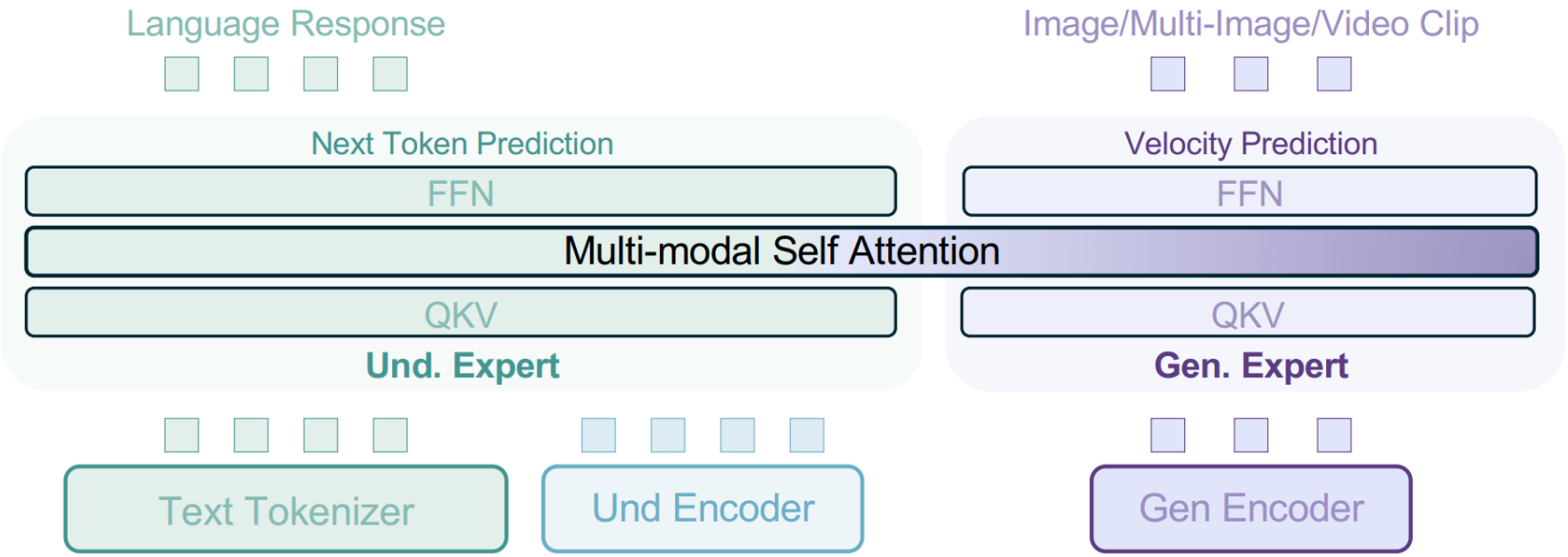
- Diffusion: OmniFlow
  - Training
    - Stage 1: Load pretrained parameters from SD3 for text2img
    - Stage 2: Training a text2audio model
    - Stage 3: Merge these two models with average weight on text branch
    - Stage 4: Finetune the unified model on multimodal dataset

End-to-End Unified Modeling

- Diffusion
  - Pros
    - High quality and generation capabilities
  - Cons
    - Slow inference speed
    - Relatively weak multimodal understanding

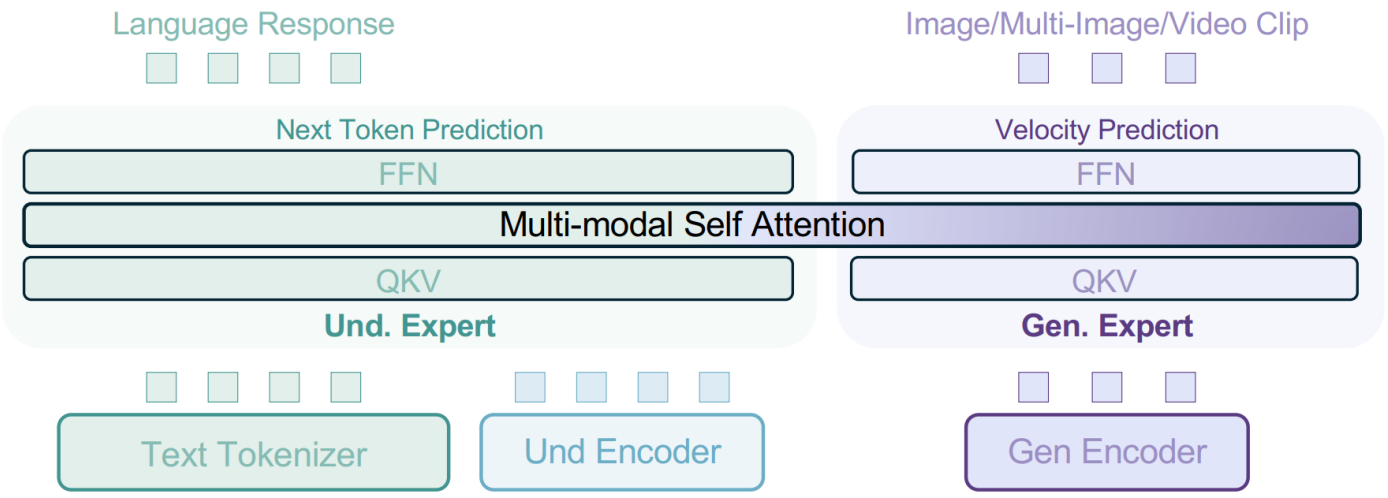
End-to-End Unified Modeling

- Autoregressive-Diffusion Hybrid: BAGEL



End-to-End Unified Modeling

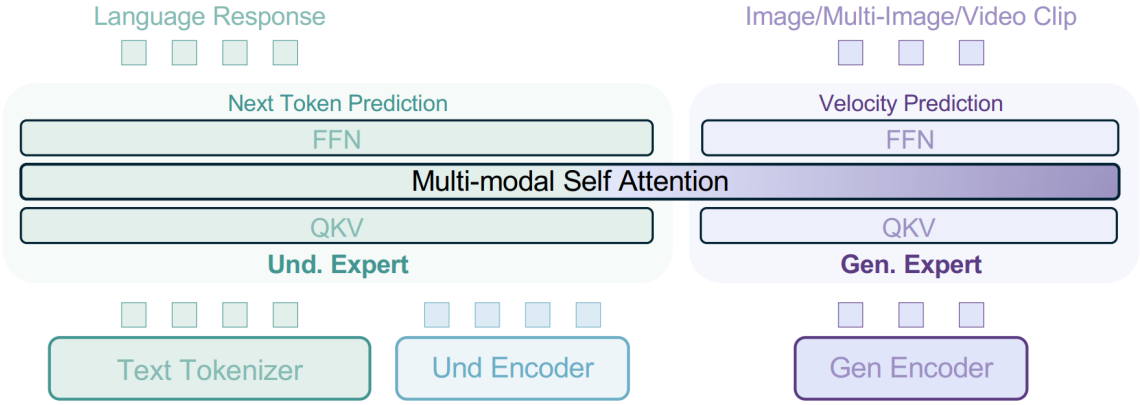
- Autoregressive-Diffusion Hybrid: BAGEL
  - Architecture
    - Two decoder-only transformer expert for Und.&Gen
    - Share the multi-modal self attention
    - Vision Tokenizer
      - ViT-based Und. Enc.
      - VAE-based Gen. Enc.
  - Backbone: Qwen2.5



End-to-End Unified Modeling

- Autoregressive-Diffusion Hybrid: BAGEL
  - Training
    - Data
    - Strategy
      - Stage 1: Alignment Und. Enc. with Qwen
      - Stage 2: Pre&Continued training
        - Clean tokens for Und.
        - Noised tokens for Gen.
      - Stage 3: Supervised Fine-tuning

Data Source	# Data (M)	# Tokens (T)
Text Data	400	0.4
Image-Text-Pair Understanding Data	500	0.5
Image-Text-Pair Generation Data	1600	2.6
Interleaved Understanding Data	100	0.5
Interleaved Generation Data: Video	45	0.7
Interleaved Generation Data: Web	20	0.4



End-to-End Unified Modeling

- Autoregressive-Diffusion Hybrid
  - Pros
    - Avoid information transmission bottlenecks
    - Enhance generative quality
    - Maintain understanding quality
  - Cons
    - Noise ingection may compromise understanding performance
    - Parameter sharing introduces conflicts

World Models

- Summary
  - The definition and taxonomy of world models
  - Representation focused world models: Multi-Modal Large Language Models
  - Unified Generation-Understanding Models: End-to-End Unified Modeling
- Next Episode Preview
  - Generation focused world models
  - Generation improvement with understanding model

Thanks!