JanusFlow: Harmonizing Autoregression and Rectified Flow for Unified Multimodal Understanding and Generation

CVPR 25'

Yiyang Ma[1,2], Xingchao Liu[1], Xiaokang Chen[1], Wen Liu[1], Chengyue Wu[1,3], Zhiyu Wu[1], Zizheng Pan[1], Zhenda Xie[1], Haowei Zhang[1], Xingkai Yu[1], Liang Zhao[1], Yisong Wang[1,4], Jiaying Liu[2], Chong Ruan[1]

[1]DeepSeek-AI, [2]Peking University, [3]The University of Hong Kong, [4]Tsinghua University

马逸扬

2025.03.30

Content

- Authors
- Background
- Method
- Experiments

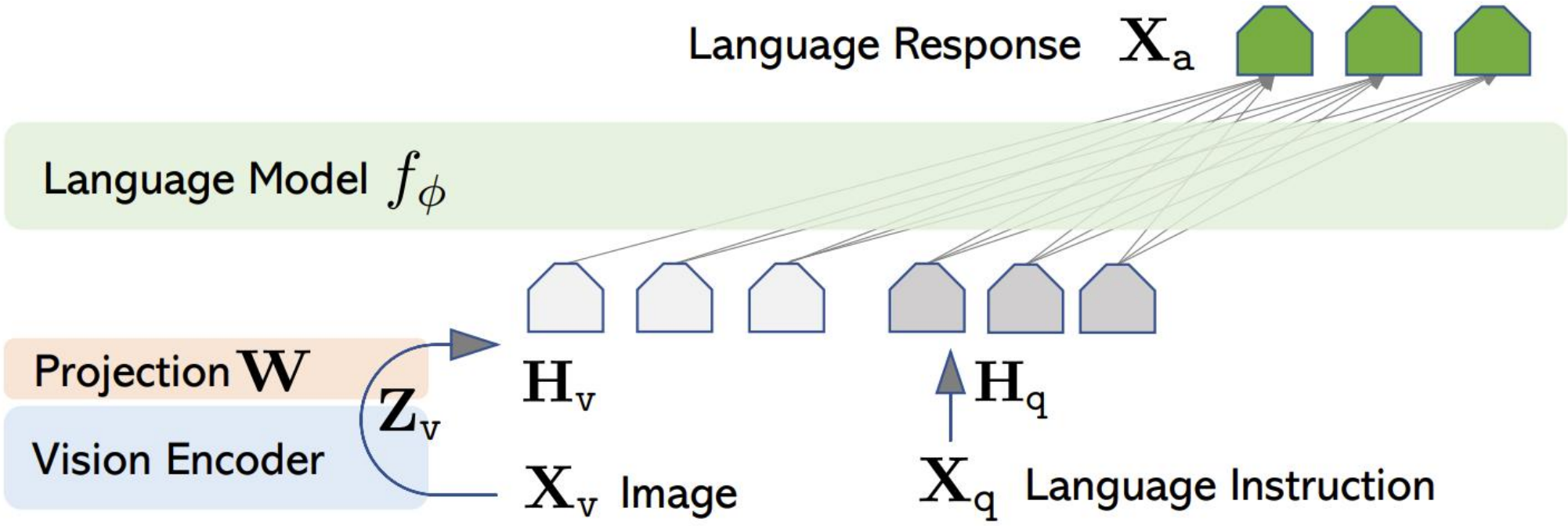
Content

- Authors
- Background
- Method
- Experiments

Background

Large Vision-Language Models

LLaVA: Combining pre-trained visual encoder and LLMs



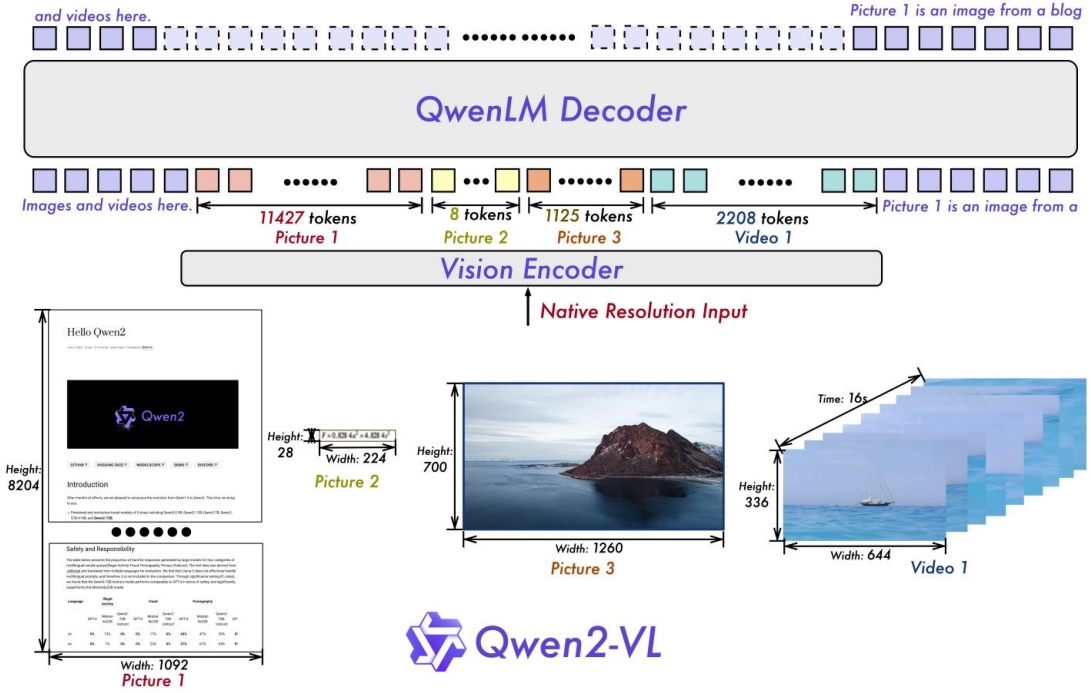
Problem: Do pre-trained models and text-only losses limit the performance?

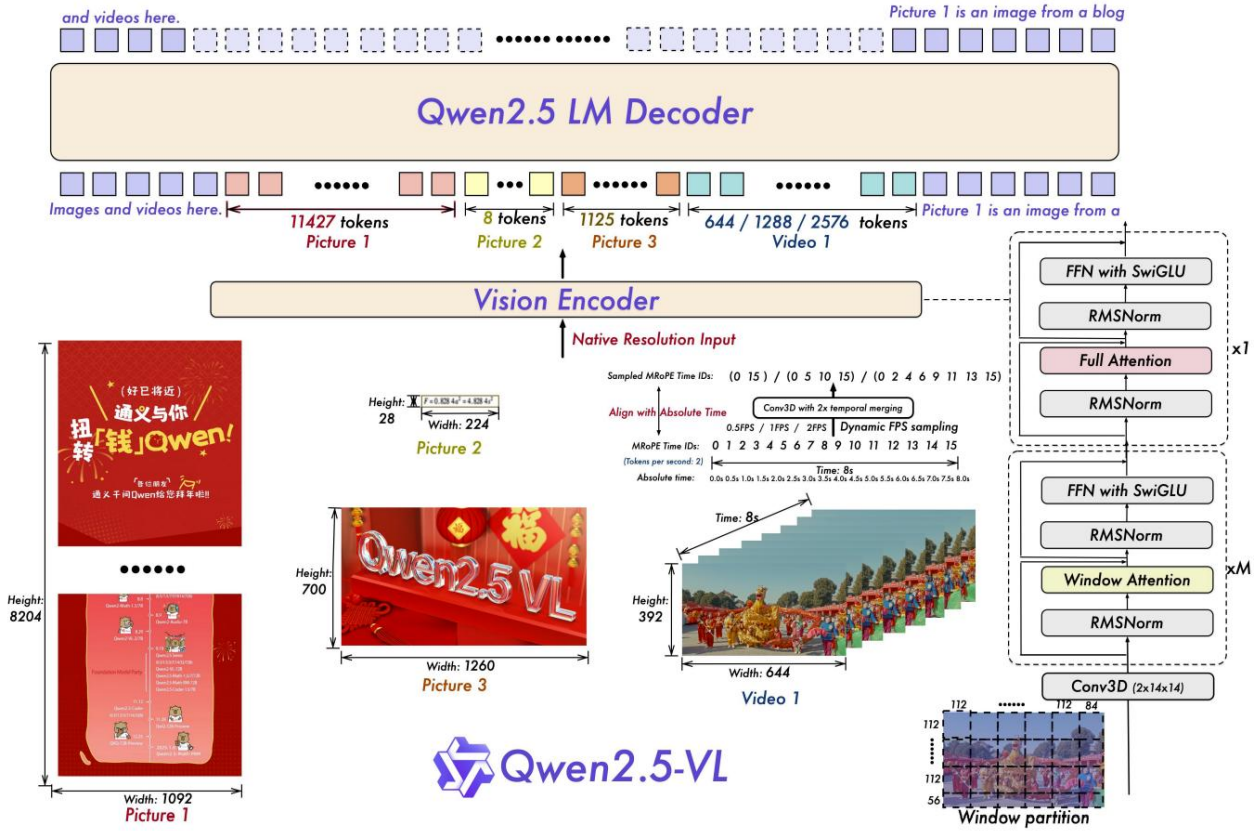
[1] Haotian Liu et al. Visual instruction tuning, NIPS23'.

Background

Large Vision-Language Models

Qwen VL series



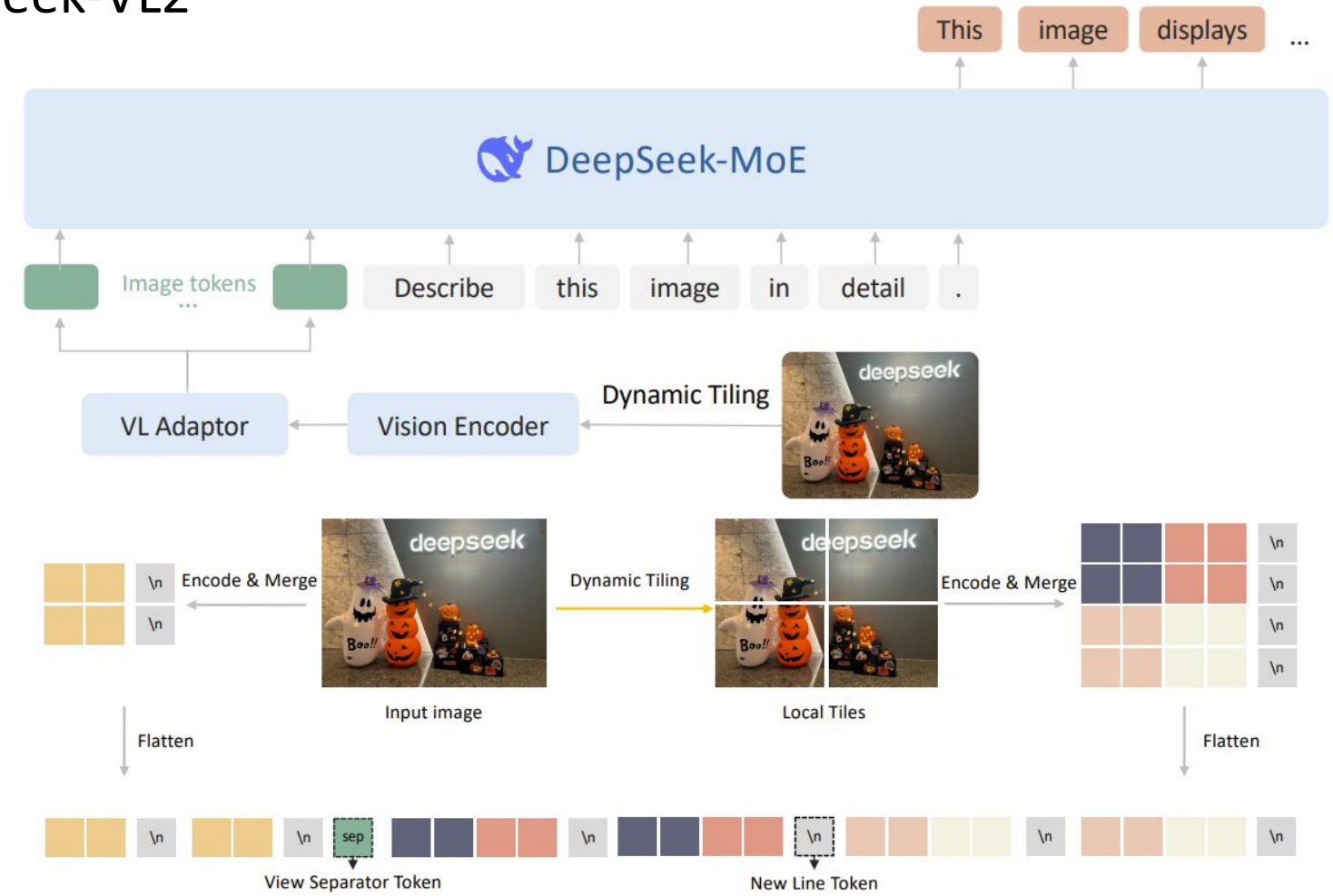


[4] Qwen Team. Qwen2-VL: Enhancing vision-language model's perception of the world at any resolution, arXiv 2409.
[5] Qwen Team. Qwen2.5-VL Technical report, arXiv 2502.

Background

Large Vision-Language Models

DeepSeek-VL2

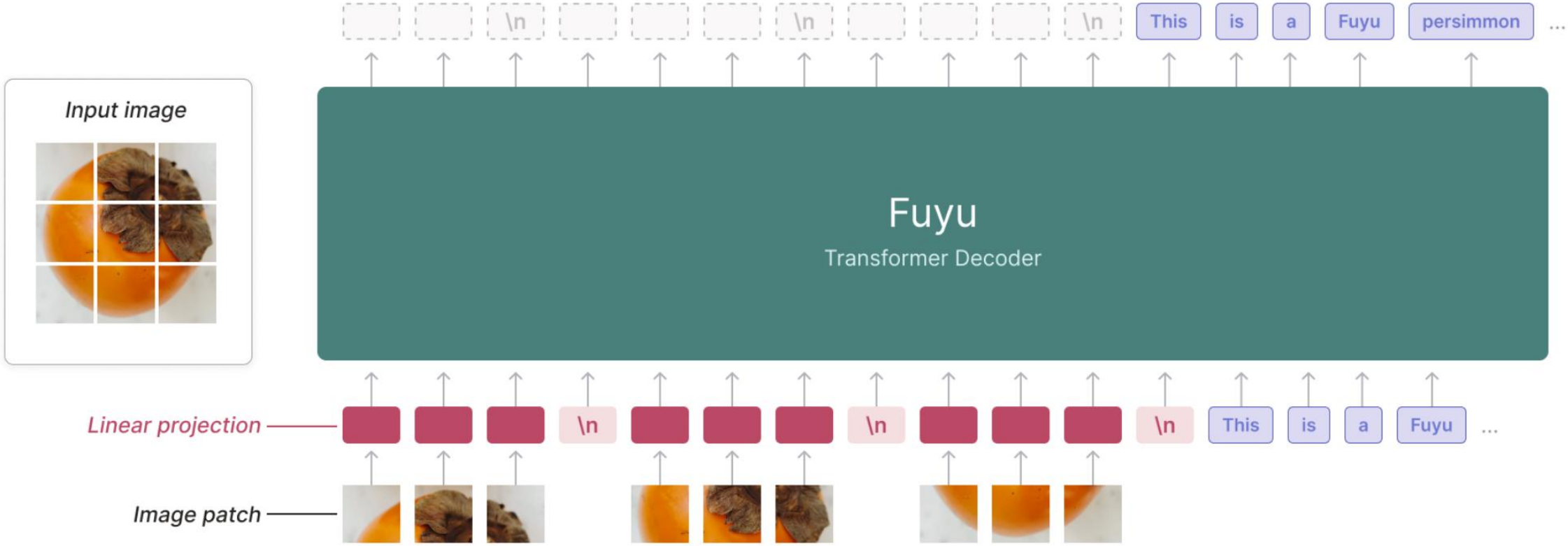


[6] DeepSeek-AI. DeepSeek-VL2: Mixture-of-experts vision-language models for advanced multimodal understanding, arXiv2412.

Background

Large Vision-Language Models

Fuyu-8B

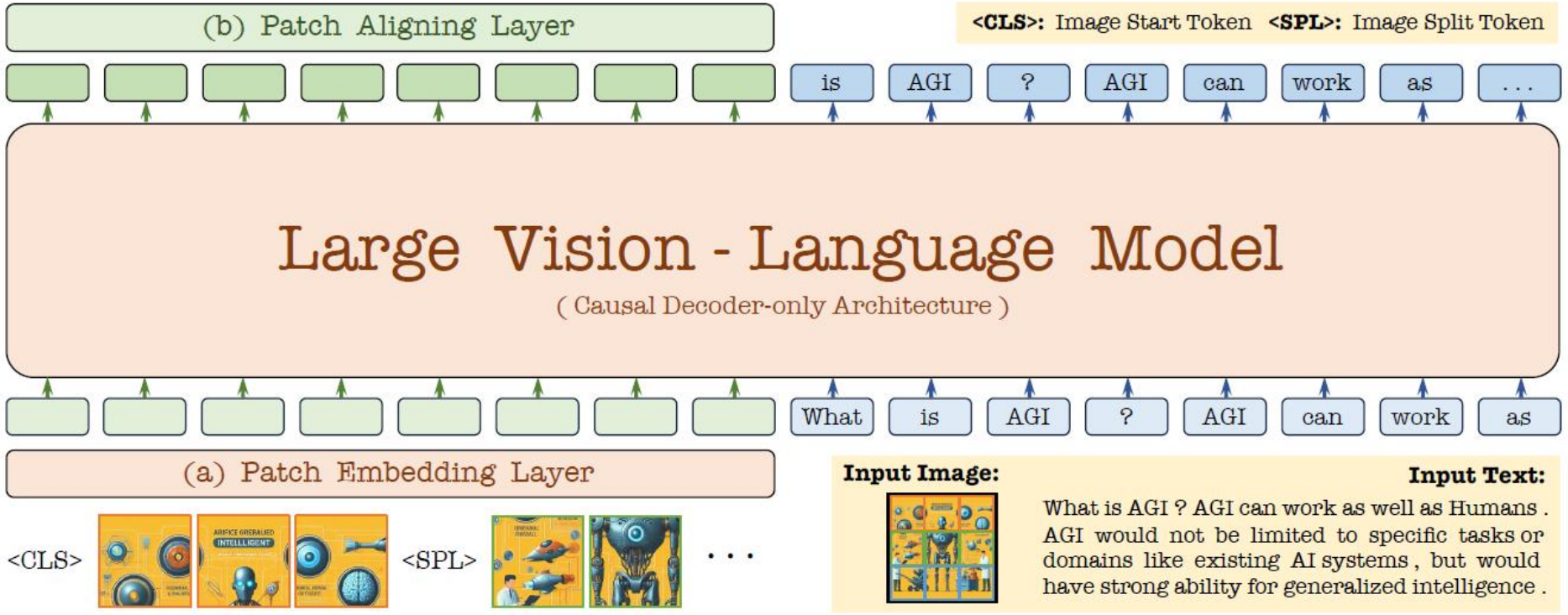


[2] Rohan Bavishi et al. Fuyu-8B: A multimodal architecture for ai agents, https://www. adept. ai/blog/fuyu-8b.

Background

Large Vision-Language Models

EVE: Encoder-free vision-language models

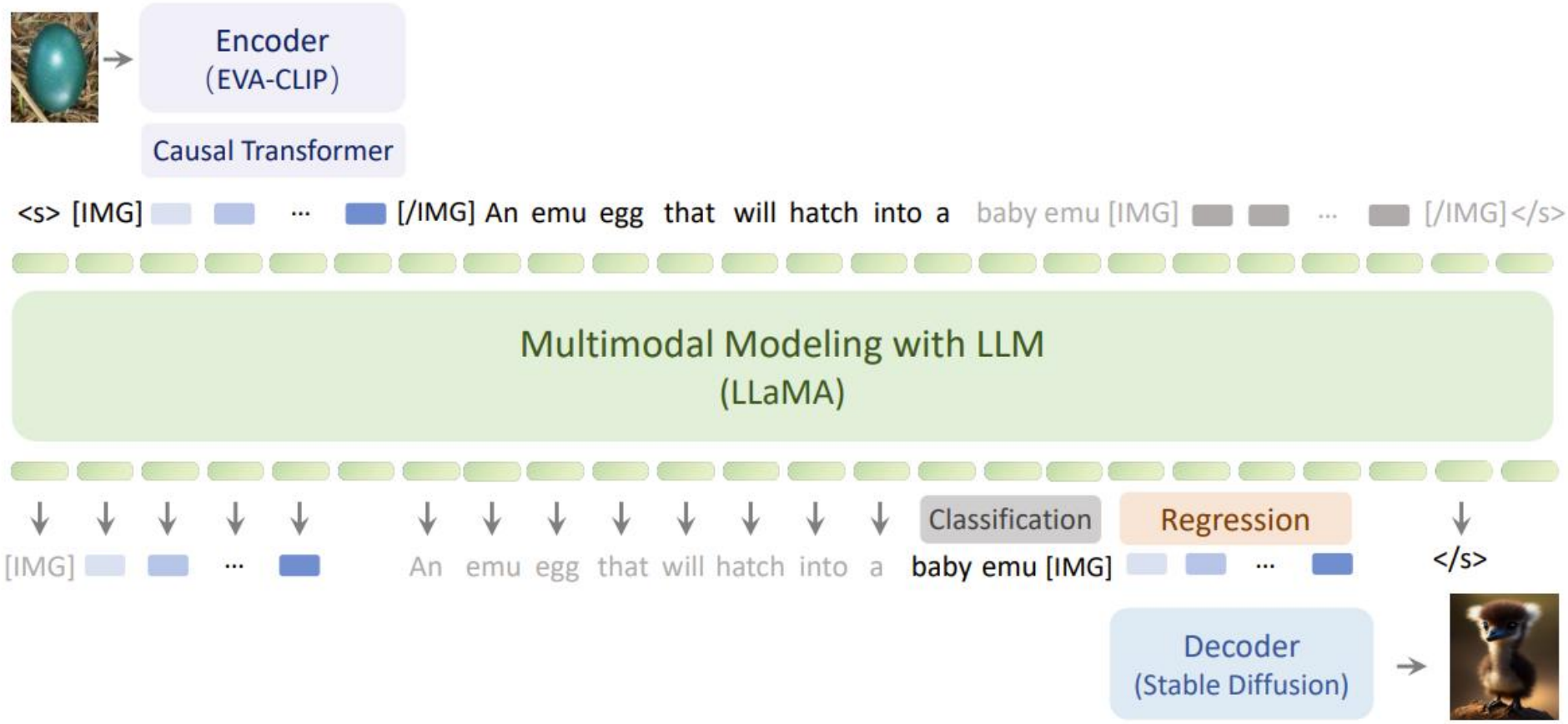


[3] Haiwen Diao et al. Unveiling encoder-free vision-language models, NIPS24'.

Background

Unified Understanding and Generation Models

Emu: Additional generative models

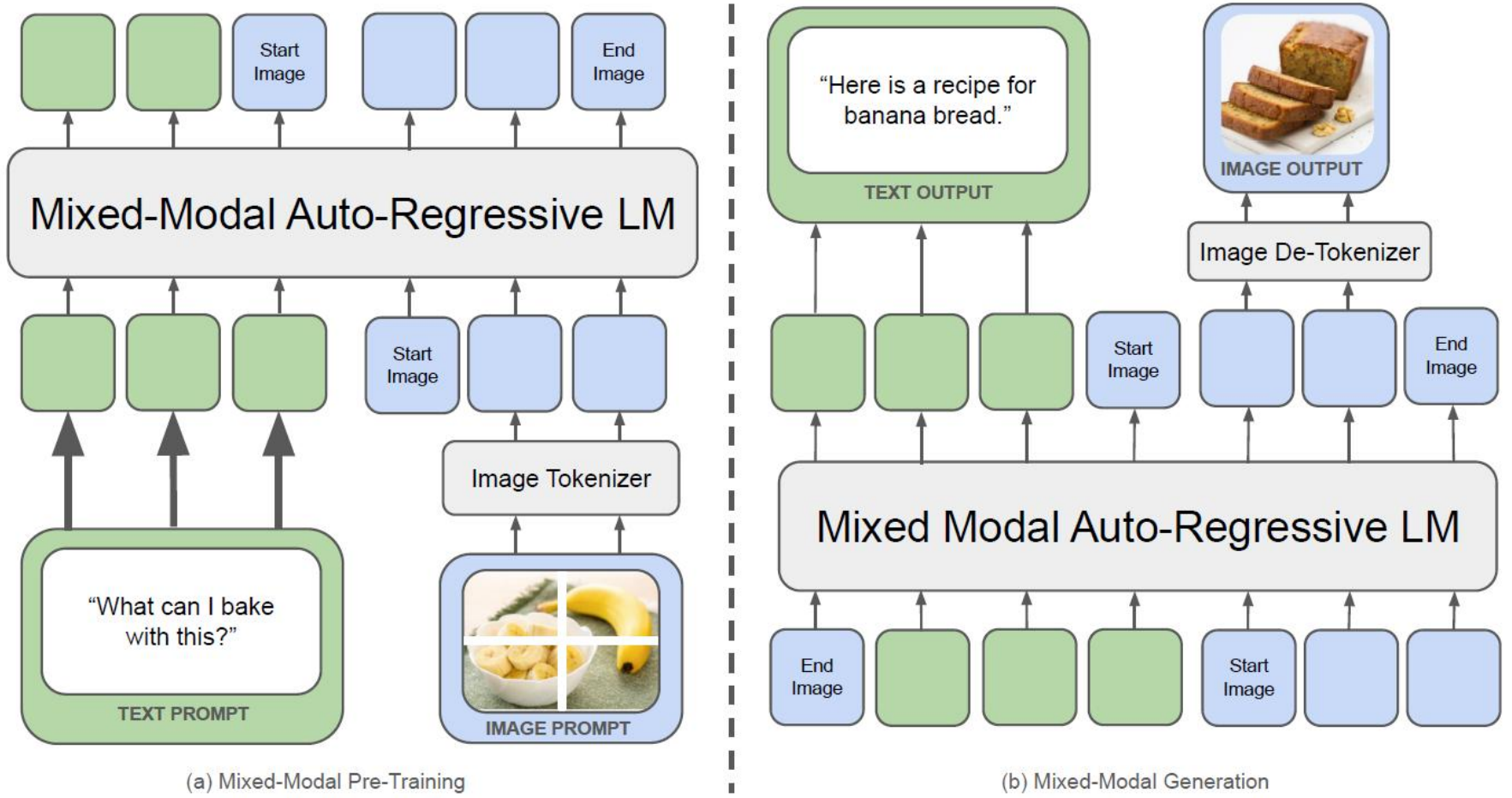


[6] Quan Sun et al. Emu: Generative pretraining in multimodality, ICLR24'.

Background

Unified Understanding and Generation Models

Chameleon: Unified AR for both tasks

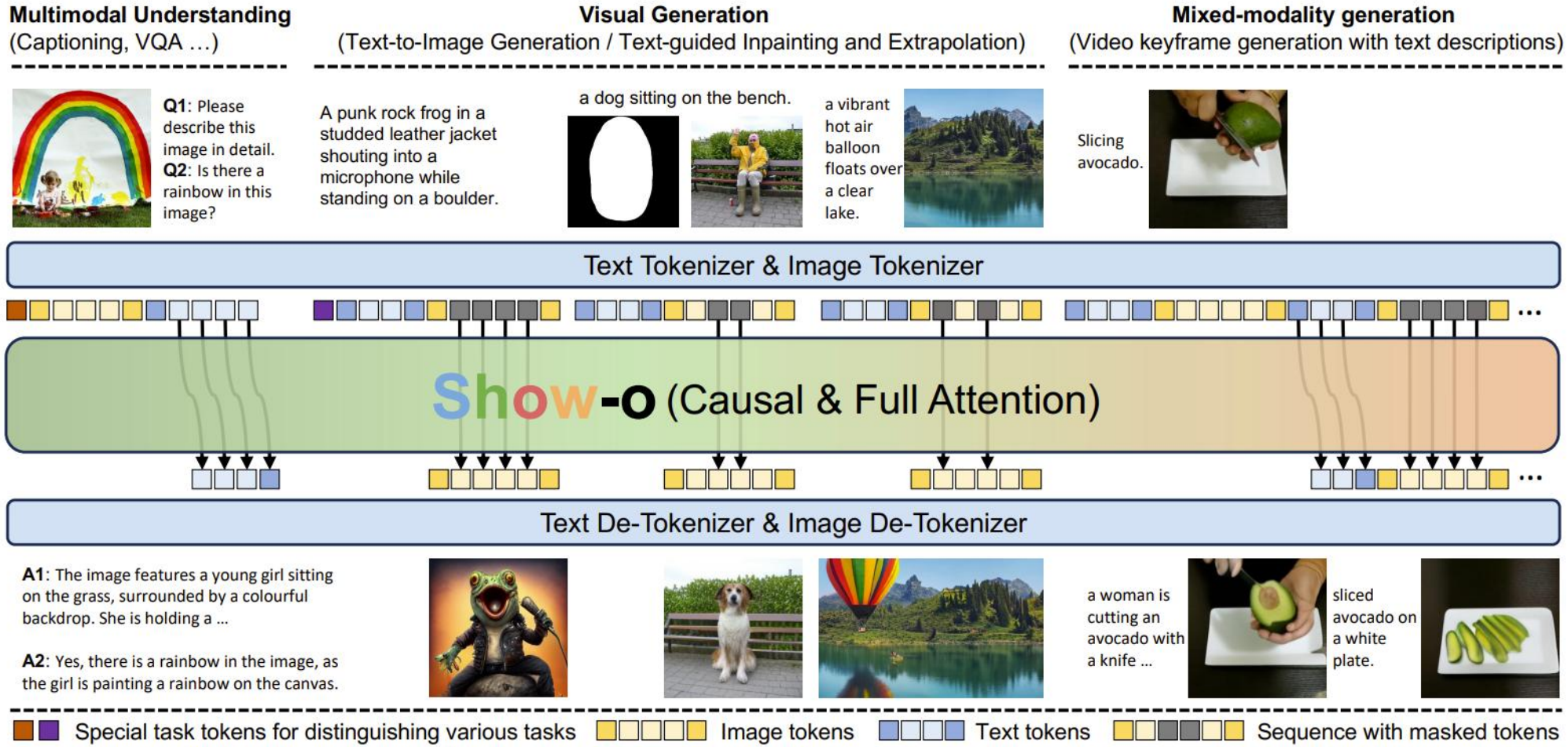


[7] Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models, arXiv2405'.

Background

Unified Understanding and Generation Models

Show-o: Combining MAGVIT and AR

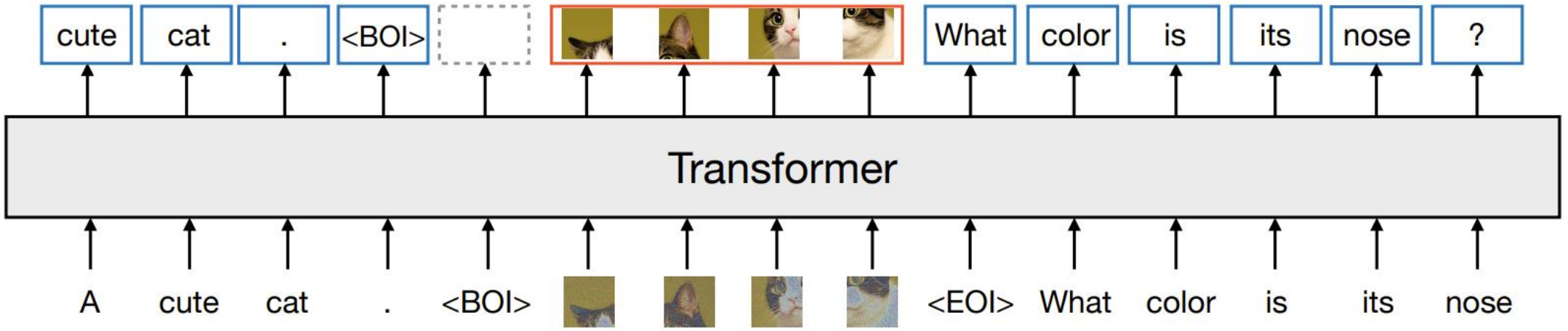


[8] Jinheng Xie et al. Show-o: One single transformer to unify multimodal understanding and generation, ICLR25'.
[9] Lijun Yu et al. Language model beats diffusion -- tokenizer is key to visual generation, ICLR24'.

Background

Unified Understanding and Generation Models

Transfusion: Combining diffusion and AR

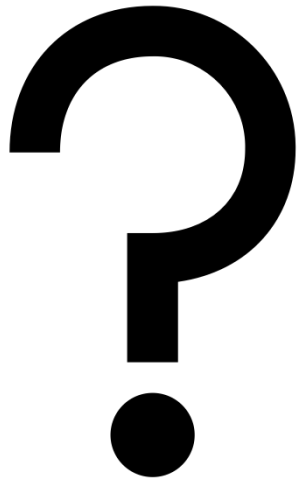


[10] Chunting Zhou et al. Transfusion: predict the next token and diffuse images with one multi-modal model, arXiv2408.

Background

Unified Understanding and Generation Models

GPT-4o



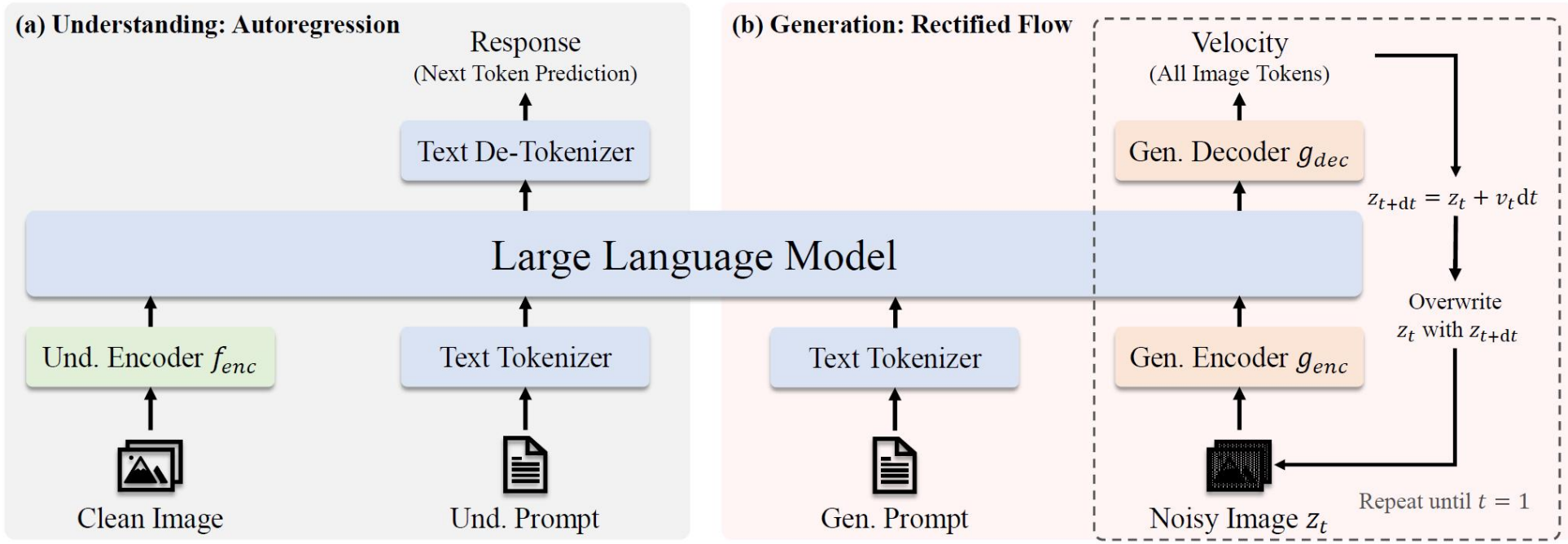
[11] OpenAI. Addendum to GPT-4o system card: Native image generation, https://cdn.openai.com/11998be9-5319-4302-bfbf-1167e093f1fb/Native_Image_Generation_System_Card.pdf.

Content

- Authors
- Background
- Method
- Experiments

Method

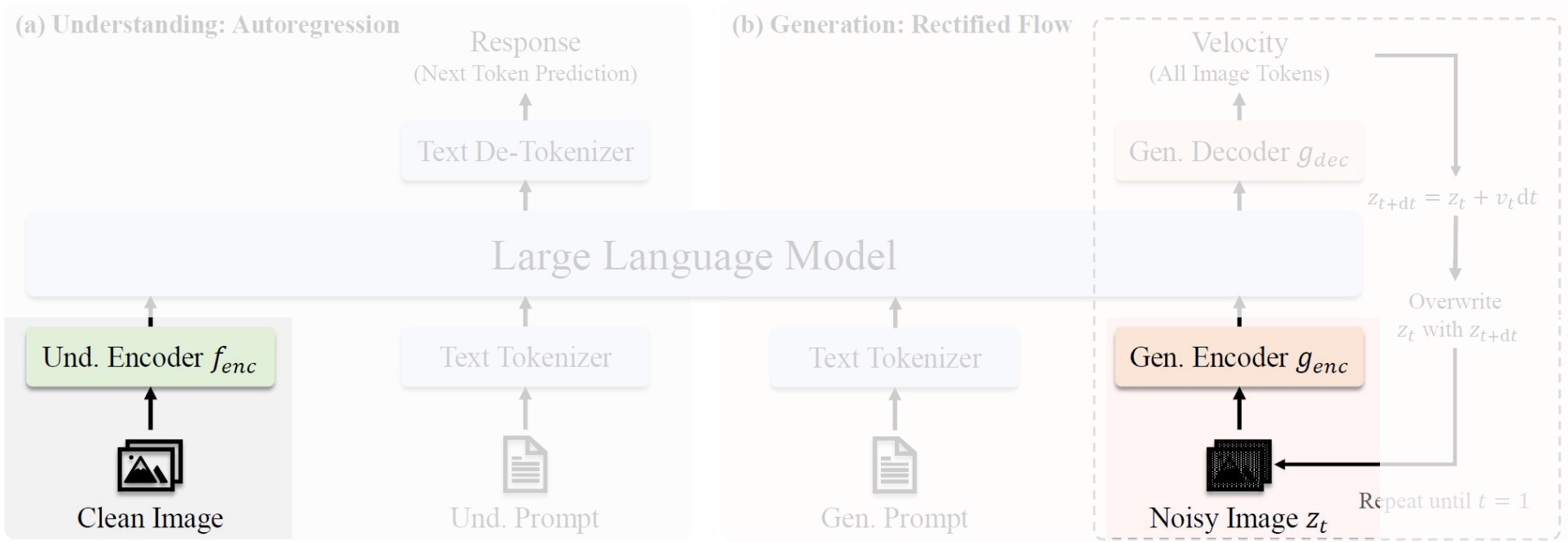
JanusFlow



- Understanding: Same as LLaVA-like models.
- Generation: Backbone as DiT.
- Semantics are directly processed by the LLM without additional generative models as decoder.

Method

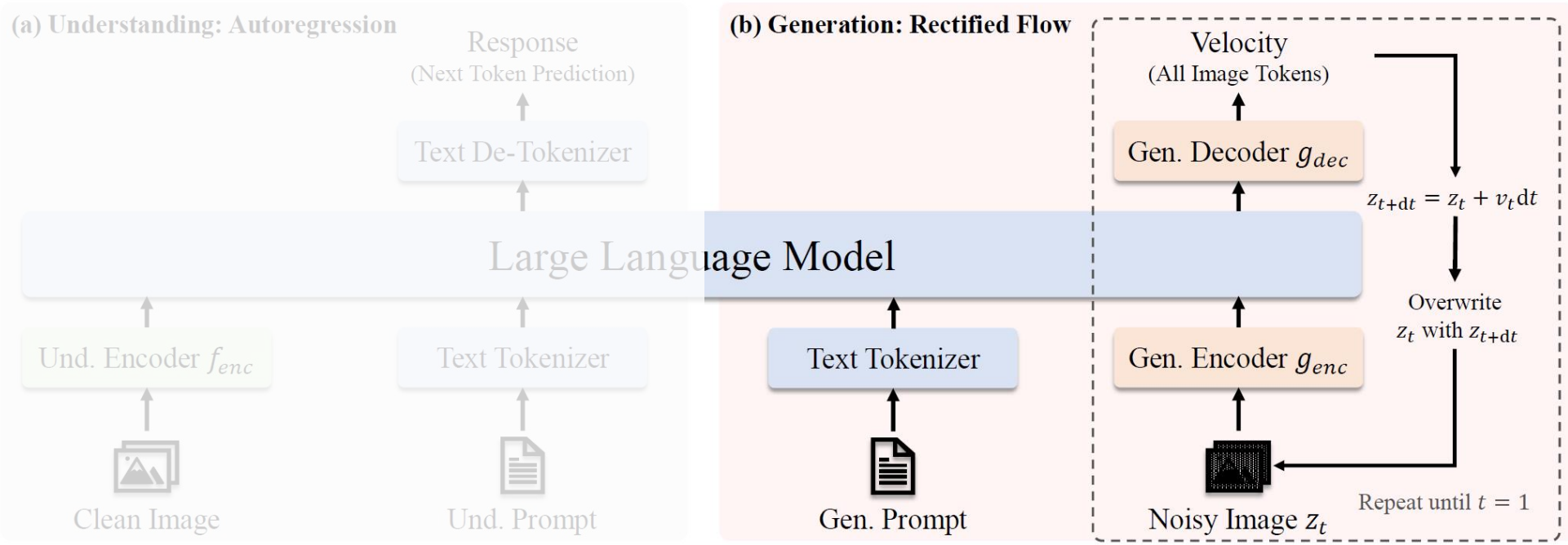
JanusFlow



- Decouple the task of understanding and generation.
- Assist the backbone to handle both of the two tasks.

Method

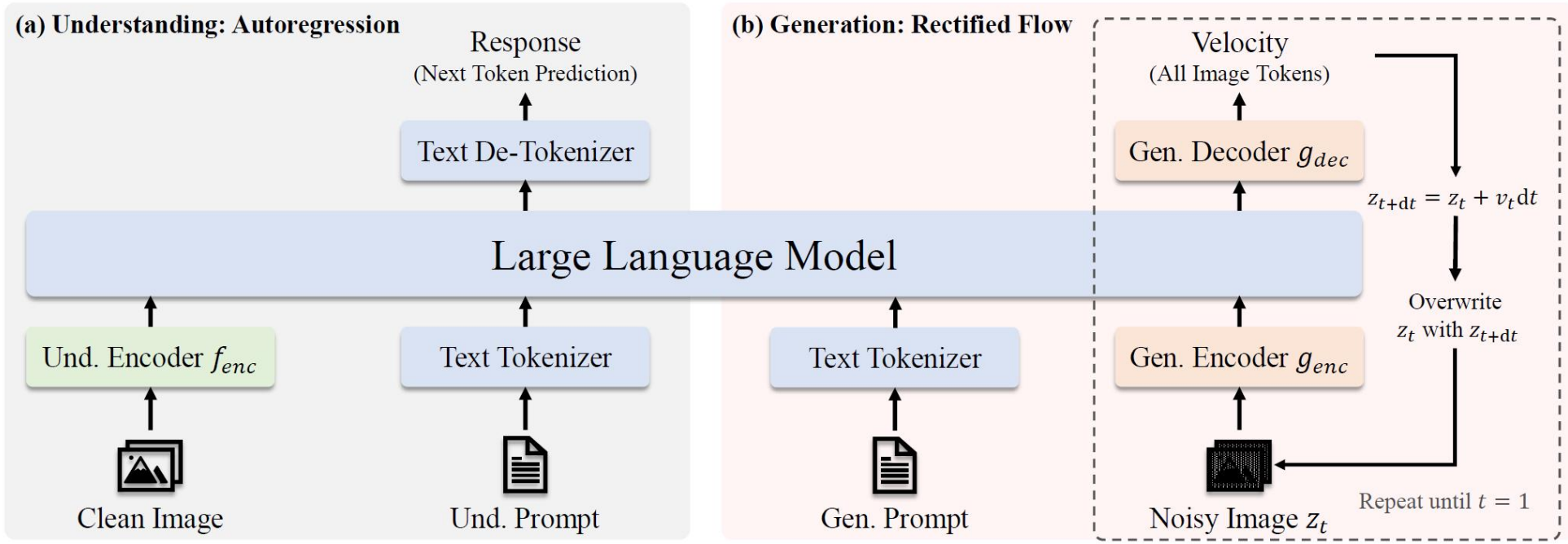
JanusFlow



- The sampling process:
  - Randomly initialize noisy image and obtain the token sequence.
  - Predict the velocity in one step and repeat the steps.

Method

JanusFlow



- Introduce the understanding features as regularizations of generation (REPA).

$$\mathcal{L}_{REPA}(\theta, \varphi) = -\mathbb{E}_{x \sim \mathcal{D}_{gen}} \left[\text{sim}\left(\texttt{stop\_grad}(f_{enc}(x^{res})), h_{\varphi}(q_{\theta}(z_t)) \right) \right]$$

Content

- Authors
- Background
- Method
- Experiments

Experiments

Implementations

- DeepSeek-LLM 1.3B as backbone.
- Understanding:
  - SigLIP-Large-Patch/16 as understanding encoder.
- Generation:
  - Employ Rectified Flow as the implementation of diffusion.
  - Employ SDXL-VAE.
  - Regularize the feature after the 6-th layer.
- Training stages shown in the table.
- Trained for 845B tokens.

	Stage 1	Stage 2	Stage 3
Learning Rate	1.0×10^{-4}	1×10^{-4}	2.0×10^{-5}
LR Scheduler	Constant	Constant	Constant
Weight Decay	0.0	0.0	0.0
Gradient Clip	1.0	1.0	1.0
Optimizer	AdamW ($\beta_1 = 0.9, \beta_2 = 0.95$)		
Warm-up Steps	2,000	0	1,000
Training Steps	10,000	380,000	26,000
Batch Size	512	512	256
Data Ratio	50 : 50 : 0	14 : 80 : 6	21 : 70 : 9

Experiments

Understanding

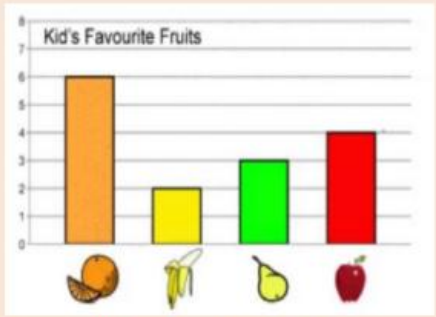
Type	Model	LLM Param	POPE	MME-P	MMB_{dev}	SEED	$VQAv2_{test}$	GQA	MMMU	MM-Vet	ChartQA	TextVQA
Und. Only	MobileVLM [12]	2.7B	84.9	1288.9	59.6	-	-	59.0	-	-	-	47.5
	MobileVLM-V2 [13]	2.7B	84.7	1440.5	63.2	-	-	61.1	-	-	-	57.5
	LLaVA-Phi [109]	2.7B	85.0	1335.1	59.8	-	71.4	-	-	28.9	-	48.6
	LLaVA [58]	7B	76.3	809.6	38.7	33.5	-	-	-	25.5	-	-
	LLaVA-v1.5 [56]	7B	85.9	1510.7	64.3	58.6	78.5	62.0	35.4	31.1	-	58.2
	InstructBLIP [15]	7B	-	-	36.0	53.4	-	49.2	-	26.2	-	50.1
	Qwen-VL-Chat [4]	7B	-	1487.5	60.6	58.2	78.2	57.5	-	-	66.3	61.5
	LLaVA-NeXT [57]	7B	-	1519.3	-	-	-	-	35.1	-	54.8	-
	Qwen2-VL [94]	7B	-	-	-	-	-	-	54.1	62.0	83.0	84.3
	IDEFICS-9B [44]	8B	-	-	48.2	-	50.9	38.4	-	-	-	25.9
	Emu3-Chat [95]	8B	85.2	-	58.5	68.2	75.1	60.3	31.6	-	68.6	64.7
	InstructBLIP [15]	13B	78.9	1212.8	-	-	-	49.5	-	25.6	-	50.7
	LLaVA-v1.5-Phi-1.5 [100]	1.3B	84.1	1128.0	-	-	75.3	56.5	30.7	-	-	-
	MobileVLM [12]	1.4B	84.5	1196.2	53.2	-	-	56.1	-	-	-	41.5
	MobileVLM-V2 [13]	1.4B	84.3	1302.8	57.7	-	-	59.3	-	-	-	52.1
Unified	Gemini-Nano-1 [89]	1.8B	-	-	-	-	62.7	-	26.3	-	53.6	62.5
	LWM [59]	7B	75.2	-	-	-	55.8	44.8	-	9.6	-	-
	VILA-U [99]	7B	85.8	1401.8	-	59.0	79.4	60.8	-	33.5	-	60.8
	Chameleon [88]	7B	-	-	-	-	-	-	22.4	8.3	-	-
	DreamLLM† [19]	7B	-	-	-	-	72.9	-	-	36.6	-	41.8
	LaVIT† [37]	7B	-	-	-	-	66.0	46.8	-	-	-	-
	Emu† [87]	13B	-	-	-	-	52.0	-	-	-	-	-
	NExT-GPT† [98]	13B	-	-	-	-	66.7	-	-	-	-	-
	Show-o [100]	1.3B	73.8	948.4	-	-	59.3	48.7	25.1	-	-	-
	Janus [97]	1.3B	87.0	1338.0	69.4	63.7	77.3	59.1	30.5	34.3	-	-
	JanusFlow (Ours)	1.3B	88.0	1333.1	74.9	70.5	79.8	60.3	29.3	30.9	64.6	55.5

Experiments

Understanding

User: What are the kinds of fruits in this picture?

JaunsFlow (Ours): The fruits in the picture are banana, strawberry, mango, persimmon, blueberry, and lime.



User: Describe this image.

JaunsFlow (Ours): Kid's Favourite Fruits is a bar graph. The x-axis shows the fruits. The y-axis shows the number of kids who like the fruit. The orange is the most popular fruit. The banana is the least popular fruit.

User: How many penguins in the picture?

JaunsFlow (Ours): There are 11 penguins in the picture.

Experiments

Generation

GenEval

Type	Method	Params	Single Obj.	Two Obj.	Count.	Colors	Pos.	Color Attri.	Overall↑
Gen. Only	LlamaGen [86]	0.8B	0.71	0.34	0.21	0.58	0.07	0.04	0.32
	LDM [77]	1.4B	0.92	0.29	0.23	0.70	0.02	0.05	0.37
	SDv1.5 [77]	0.9B	0.97	0.38	0.35	0.76	0.04	0.06	0.43
	PixArt-α [9]	0.6B	0.98	0.50	0.44	0.80	0.08	0.07	0.48
	SDv2.1 [77]	0.9B	0.98	0.51	0.44	0.85	0.07	0.17	0.50
	DALL-E 2 [76]	6.5B	0.94	0.66	0.49	0.77	0.10	0.19	0.52
	Emu3-Gen [95]	8B	0.98	0.71	0.34	0.81	0.17	0.21	0.54
	SDXL [73]	2.6B	0.98	0.74	0.39	0.85	0.15	0.23	0.55
	IF-XL [17]	4.3B	0.97	0.74	0.66	0.81	0.13	0.35	0.61
	DALL-E 3 [6]	-	0.96	0.87	0.47	0.83	0.43	0.45	0.67
Unified	Chameleon [88]	34B	-	-	-	-	-	-	0.39
	LWM [59]	7B	0.93	0.41	0.46	0.79	0.09	0.15	0.47
	SEED-X† [27]	17B	0.97	0.58	0.26	0.80	0.19	0.14	0.49
	Show-o [100]	1.3B	0.95	0.52	0.49	0.82	0.11	0.28	0.53
	Janus [97]	1.3B	0.97	0.68	0.30	0.84	0.46	0.42	0.61
	Transfusion [108]	7.3B	-	-	-	-	-	-	0.63
	JanusFlow (Ours)	1.3B	0.97	0.59	0.45	0.83	0.53	0.42	0.63

DPG-Bench

Method	Global	Entity	Attribute	Relation	Other	Overall↑
SDv1.5 [75]	74.63	74.23	75.39	73.49	67.81	63.18
PixArt-α [9]	74.97	79.32	78.60	82.57	76.96	71.11
Lumina-Next [105]	82.82	88.65	86.44	80.53	81.82	74.63
SDXL [71]	83.27	82.43	80.91	86.76	80.41	74.65
Playground v2.5 [48]	83.06	82.59	81.20	84.08	83.50	75.47
Hunyuan-DiT [54]	84.59	80.59	88.01	74.36	86.41	78.87
PixArt-Σ [10]	86.89	82.89	88.94	86.59	87.68	80.54
Emu3-Gen [91]	85.21	86.68	86.84	90.22	83.15	80.60
JanusFlow (Ours)	87.03	87.31	87.39	89.79	88.10	80.09

MJHQ-FID

Method	Params	FID↓
LWM [58]	7B	17.77
VILA-U 256 [95]	7B	12.81
VILA-U 384 [95]	7B	7.69
Show-o [96]	1.3B	15.18
Janus [93]	1.3B	10.10
JanusFlow (Ours)	1.3B	9.51

Experiments

Generation

A corgi's head depicted as an explosion of a nebula, with vibrant cosmic colors like deep purples, blues, and pinks swirling around. The corgi's fur blends seamlessly into the nebula, with stars and galaxies forming the texture of its fur. Bright bursts of light emanate from its eyes, and faint constellations can be seen in the background, giving the image a surreal, otherworldly feel.

Beautiful surreal symbolism the mesmerizing vision of a Cleopatra Queen of Egypt, mesmerizing brown eyes, black hair and ethereal features, radiating celestial aura, super high definition, true lifelike color, perfect exposure, razor sharp focus, golden ratio, soft reflections, bokeh effect, fine art photography, cinematic compositing, authentic, professional.

A lone figure in dark robes ascends worn stone steps toward a glowing light in an ancient temple entrance. Ornate arches, lush greenery, and intricate carvings adorn the scene, evoking a mystical, high-fantasy atmosphere reminiscent of works by artists like Randy Vargas, with cinematic lighting and epic storytelling.

Experiments

Generation

Massive cathedral church, battle between Heaven and hell, church on fire, 8k hyper real ultra sharp renaissance by Francisco Goya.

A handsome 24-year-old boy in the middle with sky color background wearing eye glasses, it's super detailed with anime style.

Happy dreamy owl monster sitting on a tree branch, colorful glittering particles, forest background, detailed feathers.

A man wearing Fedora hat with mafia style, realistic photography, intricate details, magical lighting, vibrant background, complex textures, rich colors, realistic style, front-facing view.

A vivid depiction of the Northern Lights dancing above the snow-covered mountains in Iceland, casting a mesmerizing glow across the sky.

A dark, high-contrast render of a psychedelic Tree of Life glowing brilliantly, illuminating swirling dust particles in a mystical, cavernous setting.

Experiments

Ablation studies

A. w/o REPA.

B and C. w/o decoupling and decoupling w/o pre-trained understanding encoder.

D and E. only understanding and only generation w/ the same framework.

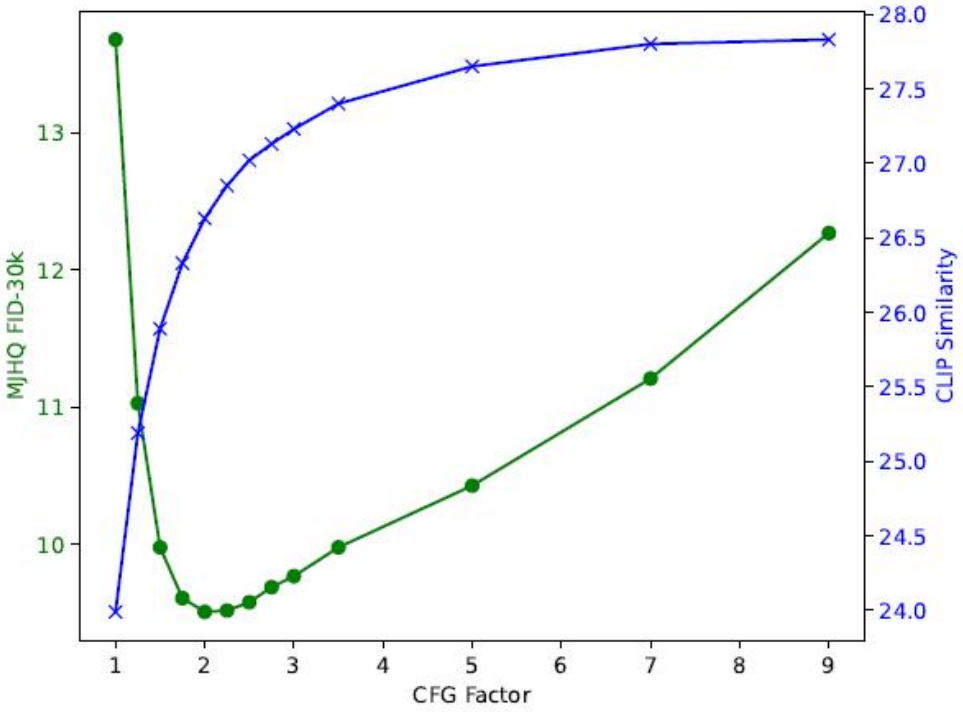
F. Final setting.

Exp. ID	Model Setting				Train. Iter.	Evaluation Benchmarks				
	REPA	Und. Modules	Gen. Modules	Type		POPE↑	VQAv2$_{val}$ ↑	GQA↑	FID↓	CLIP ↑
A	×	SigLIP	VAE†+ConvNeXt	Unified	50,000	82.40	69.62	54.43	19.84	24.94
B	✓	Shared VAE†+ConvNeXt		Unified	50,000	78.13	53.94	44.04	18.05	26.38
C	✓	VAE+ConvNeXt	VAE†+ConvNeXt	Unified	50,000	75.30	55.41	44.44	17.53	26.32
D	✓	SigLIP	-	Und. Only	13,000	85.03	69.10	54.23	-	-
E	✓	-	VAE†+ConvNeXt	Gen. Only	37,000	-	-	-	16.69	26.89
F	✓	SigLIP	VAE†+ConvNeXt	Unified	50,000	84.73	69.20	54.83	17.61	26.40

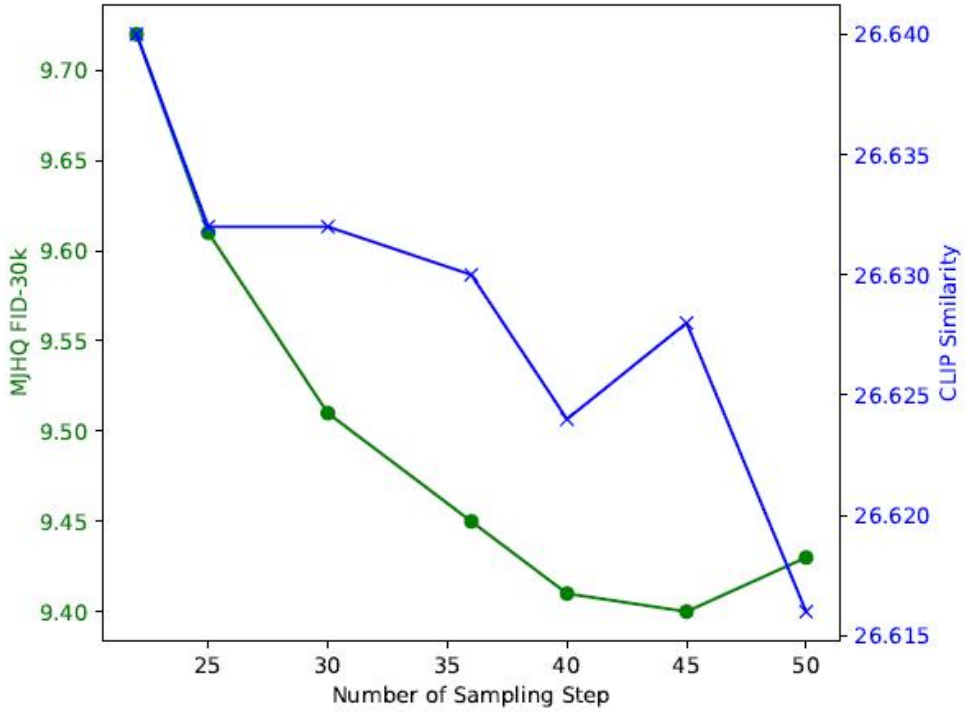
Experiments

Ablation studies

Effect of CFG and number of sampling step.



(a) Results of varying CFG Factors

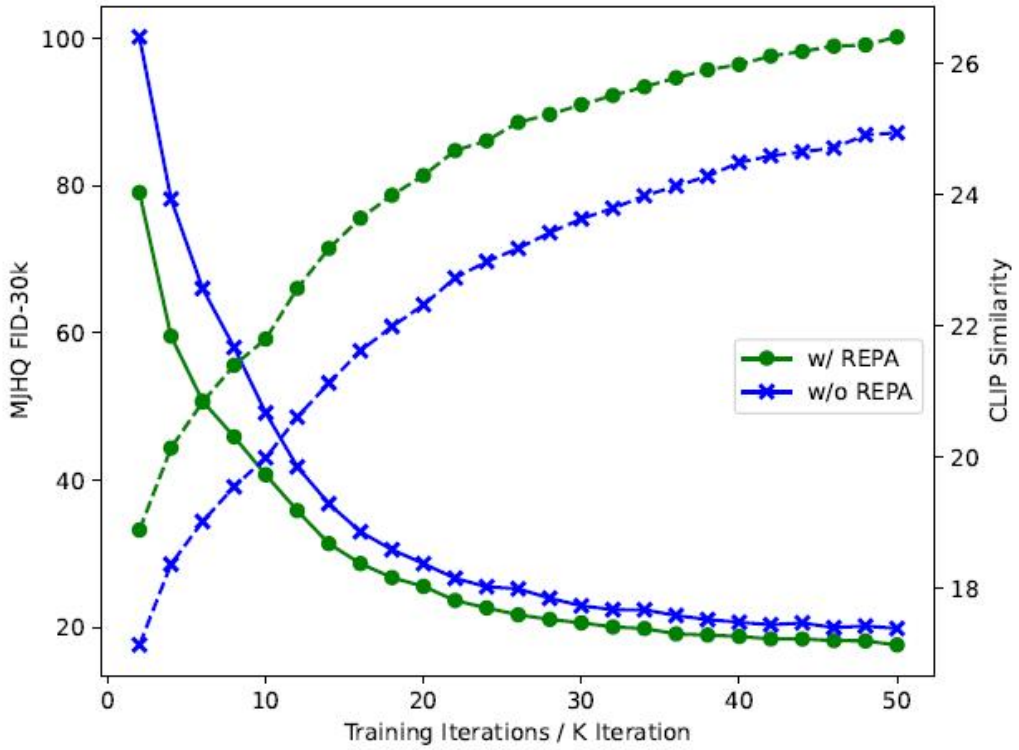


(b) Results of Varying Numbers of Sampling Steps

Experiments

Ablation studies

Effect of REPA.



Thanks for listening.

马逸扬
2025.03.30