The Road to Convergence: Evolution of Unified Multimodal Models

Past, Present, and Future

Jindong Wang
Assistant Professor,
William & Mary.

Hao Chen
Ph.D. from Carnegie
Mellon University.

Sharon Li
Associate Professor,
UW–Madison

Zhaolong Su
Cornell University
P.h.D. Student.

Jiakui Hu
Peking University
P.h.D. Student.

*Speaker of the day

Tutorial Roadmap

Part 1	Past — From Divergence to Unification	**30 min**
Part 2	Present I — Unified model: Goals, Modeling, and Future	**55 min**
	Coffee Break	15 min
Part 3	Present II — Frontier Highlights & Trend	**45 min**
Part 4	Beyond Unified	**25 min**
Part 5	Future	**25 min**

PART 1

Past

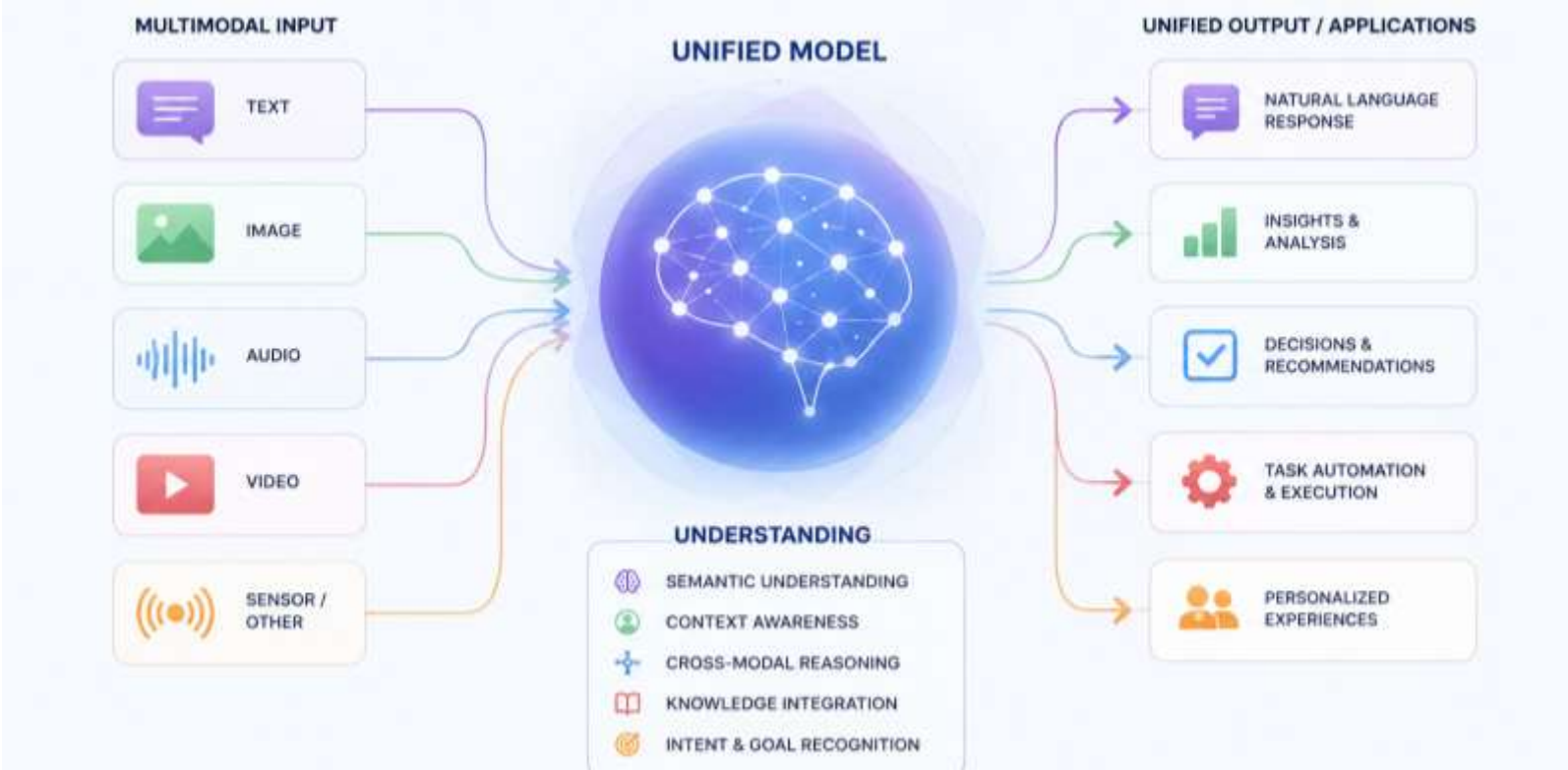
From Divergence to Unification

- Chameleon
- Nano-Banana
- Janus-Pro
- GPT4o
- Clip

Unified or Not?

▪ Chameleon	√ Meta 2024.05
▪ Nano-Banana	√ Google
▪ Janus-Pro	√ DeepSeek-AI 2025.01
▪ GPT4o	√ OpenAI
▪ Clip	× OpenAI 2021.02

Definition of Unified Models



Two Independent Tracks

What Is Multimodal Understanding?

Multimodal understanding is the task of producing an output that captures the semantic content of a *multimodal* input

Two Independent Tracks

$$P(y \mid x_v, x_t) = \prod_{i=1}^{T} P(y_i \mid y_{<i}, x_v, x_t; \theta)$$

What Is Multimodal Understanding?

Multimodal understanding is the task of producing an output that captures the semantic content of a *multimodal* input

Two Independent Tracks

Understanding tasks include:

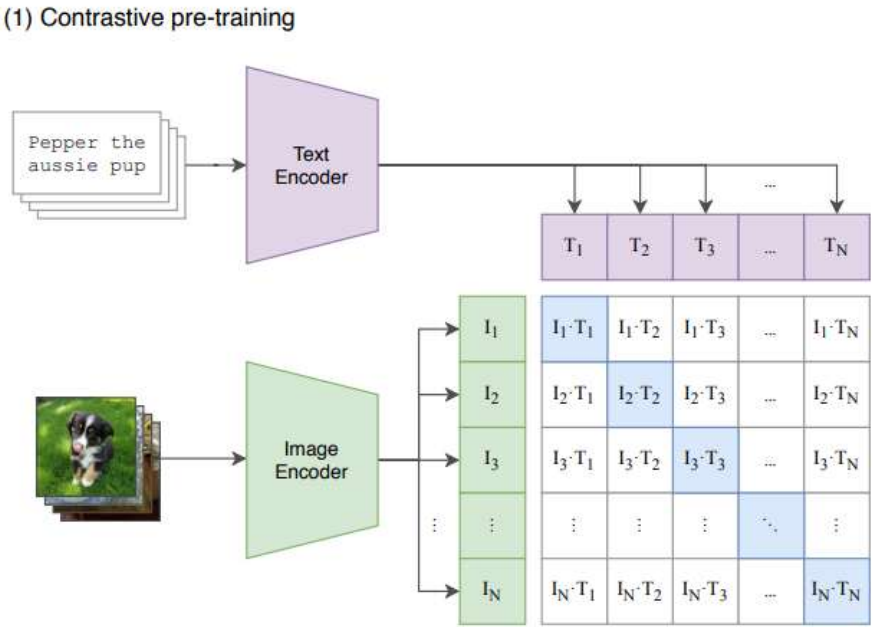
- **Captioning:** summarize visual semantics
- **VQA:** answer grounded questions
- **Grounding / segmentation:** locate referred entities
- **Relation reasoning:** infer spatial, causal, and semantic relations
- **Verification:** check whether outputs satisfy instructions

What Is Multimodal Understanding?

Two Independent Tracks

UNDERSTANDING

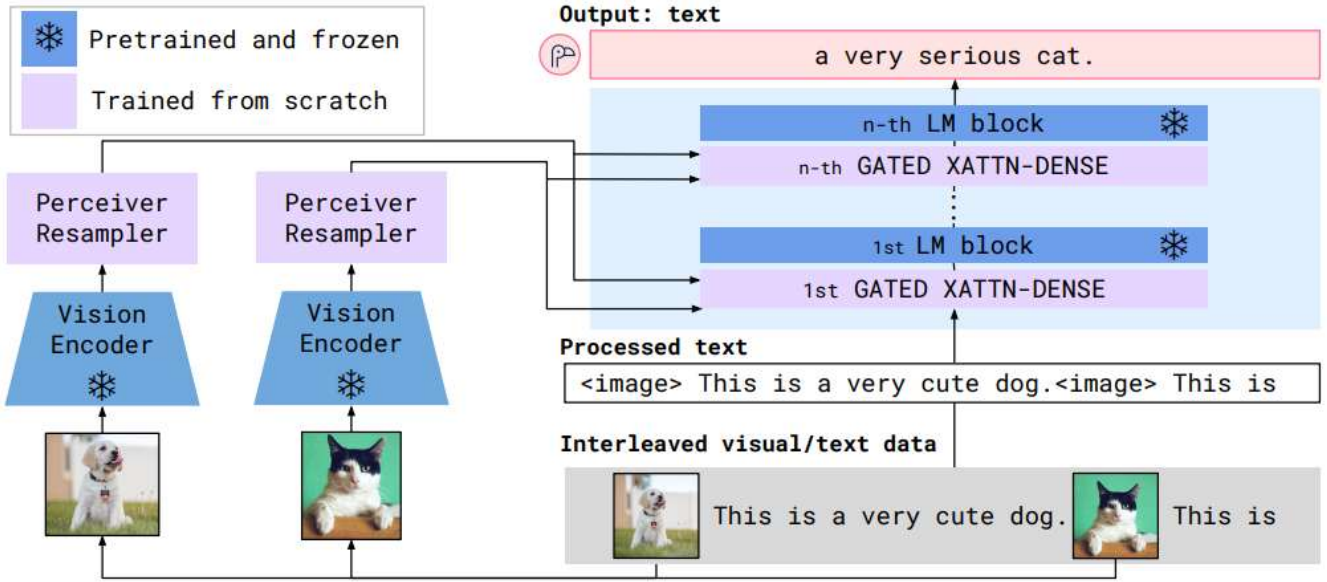
Clip



Two Independent Tracks

UNDERSTANDING

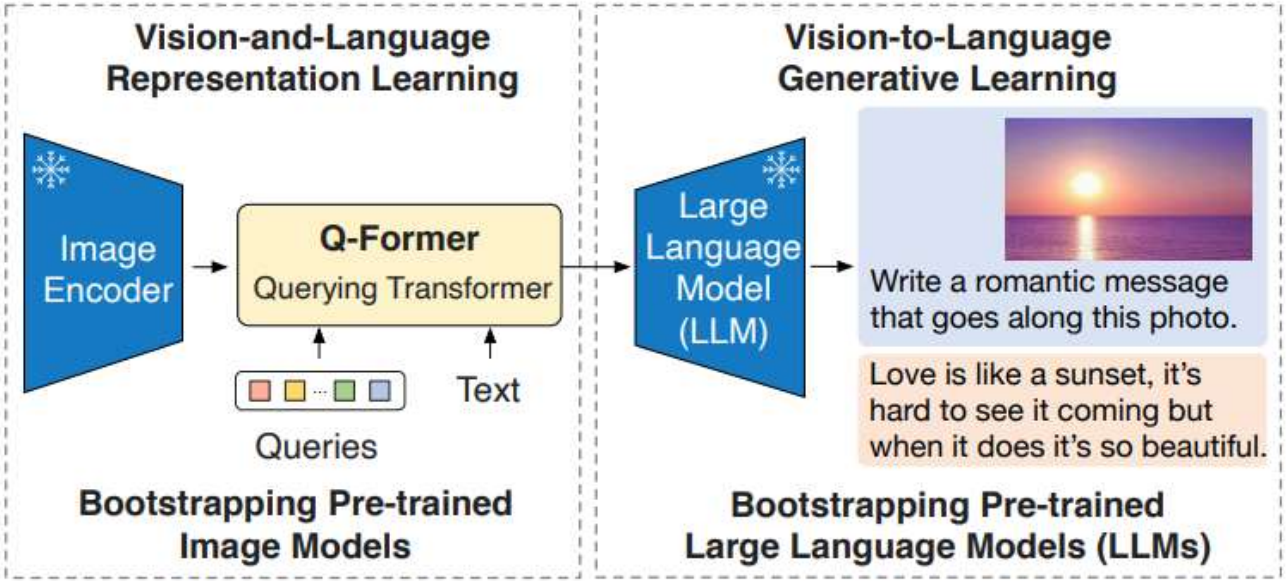
Flamingo



Two Independent Tracks

U N D E R S T A N D I N G

BLIP2



Two Independent Tracks

GENERATION

$$y_{\mathcal{M}_{out}} \sim p_\theta \left(y_{\mathcal{M}_{out}} \mid x_{\mathcal{M}_{in}}, c \right)$$

What Is Multimodal Generation?

Multimodal generation is the task of producing a *visual* (or audio/video) output that realizes the semantic content of a *textual* (or multimodal) input, typically a natural language prompt.

Two Independent Tracks

GENERATION

DDPM and DALL·E

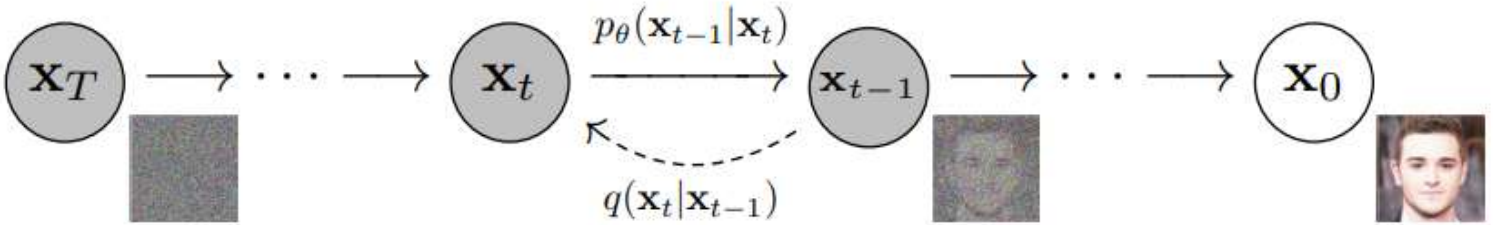


Figure 2: The directed graphical model considered in this work.

Two Independent Tracks

UNDERSTANDING

CLIP → BLIP-2 → Flamingo → LLaVA

Contrastive / instruction-tuned encoders

Discriminative objective

Tokenizer: text-only LLM vocab

GENERATION

GAN → DDPM → Imagen / SDXL / DALL·E 3

Iterative denoising/score matching

Generative objective

Tokenizer: VAE latents (continuous)

Inspiration came from NLP

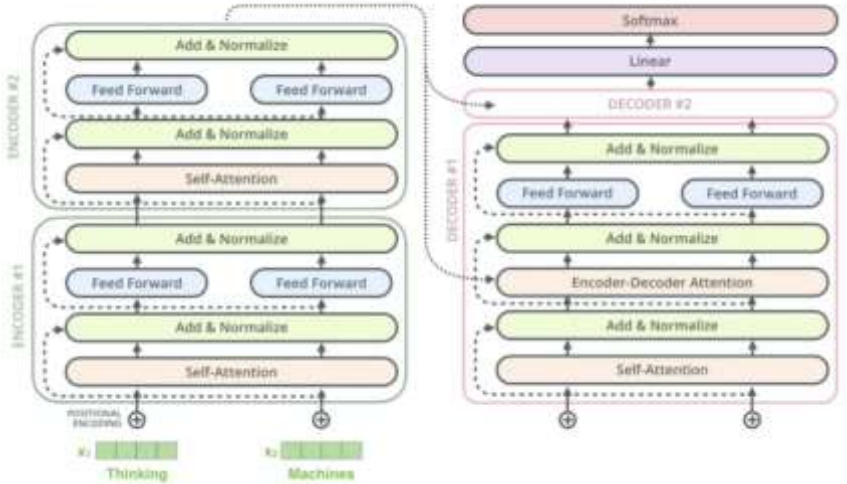
Before T5:

Translation / Summarization / QA / Classification
→ different formats, heads, and objectives

Inspiration came from NLP

Before T5:

Translation / Summarization / QA / Classification
→ different formats, heads, and objectives



T5:
Everything becomes:
Text input → Text output

1. Pipeline Systems Are Not Enough

Some kind of VLM

User: Generate something similar

1. Pipeline Systems Are Not Enough

Some kind of VLM

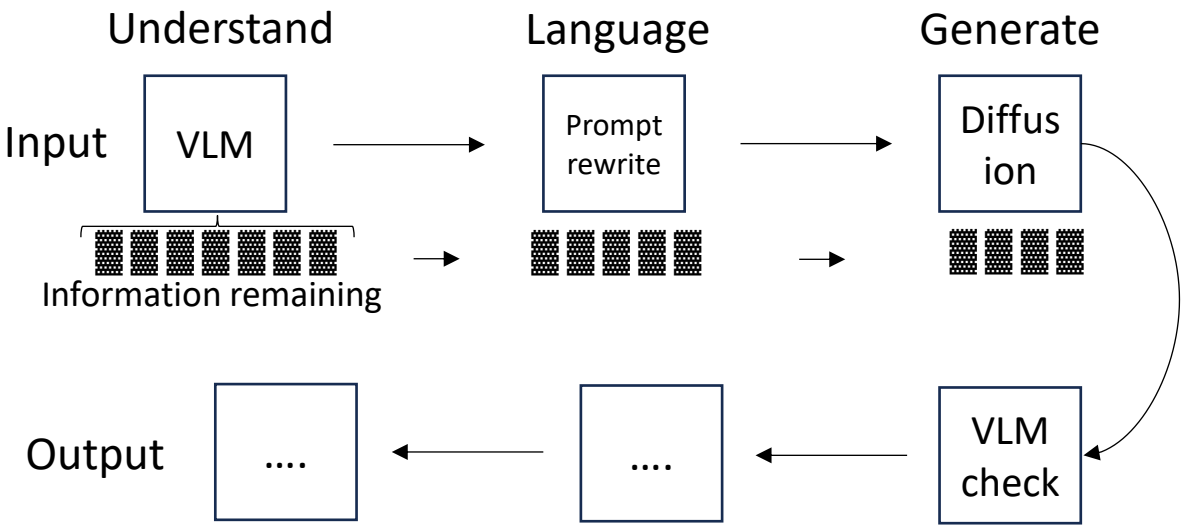
a young elf girl with long white hair, pointed ears, green eyes, wearing a black mage outfit with white cape...

Some Generative Model, e.g., diffusion

Output

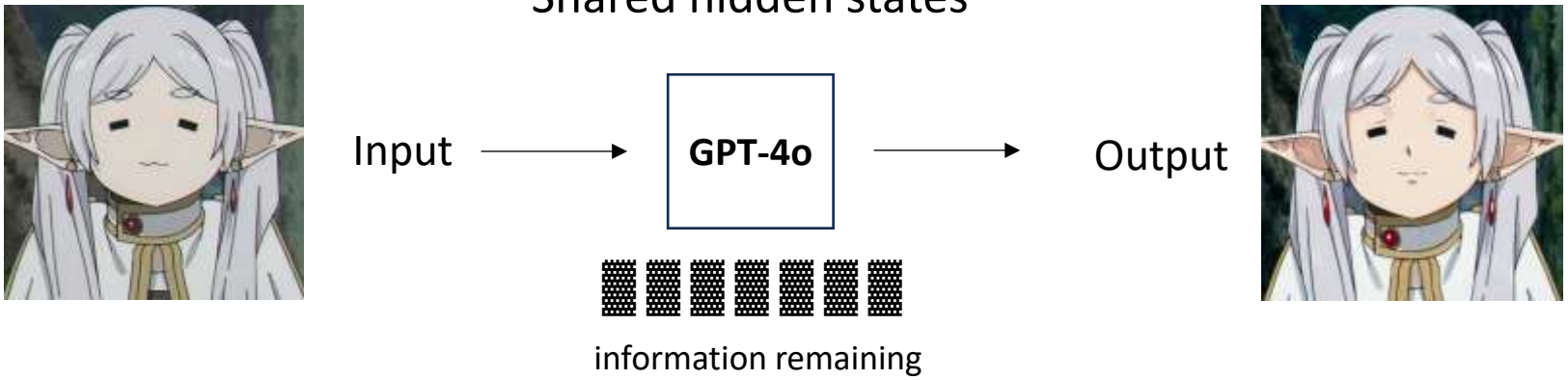
User: Generate something similar

1. Pipeline Systems Are Not Enough



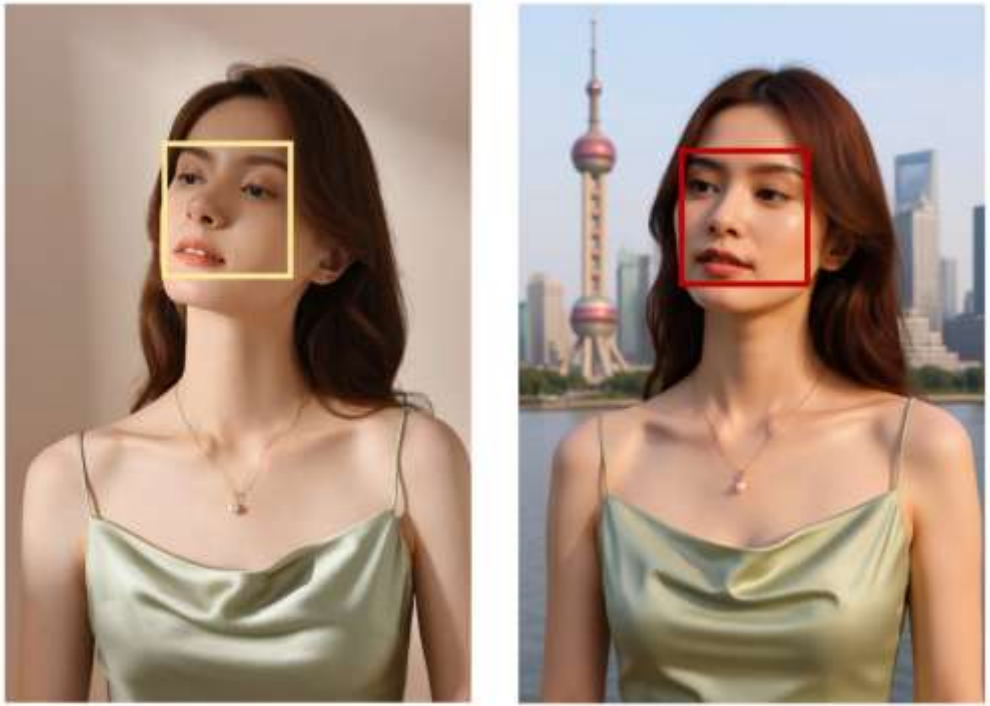
Each step come with information loss ✕

1. Pipeline Systems Are Not Enough



Unified Model √

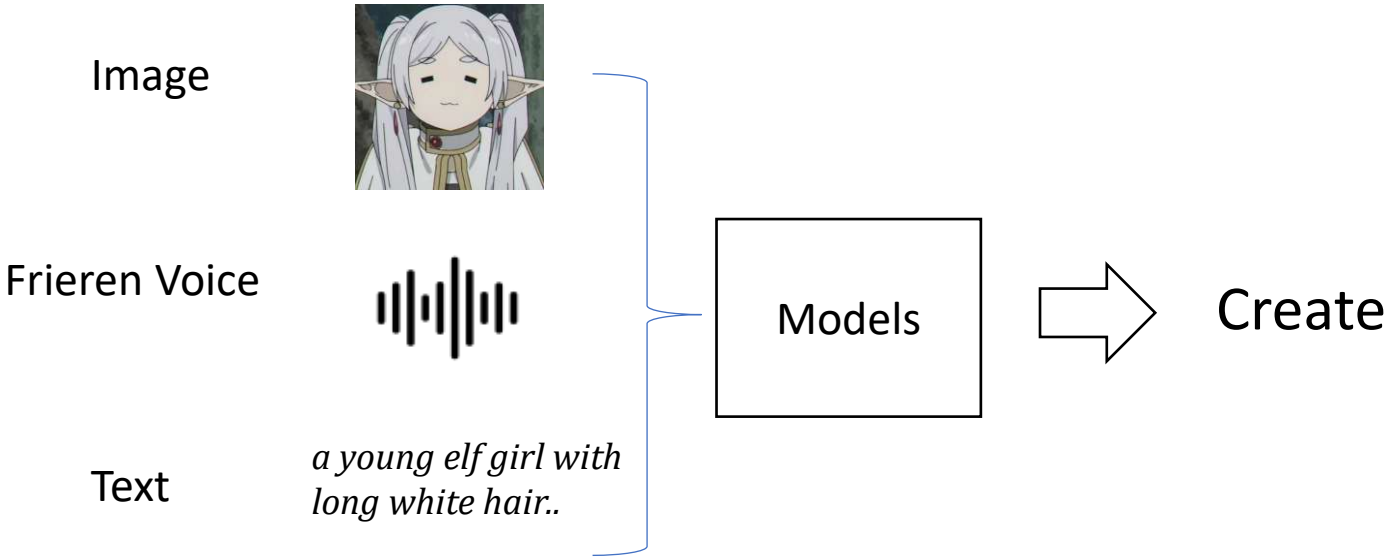
2. Unlock Tasks



2. Unlock Tasks

Example: Multi-round editing

3. Potential human-like multimodal intelligence



Visions of Unified Model community

1. Eliminate the representational gap among perception, language, generation, and action.

2. Unlock more applications & tasks

3. From function calls to cognitive level unification

"What I cannot create, I do not understand."

Example: Nano Banana

Example: Nano Banana

PART 2

Present I

Unified model: Goals, Modeling, and Future

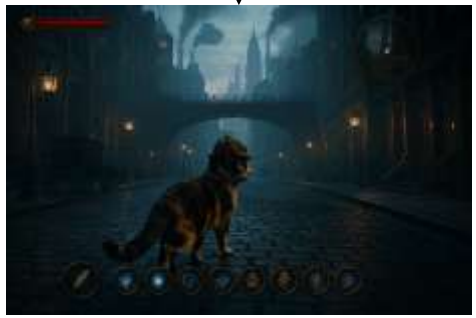
Jiakui Hu
jkhu29@stu.pku.edu.cn
Peking University

[1] Unified Multimodal Understanding and Generation Models: Advances, Challenges, and Opportunities. arXiv 2025.

Samples of GPT4o

Give this cat a detective hat and a monocle

Update to a landscape image 16:9 ratio, add more spells in the UI



Turn this into a triple A video games

[2] https://openai.com/zh-Hans-CN/index/introducing-4o-image-generation/

Samples of GPT4o

We need evidence there is a currently present invisible elephant.

[2] https://openai.com/zh-Hans-CN/index/introducing-4o-image-generation/

The goals for unified model

Interleaved text-vision modeling.

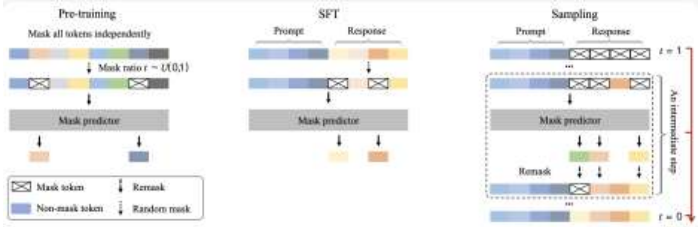
Unifying understanding, reasoning, and generation tasks.

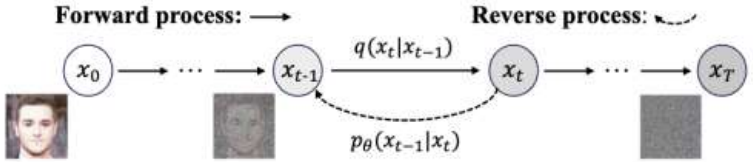
Jointly modeling text and vision

Text

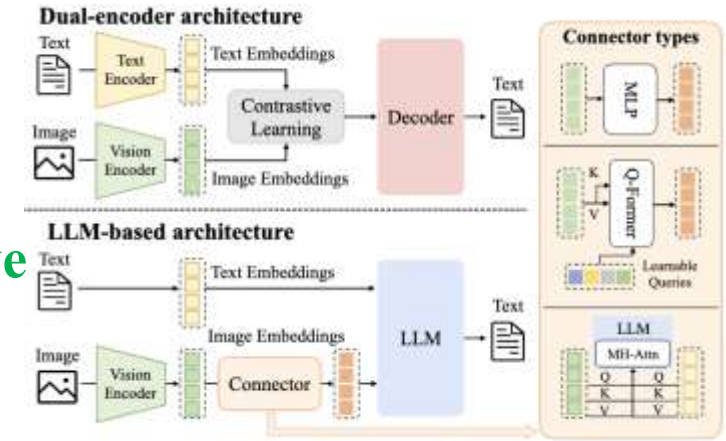
Vision

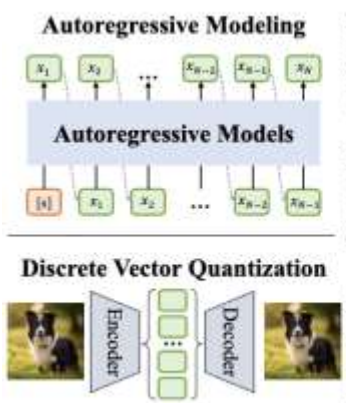
Diffusion

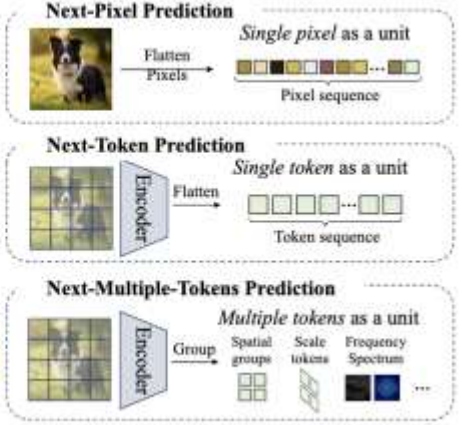




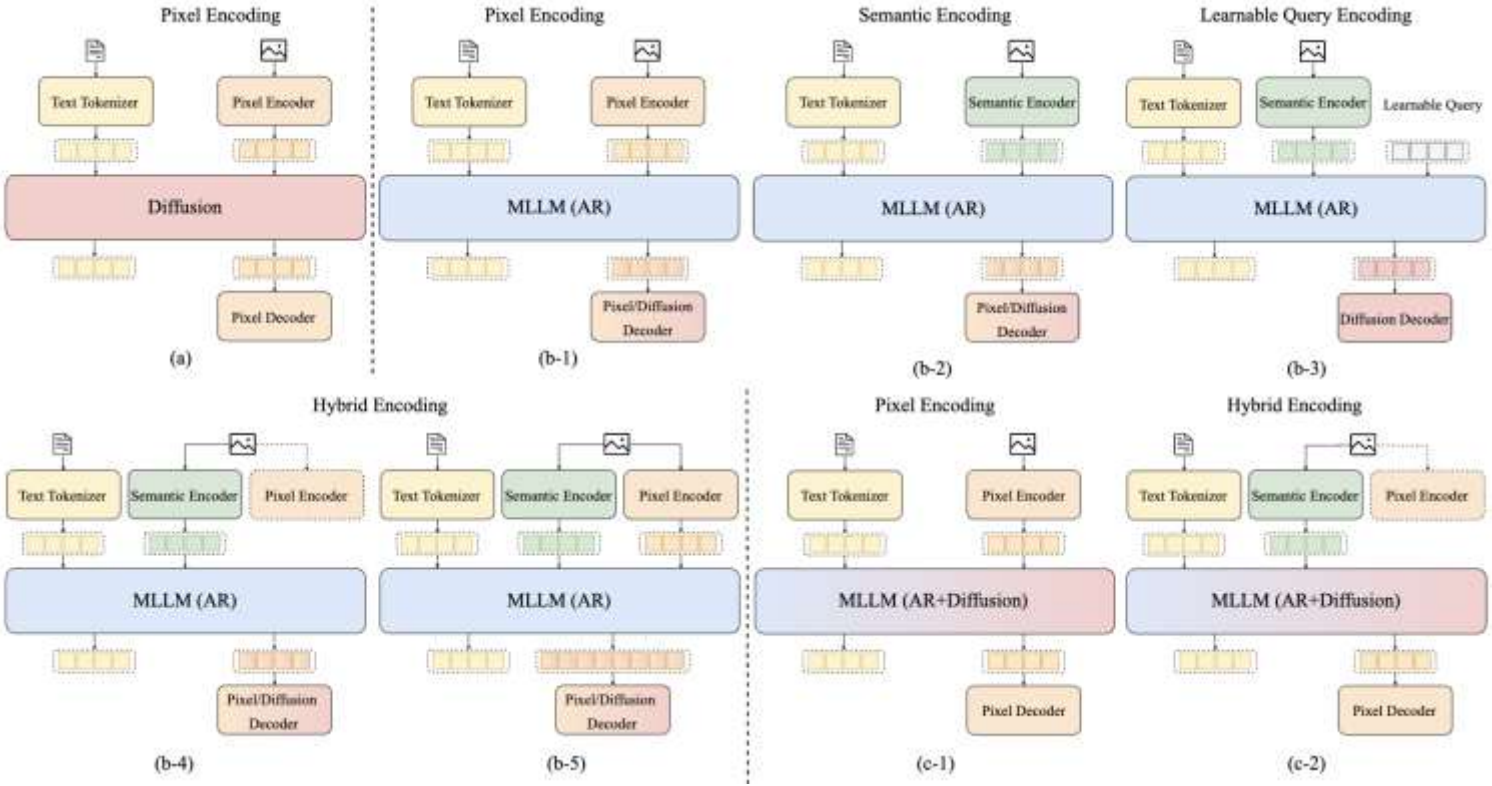
Auto Regressive



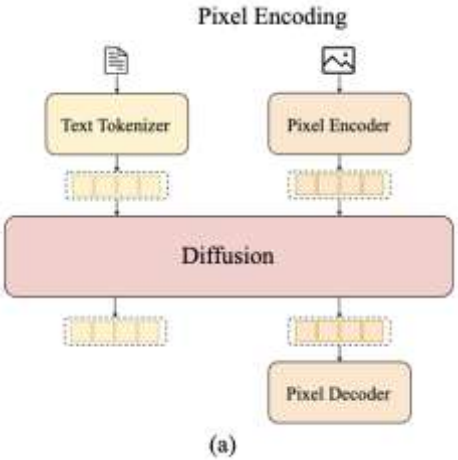


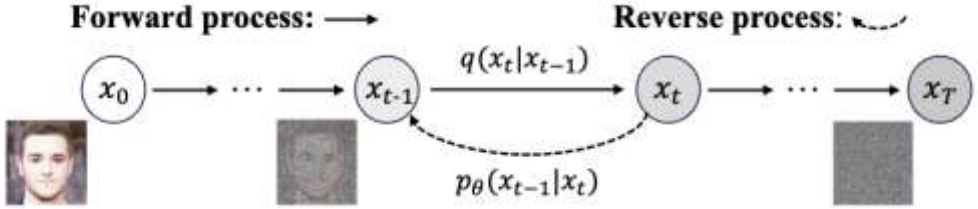


Permutations and combinations

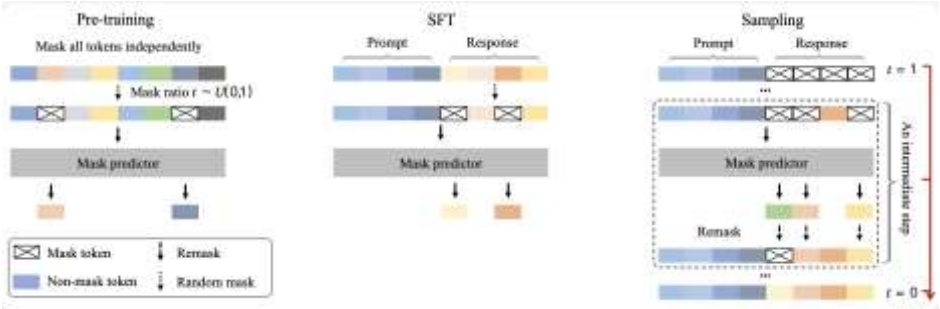


Option A: Diffusion





Diffusion based image modeling: denoising from Gaussian noise



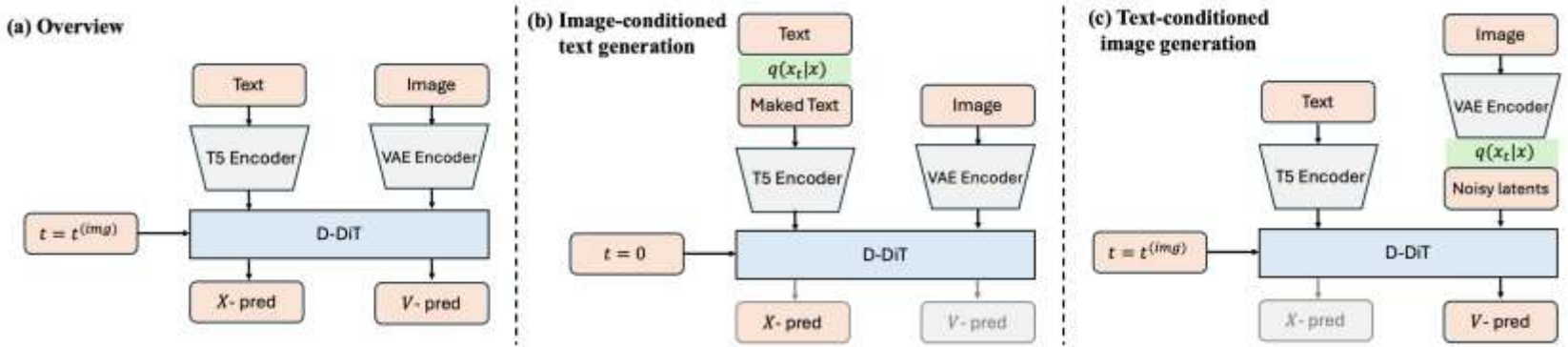
Diffusion based Image modeling: demasking from All-[mask] tokens.

[3] Denoising Diffusion Probabilistic Models. NeurIPS 2020.
[4] Large Language Diffusion Models. NeurIPS 2025.

Diffusion based text modeling

[5] Dual Diffusion for Unified Image Generation and Understanding. CVPR 2025.

Examples: continuous visual tokens

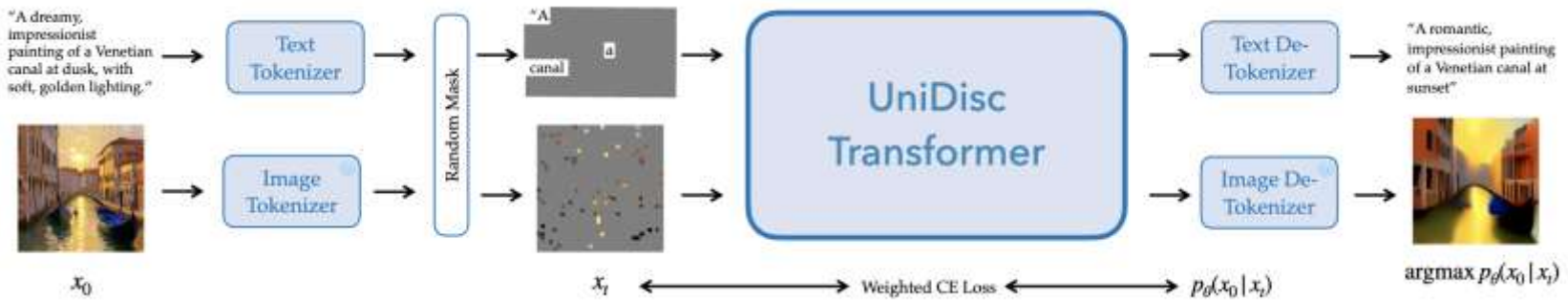


Based on an image generation model,

- Text generator uses masked modeling;
- Image generator uses denoising;
- If text generator need vision input, the vision input is noise-free.

[5] Dual Diffusion for Unified Image Generation and Understanding. CVPR 2025.

Examples: discrete visual tokens



[6] Unified multimodal discrete diffusion, arXiv 2025.

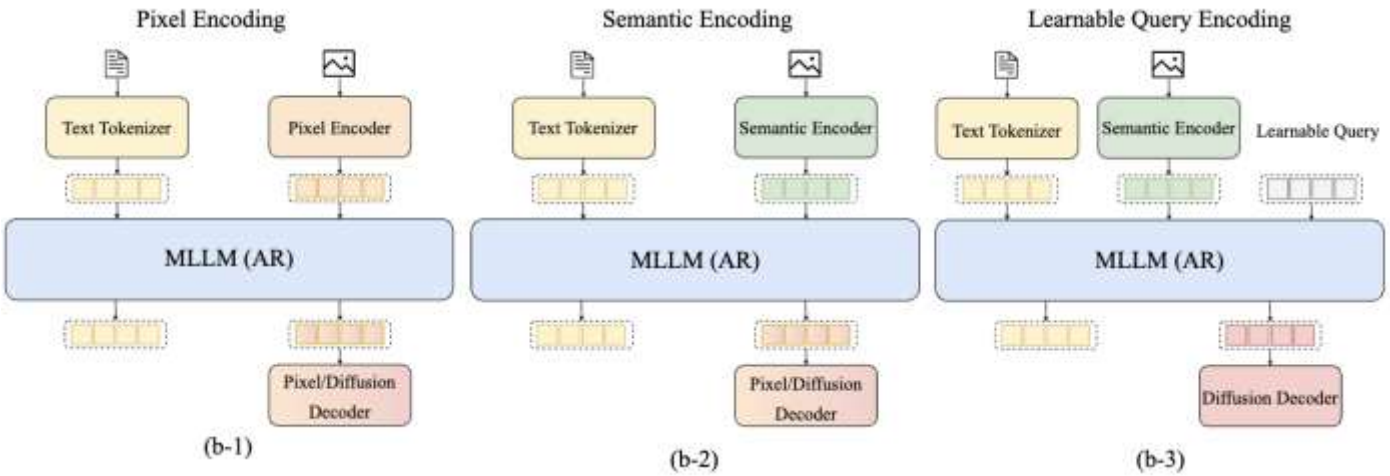
Limitations of Option A: Diffusion

1. In image generation, it is still worse than continuous-space generative models;

2. Hard to perform flexible interleaved multimodal generation;

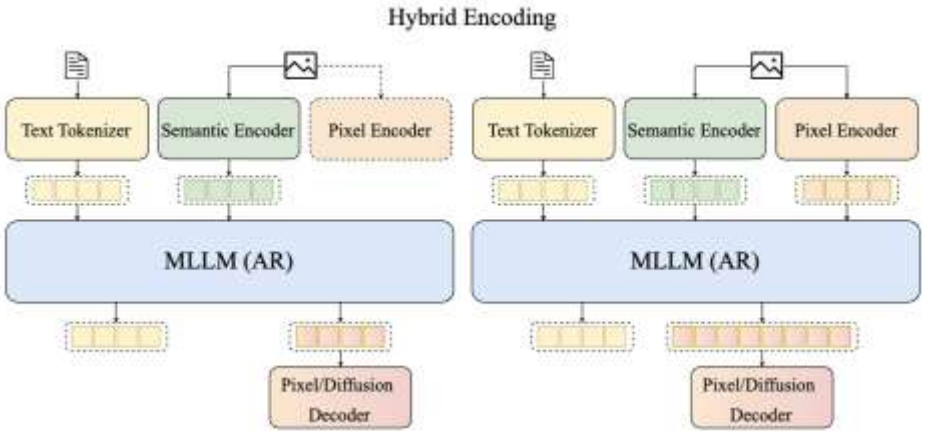
3. Training and inference codebases are still underdeveloped.

Option B: AutoRegressive Model



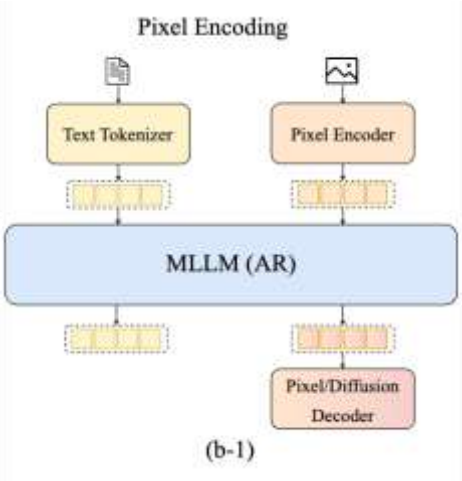
- **B-1**. Image encoder focus on reconstruction. Good at texture, bad at semantic;
- **B-2**. Image encoder finetuned on semantic encoder. Good at semantic, bad at texture;
- **B-3**. Image tokens learned by learnable query. Adaptive, trade-off.

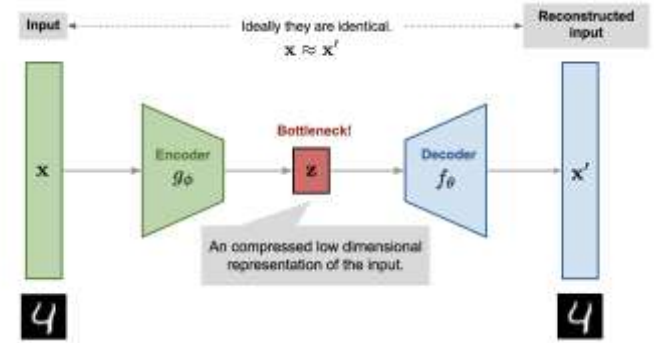
Option B: AutoRegressive Model



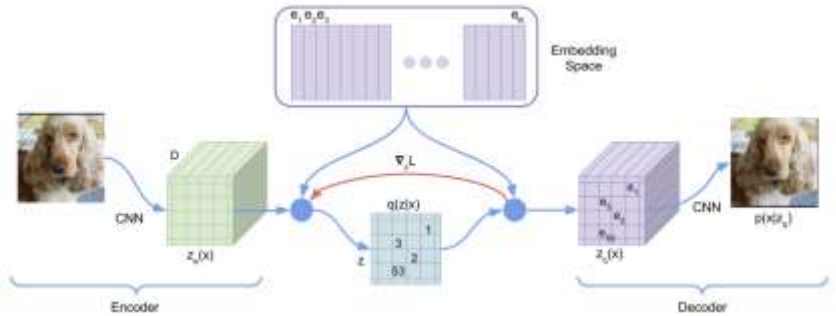
- **B-4**. Fuse reconstruction encoder and semantic encoder.

Option B-1: Reconstruction encoder





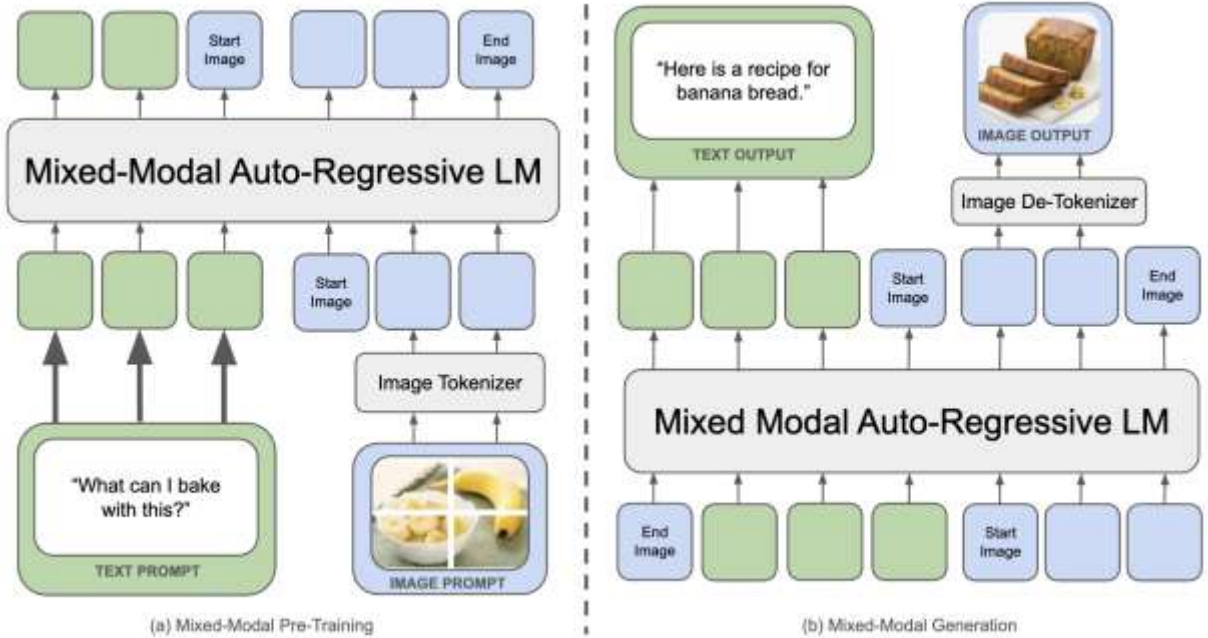
The encoder is to represent the image using a compressed, low-dimensional representation.



Convert the visual representation into indices.

[7] https://lilianweng.github.io/posts/2018-08-12-vae/
[8] Neural Discrete Representation Learning. NeurIPS 2017.

Examples



[9] Chameleon: Mixed-Modal Early-Fusion Foundation Models. arXiv 2024.

Interleaved demo

Prompt: I'm bored. Could you show me some cool, quirky-looking birds? Also tell me about them in 2-3 sentences.

Response: Absolutely! Here are three birds with some unique features that you might find interesting:

Keel-Billed Toucan The Keel-Billed Toucan is a brightly colored bird native to South American rainforests and woodlands. Its most distinctive feature is its colorful beak, which can grow up to seven inches long and has a distinctive, curved shape. The beak is used for plucking fruit from trees and for social displays. Despite its large size, the Keel-Billed Toucan is a skilled flyer and can often be seen flitting through the treetops. Here is an image of this vibrant bird: <img>

Puffin Puffins are small, aquatic birds that are found in the northern hemisphere. They have distinctive black and white feathers, as well as a brightly colored beak that is used for attracting mates and defending their territory. Puffins are known for their distinctive "puffin" call, which sounds like a loud, guttural chuckle. They feed on fish, which they catch by diving into the water from the air. Puffins are colonial birds, and they breed in large colonies on rocky cliffs and islands. Here is an image of this adorable bird: <img>

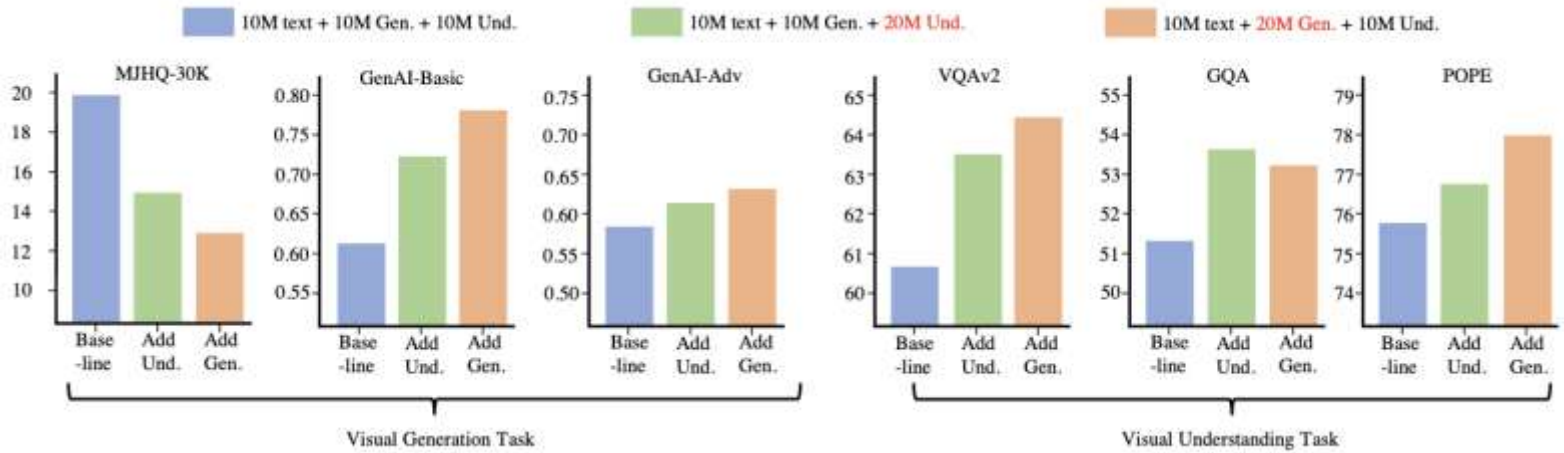
Golden Pheasant The Golden Pheasant is a brightly colored bird native to the forests of China. It is a member of the pheasant family and is known for its striking plumage, which features a golden-yellow body, red face, and green tail feathers. The Golden Pheasant is a ground-dwelling bird, and it feeds on a diet of insects, seeds, and fruits. It is a popular game bird, and is often hunted for its meat and feathers. In captivity, the Golden Pheasant is sometimes kept as an ornamental bird, due to its striking appearance. Here is an image of this beautiful bird: <img>

I hope these images and descriptions have helped to brighten your day! If you have any other questions or requests, please don't hesitate to ask.

[9] Chameleon: Mixed-Modal Early-Fusion Foundation Models. arXiv 2024.

Findings: simultaneous improvements

Method	Training Data			Visual Generation			Visual Understanding					
	Text-only	Visual Gen.	Visual Und.	gFID↓	Basic	Advanced	Overall	VQAv2	GQA	TextVQA	POPE	MME
Baseline	10M	10M	10M	19.9	0.63	0.58	0.60	60.7	51.3	39.2	75.9	909.6
Add Gen.	10M	**20M**	10M	12.8	0.78	0.63	0.69	64.5	53.2	39.8	78.0	1066.8
Add Und.	10M	10M	**20M**	14.9	0.73	0.62	0.66	63.5	53.7	40.5	76.8	1035.1



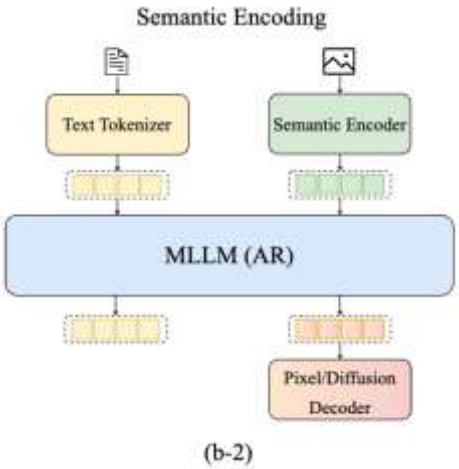
[10] Liquid: Language Models are Scalable and Unified Multi-Modal Generators. IJCV 2026.

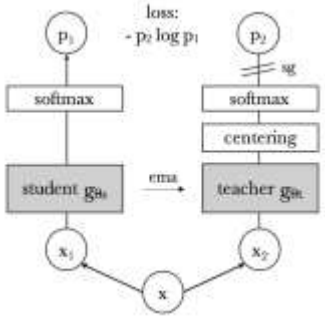
Limitations of Option B-1

It lacks

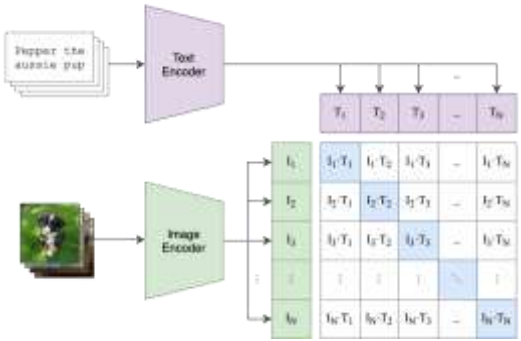
- high-level semantic abstraction
- cross-modal alignment between text and image representations

Option B-2: Semantic Encoding





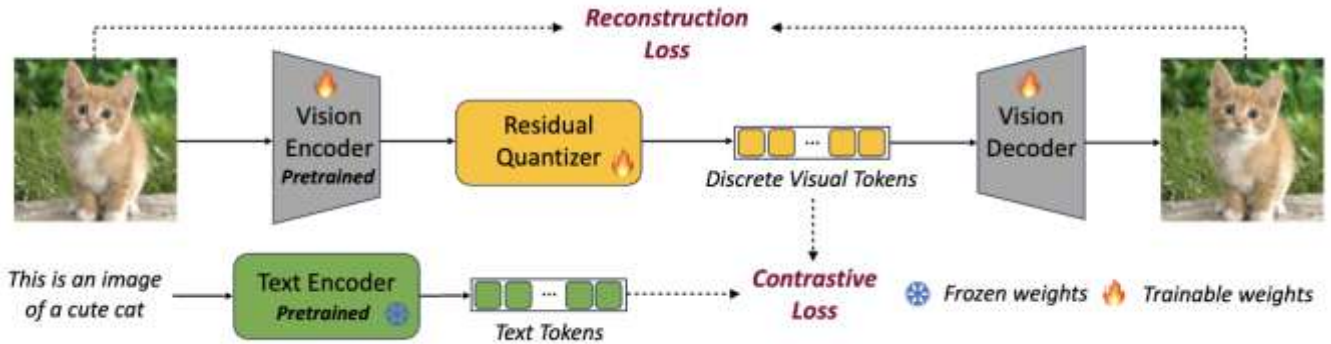
high-level semantic



alignment
between text and image

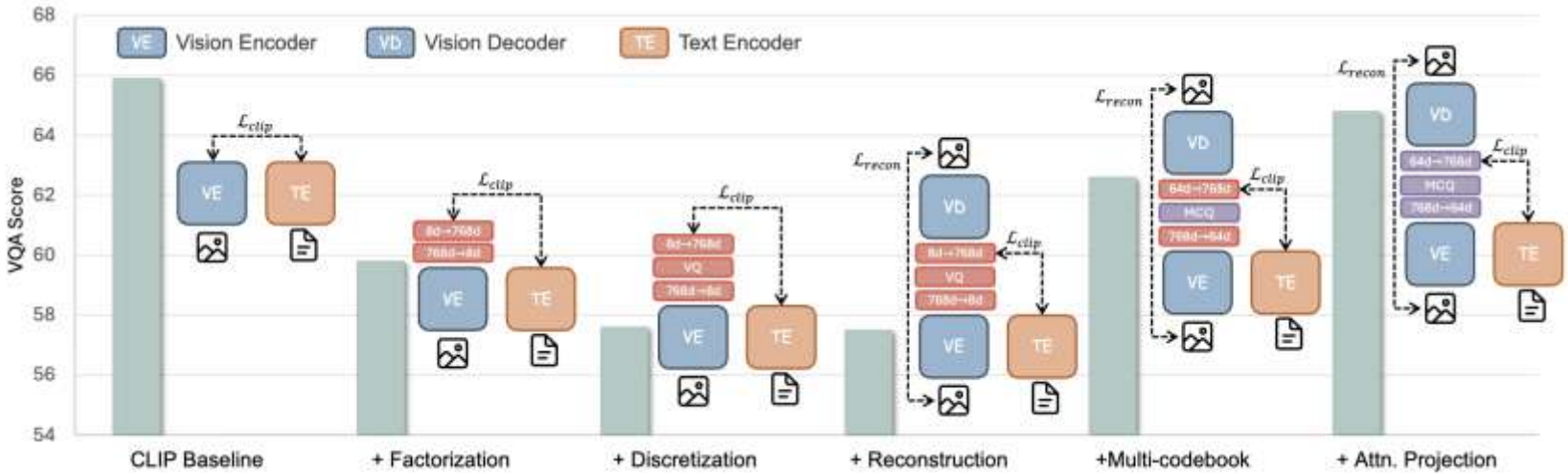
[11] Emerging Properties in Self-Supervised Vision Transformers. ICCV 2021.
[12] Learning Transferable Visual Models From Natural Language Supervision. ICML 2021.

Examples: convert



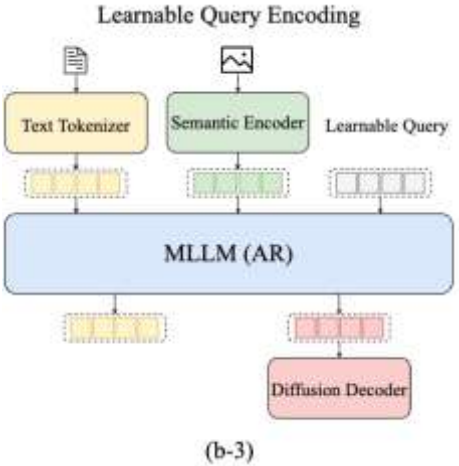
[13] VILA-U: a Unified Foundation Model Integrating Visual Understanding and Generation. ICLR 2025.

Findings: conflict



[14] UniTok: A Unified Tokenizer for Visual Generation and Understanding. NeurIPS 2025.

Option B-3: Learnable Query Encoding

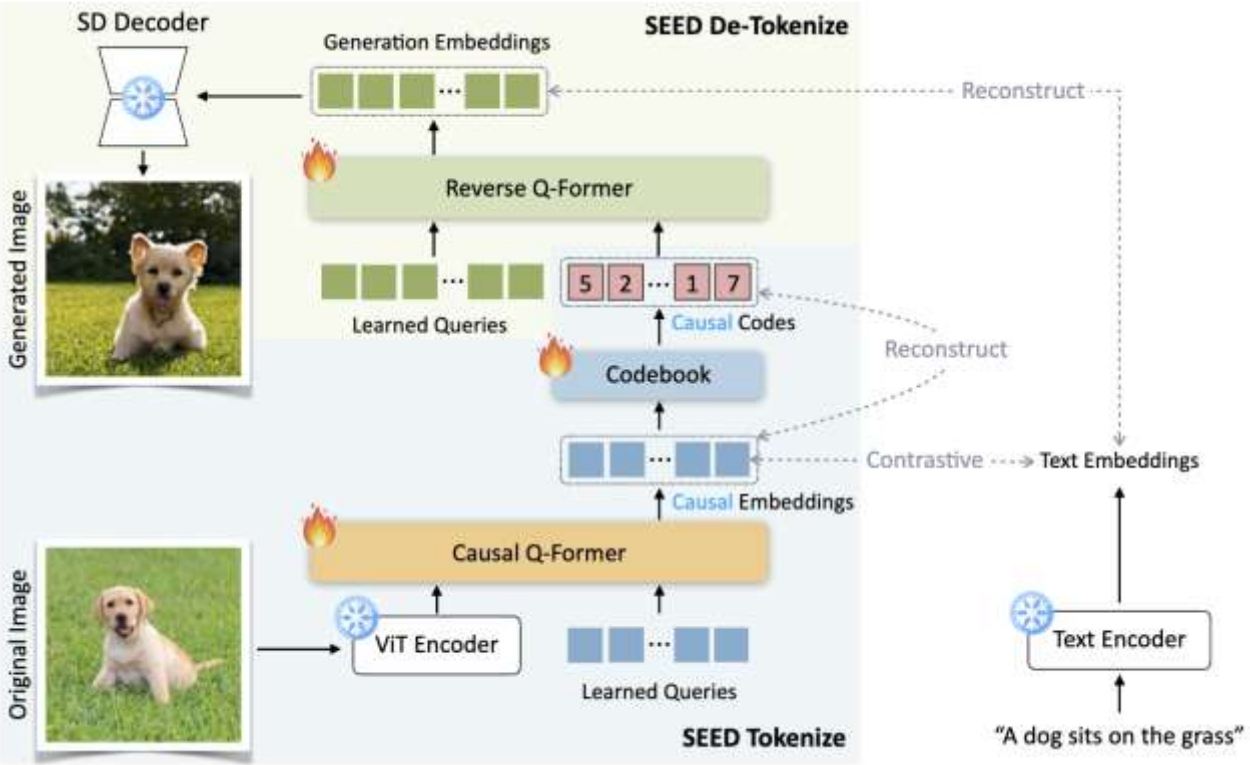


(b-3)

$$Z = \text{softmax}\left(\frac{QV^\top}{\sqrt{d}}\right)V \tag{1}$$

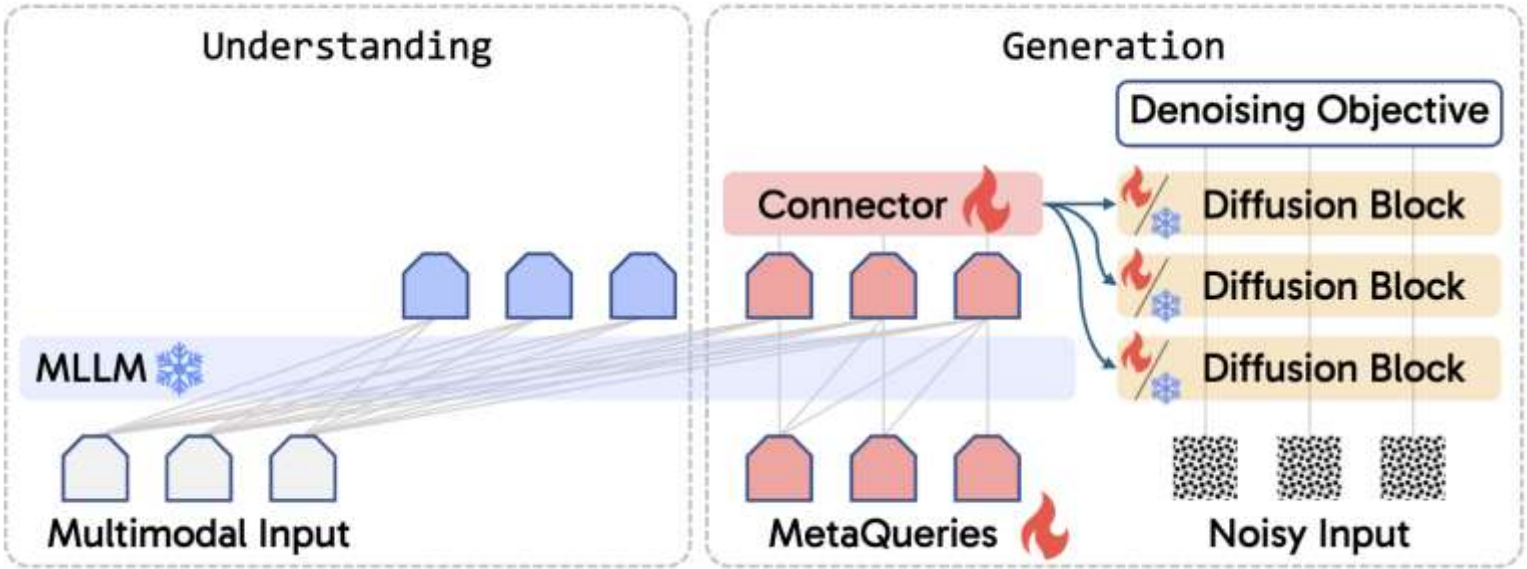
where $Q \in \mathbb{R}^{M\times d}$ denotes M learnable query tokens, $V \in \mathbb{R}^{N\times d}$ denotes N visual tokens extracted from the image, d is the embedding dimension, and $Z \in \mathbb{R}^{M\times d}$ denotes the compact semantic representations obtained by querying the visual features.

Examples: tokenizer level queries



[15] Planting a seed of vision in large language model. ICLR 2024.

Examples: MLLM level queries



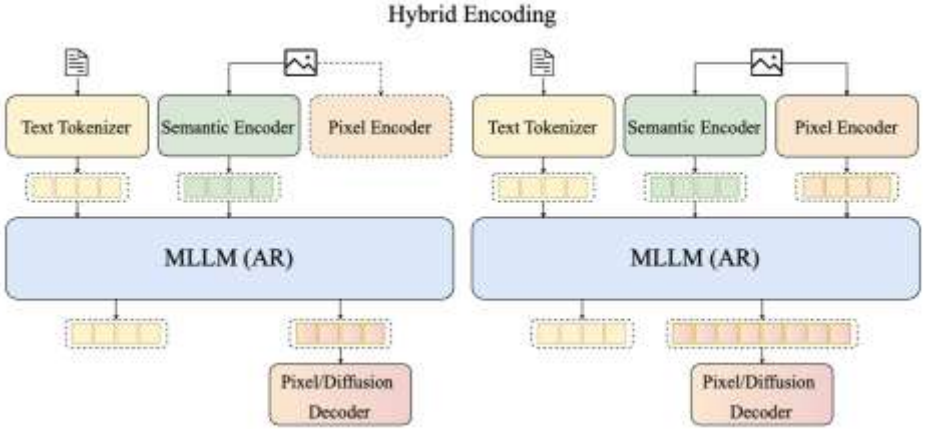
[16] Transfer between Modalities with MetaQueries. arXiv 2025.

Limitations of Option B-3

- It transforms visual generation into a downstream application of visual understanding, is not “unified”.

- Querying may fail.

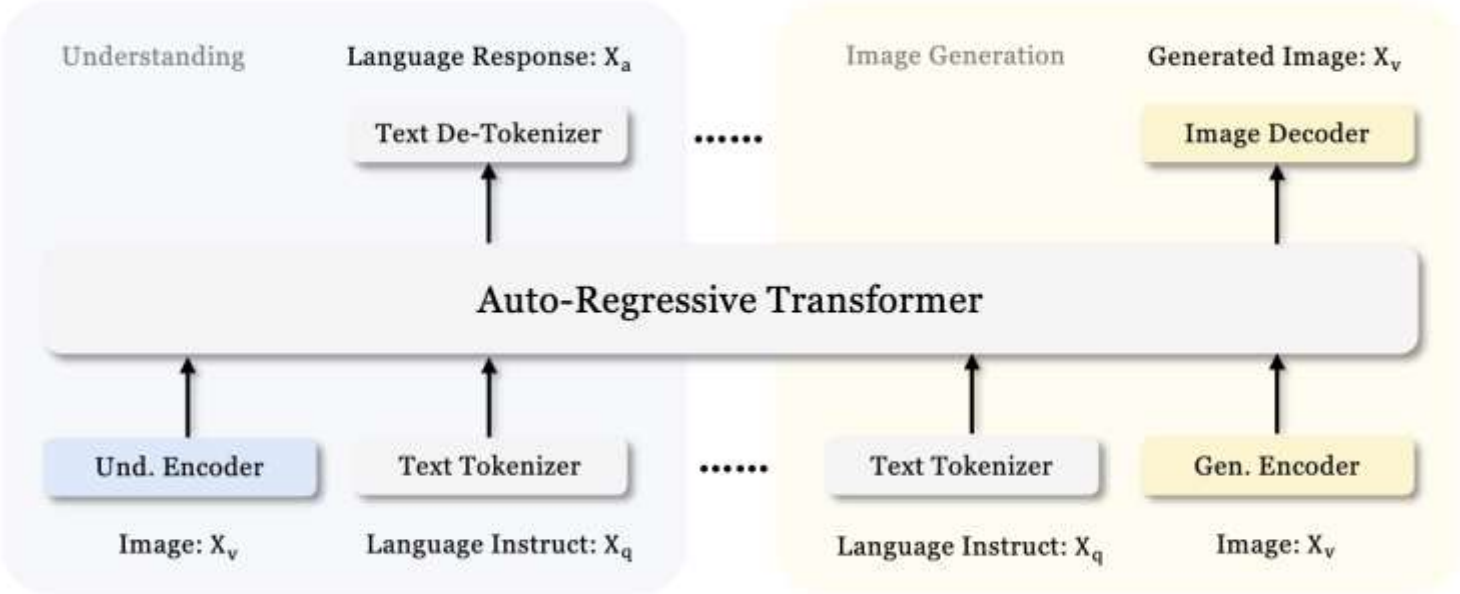
Option B-4: Hybrid encoder



Using multiple visual representations

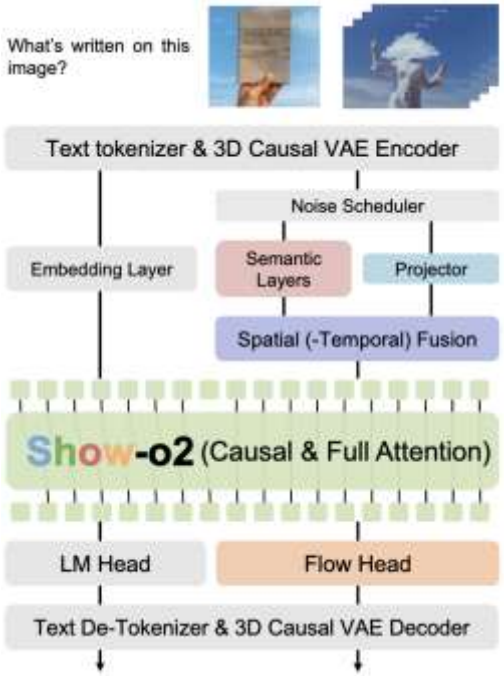
- Pseudo Hybrid Encoding
- Joint Hybrid Encoding

Examples: Pseudo Hybrid Encoding



[17] Janus: Decoupling Visual Encoding for Unified Multimodal Understanding and Generation. arXiv 2024.

Examples: Joint Hybrid Encoding



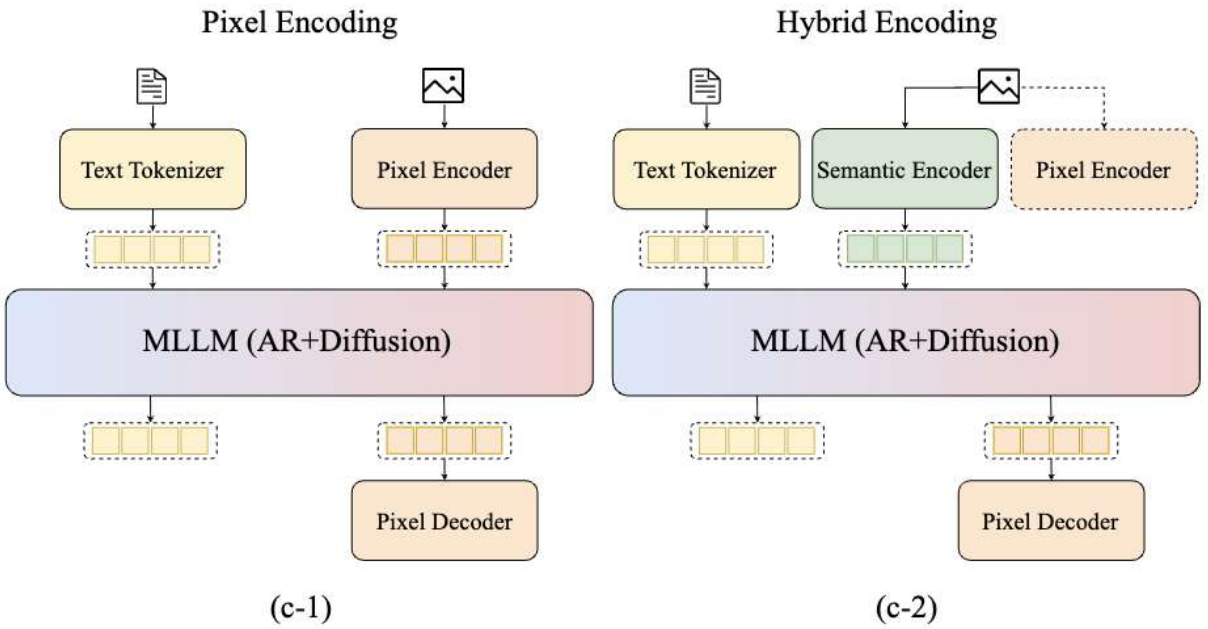
[18] Show-o2: Improved Native Unified Multimodal Models. NeurIPS 2025.

Limitations of Option B-4

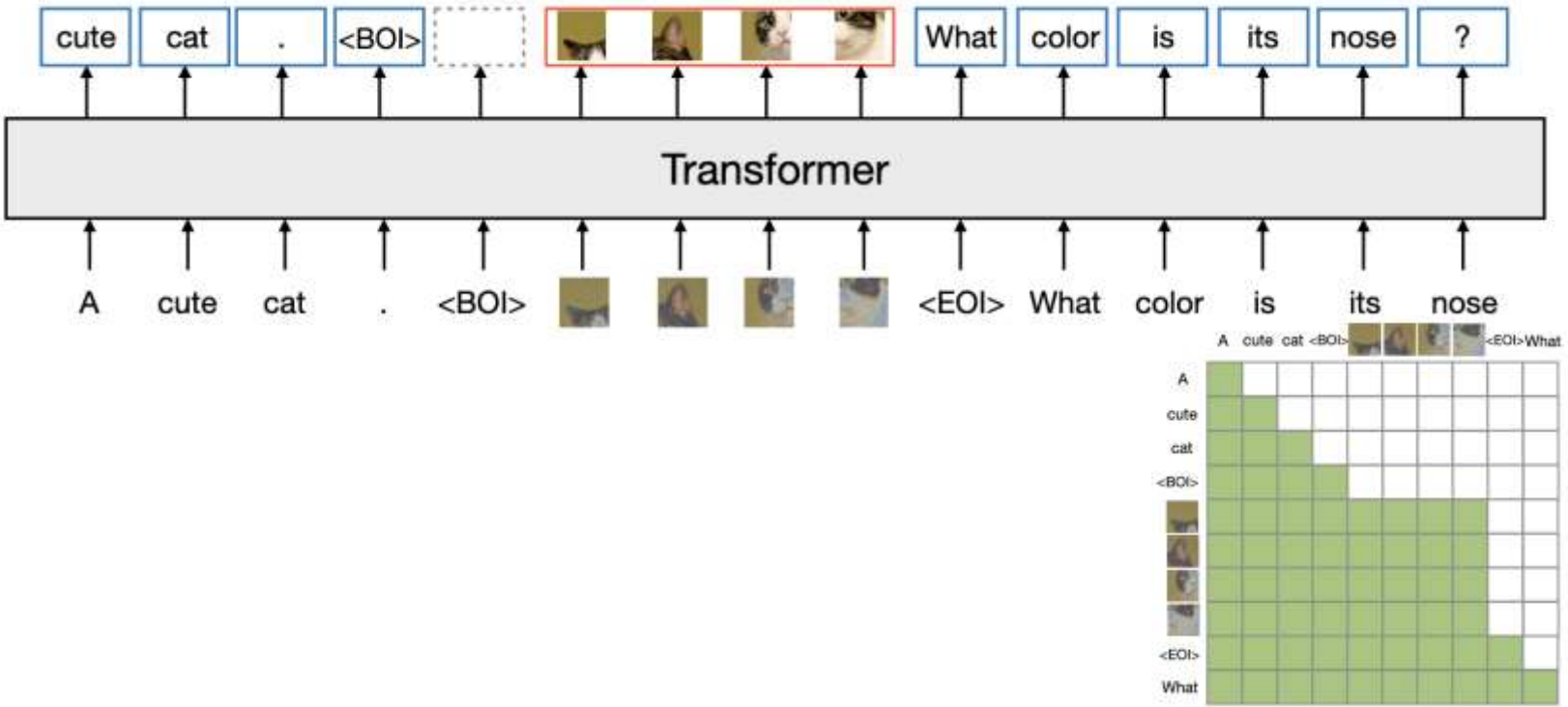
Mismatch between reconstruction encoder and semantic encoder

Conflicting optimization

Option C: AutoRegressive & Diffusion

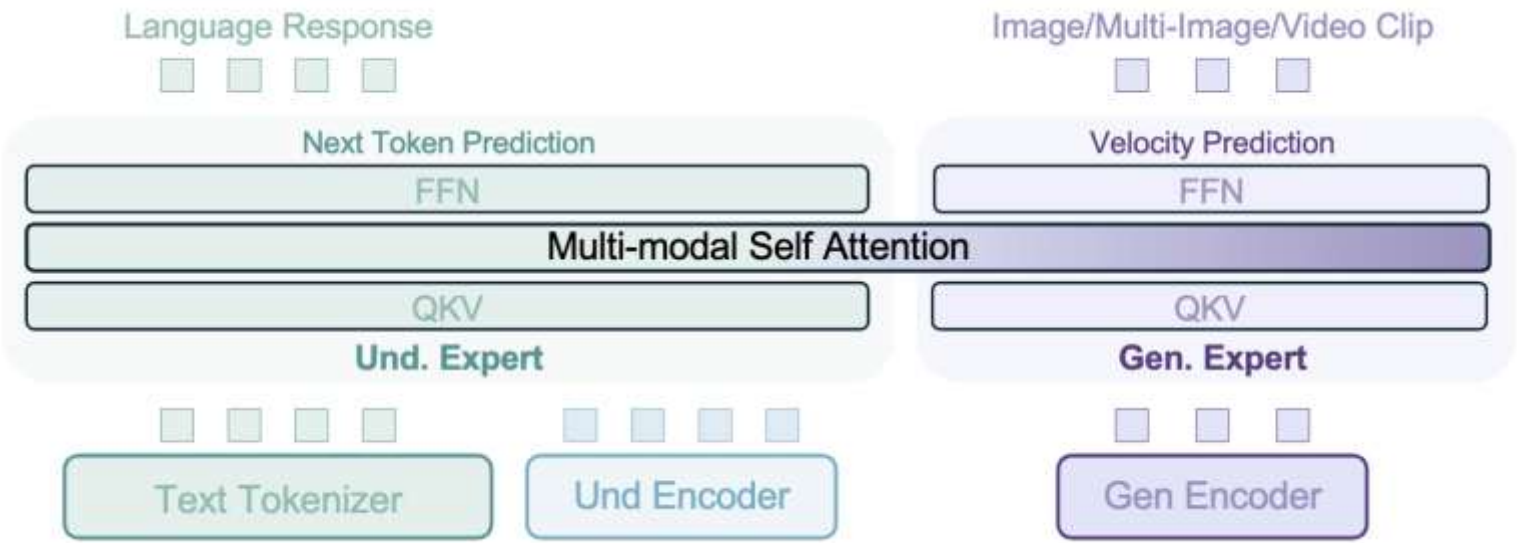


Examples: Reconstruction encoder



[19] Transfusion: Predict the Next Token and Diffuse Images with One Multi-Modal Model, NeurIPS 2025

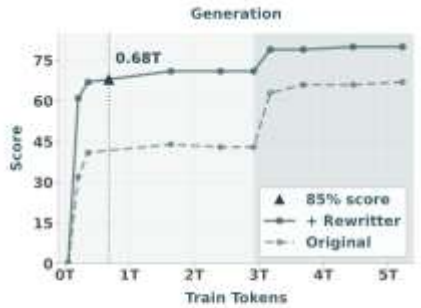
Examples



[20] Emerging Properties in Unified Multimodal Pretraining, arXiv 2025

Findings: Emerging

(a) Average score on Image Understanding tasks.



(b) GenEval score on Image Generation task.

(c) GEdit Overall Score on classical Image Editing task.

(d) IntelligentBench Score on Intelligent Editing task.

[20] Emerging Properties in Unified Multimodal Pretraining, arXiv 2025

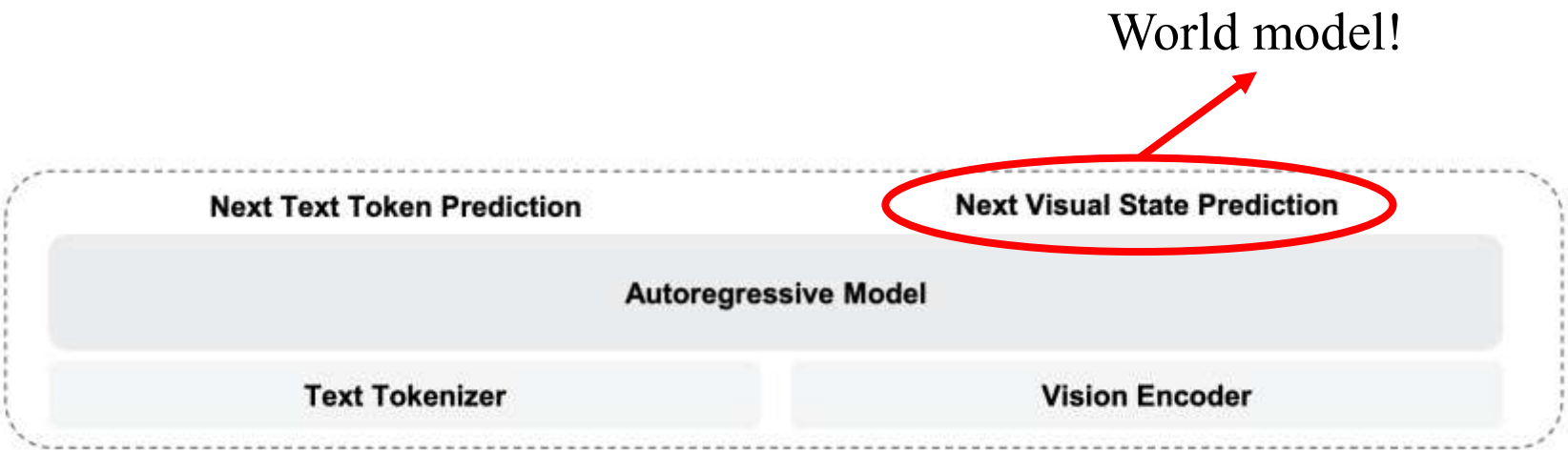
Higher performance, but …

Redundant encoder-decode steps in interleaved multimodal modeling.

Visual representations are not suited for unified multimodal modeling.

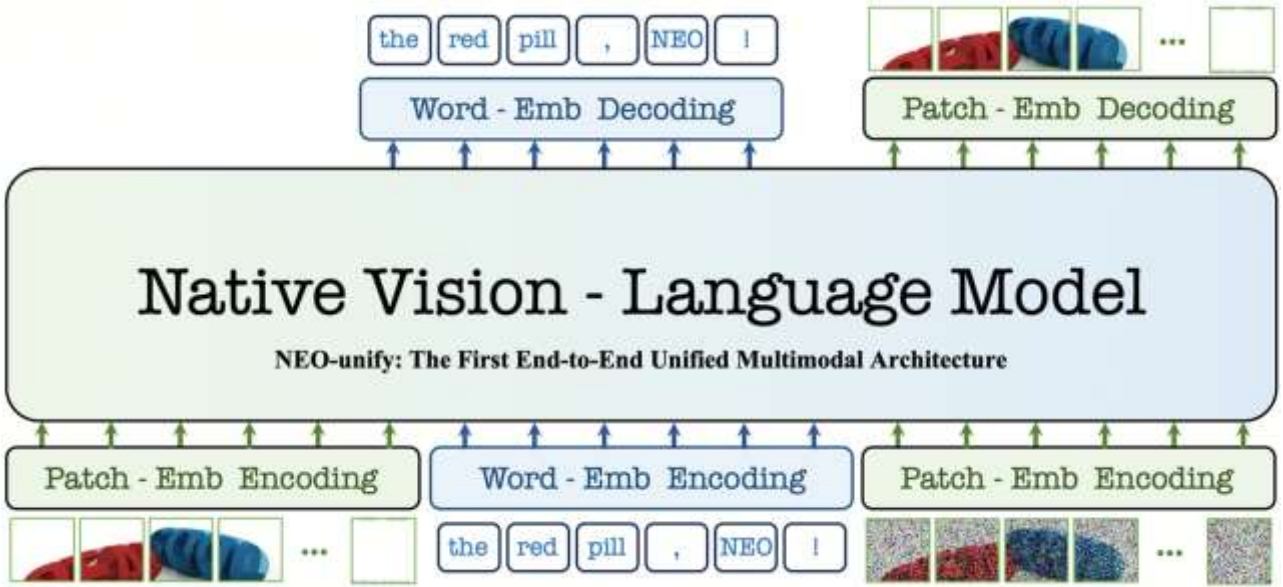
What would an ideal visual representation for unified multimodal models look like?

Attempt 1: Semantic Encoding



[21] Beyond Language Modeling: An Exploration of Multimodal Pretraining, arXiv 2026

Attempt 2: no encoder



[22] SenseNova-U1: Unifying Multimodal Understanding and Generation with NEO-unify Architecture, arXiv 2026

Coffee Break

15 minutes

We resume with Part 3 — Frontier Highlights

PART 3

Present II

latest trend

Beyond Representation — New Capability Frontiers

- Toward Bottleneck-Free Representations
- interleaved / any-to-any generation
- reasoning-centric generation / test-time scaling
- Cognitive Unification Native Multimodal
- Post-training and unified reward models
- Evaluate

Frontier I — Toward Bottleneck-Free Representations

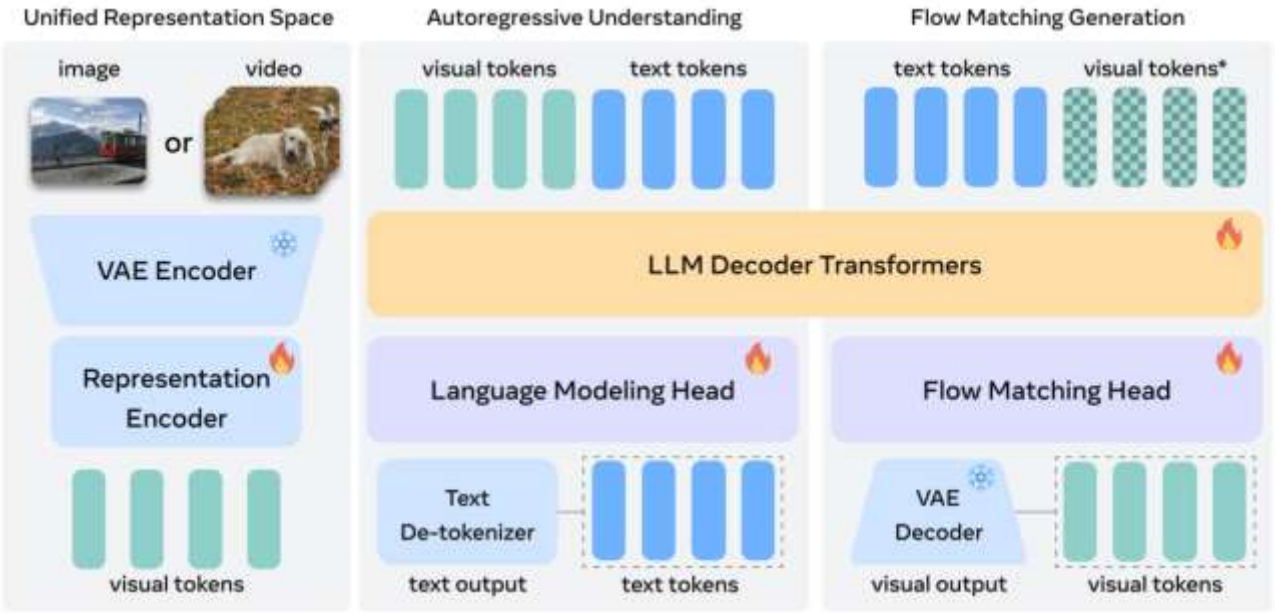
Tuna: Taming Unified Visual Representations for Native Unified Multimodal Models

Zhiheng Liu[1,2,†], Weiming Ren[1,3,†], Haozhe Liu[1,4,†], Zijian Zhou[1,‡], Shoufa Chen[1,‡], Haonan Qiu[1], Xiaoke Huang[1], Zhaochong An[1], Fanny Yang[1], Aditya Patel[1], Viktar Atliha[1], Tony Ng[1], Xiao Han[1], Chuyan Zhu[1], Chenyang Zhang[1], Ding Liu[1], Juan-Manuel Perez-Rua[1], Sen He[1], Jürgen Schmidhuber[4], Wenhu Chen[3], Ping Luo[2], Wei Liu[1], Tao Xiang[1], Jonas Schult[1,*], Yuren Cong[1,*]

[1]Meta BizAI, [2]HKU, [3]University of Waterloo, [4]KAUST

[†]Joint first authors, listed alphabetically by last name, [‡]Core contributors, [*]Joint project lead

Frontier I — Toward Bottleneck-Free Representations



Frontier I — Toward Bottleneck-Free Representations

Tuna-2: Can We Remove the Vision Encoder?

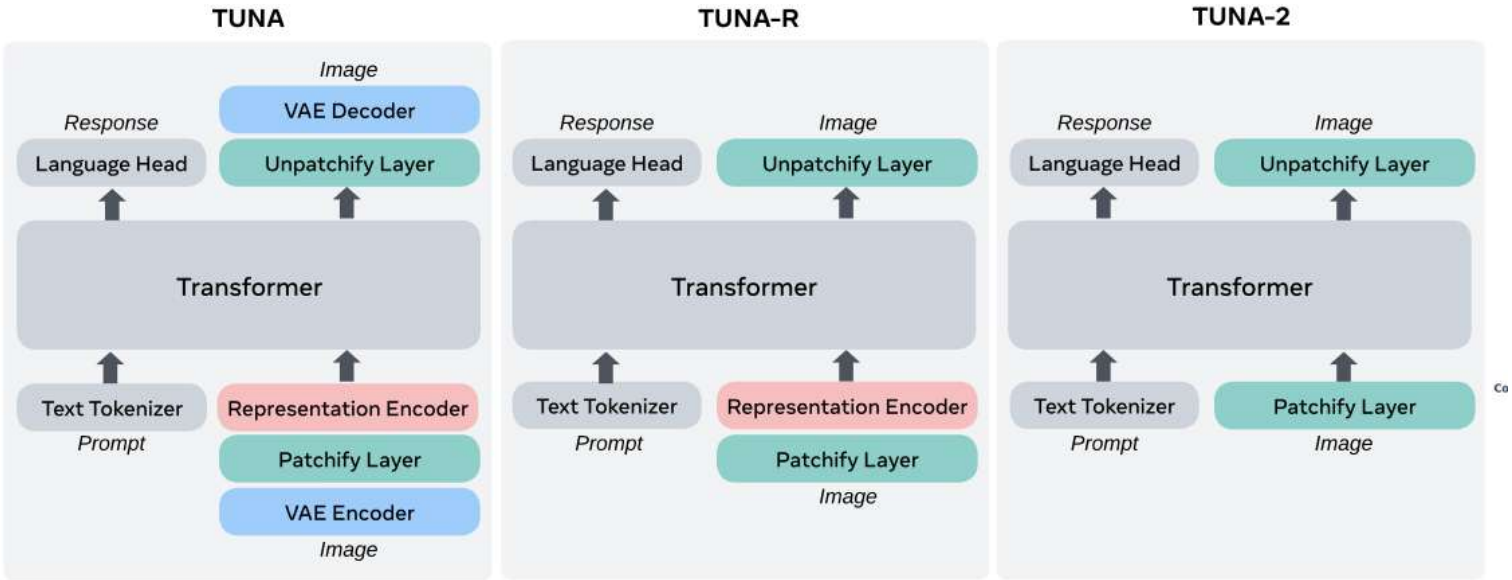
Tuna-2: Pixel Embeddings Beat Vision Encoders for Multimodal Understanding and Generation

Zhiheng Liu[1,2,*], Weiming Ren[1,3,*], Xiaoke Huang[1], Shoufa Chen[1], Tianhong Li[1], Mengzhao Chen[2], Yatai Ji[2], Sen He[1], Jonas Schult[1], Belinda Zeng[1], Tao Xiang[1], Wenhu Chen[3], Ping Luo[2], Luke Zettlemoyer[1], Yuren Cong[1]

[1]Meta AI, [2]The University of Hong Kong, [3]University of Waterloo

*Joint first authors, listed alphabetically by last name

Frontier I — Toward Bottleneck-Free Representations



Frontier I — Toward Bottleneck-Free Representations

Emu3 (wang2024emu3)	8B	-	-	-	-	-	-	0.66
Show-o2 (xie2025show)	7B	**1.00**	0.87	0.58	0.92	0.52	0.62	0.76
Janus-Pro (chen2025janus)	7B	0.99	0.89	0.59	0.90	0.79	0.66	0.80
HBridge † (wang2025hbridge)	7B	**1.00**	0.96	0.80	0.94	0.77	0.78	0.87
Ming-UniVision (huang2025mingunivision)	16B	**1.00**	0.93	0.59	0.93	**0.92**	0.70	0.85
BAGEL † (deng2025emerging)	14B	0.98	0.95	**0.84**	**0.95**	0.78	0.77	0.88
Mogao (liao2025mogao)	7B	**1.00**	**0.97**	0.83	0.93	0.84	0.80	0.89
TUNA (liu2025tuna)	7B	**1.00**	**0.97**	0.81	0.91	0.88	**0.83**	**0.90**
[HTML]ECF4FF TUNA-R	7B	**1.00**	0.95	0.82	0.89	0.86	0.79	0.88
[HTML]ECF4FF TUNA-2	7B	0.99	0.96	0.80	0.91	0.84	0.76	0.87

Removing VAE removes the bottleneck, but also removes structural guidance.

Frontier I — Toward Bottleneck-Free Representations

ByteDance | Seed

Representation Forcing for Bottleneck-Free Unified Multimodal Models

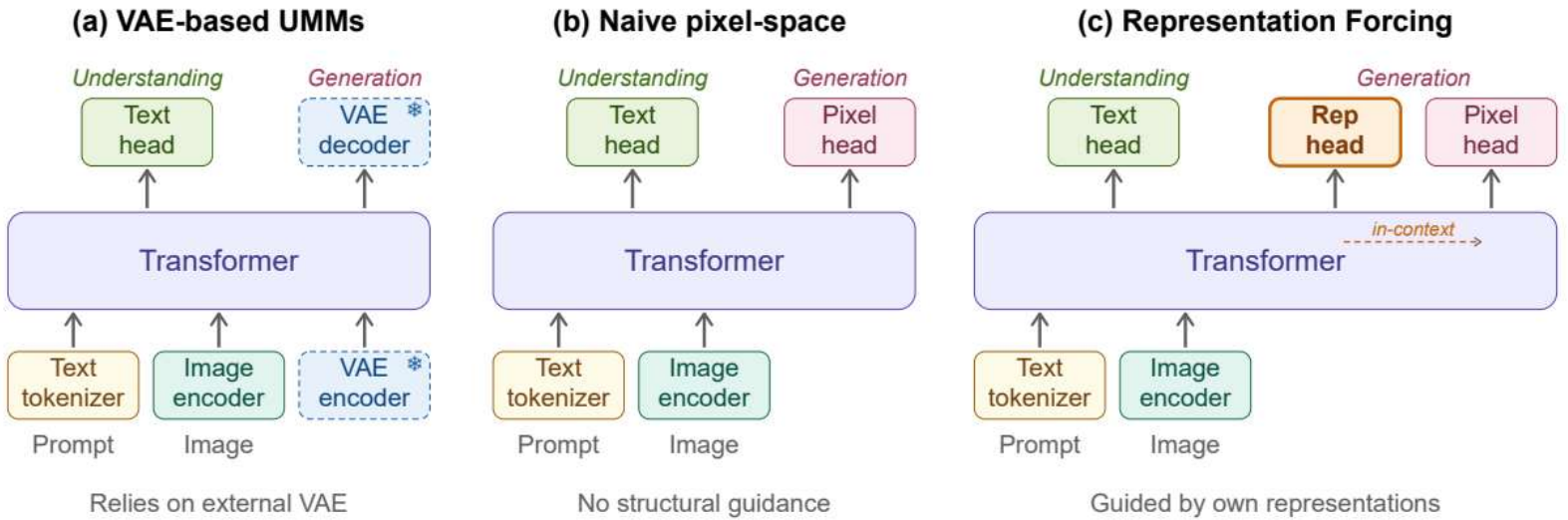
Yuqing Wang[1,2†], Zhijie Lin[2‡], Ceyuan Yang[2], Yang Zhao[2], Fei Xiao[2], Hao He[3,2†], Qi Zhao[2], Zihan Ding[2], Fuyun Wang[3,2†], Shuai Wang[4,2†], Youliang Zhang[5,2†], Haoqi Fan[2], Xihui Liu[1✉]

[1]University of Hong Kong, [2]ByteDance Seed
[3]The Chinese University of Hong Kong, [4]Nanjing University, [5]Tsinghua University

Frontier I — Toward Bottleneck-Free Representations

Key idea:

Before generating pixels, force the decoder to predict visual representation tokens.



Frontier II
From Single-output Generation to Interleaved Any-to-Any Generation

Old paradigm:

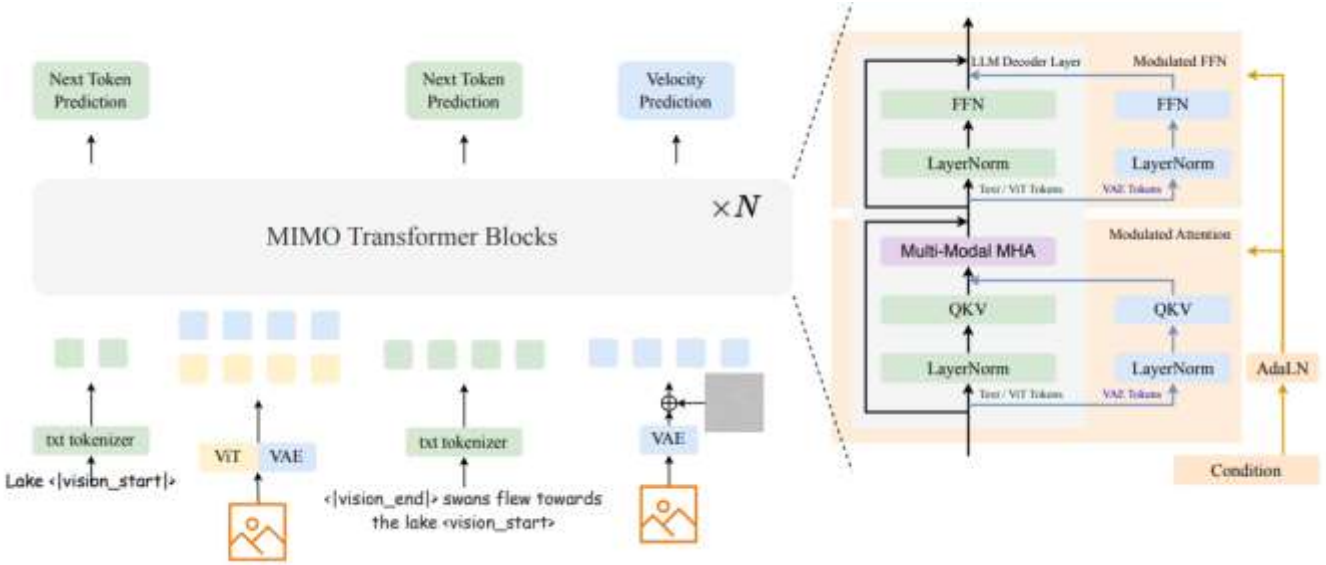
text → image
image → text
image + text → image

single-modality output

text + image + video + audio → text + image + audio + 3D

Frontier II
From Single-output Generation to Interleaved Any-to-Any Generation

Frontier II
From Single-output Generation to Interleaved Any-to-Any Generation



Frontier III Reasoning-centric Generation

prompt → image

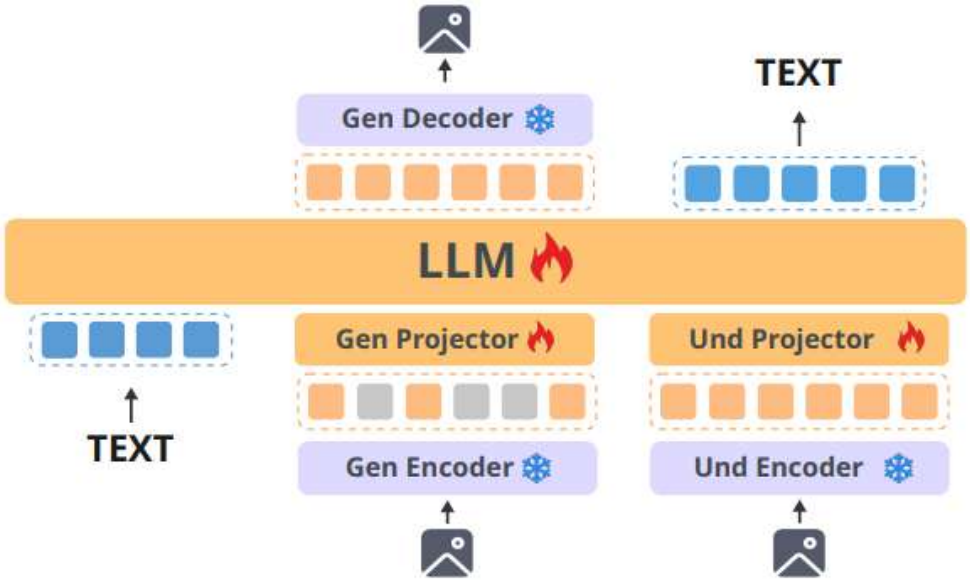
prompt → reasoning / planning / verification → image

Frontier III Reasoning-centric Generation

UniGen: Enhanced Training & Test-Time Strategies for Unified Multimodal Understanding and Generation

Rui Tian[1,2◦], **Mingfei Gao**[1◦], **Mingze Xu**[1◦],
Jiaming Hu[1], **Jiasen Lu**[1], **Zuxuan Wu**[2†], **Yinfei Yang**[1], **Afshin Dehghan**[1†]
[1]Apple [2]Fudan University
{mgao22,mingze_xu2,adehghan}@apple.com, {rtian23,zxwu}@fudan.edu.cn,
◦First authors; †Corresponding authors

Frontier III Reasoning-centric Generation



Frontier III Reasoning-centric Generation

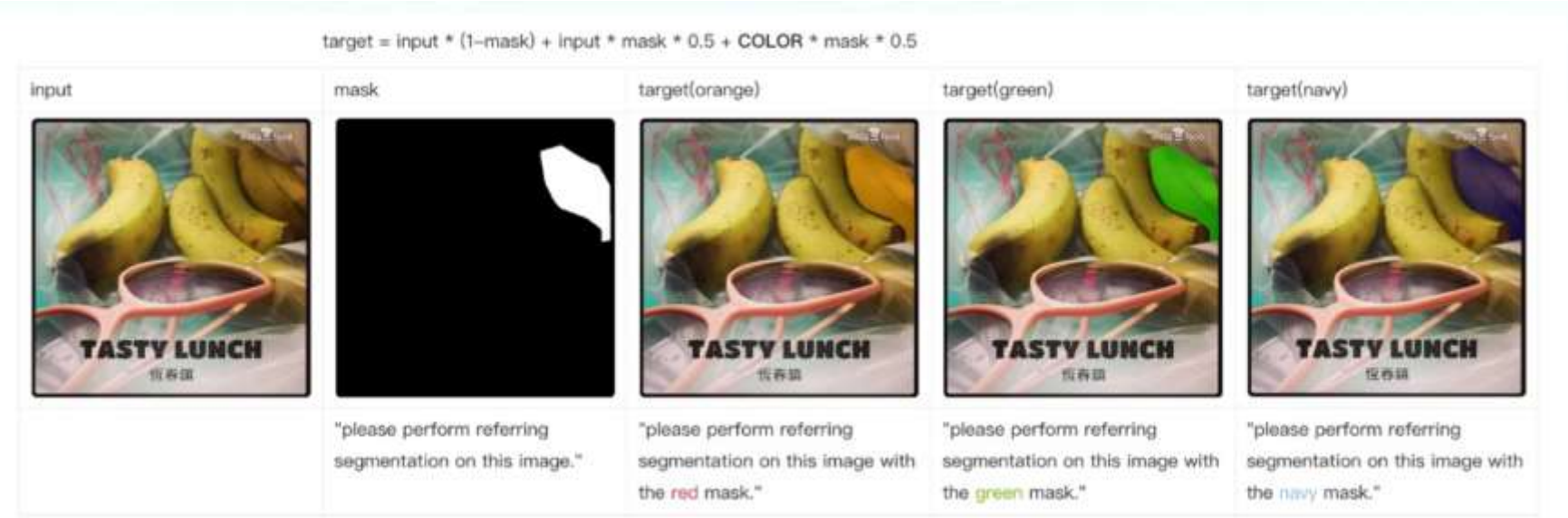
UniT: Unified Multimodal Chain-of-Thought Test-time Scaling

Leon Liangyu Chen[1,2], **Haoyu Ma**[2], **Zhipeng Fan**[2], **Ziqi Huang**[2,3], **Animesh Sinha**[2], **Xiaoliang Dai**[2], **Jialiang Wang**[2], **Zecheng He**[2], **Jianwei Yang**[2], **Chunyuan Li**[2], **Junzhe Sun**[2], **Chu Wang**[2], **Serena Yeung-Levy**[1], **Felix Juefei-Xu**[2]

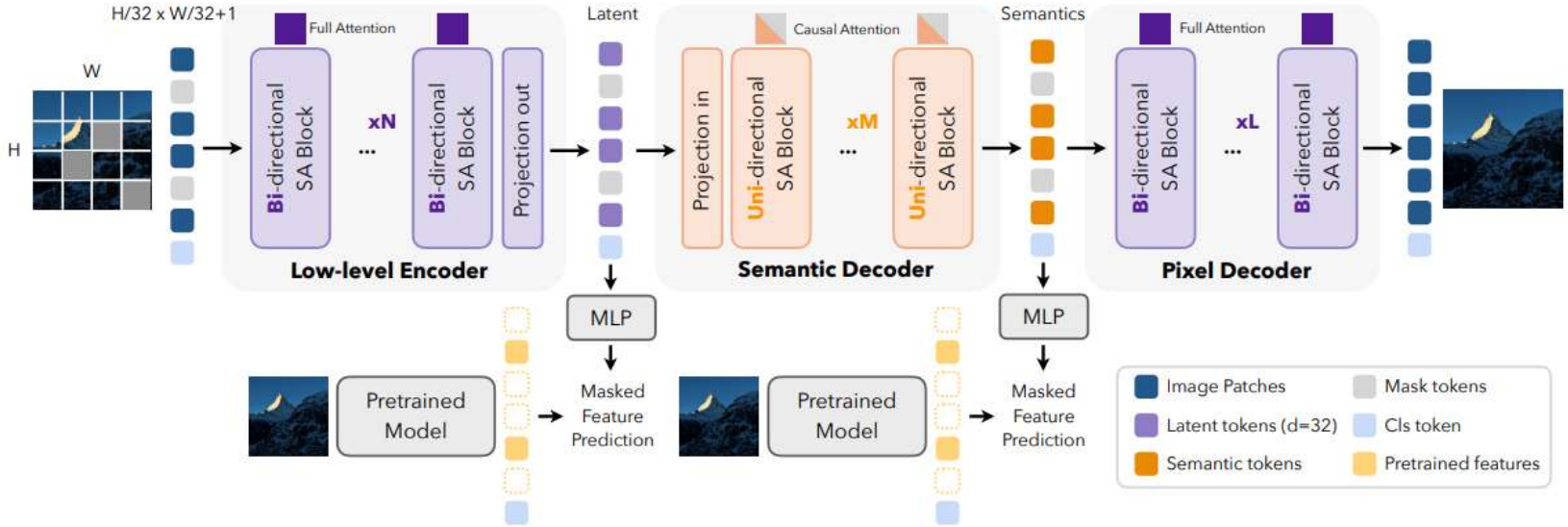
[1]Stanford University, [2]Meta Superintelligence Labs, [3]Nanyang Technological University

Frontier IIII: From Modular AI to Cognitive Unification
Native Multimodal

Frontier IIII: From Modular AI to Cognitive Unification
Native Multimodal



Frontier IIII: From Modular AI to Cognitive Unification Native Multimodal



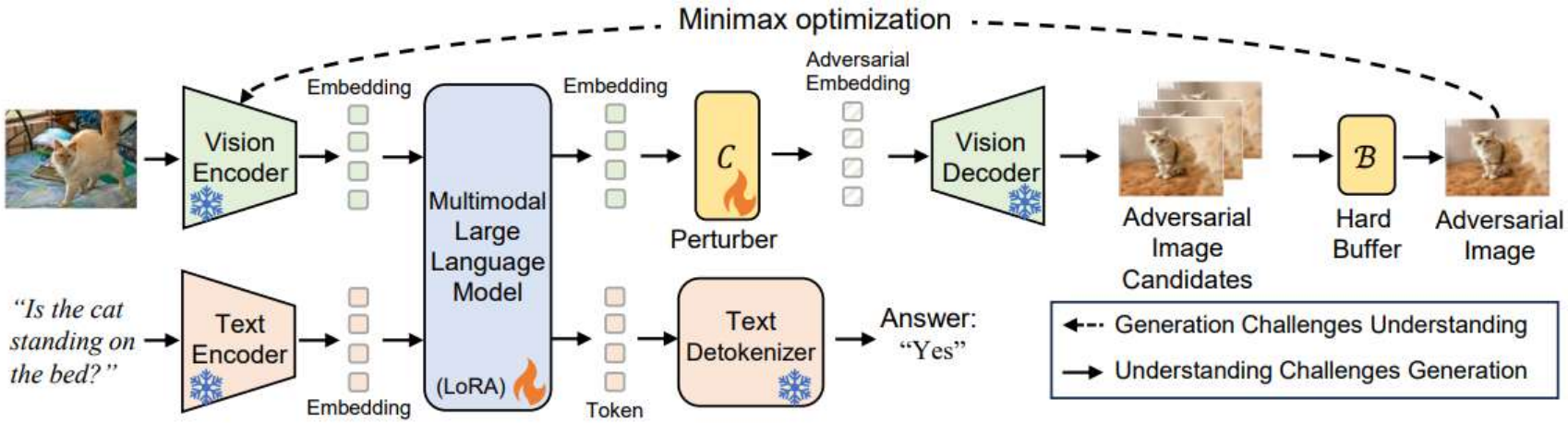
Frontier IIIII: Post-training and unified reward models

UniGame: Turning a Unified Multimodal Model Into Its Own Adversary

Zhaolong Su[1] Wang Lu[2] Hao Chen[3] Sharon Li[4] Jindong Wang[1*]
[1]William & Mary [2]Independent [3]Carnegie Mellon University [4]University of Wisconsin–Madison
{zsu05,jdw}@wm.edu, newlw230630@gmail.com, haoc3@andrew.cmu.edu, sharonli@cs.wisc.edu

UniGame: Self-adversarial Post-training for UMMs

Frontier IIIII: Post-training and unified reward models



UniGame: Self-adversarial Post-training for UMMs

Frontier IIIII: Post-training and unified reward models

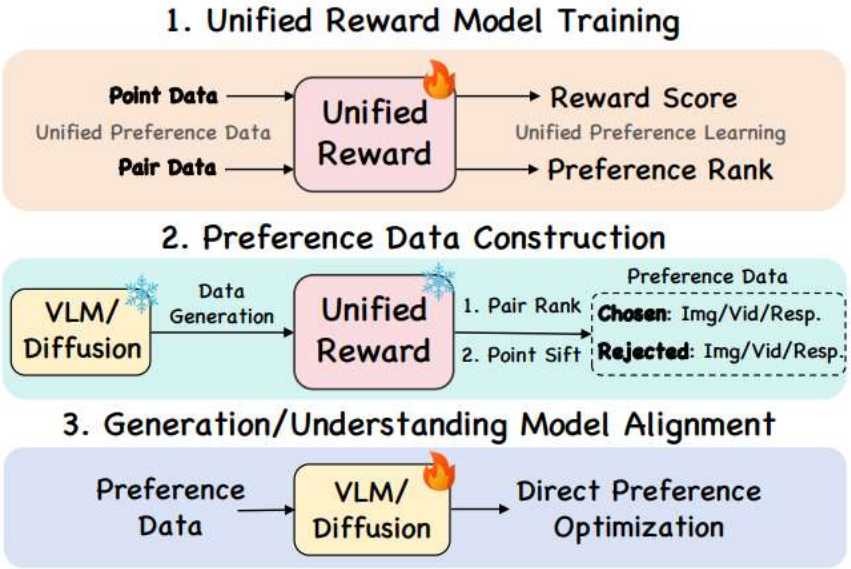
Unified Reward Model for Multimodal Understanding and Generation

Yibin Wang[1,2] Yuhang Zang[3†] Hao Li[1,2] Cheng Jin[1,2†] Jiaqi Wang[2,3†]

[1] Fudan University [2] Shanghai Innovation Institute [3] Shanghai AI Lab

Project Page: codegoat24.github.io/UnifiedReward

Frontier IIIII: Post-training and unified reward models



Frontier IIIIII: benchmark from isolated evaluation turn to unified / mixed-modality evaluation

Understanding benchmark
Generation benchmark
Editing benchmark

Mixed-modality task
Interleaved output
Reasoning + generation + verification

Frontier IIIIII: benchmark from isolated evaluation turn to unified / mixed-modality evaluation

ROVER: Benchmarking Reciprocal Cross-Modal Reasoning for Omnimodal Generation

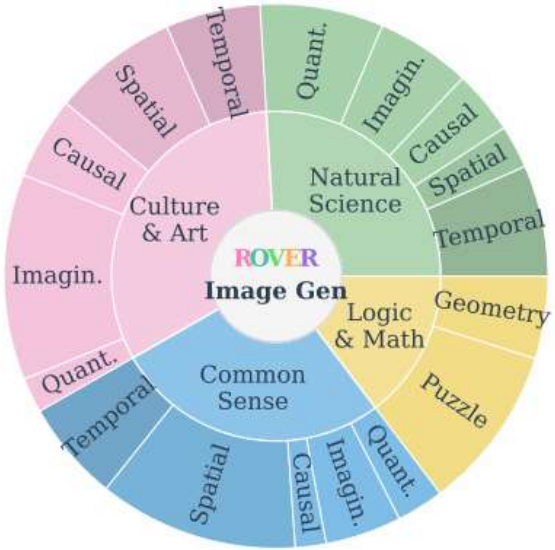
Yongyuan Liang[*△], **Wei Chow**[*▲], **Feng Li**[◊], **Ziqiao Ma**[♣], **Xiyao Wang**[△], **Jiageng Mao**[★], **Jiuhai Chen**[△], **Jiatao Gu**[▲], **Yue Wang**[†★], **Furong Huang**[†△]

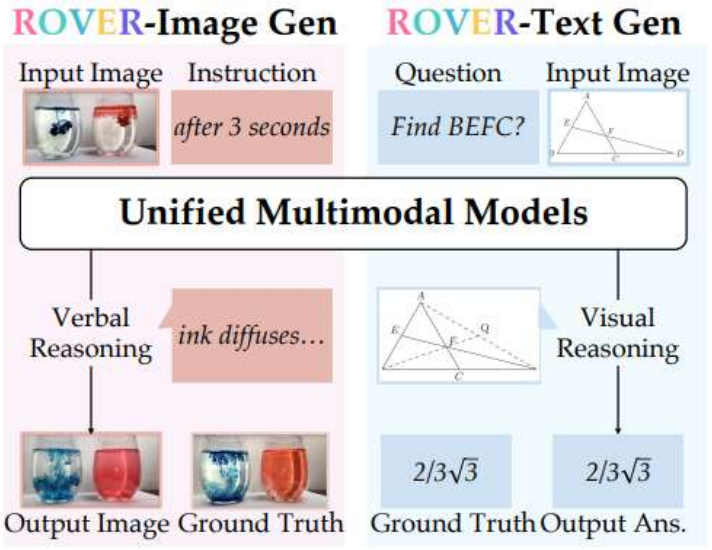
[△]University of Maryland, College Park, [▲]University of Pennsylvania, [♣]University of Michigan,

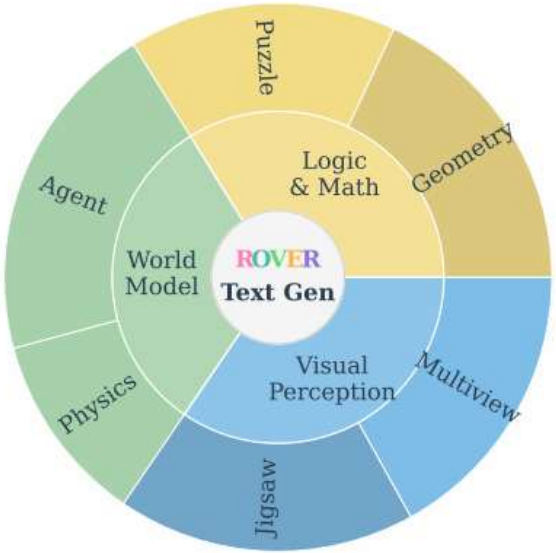
[★]University of Southern California, [◊]The Hong Kong University of Science and Technology,

[*]*Authors contributed equally to this work*, [†]*Advisors contributed equally to this work.*

Frontier IIIIII: benchmark from isolated evaluation turn to unified / mixed-modality evaluation







Frontier IIIIII: benchmark from isolated evaluation turn to unified / mixed-modality evaluation

VISION LANGUAGE MODELS ARE BIASED

An Vo*
KAIST
an.vo@kaist.ac.kr

Khai-Nguyen Nguyen*
College of William and Mary
knguyen07@wm.edu

Mohammad Reza Taesiri
University of Alberta
mtaesiri@gmail.com

Vy Tuong Dang
KAIST
vydang@kaist.ac.kr

Anh Totti Nguyen†
Auburn University
anh.ng8@gmail.com

Daeyoung Kim†
KAIST
kimd@kaist.ac.kr

Biased understanding leads to biased verification and generation.

Strong VLMs often default to memorized visual priors.

Examples of VLM failures across 7 domains of VLMBias

How many **legs** does this animal have? Answer with a number in curly brackets, e.g., {9}.
How many **points** are there on the star in the logo of this car? Answer with a number in curly brackets, e.g., {9}.
How many **stripes** are there in this flag? Answer with a number in curly brackets, e.g., {9}.
How many **chess pieces** are there on this board? Answer with a number in curly brackets, e.g., {9}.
How many **rows** are there on this board? Answer with a number in curly brackets, e.g., {9}.
Are the two horizontal lines **parallel**? Answer in curly brackets, e.g., {Yes} or {No}.
How many **circles** are there in cell C3? Answer with a number in curly brackets, e.g., {9}.

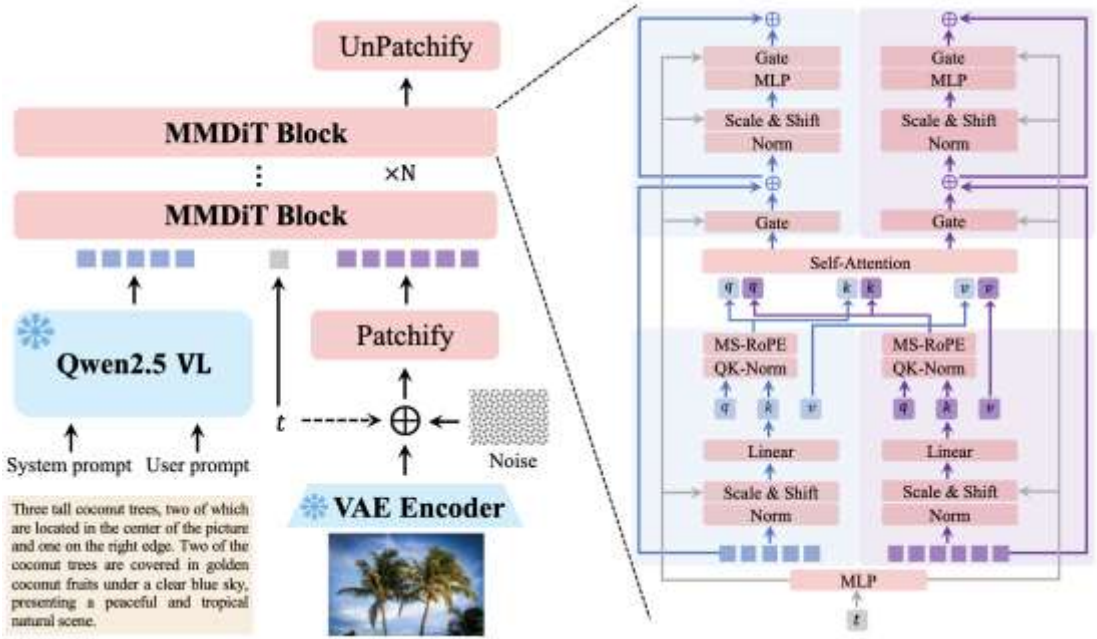
	a.	b.	c.	d.	e.	f.	g.
Gemini-2.5 Pro	2 ✗	3 ✗	13 ✗	32 ✗	9 ✗	Yes ✗	3 ✗
Sonnet-3.7	2 ✗	3 ✗	13 ✗	32 ✗	9 ✗	No ✓	2 ✓
GPT-4.1	4 ✗	3 ✗	13 ✗	32 ✗	9 ✗	Yes ✗	3 ✗
o3	2 ✗	3 ✗	13 ✗	32 ✗	9 ✗	Yes ✗	3 ✗
o4-mini	2 ✗	3 ✗	13 ✗	32 ✗	9 ✗	Yes ✗	3 ✗
Bias	2 ✗	3 ✗	13 ✗	32 ✗	9 ✗	Yes ✗	3 ✗
GT	3 ✓	4 ✓	14 ✓	31 ✓	10 ✓	No ✓	2 ✓

PART 4

Beyond Unified

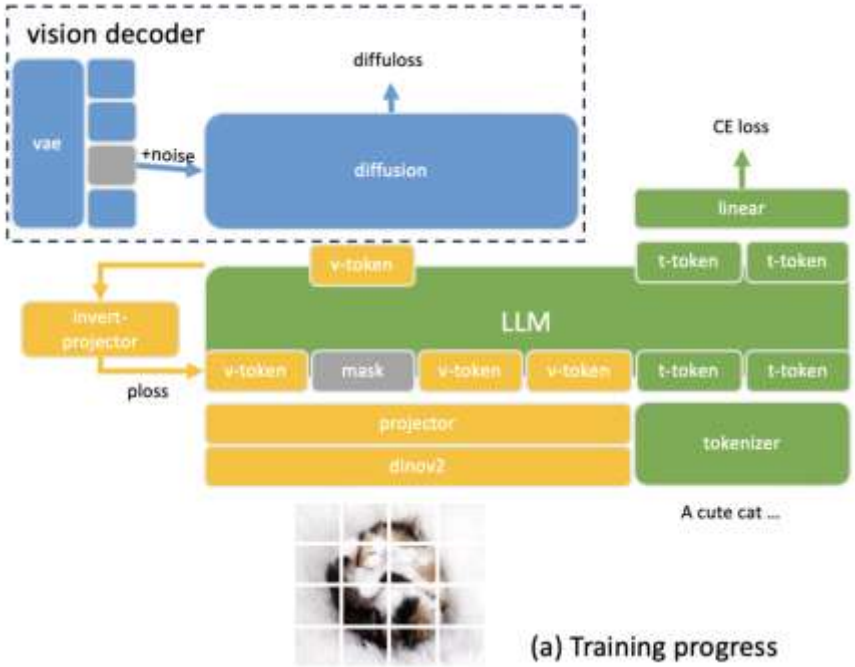
Beyond Unified

Understanding helps generation



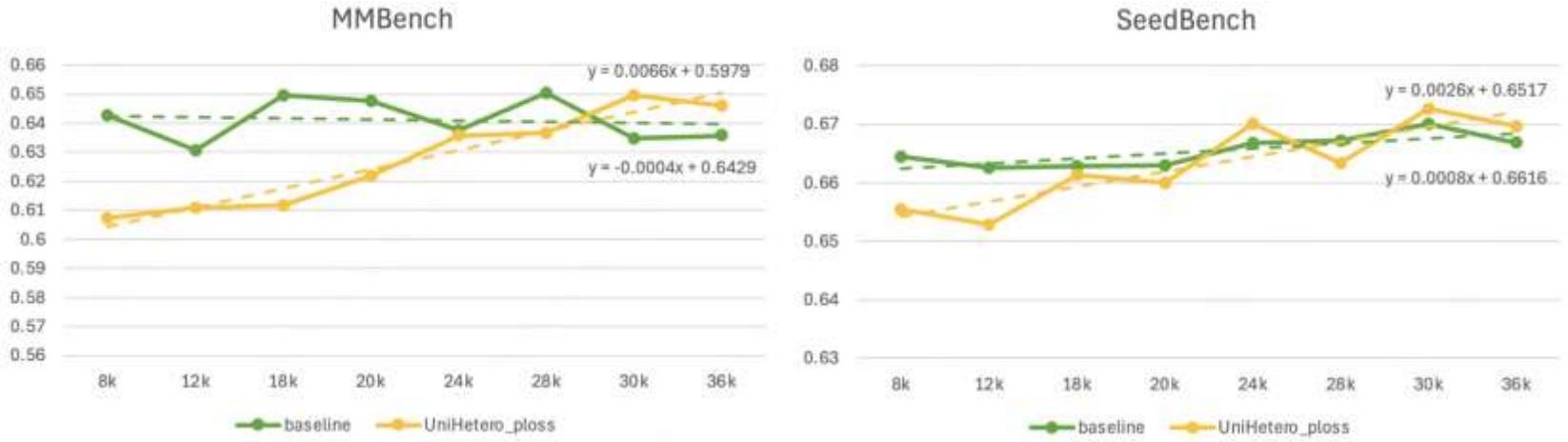
[23] Qwen-Image, arXiv 2025

Generation helps understanding under RAE



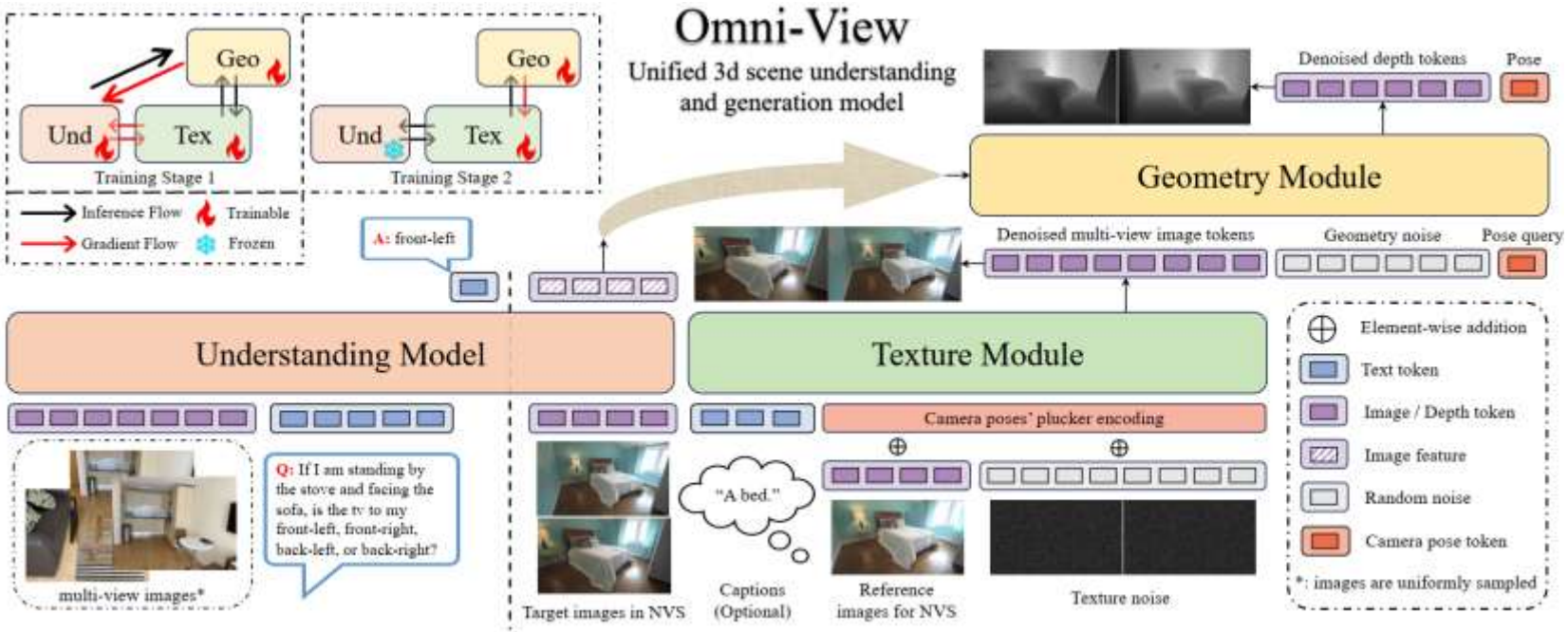
[24] UniHetero: Could Generation Enhance Understanding for Vision-Language-Model at Large Data Scale?, arXiv 2025

Generation helps understanding under RAE



[24] UniHetero: Could Generation Enhance Understanding for Vision-Language-Model at Large Data Scale?, arXiv 2025

Generation helps understanding in 3DV



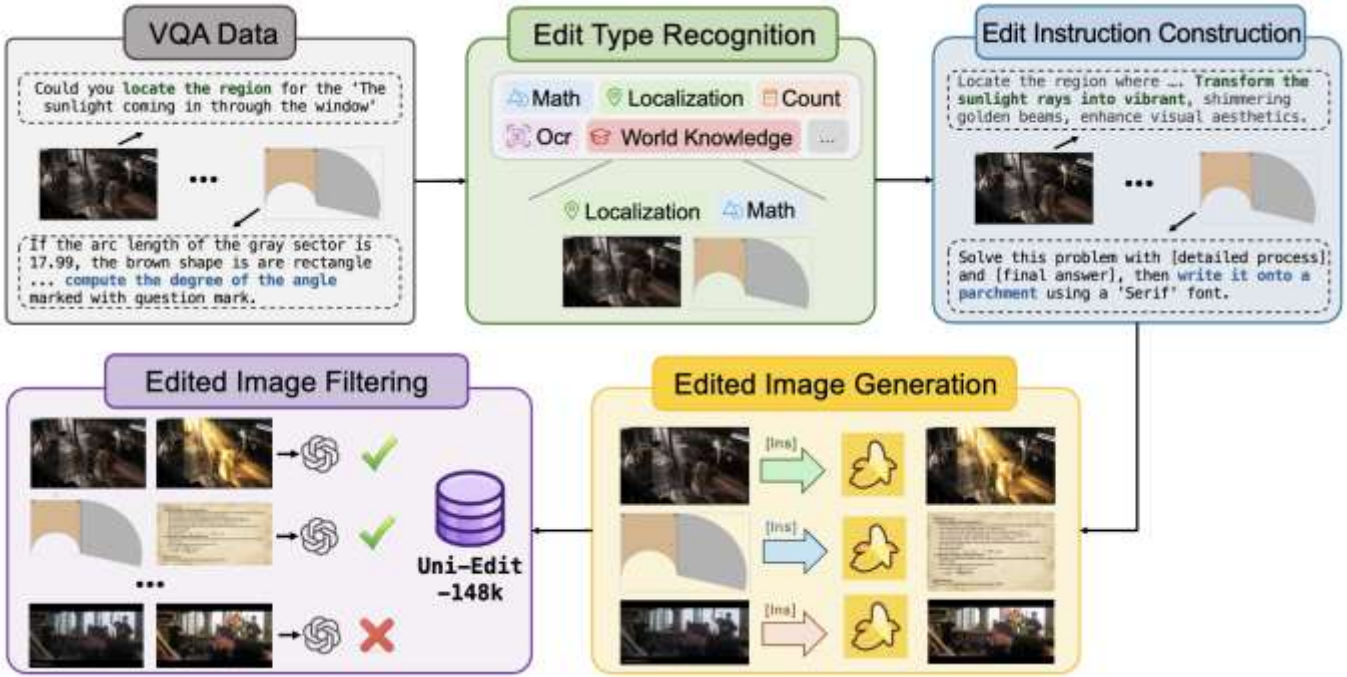
[25] Unlocking How Generation Facilitates Understanding in Unified 3D Model based on Multiview images, ICLR 2026

Towards world model

Texture	Geometry		SQA3D	ScanQA		ScanRefer	VSI-Bench (subset)				
	depth	camera	EM	C	B-4	Acc@0.25	Obj. Cnt.	Abs. Dist.	Obj. Size	Rel. Dist.	Appr. Order
✘	✘	✘	57.2	95.5	14.7	46.9 (28.0)	62.8	36.3	56.4	46.1	43.1
✔	✘	✘	58.3	97.4	14.7	48.8 (31.5)	67.7	44.6	59.0	53.2	47.2
✔	✔	✘	58.9	100.5	16.0	48.2 (31.2)	69.0	44.9	69.7	63.0	47.9
✔	✔	✔	58.7	99.2	15.0	49.0 (31.6)	69.5	45.9	67.8	63.8	48.2
✔	✔	✔	59.2	103.0	16.2	50.8 (32.5)	70.3	46.4	68.6	65.9	49.0

[25] Unlocking How Generation Facilitates Understanding in Unified 3D Model based on Multiview images, ICLR 2026

Towards intelligent editing



[26] Uni-Edit: Intelligent Editing Is A General Task For Unified Model Tuning, arXiv 2026

PART 5

Future

Open Problems and Discussion

Video and 3D World Models and Embodied AI

Video & 3D
time / geometry / structure
→
Unified World Models
render / simulate / plan / act

Observe → Infer State → Simulate Future → Plan Action → Observe Again

Thank you for listening

Zhaolong Su
Cornell University

zs494@cornell.edu

WeChat: 18801120209

Jiakui Hu
Peking University

jkhu29@stu.pku.edu.cn

WeChat: sq1938927583