

Multimodal Foundation Models 3

Advanced multimodal models

Sangdoo Yun, Jin-Hwa Kim

Please don't hesitate to ask questions!
Your questions help everyone (including me) learn better.

Today's lecture

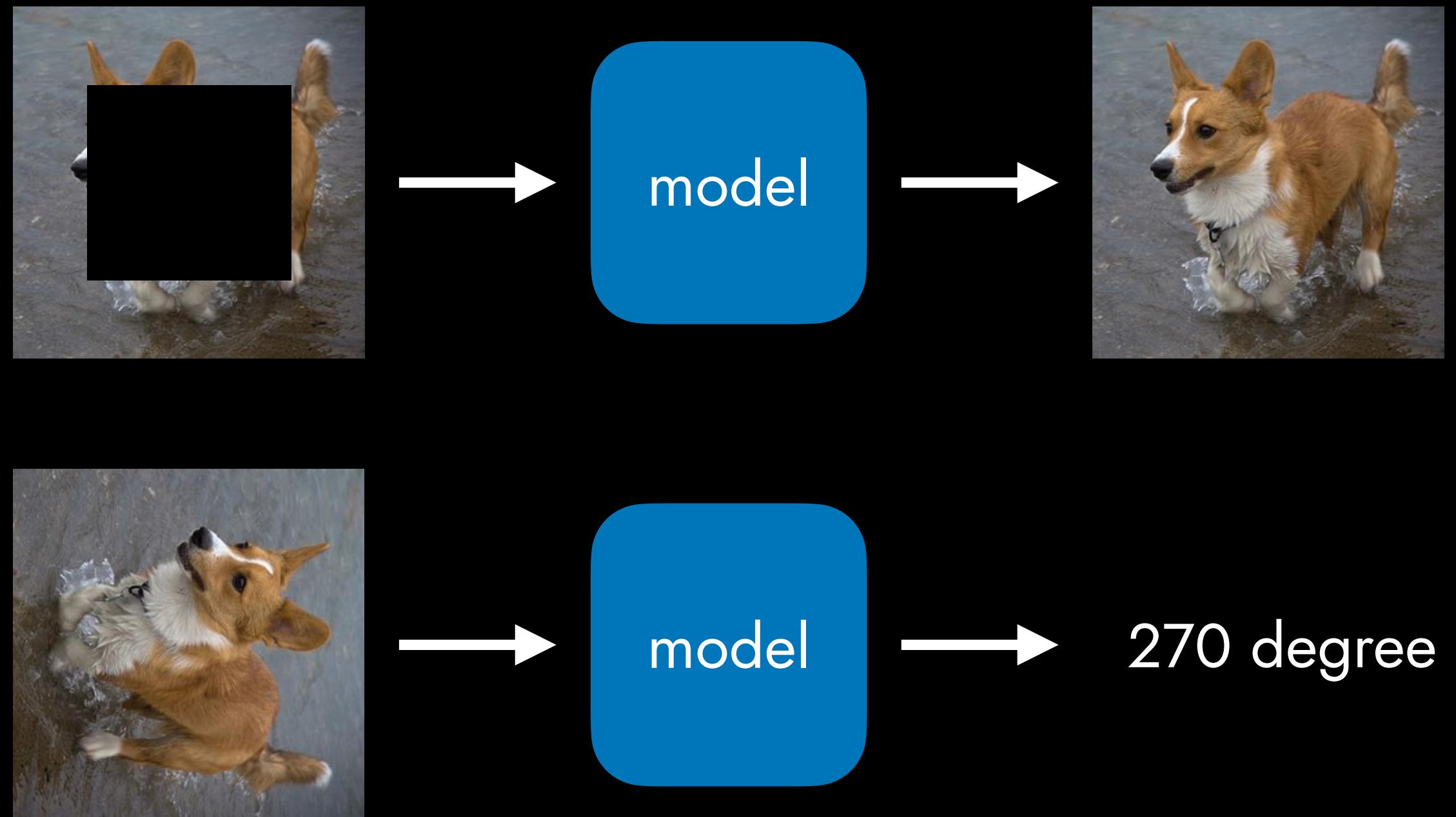
- Recap the last lectures on multimodal models
- Recent research trends in multimodal models
 - Reasoning models
 - Vision-language-action models
 - Omni (any-to-any) models

(Recap)

Self-supervised learning

Self-supervised learning

- *Generate artificial labels* and train models to *predict the generated labels*.
(Recap. Lecture 3)

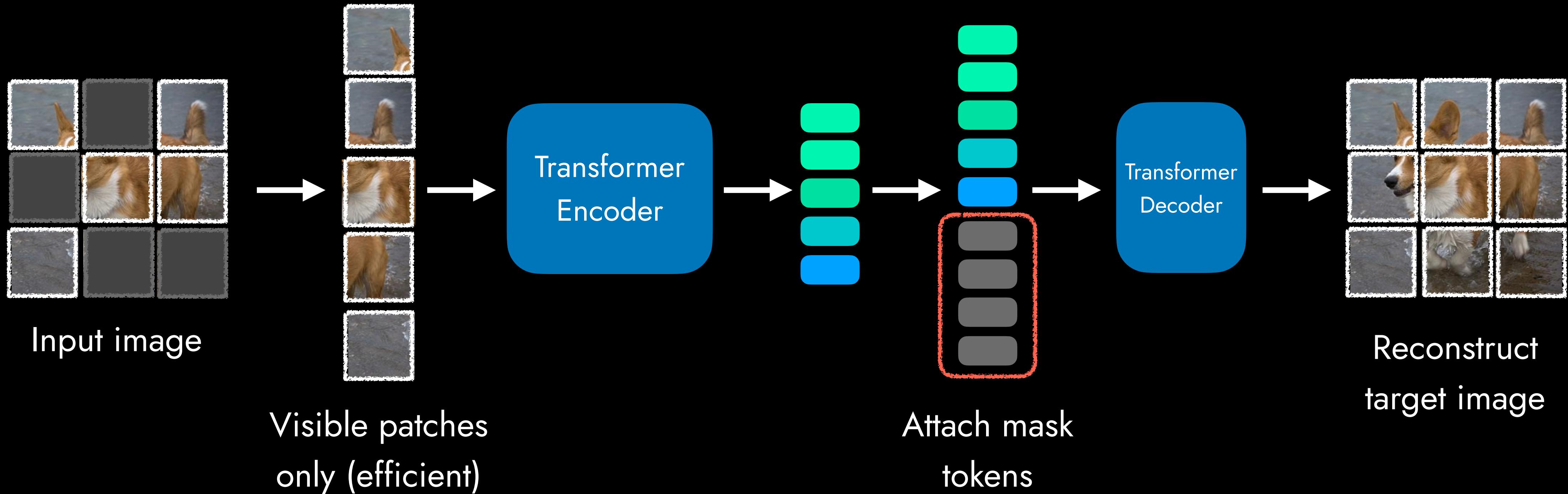


Pathak et al., "Context Encoders: Feature Learning by Inpainting", CVPR 2016.

Gidaris et al., "Unsupervised Representation Learning by Predicting Image Rotations", ICLR 2018.

Masked image modeling

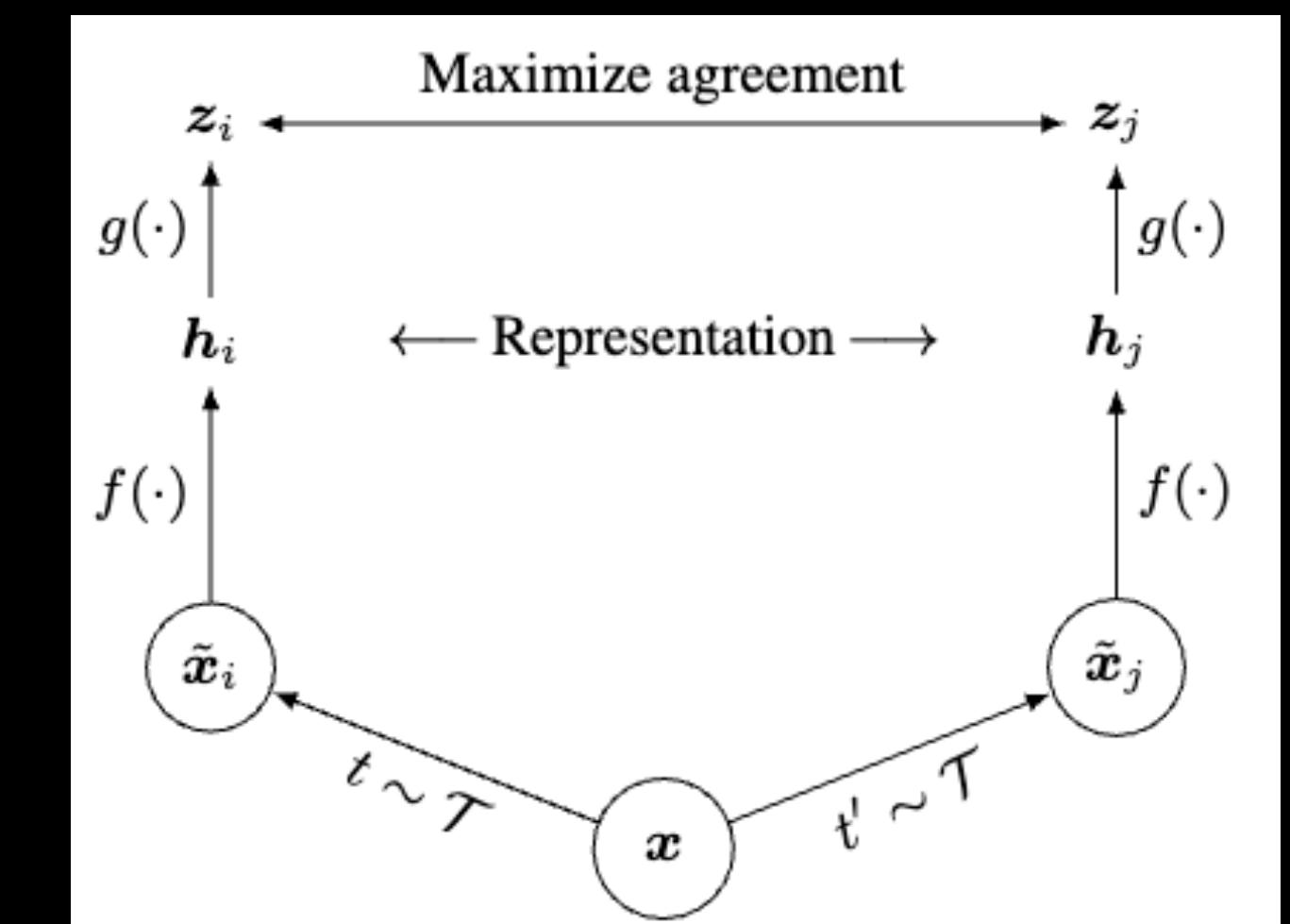
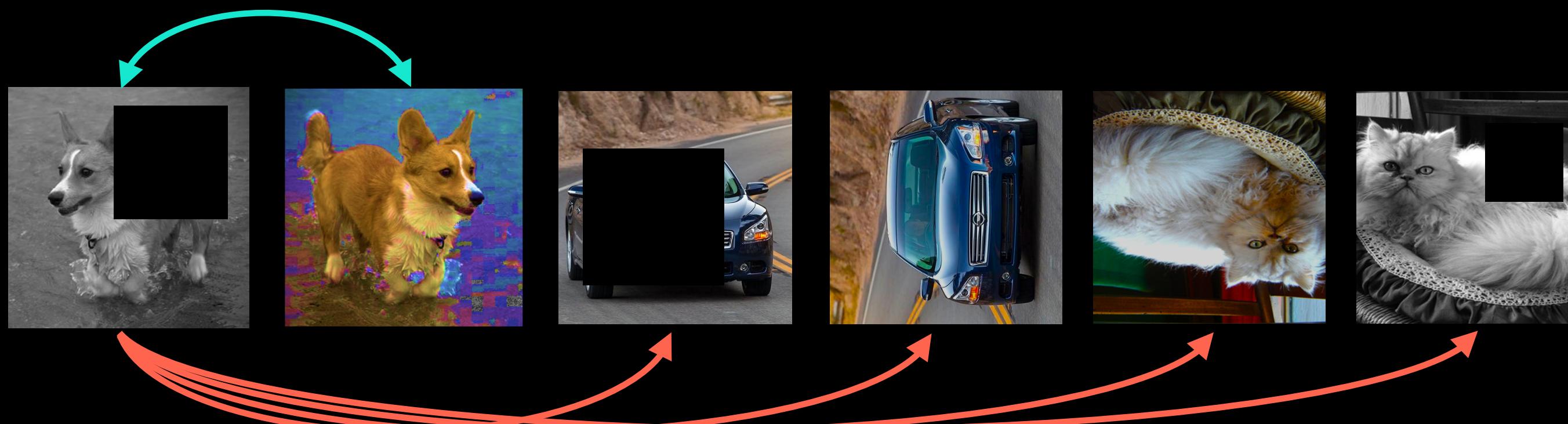
- Masked Auto-Encoder (MAE) (He et al., 2022)



Inter-sample prediction

- Beyond sample-wise prediction, learn the *relationship* between samples
- SimCLR
 - Given one positive pair, other $2(N-1)$ samples are negative

InfoNCE loss: $\mathcal{L}_{\text{SimCLR}}^{(i,j)} = - \log \frac{\exp(\text{sim}(\mathbf{z}_i, \mathbf{z}_j)/\tau)}{\sum_{k=1}^{2N} \mathbf{1}_{k \neq i} \exp(\text{sim}(\mathbf{z}_i, \mathbf{z}_k)/\tau)}$

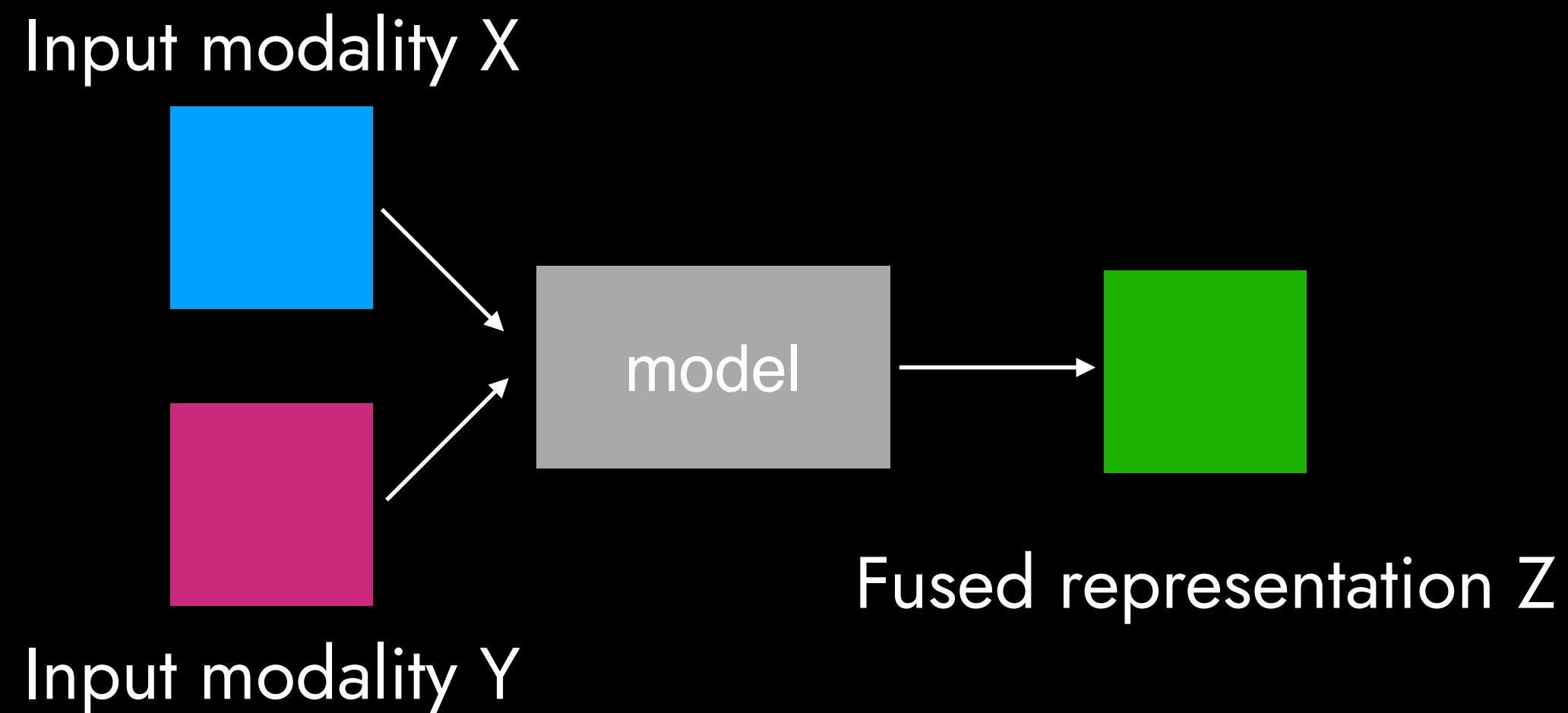


(Recap)

Multimodal pre-training

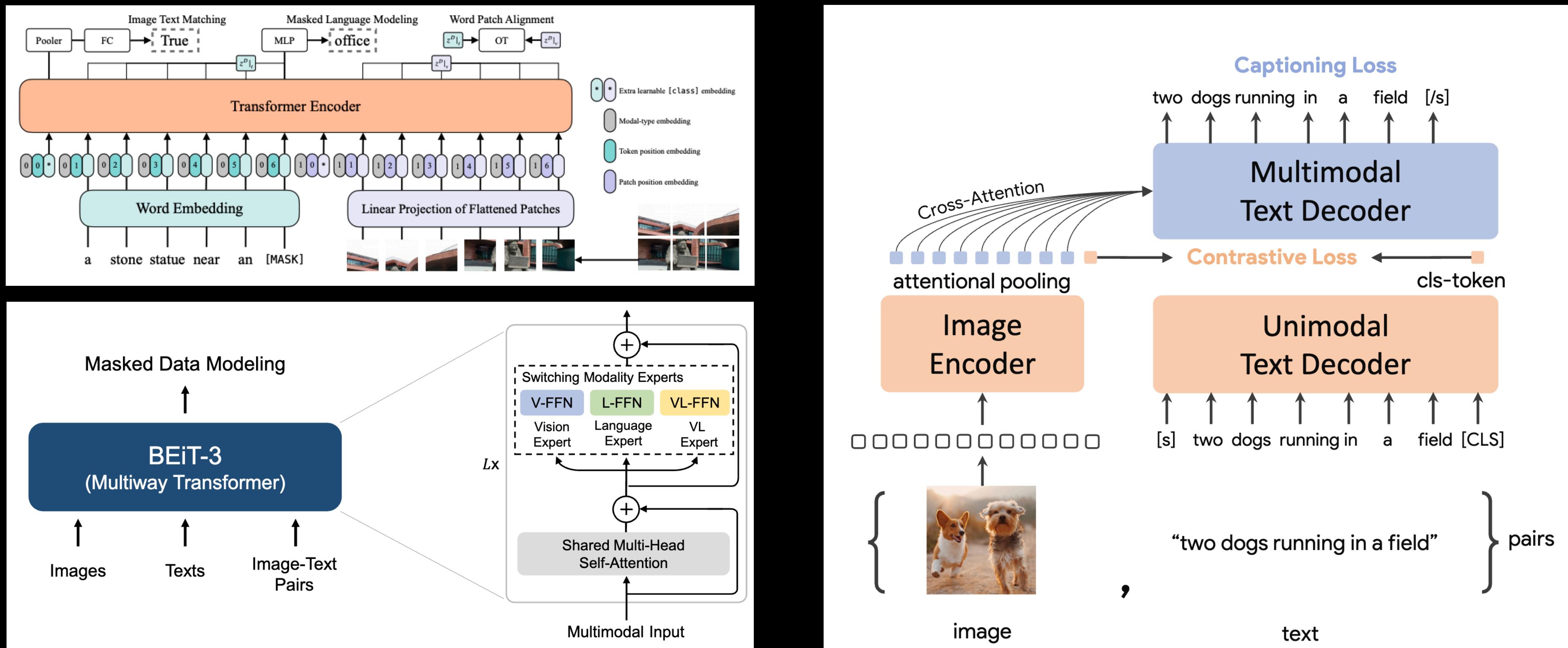
Multimodal representation

- Multimodal representation (recap. Lecture 3 & 4)
 - Fused representation



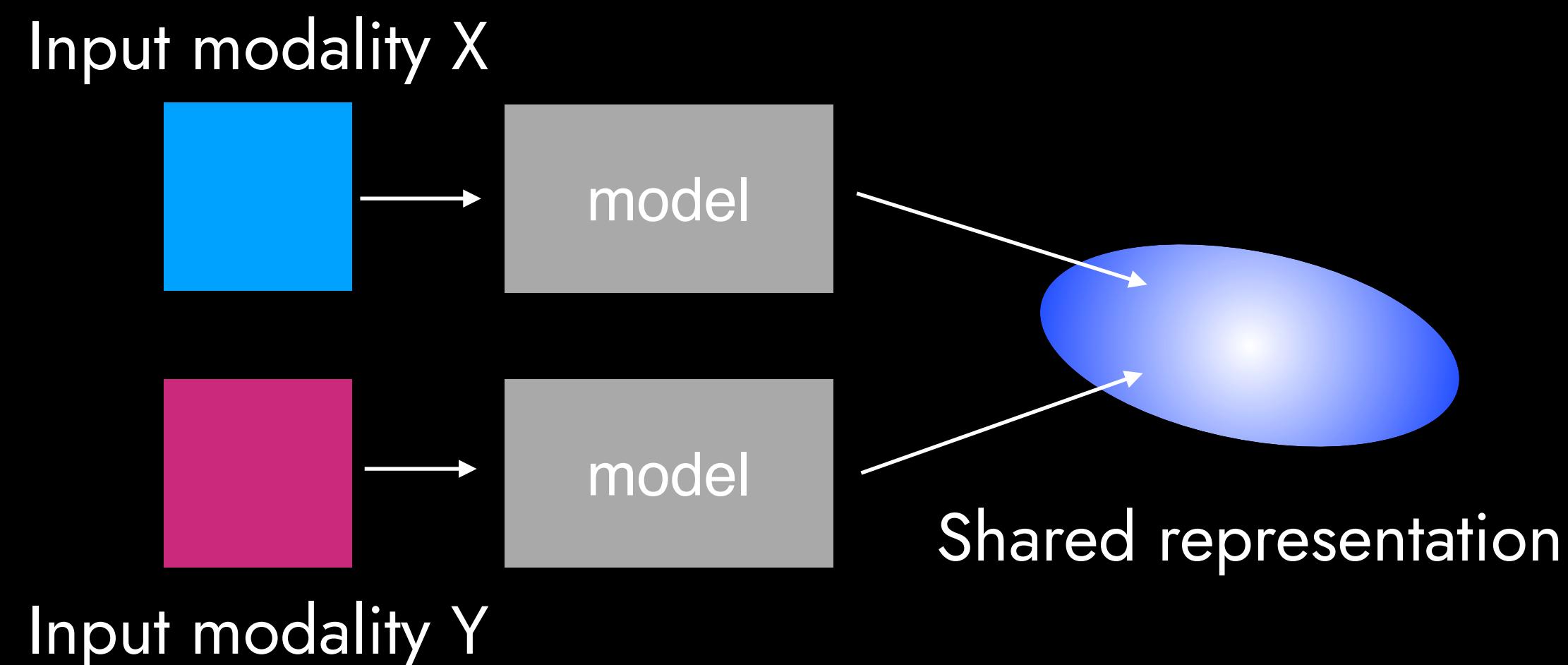
Multimodal representation

- Multimodal representation (recap. Lecture 3 & 4)
 - Fused representation



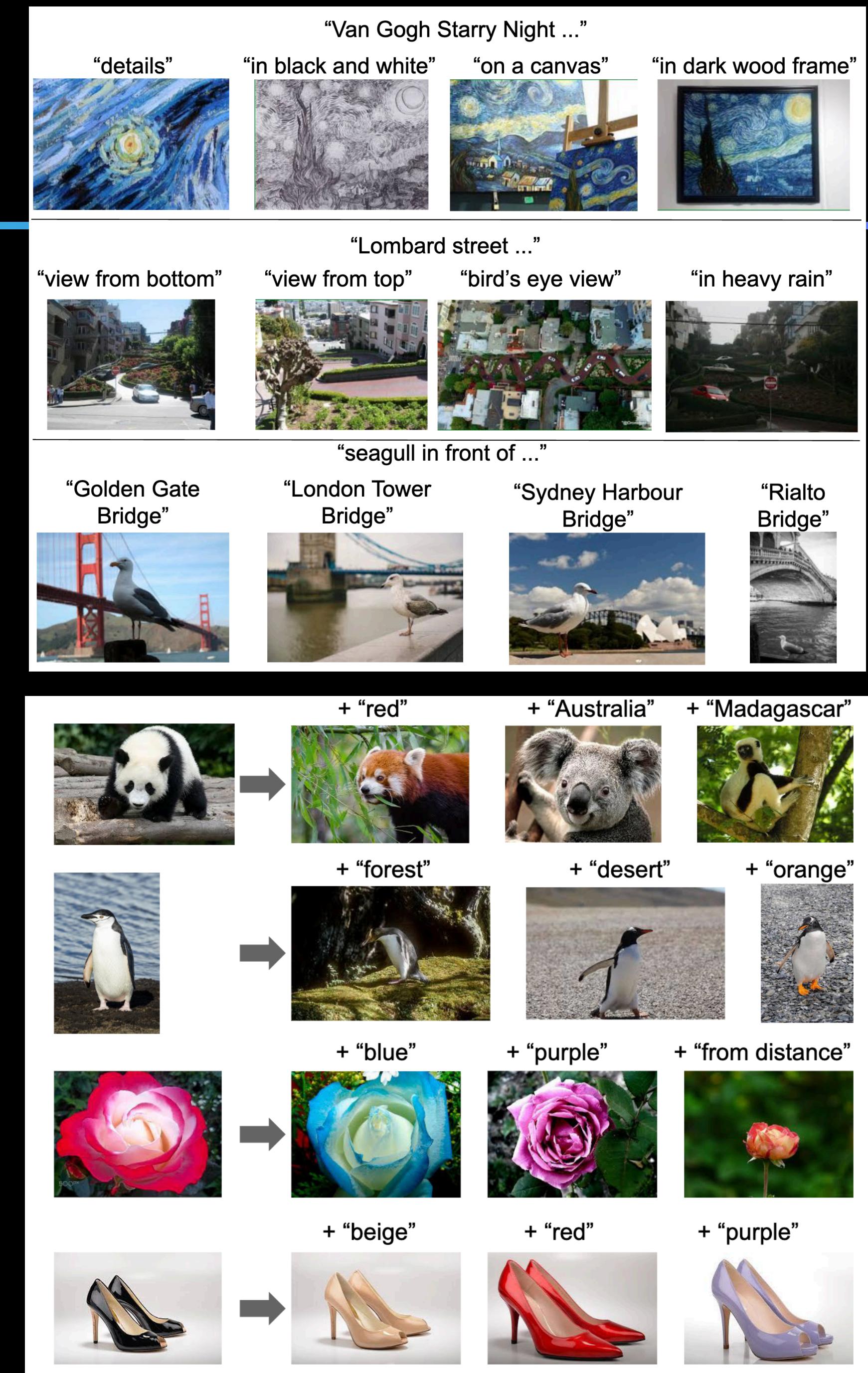
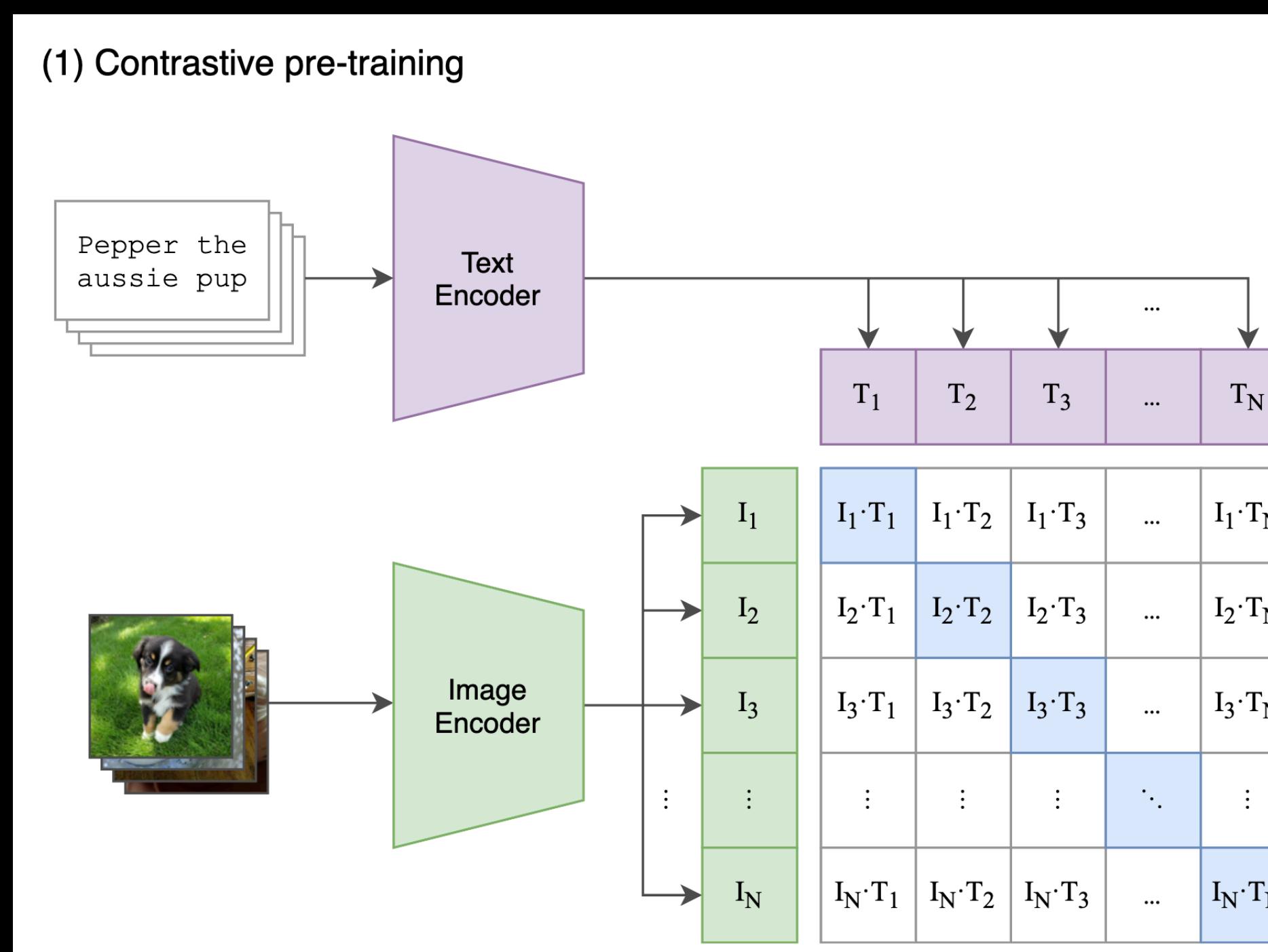
Multimodal representation

- Multimodal representation (recap. Lecture 3 & 4)
 - Shared representation



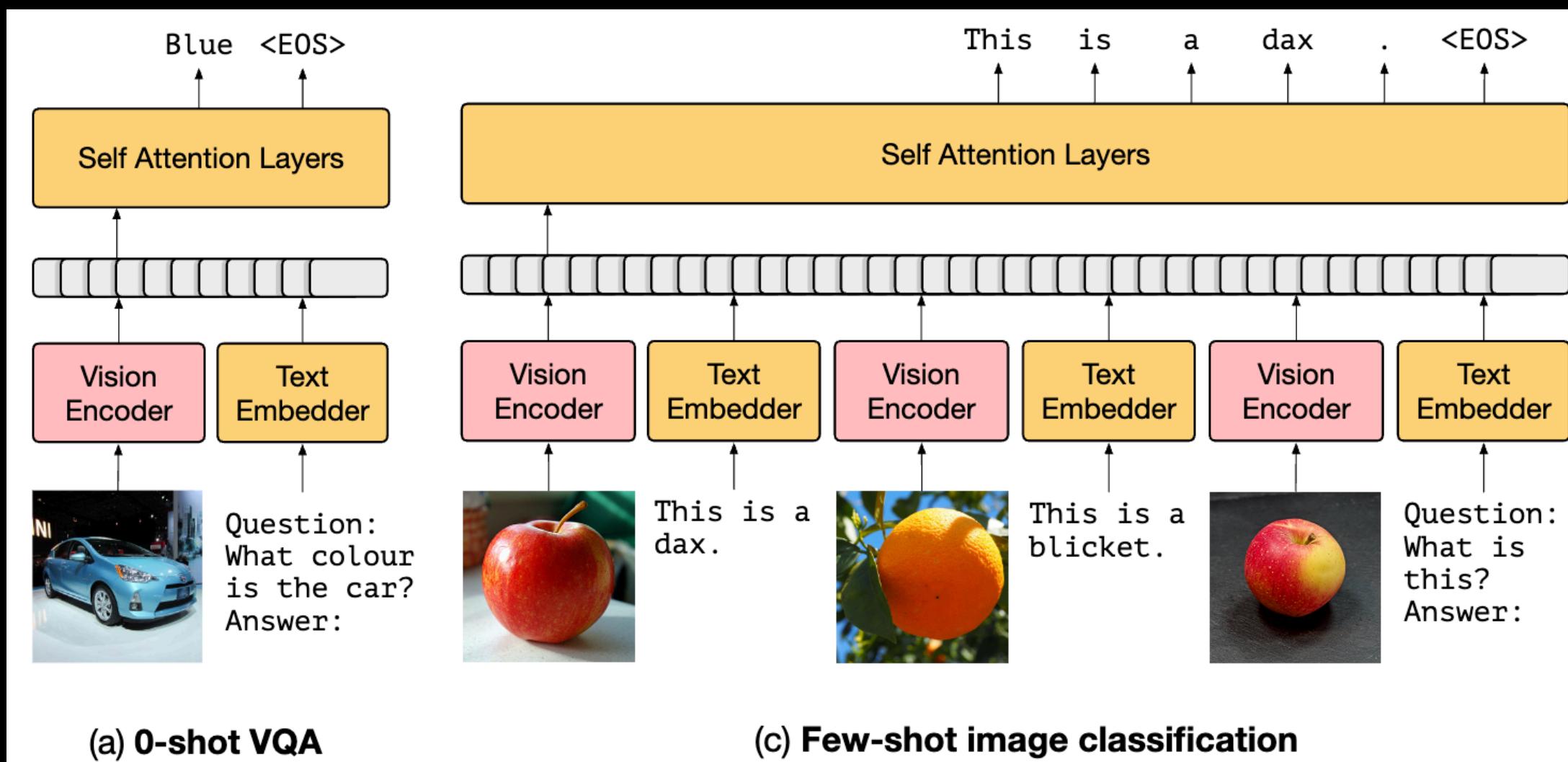
Multimodal representation

- Multimodal representation (recap. Lecture 3 & 4)
 - Shared representation



Multimodal representation

- Few-shot (zero-shot) scenarios (recap. Lecture 4)



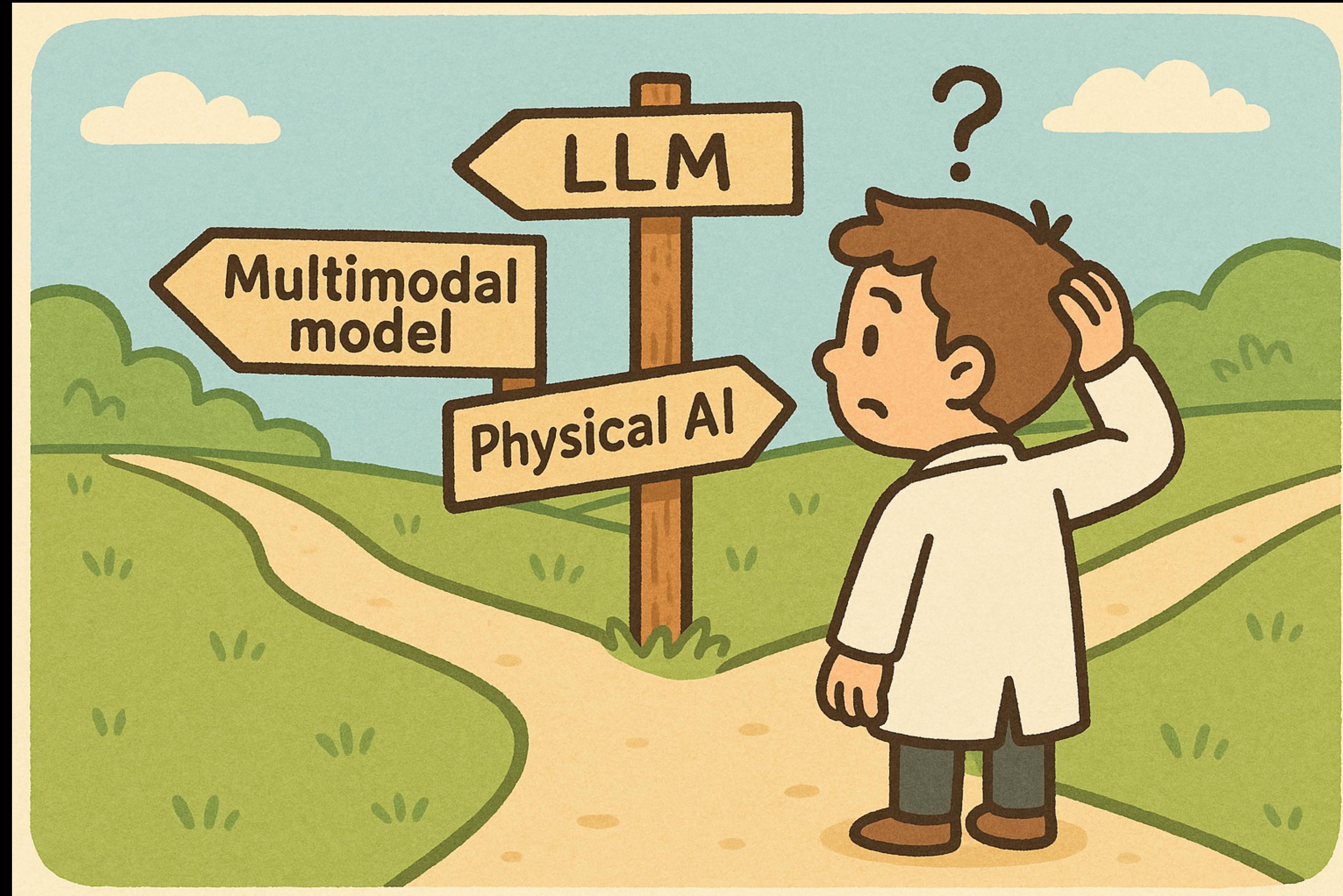
Where they are going?

- Unified architecture/objectives
- Scaling-up models
- Scaling-up data

Multimodal Foundation Model

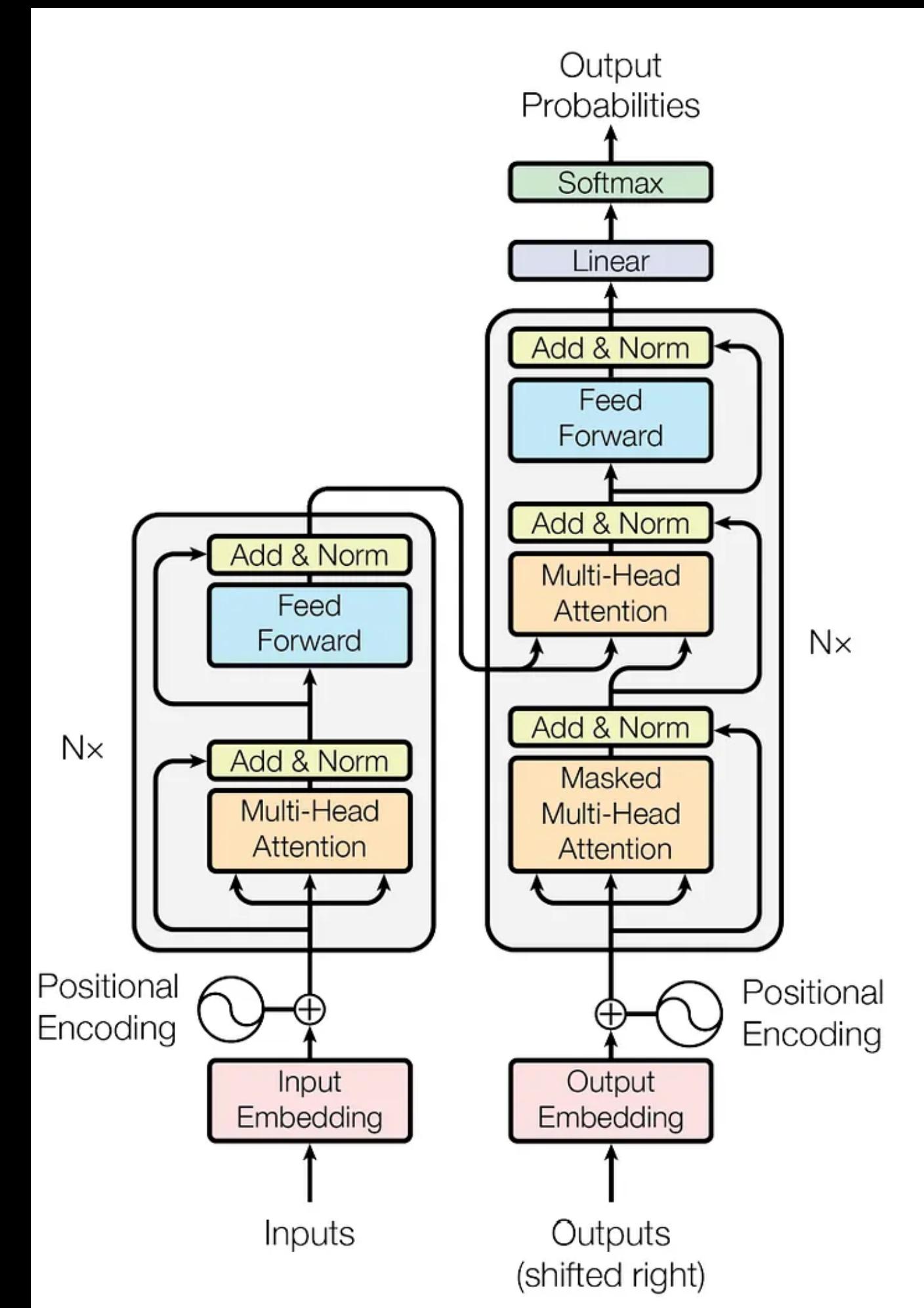
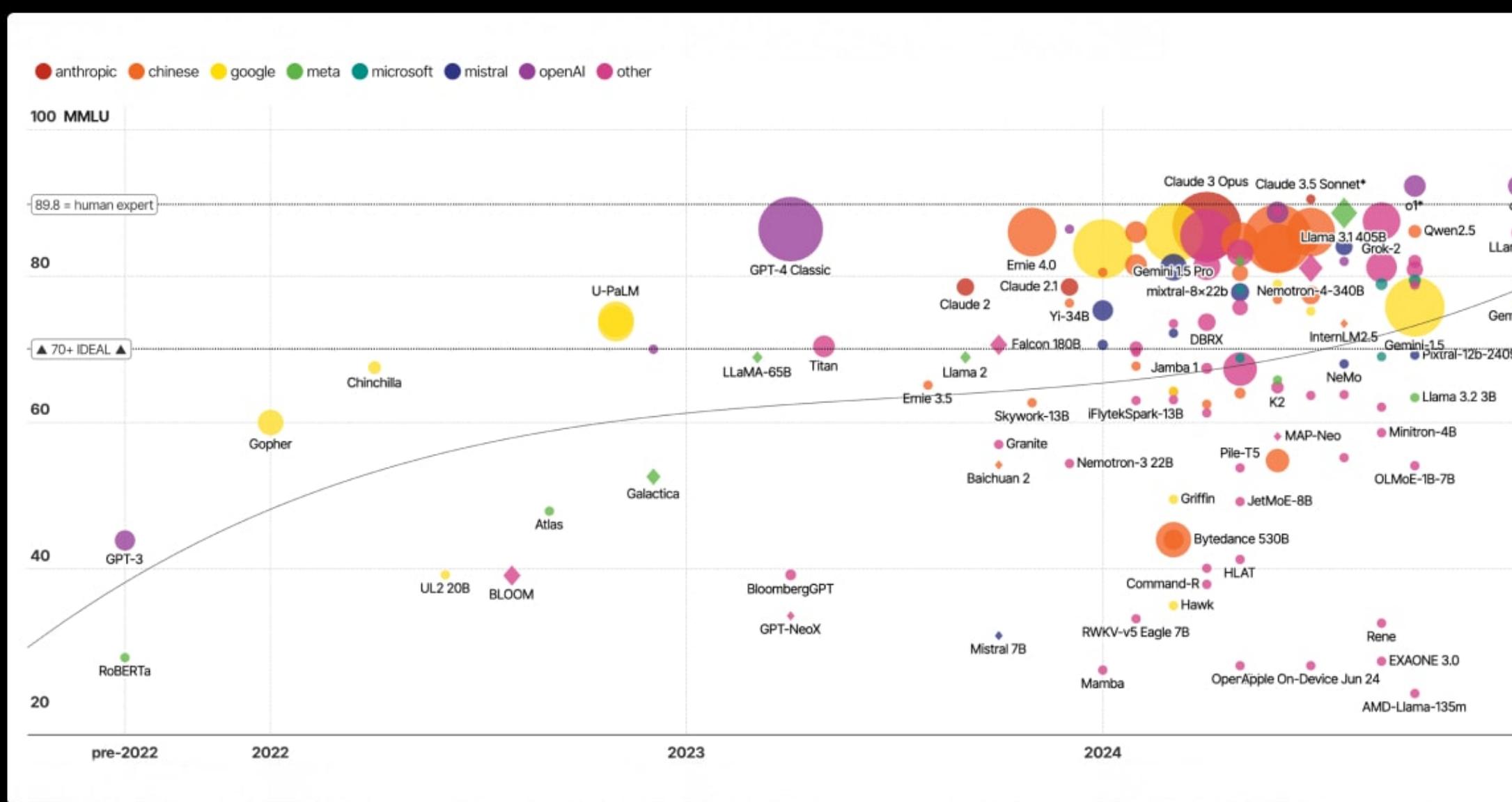
Cutting edges

Where are we going?



Scaling is all you need?

- More parameters
 - AlexNet, ResNet, Transformers, ...
- More data
 - ImageNet, Laion, Datacomp, massive web data, ...



Are we done with AI?



Pre-training as we know it will end

Compute is growing:

- Better hardware
- Better algorithms
- Larger clusters

Data is not growing:

- We have but one internet
- The fossil fuel of AI

Internet. We have, but one Internet. You could even say you can even go as far as to
That data is the fossil fuel of AI. It was like, created somehow. And now we use it.

- Probably, No.
- We still need,
 - *Efficient algorithms*
 - *Multi-modality*
 - *Reliability*

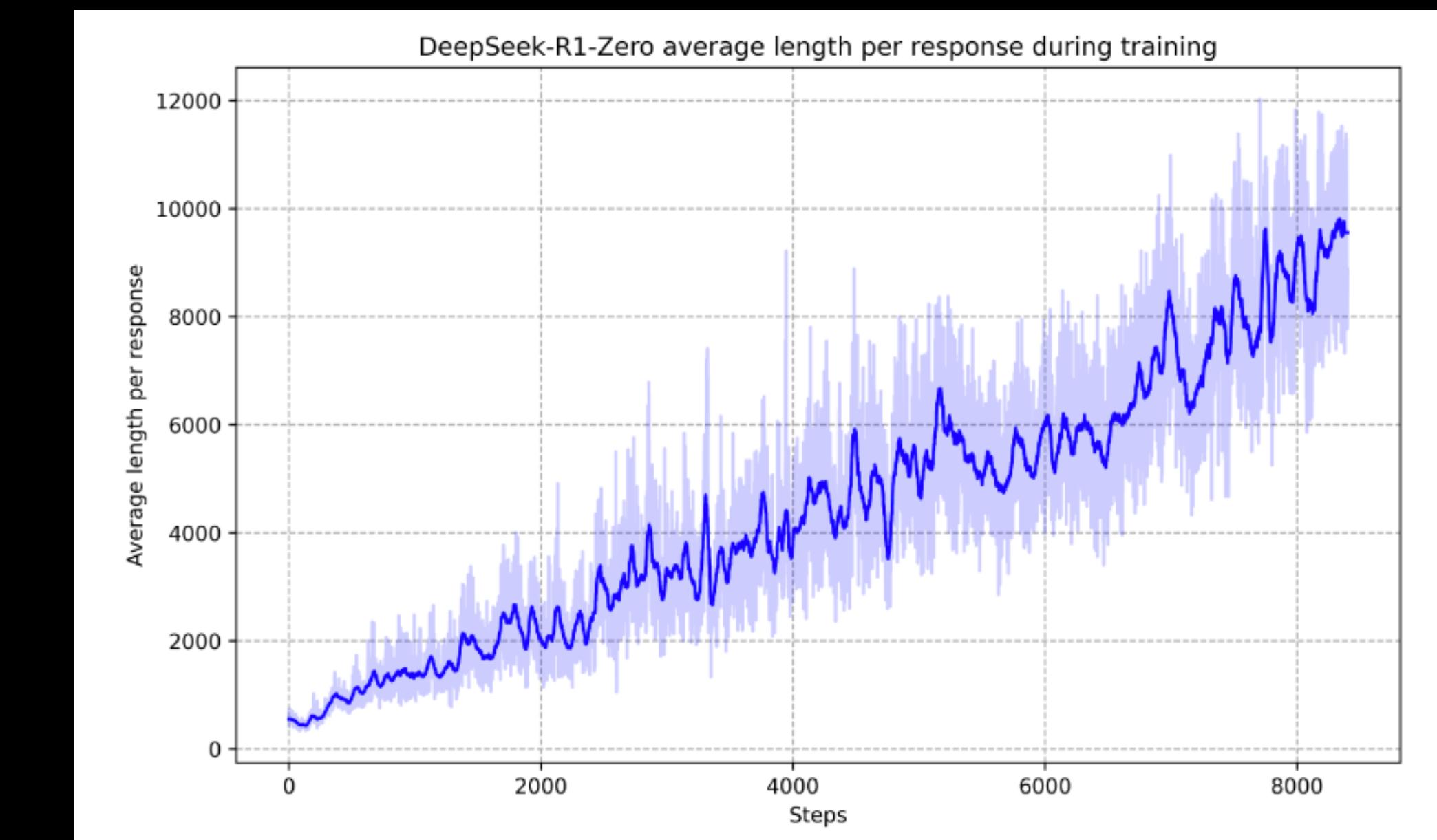
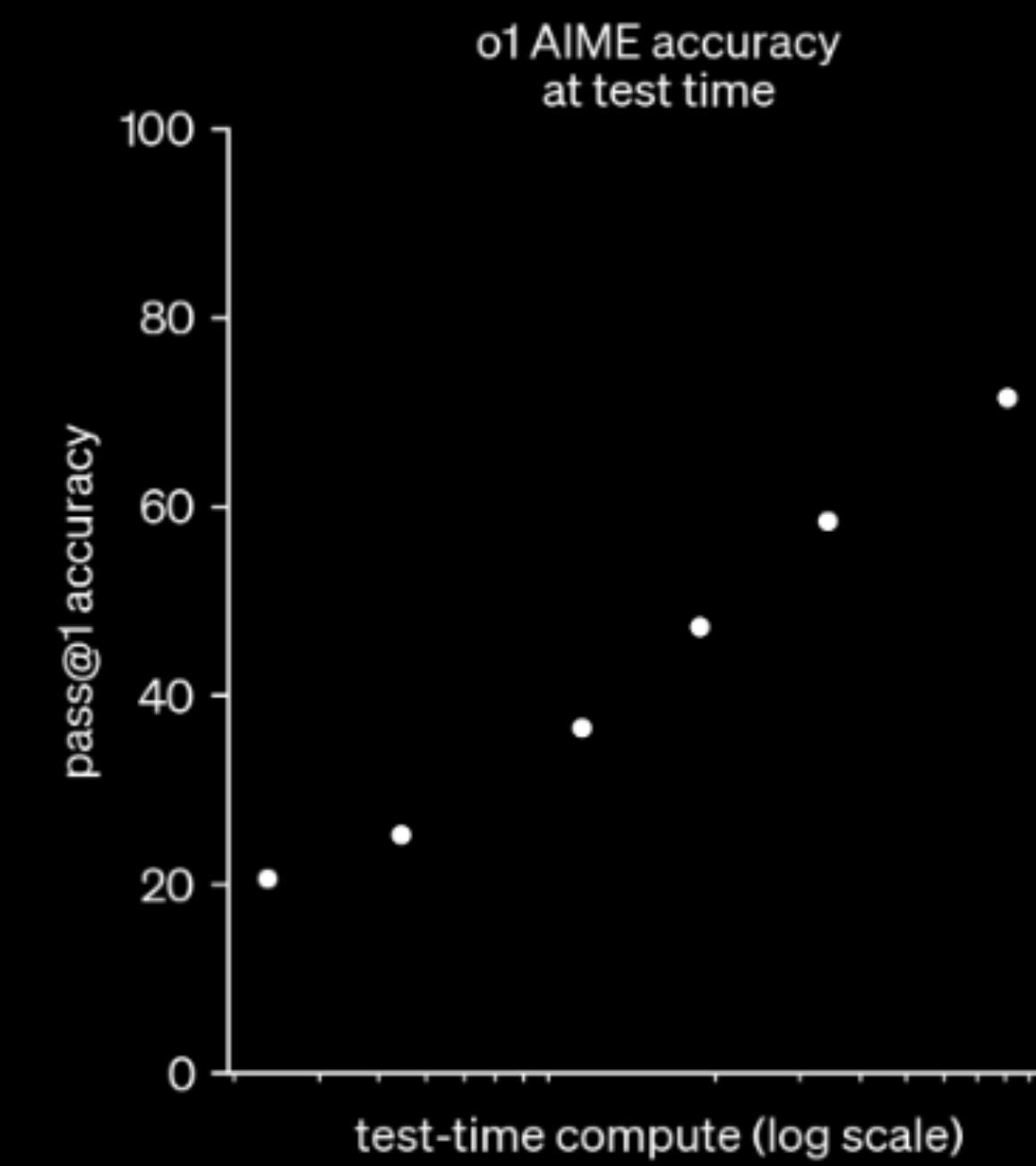
Topics

- Reasoning models
- Vision-language-action models
- Omni (any-to-any) models

Reasoning models

Test-time scaling

- Reasoning models: clear trend with increased test-time computation



Test-time scaling

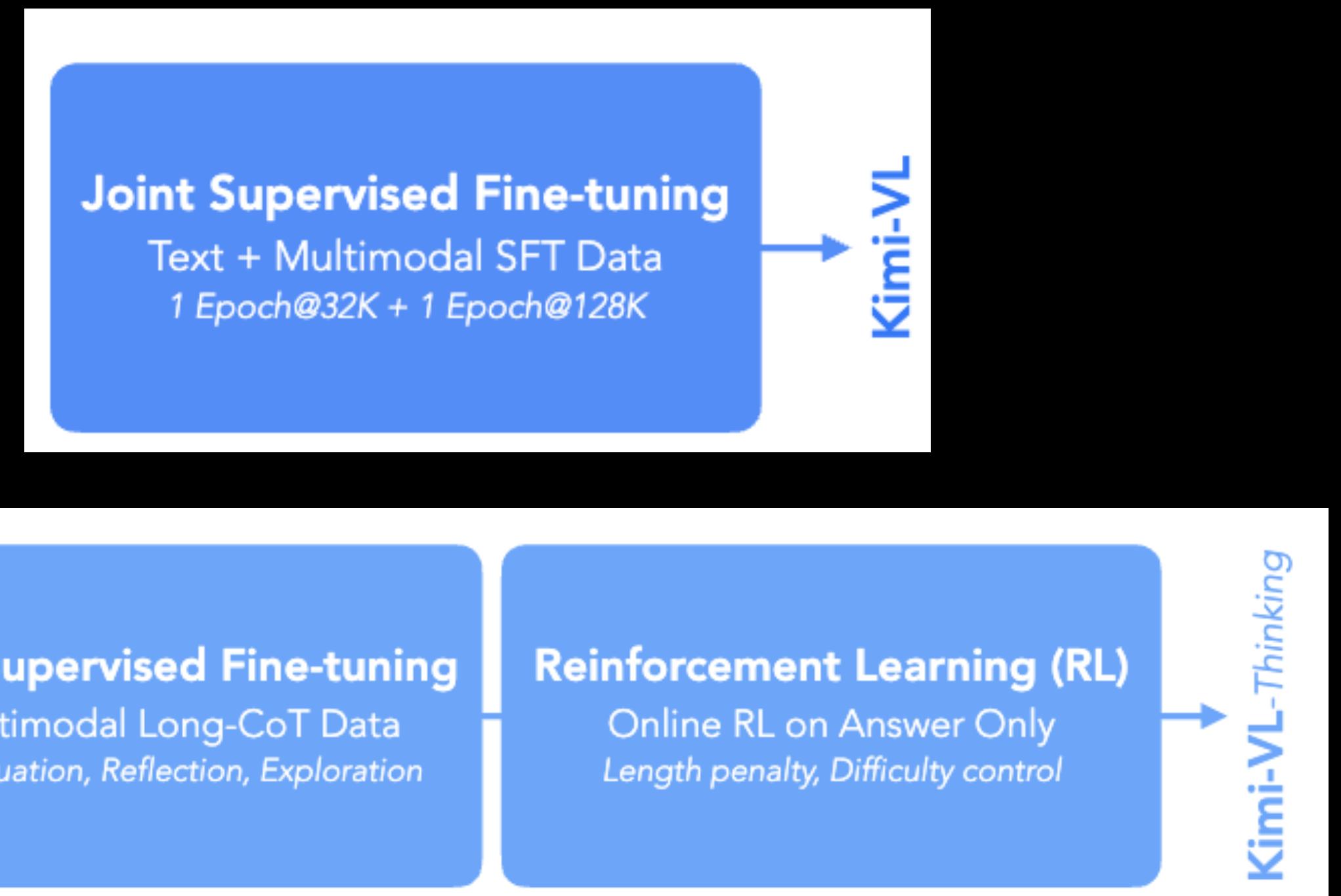
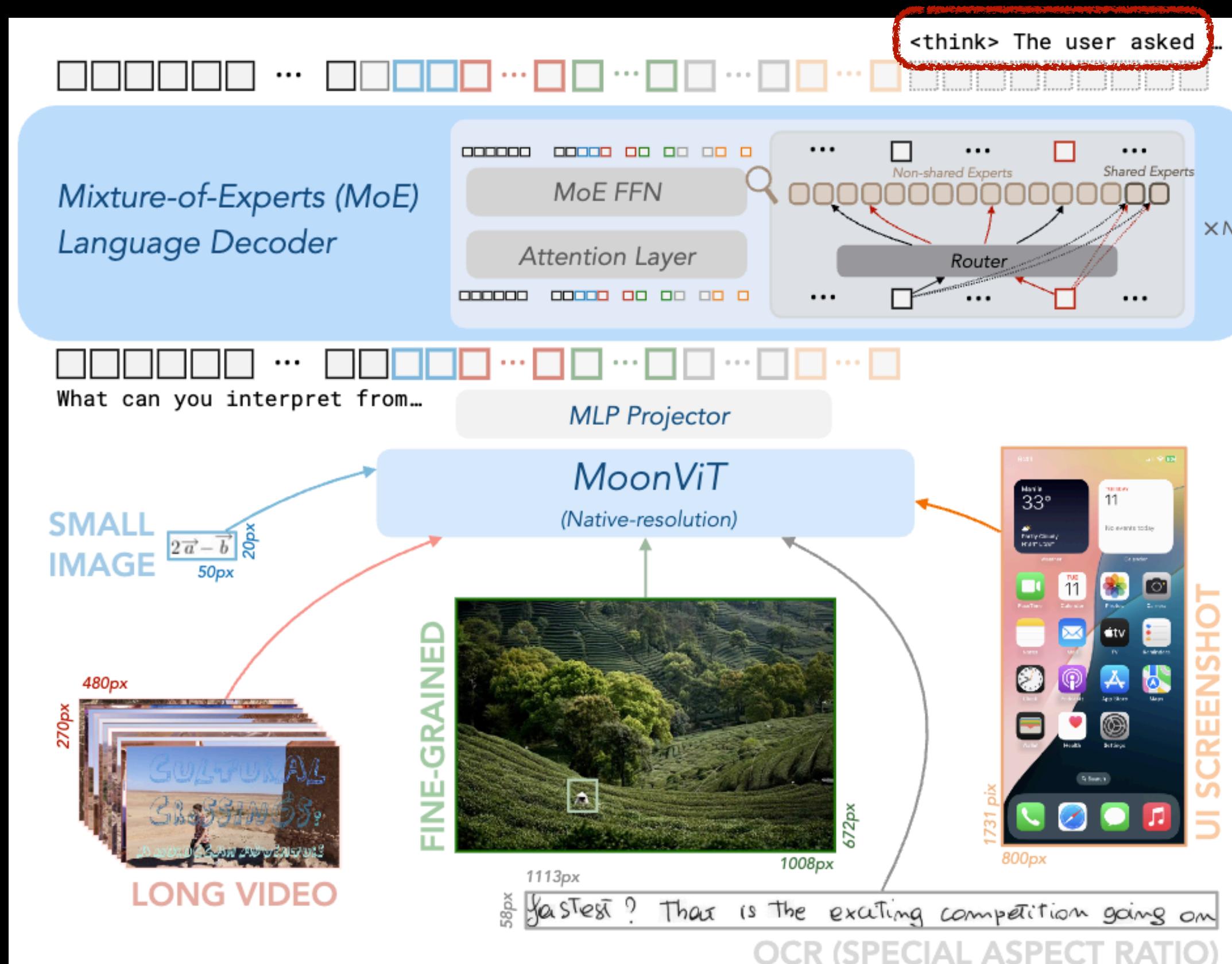
- Reasoning models show impressive performance

RESEARCH

Advanced version of Gemini with Deep Think officially achieves gold-medal standard at the International Mathematical Olympiad

21 JULY 2025

Test-time scaling in VLMs (Kimi-VL)



Leverage strong test-reasoning models to synthesize multimodal CoT data

Test-time scaling in VLMs (Kimi-VL)

Instruction

Which of the four subgraphs in the figure is taken at the same place as subgraph 1?

Answer: The image 4 is taken at the same place as subgraph 1.

Response

The first picture (image 1) depicts a dense urban area with a mix of buildings, green spaces, and a distinctive circular structure (possibly a dome or observatory). To determine which image was taken at the same place as subgraph 1, I'll analyze the other three images:

- Image 4 shows a dense urban area with a mix of buildings and green spaces, similar to image 1.
- Both image 3 and image 4 show a large building complex with a prominent dome, matching the circular structure in image 1.
- Image 2 is taken from a higher altitude, showing a broader view of the urban setting.
- Conclusion: Both image 3 and image 4 feature the same building complex with a dome, making them potential matches for subgraph 1.

Instruction

Please infer step by step who this manuscript belongs to and what it records

Response

<Think> Alright, I need to figure out who this manuscript belongs to and what it records.

First, I'll look at the visual characteristics of the manuscript. The handwriting is cursive, and the paper looks old and has a yellowish tint, indicating it might be from a previous century. The content appears to be handwritten notes, possibly from the 20th century.

Next, I'll examine the content of the manuscripts. Both images contain mathematical equations. The equations seem to be related to gravitational fields or similar physical phenomena.

Looking at the first image (Image 1), I can see equations involving "Einheitsvektor" (unit vector in German) suggests this might be related to a specific problem in physics.

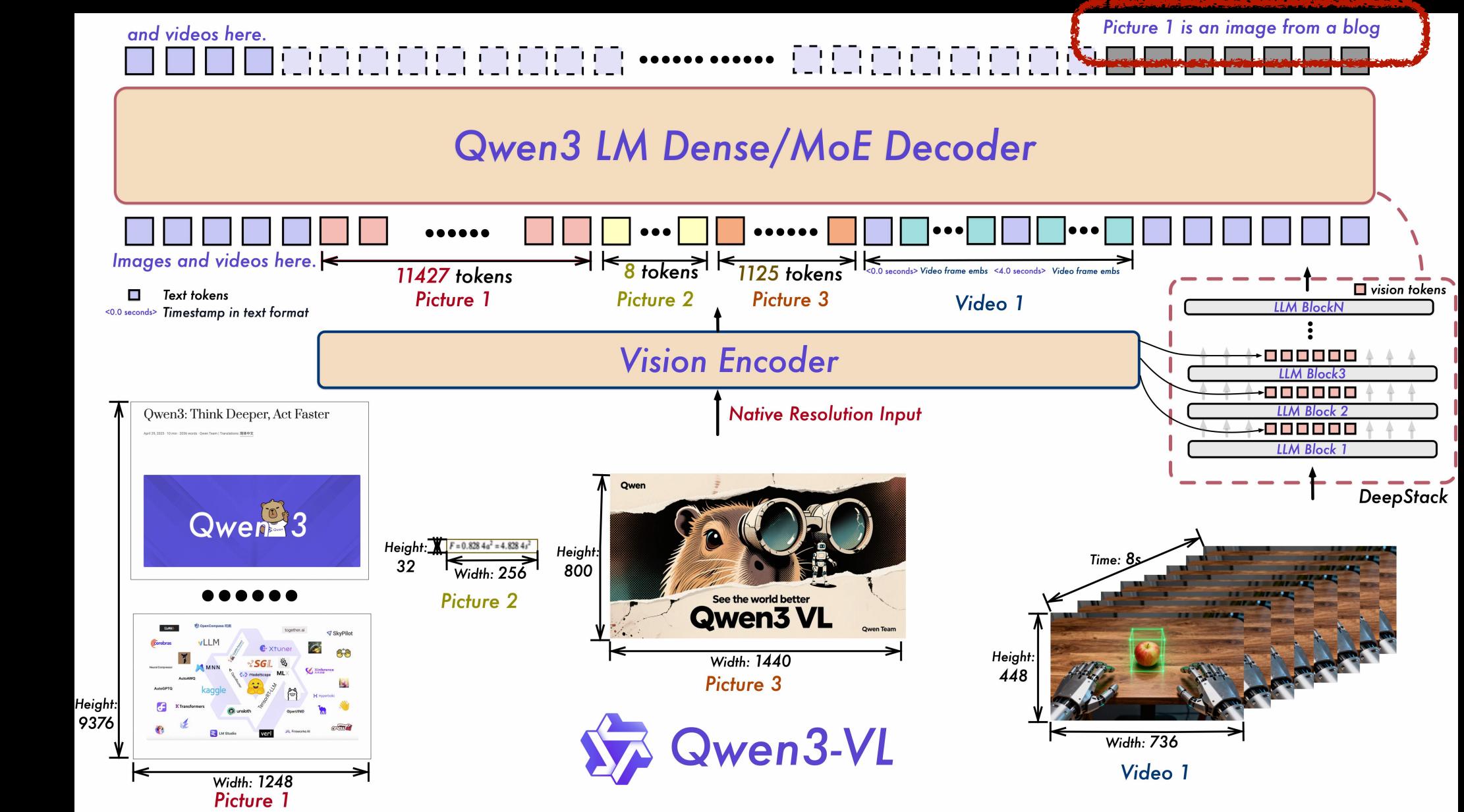
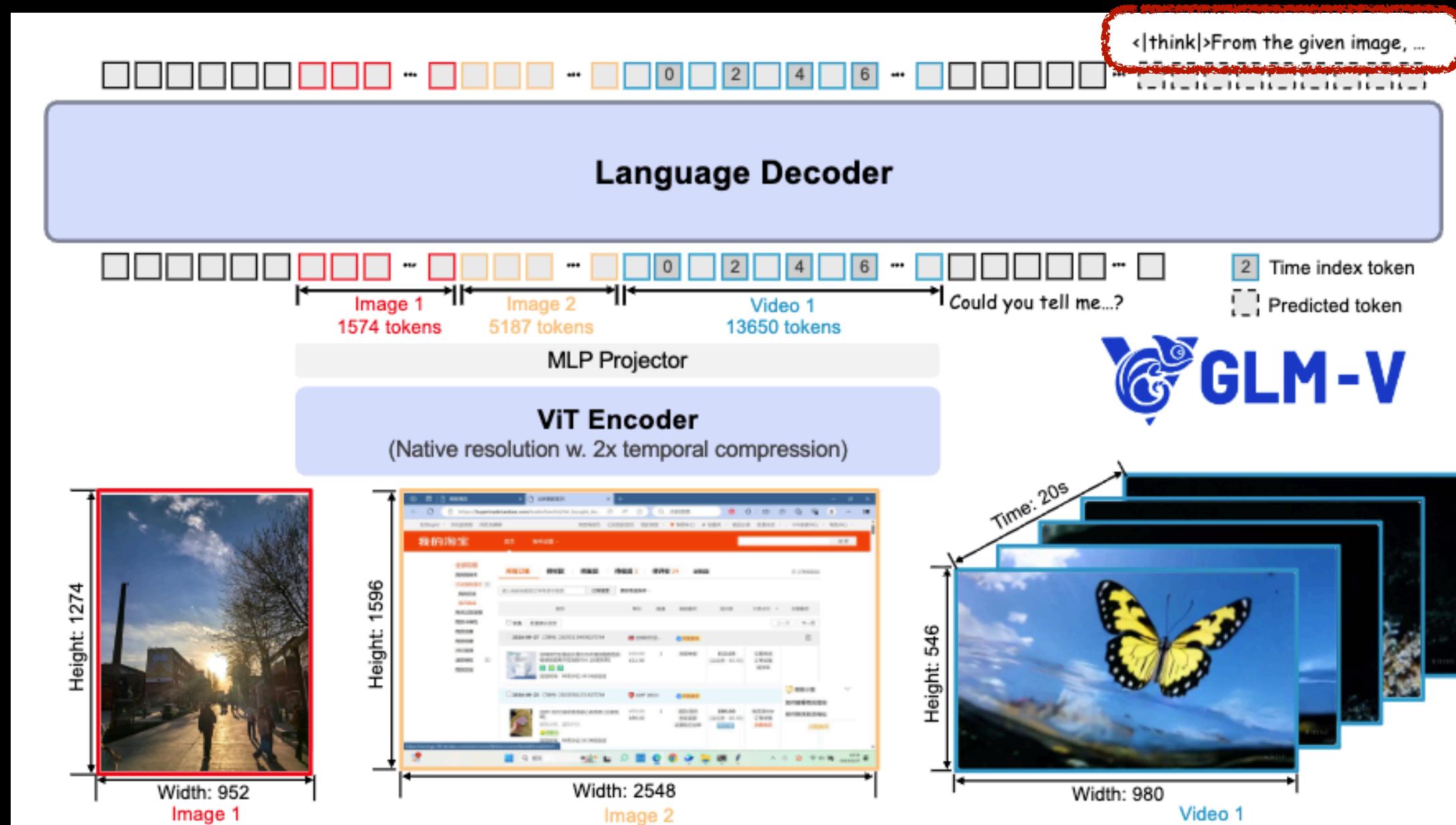
The second image (Image 2) contains more numerical calculations and algebraic manipulations. The equations include integrals, derivatives, and algebraic manipulations.

Now, considering the writing style and the content, I notice that the "Einheitsvektor" and "Gln" (which could be an abbreviation for "Gleichung") are common terms in physics, particularly in the context of relativity.

The first image seems to be a page from a mathematical manuscript, likely related to relativity, particularly in the context of gravitational fields. The second image is likely a different section of the same manuscript.

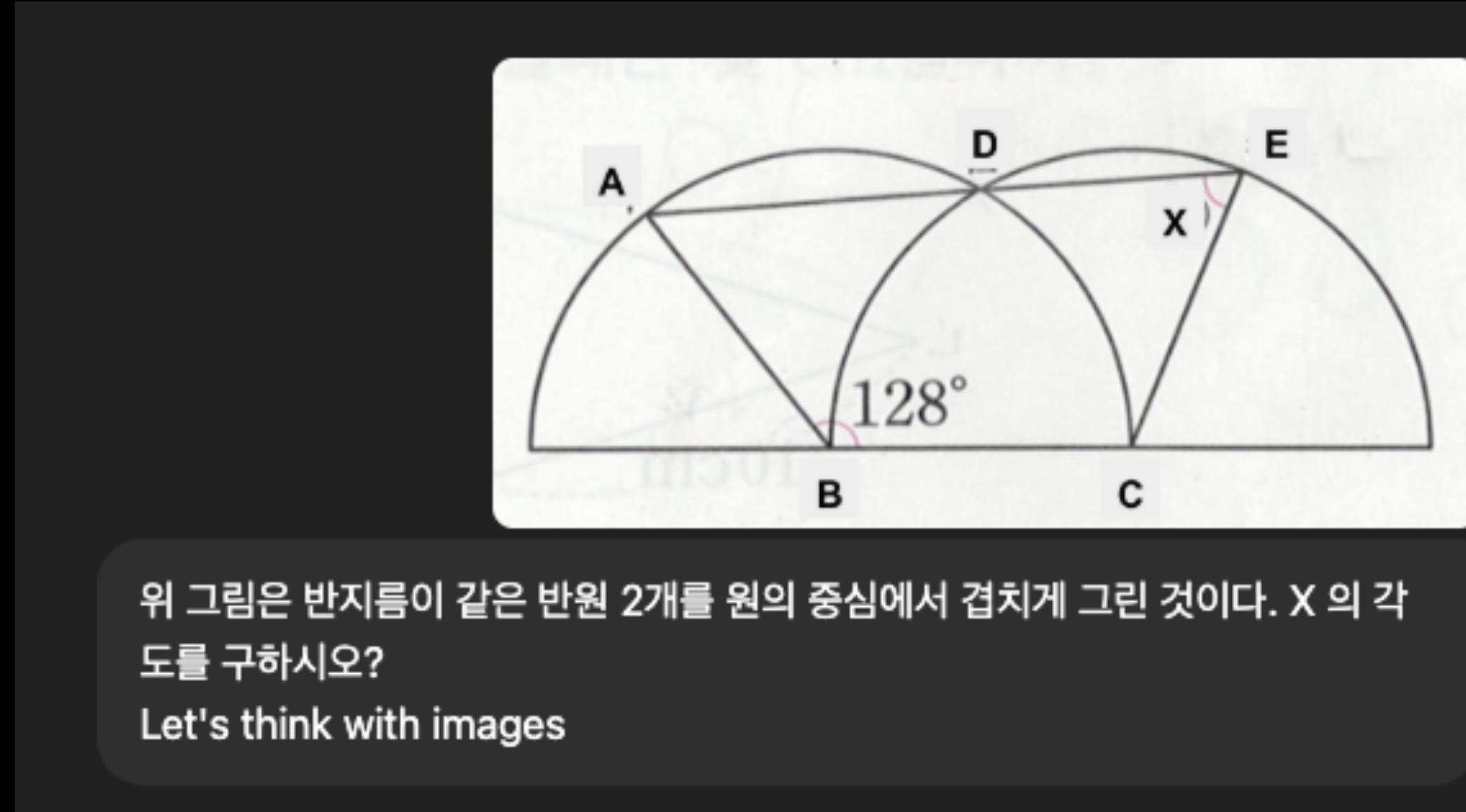
VLM + Thinking

- The same paradigm for multimodal thinking models 🤔🤔



Can they solve this?

ChatGPT-5-thinking



생각 중

문제의 기하학적 분석

```
from PIL import Image, ImageOps, Image
img=Image.open('/mnt/data/fd696dbb-fc2f-43e0-8333-1a2a2a2a2a2a.jpg')
img.size
(478, 226)
```

- Considering geometric distances
I'm trying to figure out the intersection of two circles, each with radius r , centered at different midpoints along a line segment like AB and BC. The calculation of the separation distance between centers seems to be $2r$, and I'm working through how the midpoints line up.
- Reassessing semicircle intersections
So, if the semicircle radius is R , then the diameter is $2R$, and $AB=BC=2R$. The centers of the circles are separated by $2R$, which means they just touch at point B. But wait—these are full circles, not semicircles, and they meet at exactly two points. Hmm... I need to clarify whether the arcs are overlapping correctly.
- Clarifying the arc intersection
Okay, I think I need to refine the interpretation here. The phrase about overlapping at centers suggests that the top arc of one semicircle passes through the other's center. If the centers are at the midpoints of AB and BC, maybe the intersection is happening at point B. I need to check the distances and radius assumptions again.
- Interpreting the diagram
It looks like the centers of the semicircles, O₁ and O₂, are at the base points. The intersection seems to occur near the central vertical of the diagram, with an angle of 128° marking a key spot near point B. I'm thinking of using properties of inscribed angles, where an angle subtended by a diameter is 90°.

After 5min thinking...
wrong answer +
wrong chain-of-thought

Why it failed when thinking with images?

- Let's think 🤔
- Reasoning (thinking) is mainly handled by text
- Text reasoning models are pre-trained with reasoning-rich text data (e.g., QAs, debates, ...)
- Multimodal models (relying on LLMs) do not have such visual-thinking-ish pre-training data.
- Multimodal models lack visual generation ability

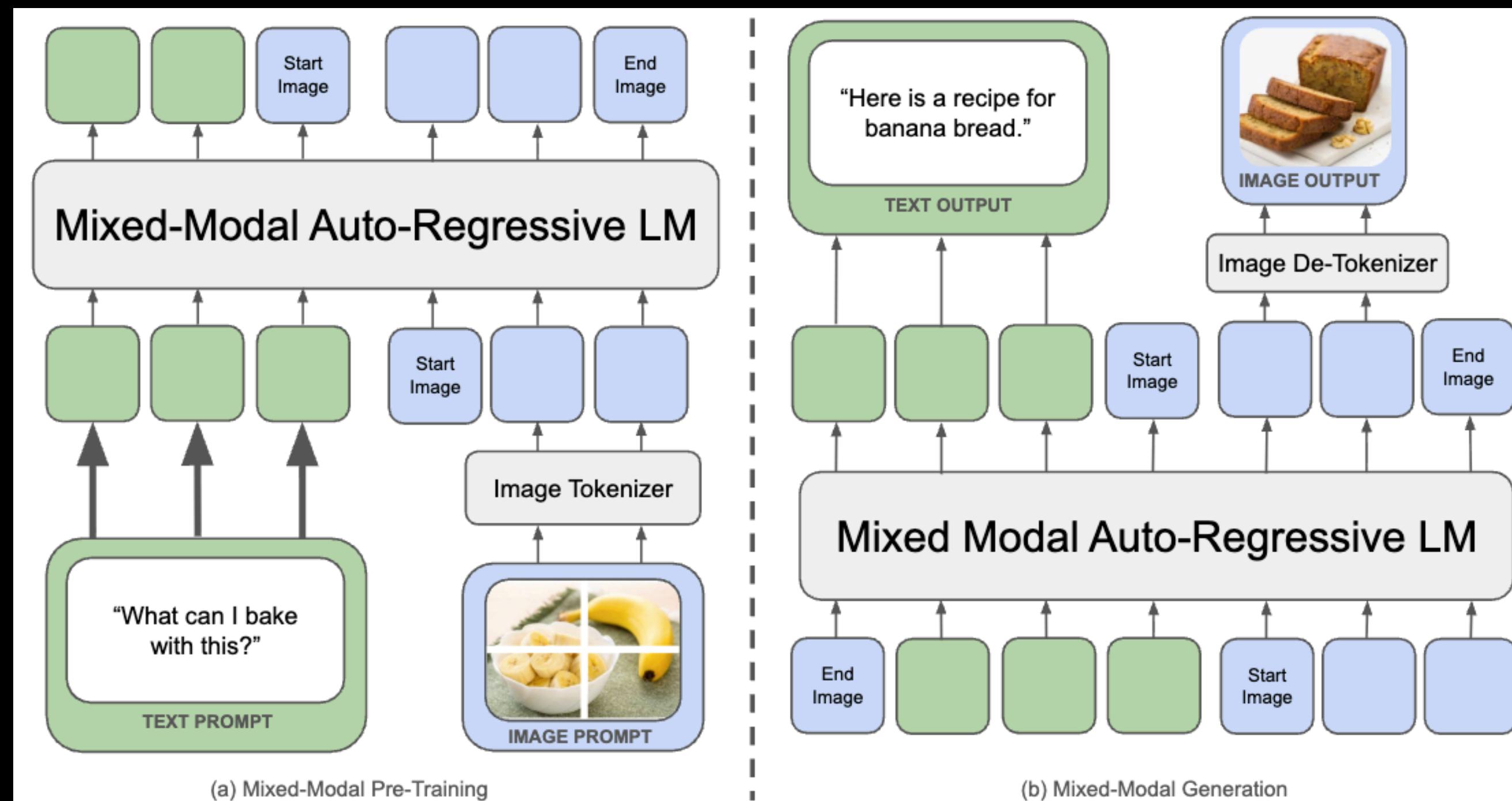
*Omni models
with unified architecture*

Omni-models

- Any-to-any model
- Emerging Trends from 2024

Chameleon (2024.05)

- Chameleon: Mixed-Modal Early-Fusion Foundation Models (Meta)

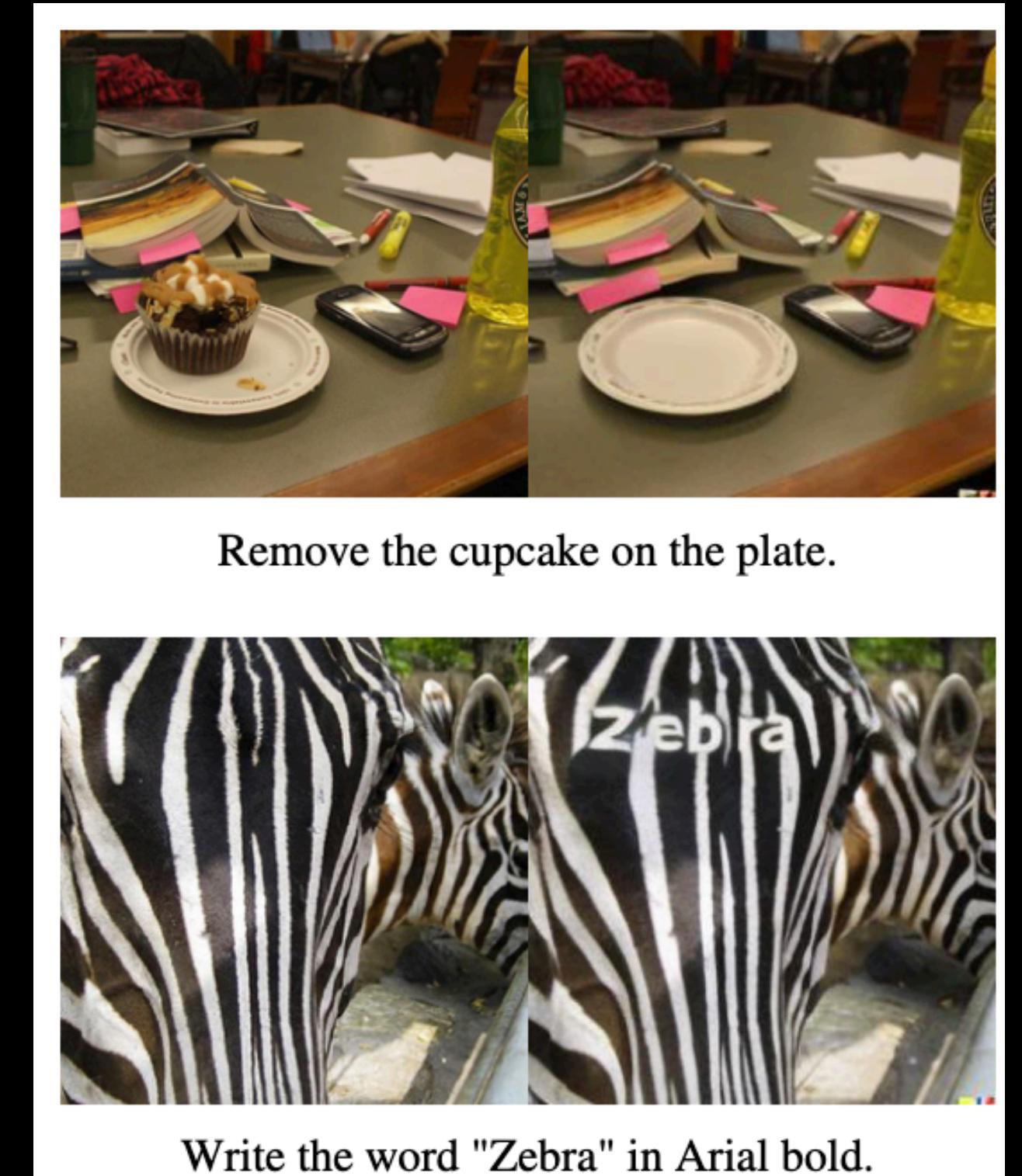
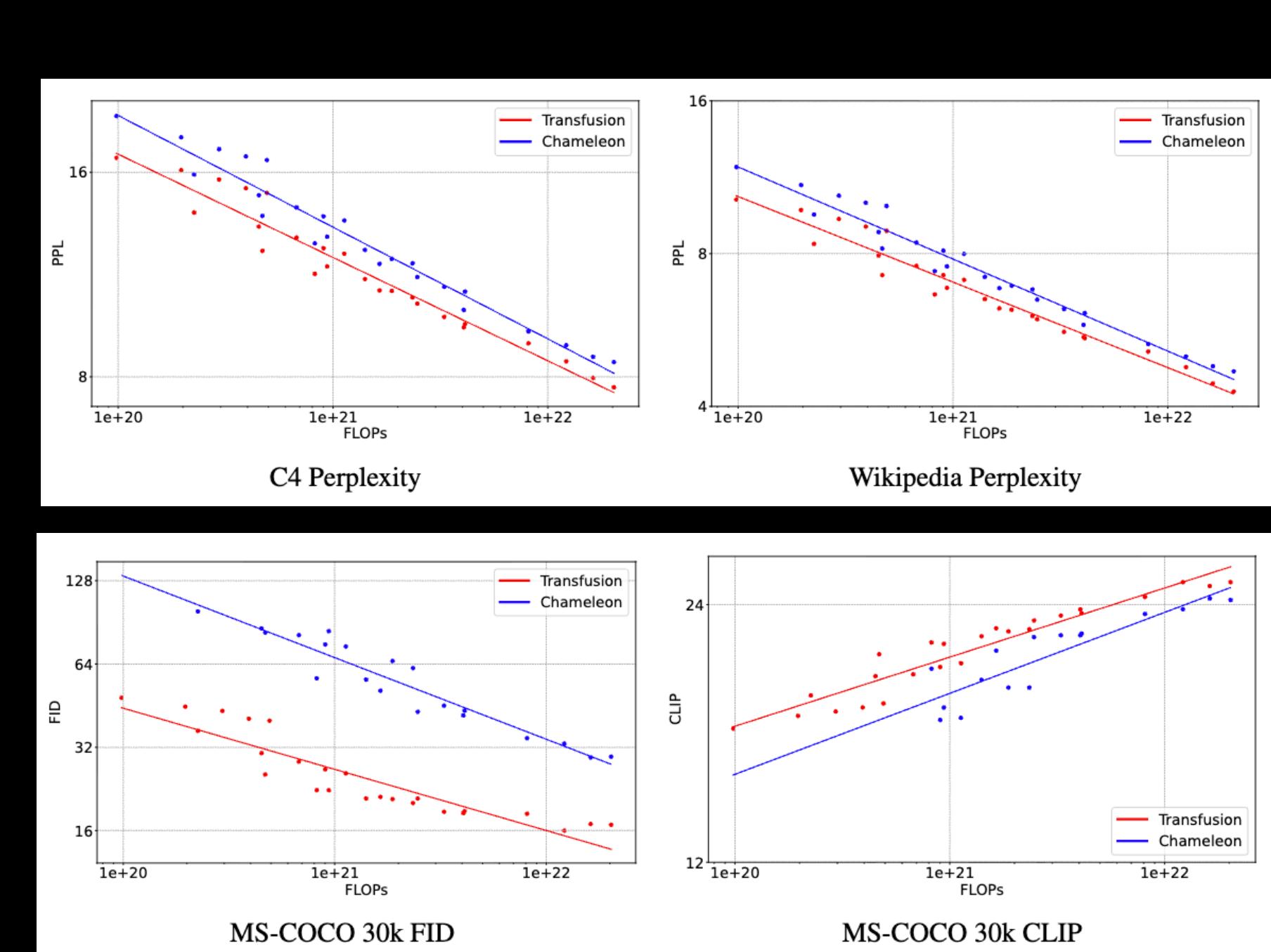
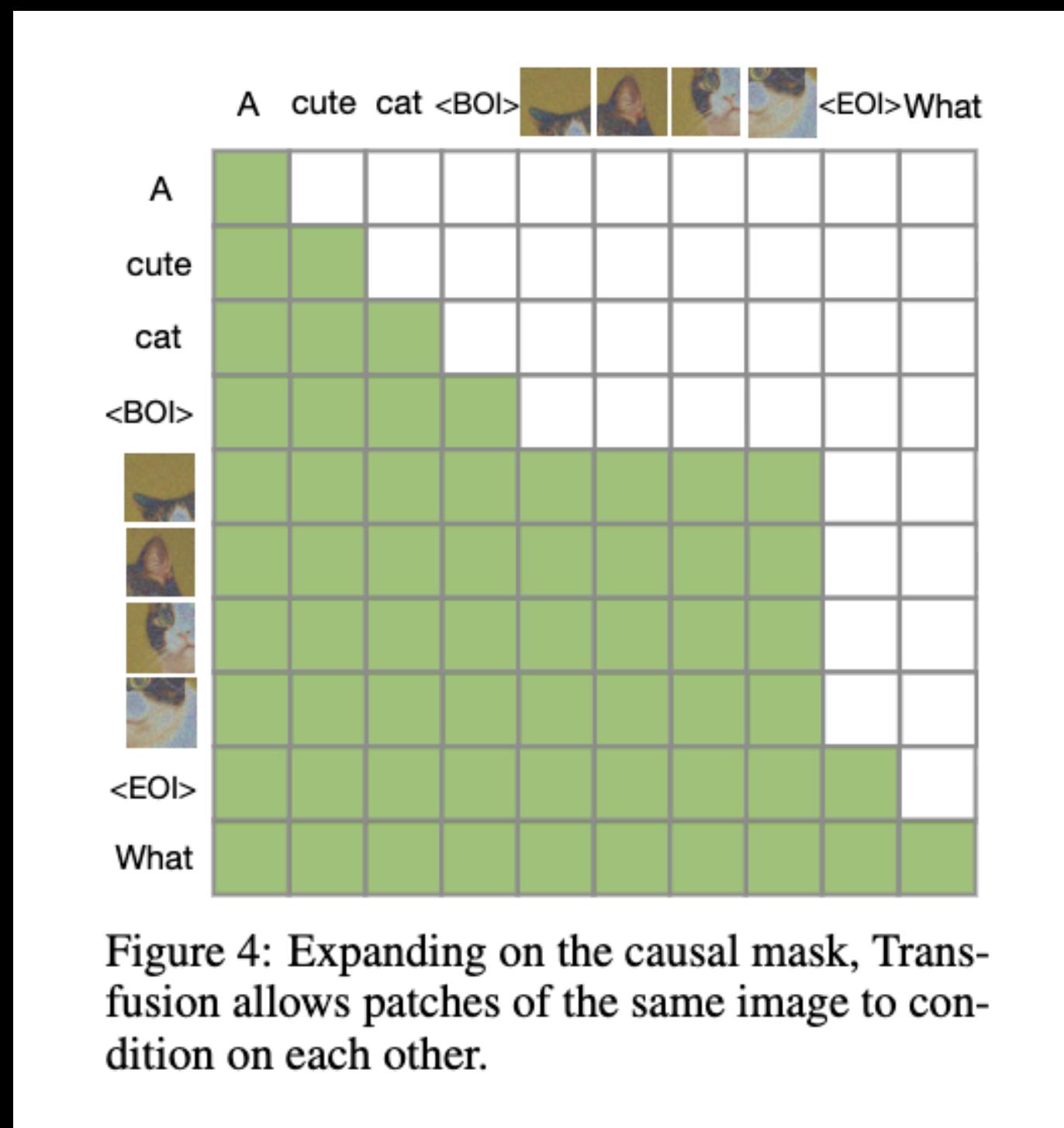


Arch: Unified auto-regressive Transformer

Data: 3T (text-only), 1.5T (image-text),
400B (interleaved) tokens

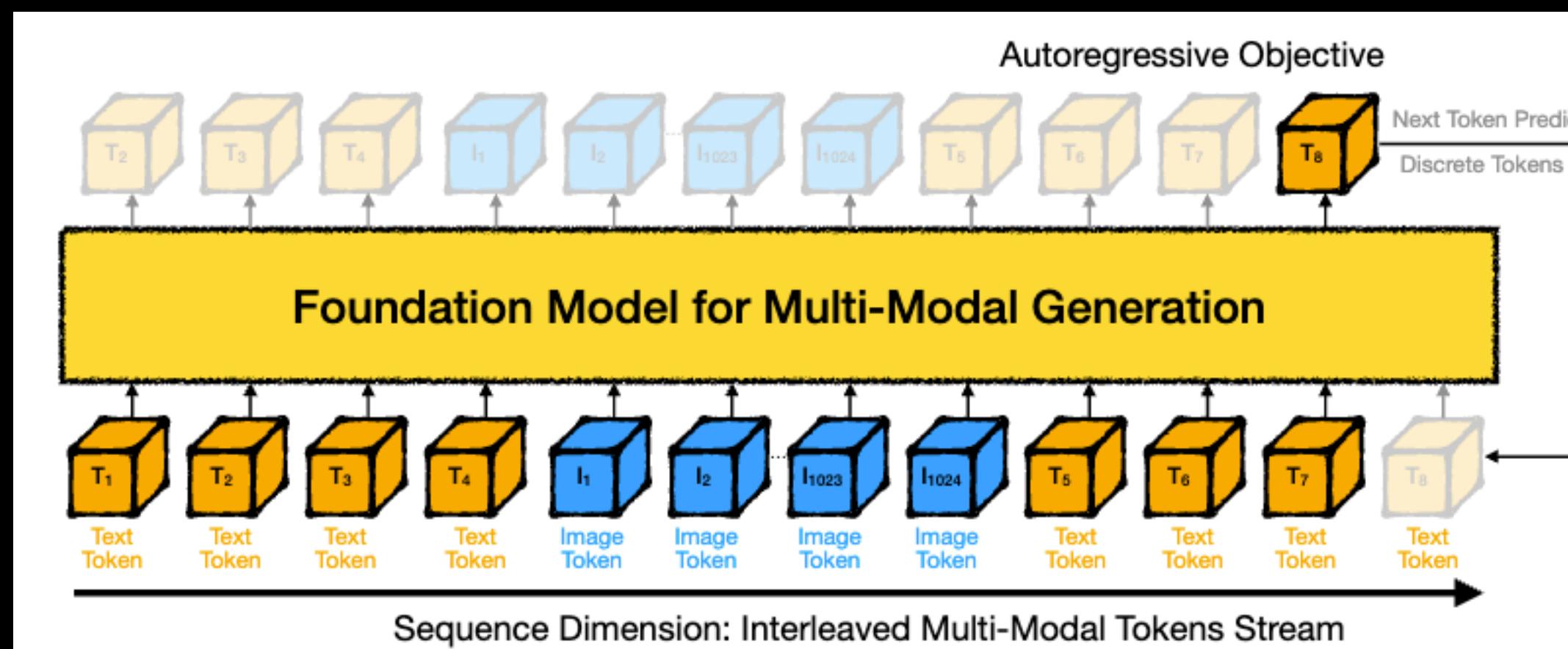
Transfusion (2024.08)

- Transfusion: Predict the Next Token and Diffuse Images with One Multi-Modal Model (Meta)
- Unified transformer, but switching auto-regressive mode & diffusion process



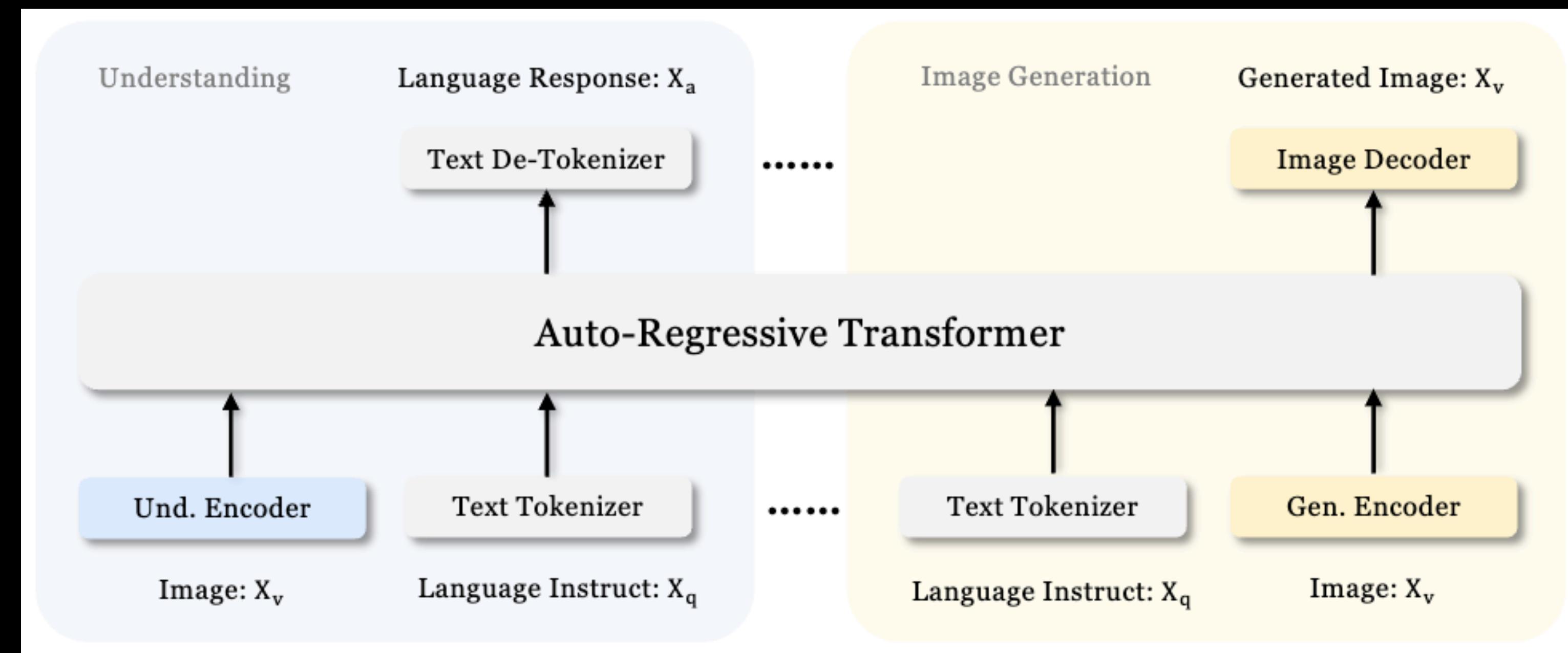
Mixture-of-Transformers (2024.11)

- Mixture-of-Transformers: A Sparse and Scalable Architecture for Multi-Modal Foundation Models (Meta)
- Unified architecture, but use *different weights* (qkv_proj, MLP, ...) *for each modality*
- *Attention is shared*



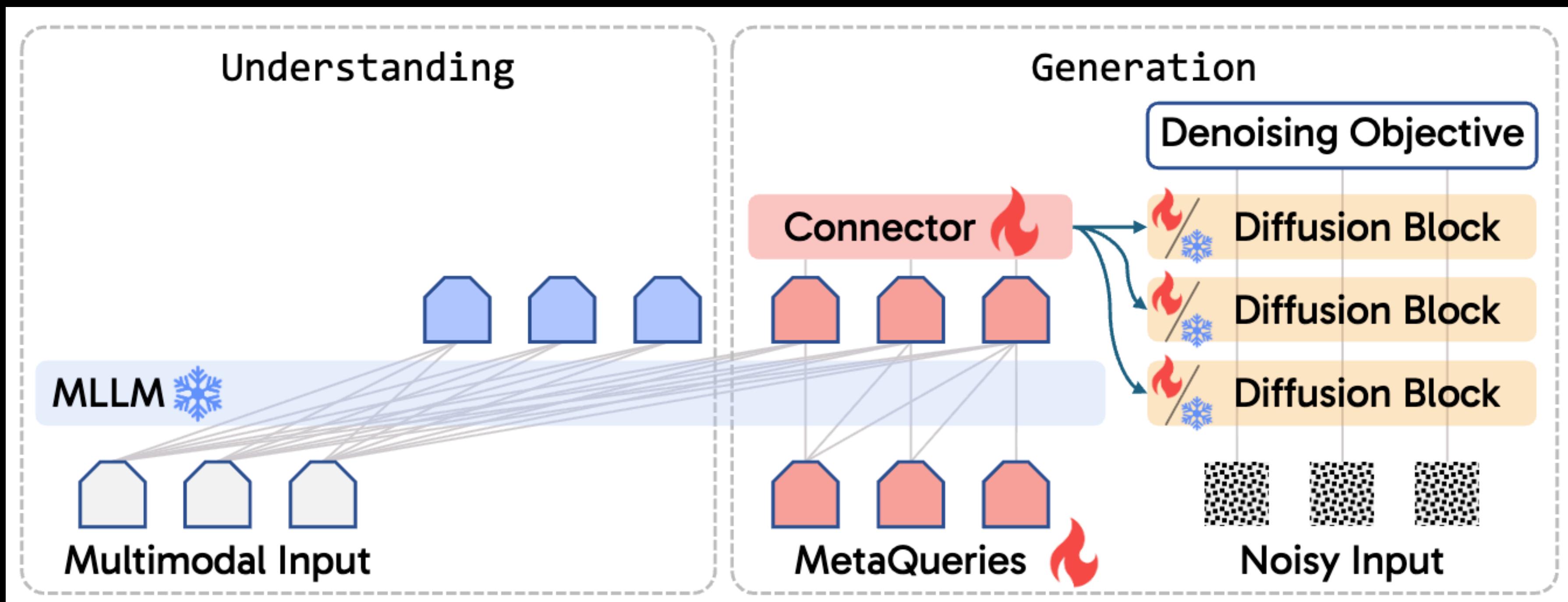
Janus-Pro (2025.01)

- Janus-Pro: Unified Multimodal Understanding and Generation with Data and Model Scaling (DeepSeek-AI)
- *Separated two visual encoders* — one for “understanding”, another for “generating”
- Why? Image generation quality



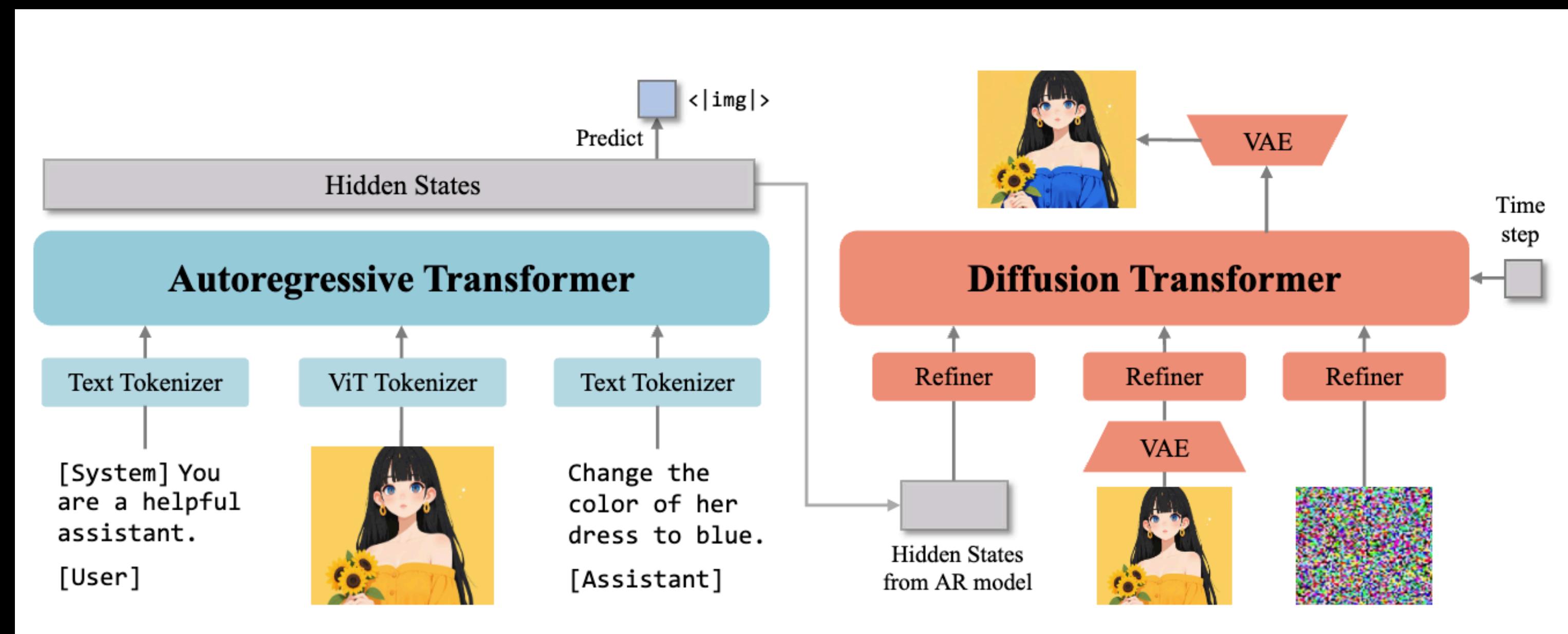
MetaQuery (2025.04)

- Transfer between Modalities with MetaQueries (Meta & NYU)
- Frozen VLM* → maintaining understanding ability



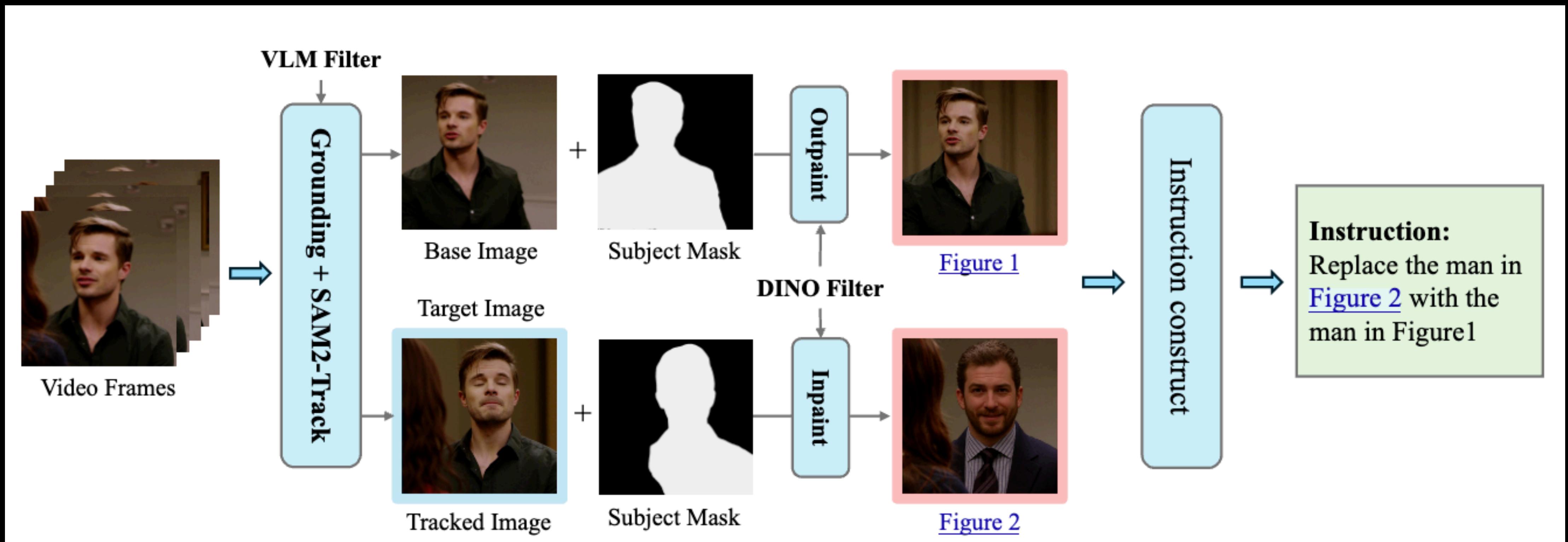
OmniGen2 (2025.06)

- OmniGen2: Exploration to Advanced Multimodal Generation (BAAI)
- Cascaded architecture: Auto-regressive transformer → Diffusion transformer
- Also use two visual encoders



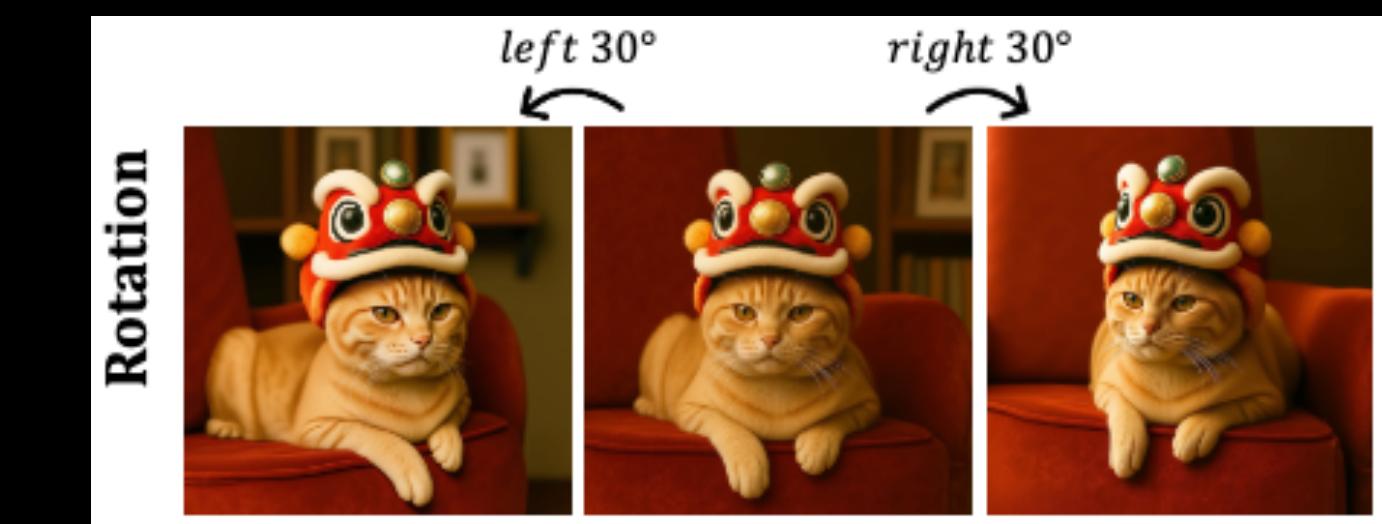
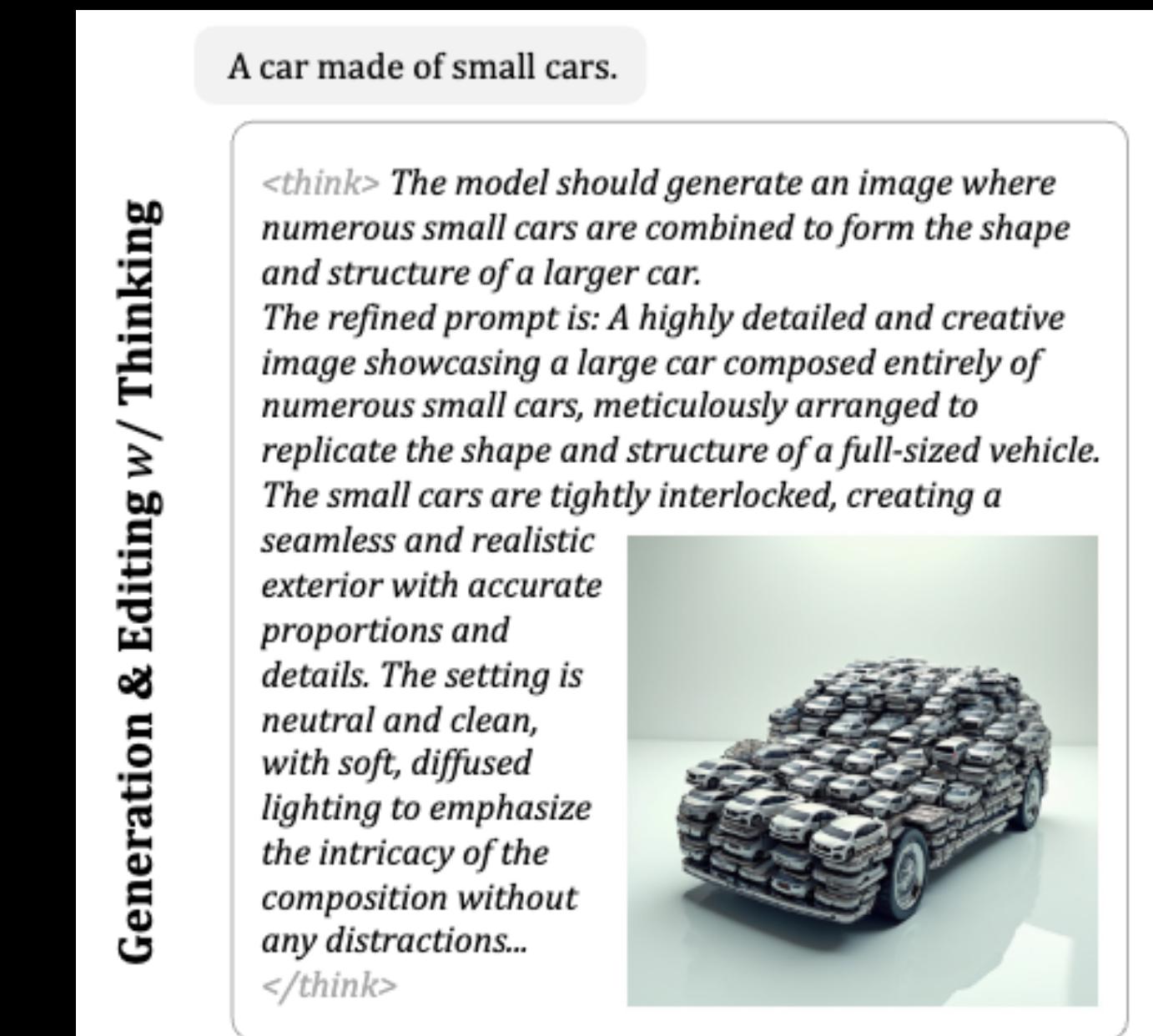
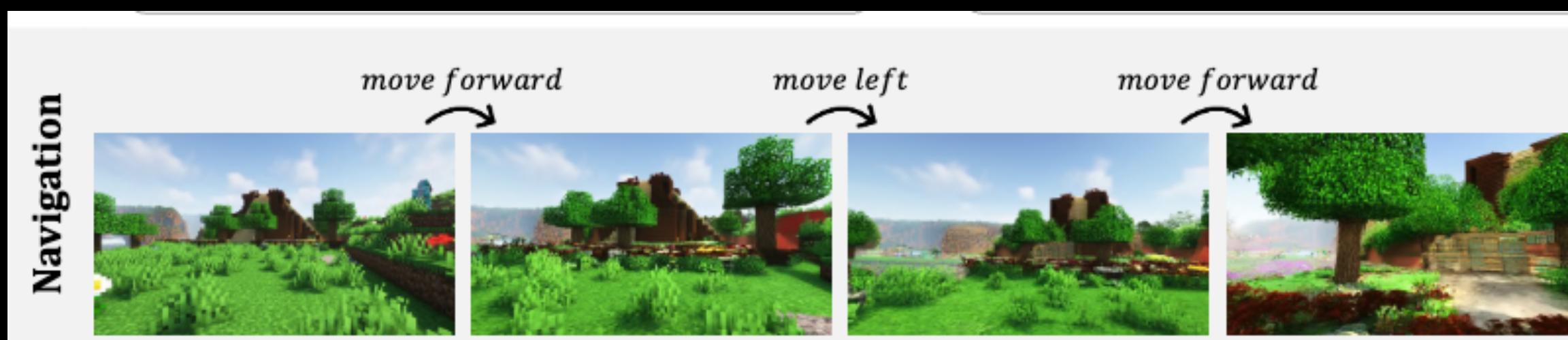
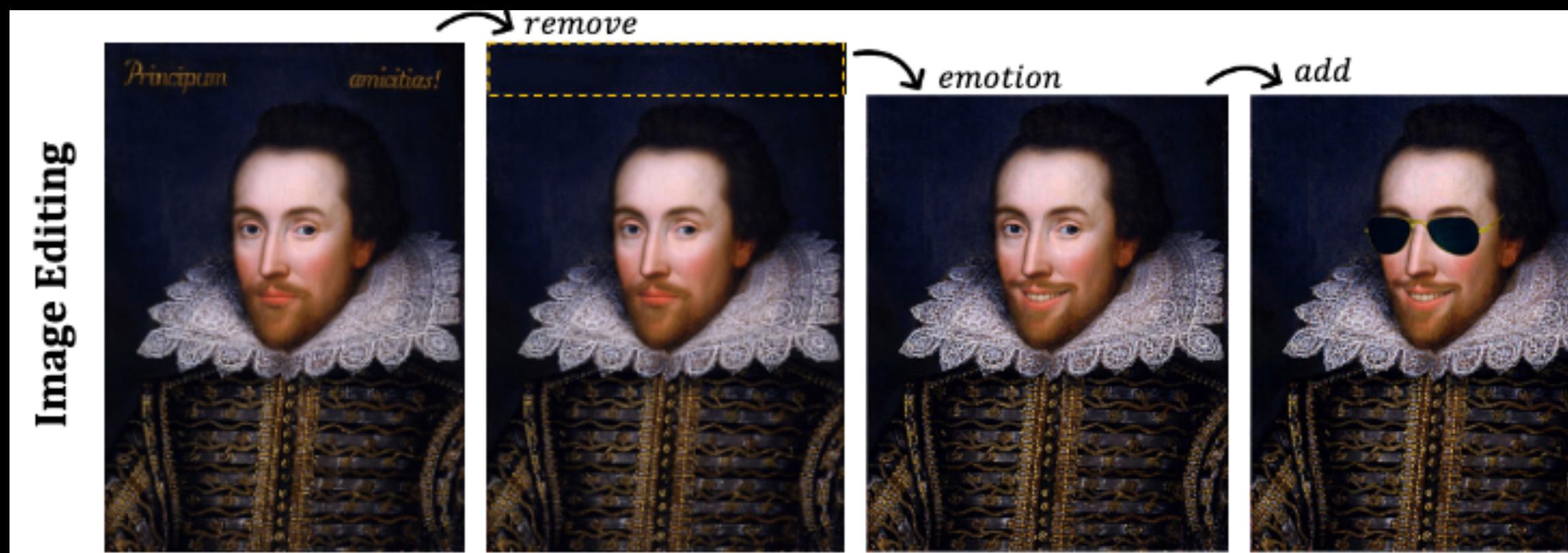
OmniGen2 (2025.06)

- Dataset for in-context generation



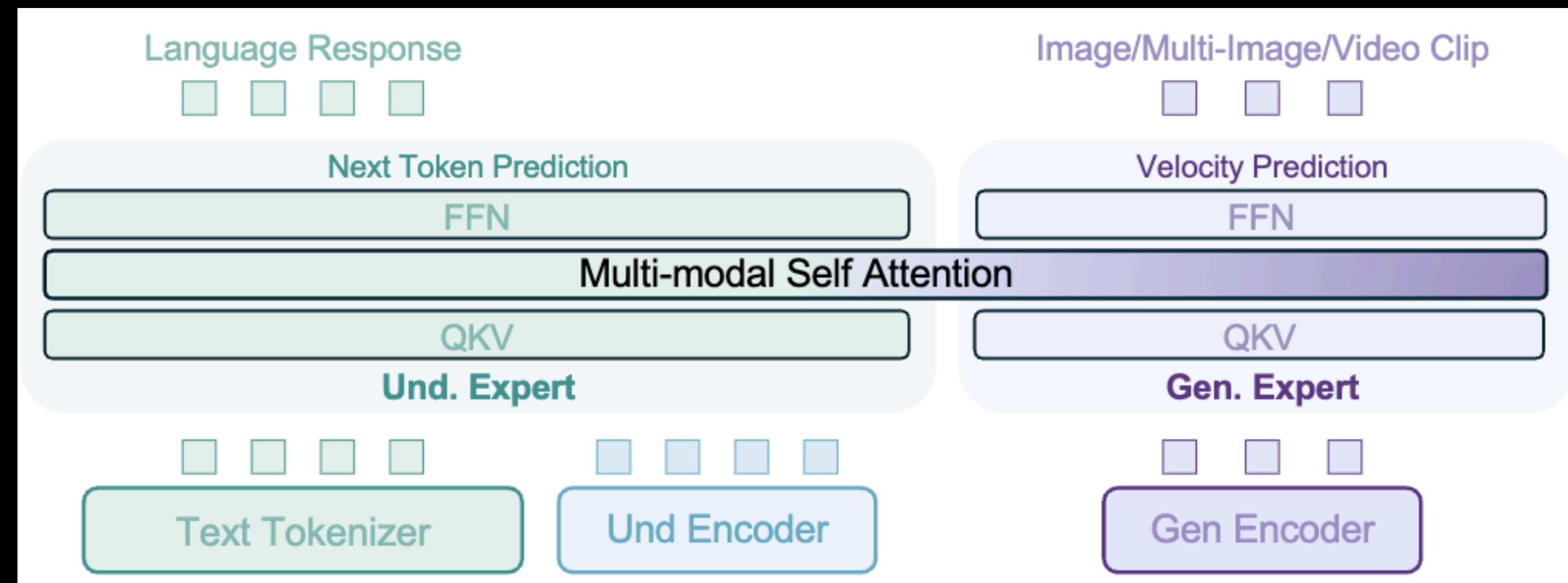
BAGEL (2025.05)

- Emerging Properties in Unified Multimodal Pretraining (ByteDance)
- Use Mixture-of-transformers



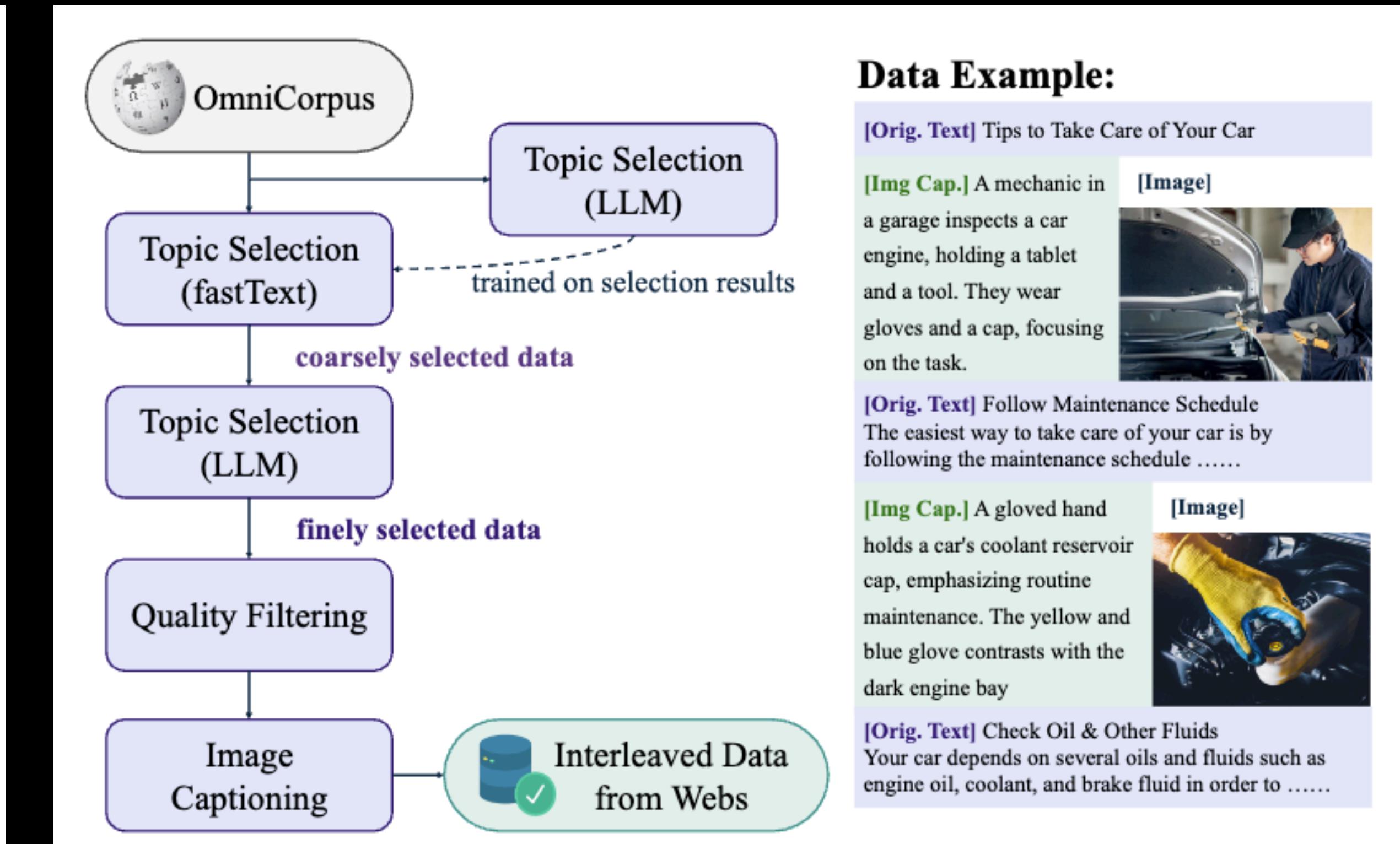
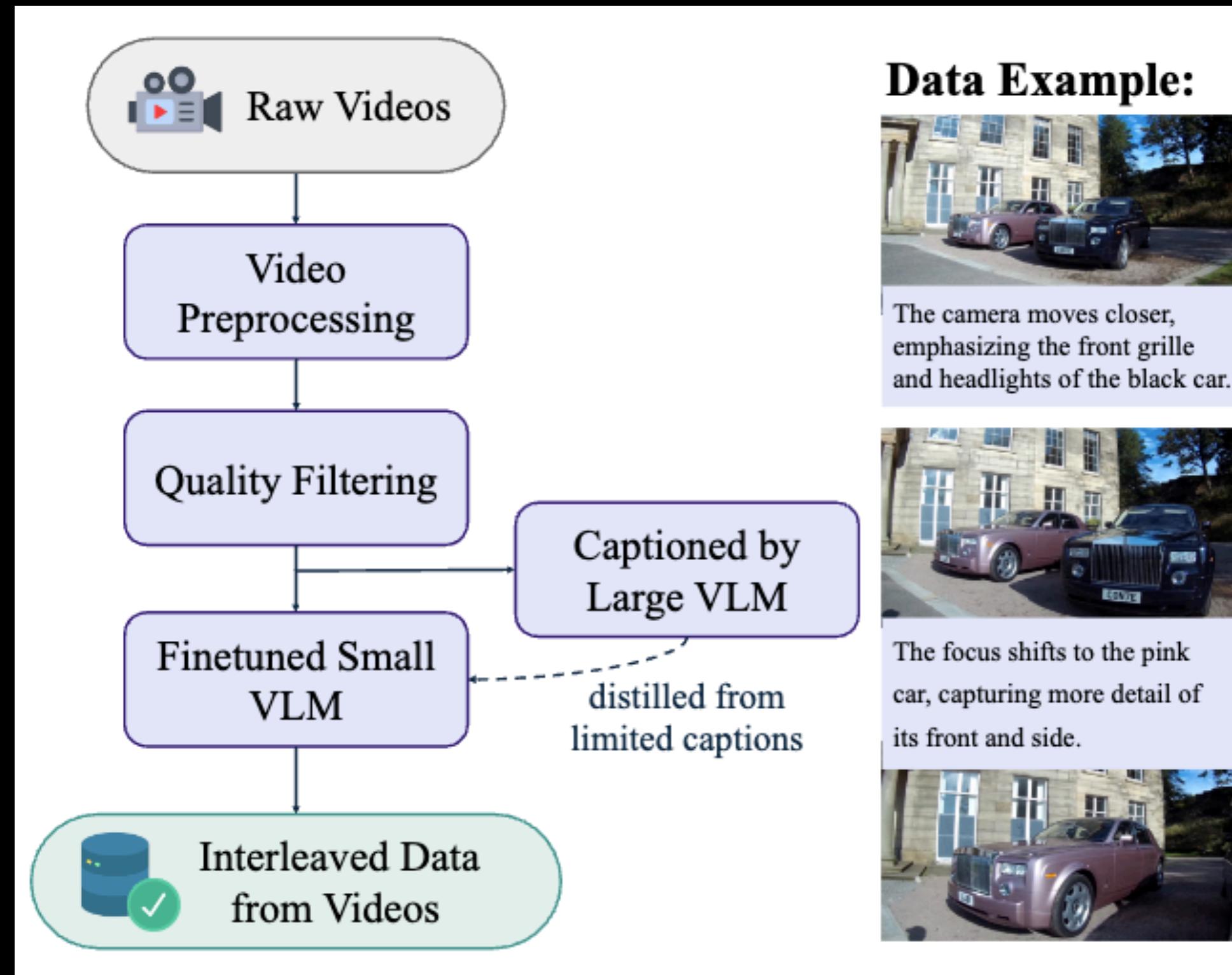
BAGEL (2025.05)

- Emerging Properties in Unified Multimodal Pretraining (ByteDance)
- Use *Mixture-of-transformers* and *separated* vision encoders

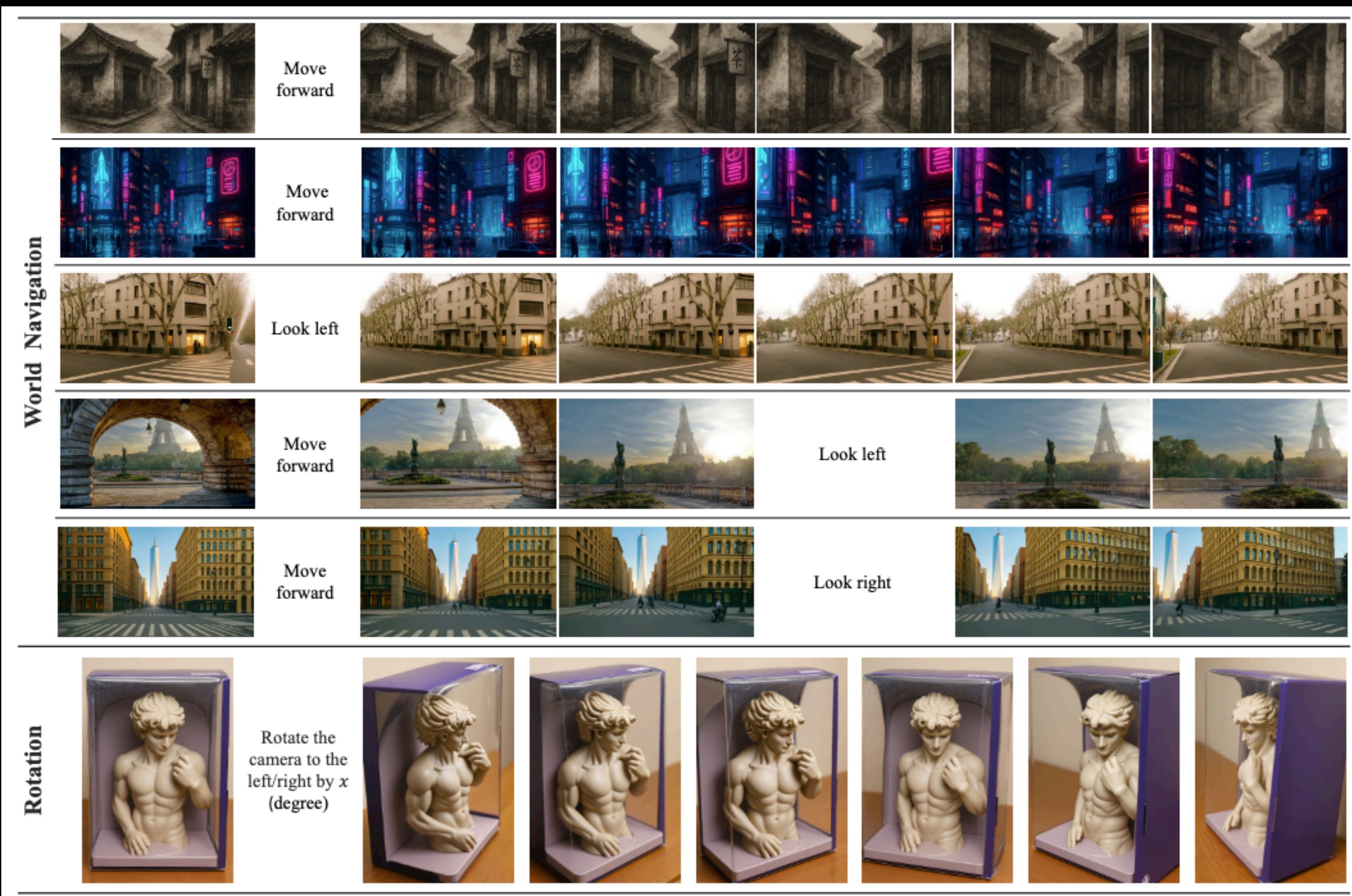


BAGEL (2025.05)

- Interleaved dataset construction



BAGEL (2025.05)



BAGEL (2025.05)

- For those who are interested in detail (how to switch understanding/generation modes):

Understanding

```
In [ ]: inference_hyper=dict(
    max_think_token_n=1000,
    do_sample=False,
    # text_temperature=0.3,
)

In [ ]: image = Image.open('test_images/meme.jpg')
prompt = "Can someone explain what's funny about this meme??"

display(image)
print(prompt)
print('*'*10)
output_dict = inferencer(image=image, text=prompt, understanding_output=True, **inference_hyper)
print(output_dict['text'])
```

Edit with Think

```
In [ ]: inference_hyper=dict(
    max_think_token_n=1000,
    do_sample=False,
    # text_temperature=0.3,
    cfg_text_scale=4.0,
    cfg_img_scale=2.0,
    cfg_interval=[0.0, 1.0],
    timestep_shift=3.0,
    num_timesteps=50,
    cfg_renorm_min=0.0,
    cfg_renorm_type="text_channel",
)

In [ ]: image = Image.open('test_images/octupusy.jpg')
prompt = 'Could you display the sculpture that takes after this design?'

display(image)
print('*'*10)
output_dict = inferencer(image=image, text=prompt, think=True, **inference_hyper)
print(output_dict['text'])
display(output_dict['image'])
```

```
257
258     if understanding_output:
259         gen_text = self.gen_text(gen_context, do_sample=do_sample, temperature=text_temperature, max_length=
260         output_list.append(gen_text)
261
262     else:
263         if think:
264             gen_text = self.gen_text(gen_context, do_sample=do_sample, temperature=text_temperature, max_len
265             gen_context = self.update_context_text(gen_text, gen_context)
266             output_list.append(gen_text)
267
268         img = self.gen_image(
269             image_shapes,
270             gen_context,
271             cfg_text_precontext=cfg_text_context,
272             cfg_img_precontext=cfg_img_context,
273
274             cfg_text_scale=cfg_text_scale,
275             cfg_img_scale=cfg_img_scale,
276             cfg_interval=cfg_interval,
277             timestep_shift=timestep_shift,
278             num_timesteps=num_timesteps,
279             cfg_renorm_min=cfg_renorm_min,
280             cfg_renorm_type=cfg_renorm_type,
281             enable_taylorseer=enable_taylorseer,
282
283         )
284
285         output_list.append(img)
286
```

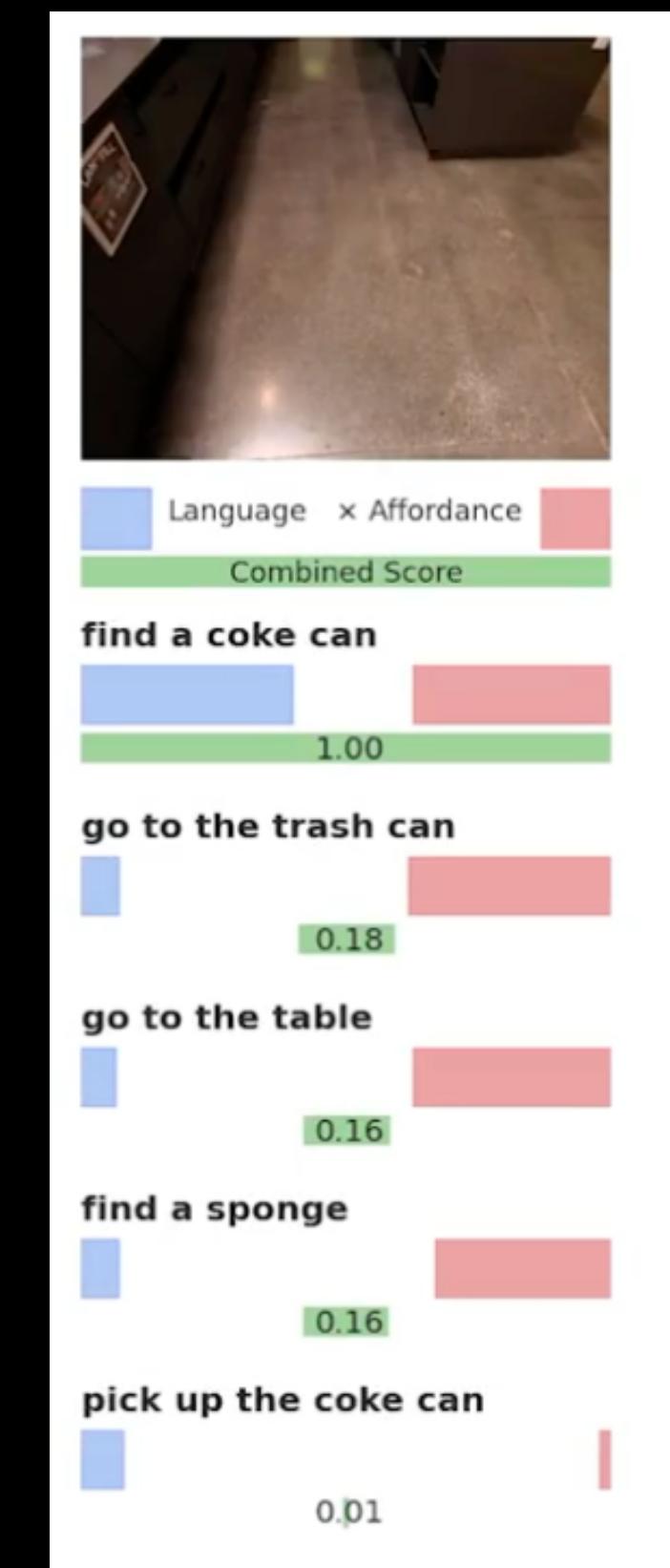
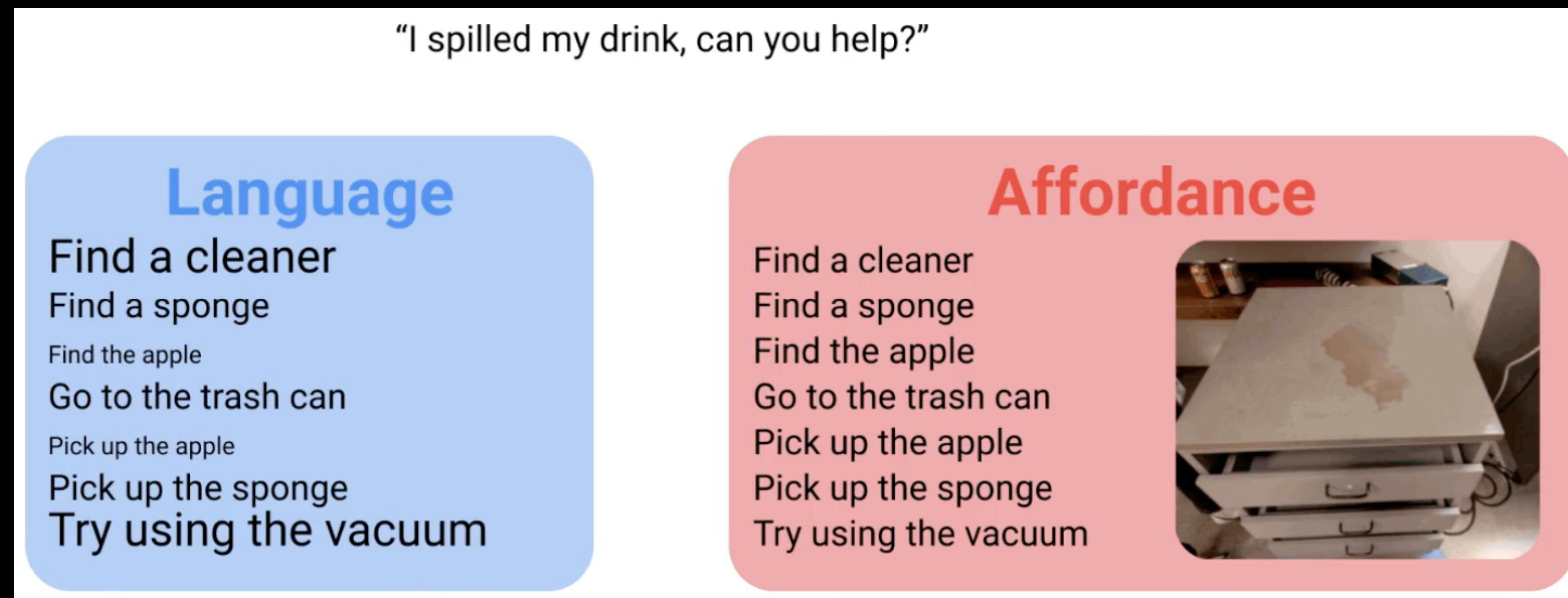
Next steps?

- Improve unified architecture
- Multimodal reasoning (interleaved) data construction to Pre-training level

Vision-Language-Action models (VLA)

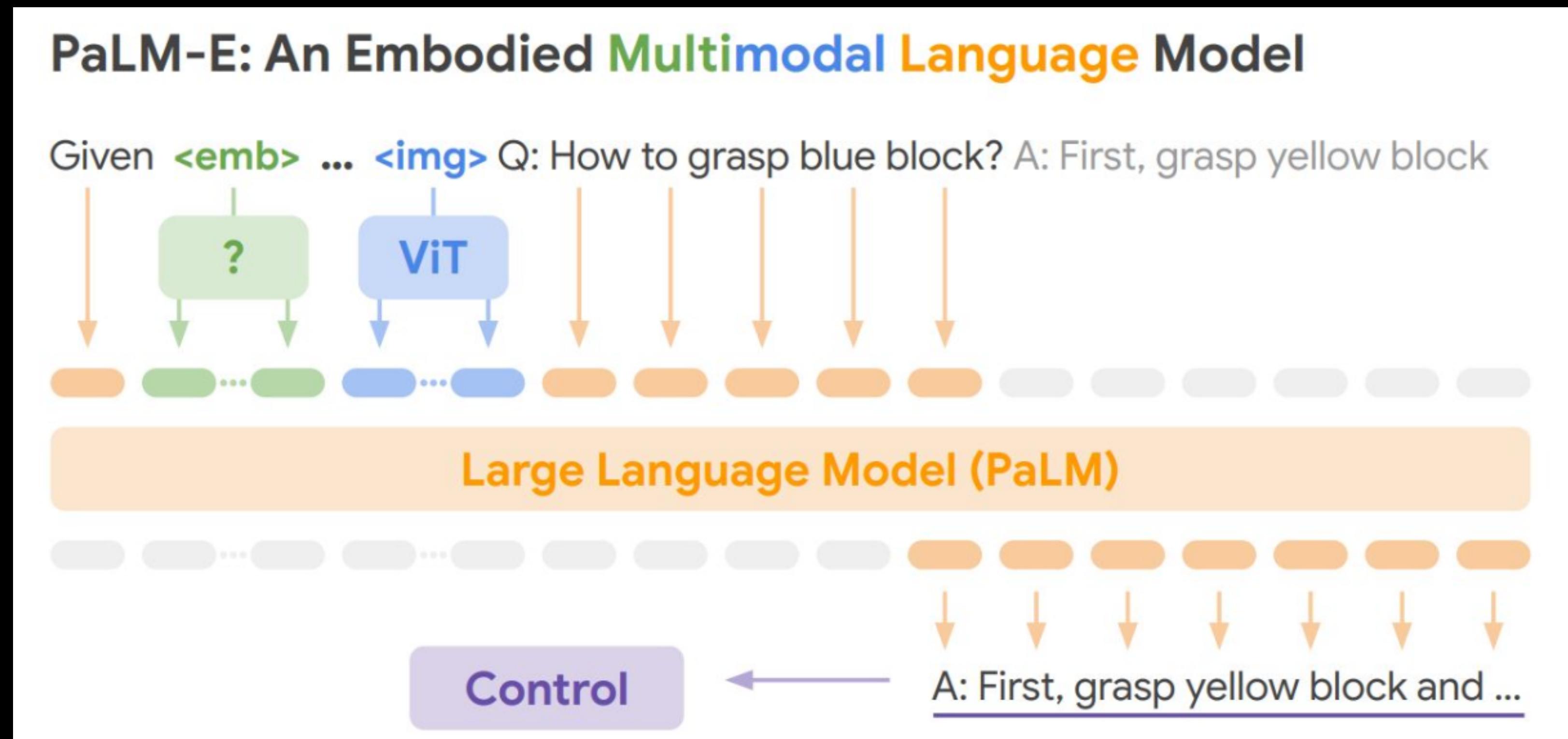
VLM/LLM as a planner

- Do As I Can, Not As I Say: Grounding Language in Robotic Affordances



VLM/LLM as a planner

- PaLM-E: An Embodied Multimodal Language Model

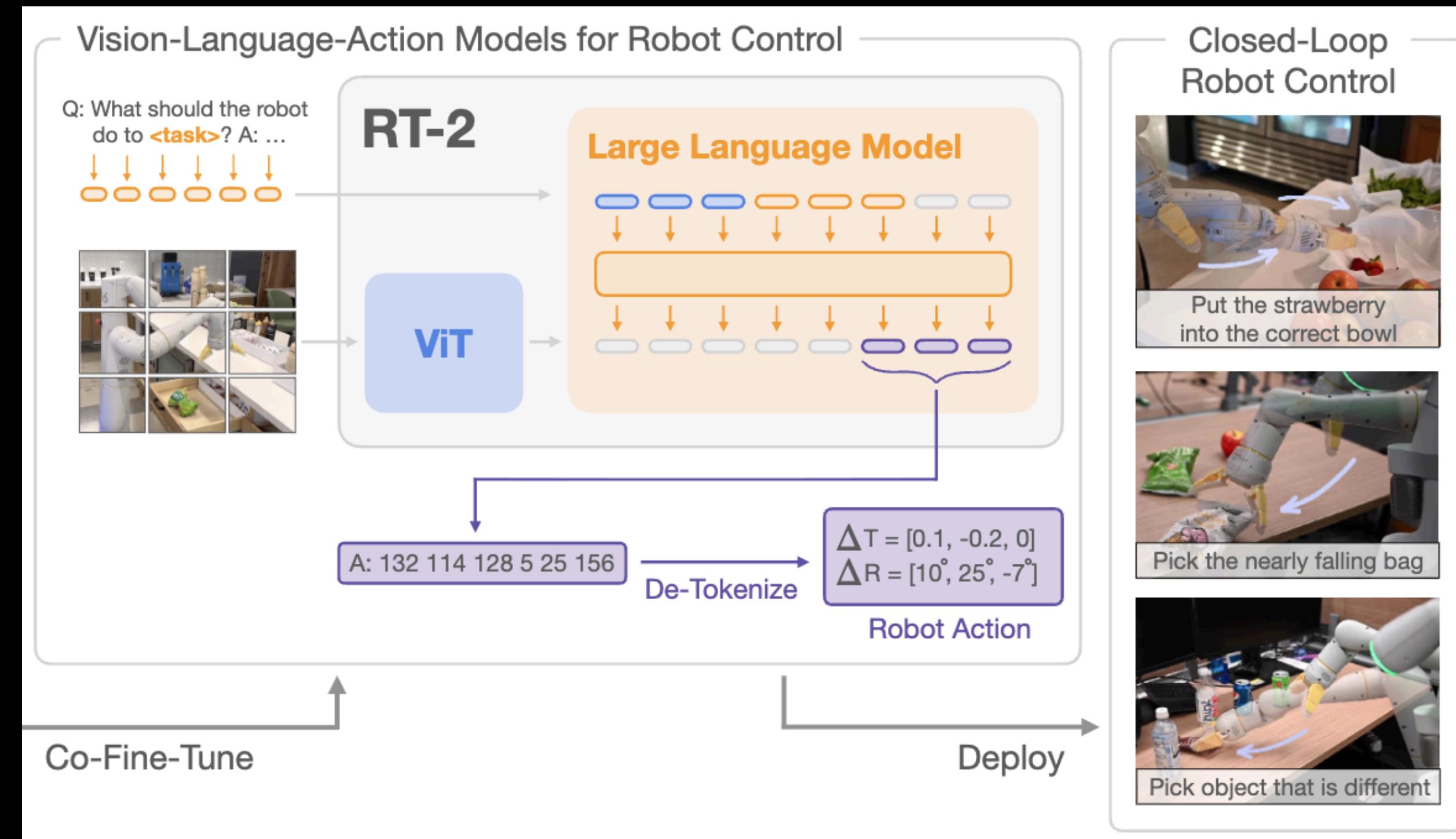


Vision-language-action (VLA)

- *Robot actions as a foreign language*

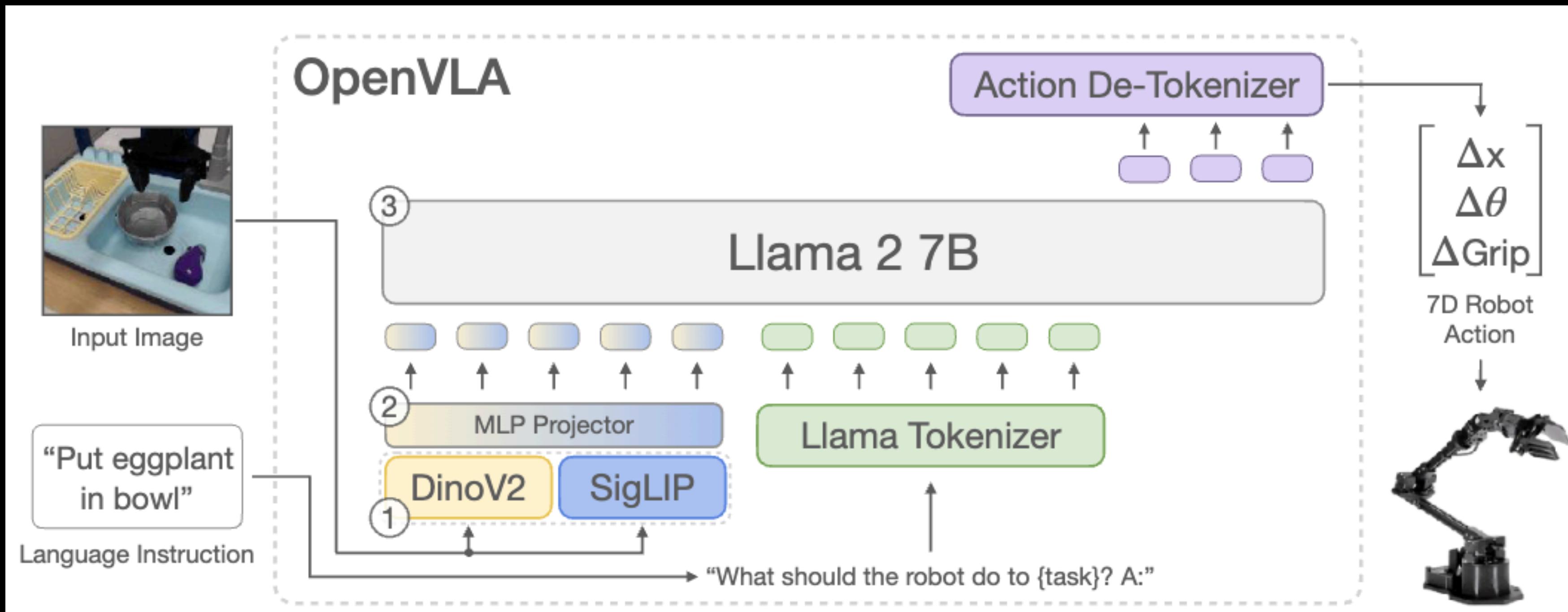
RT-2

- RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control



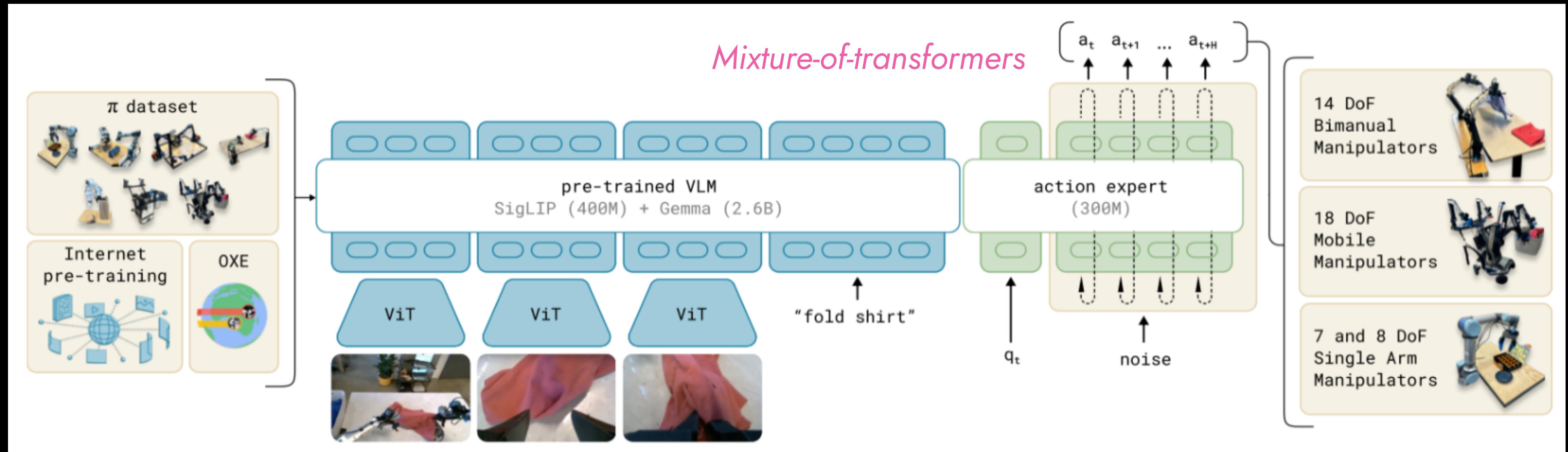
OpenVLA

- Improved with better vision encoder



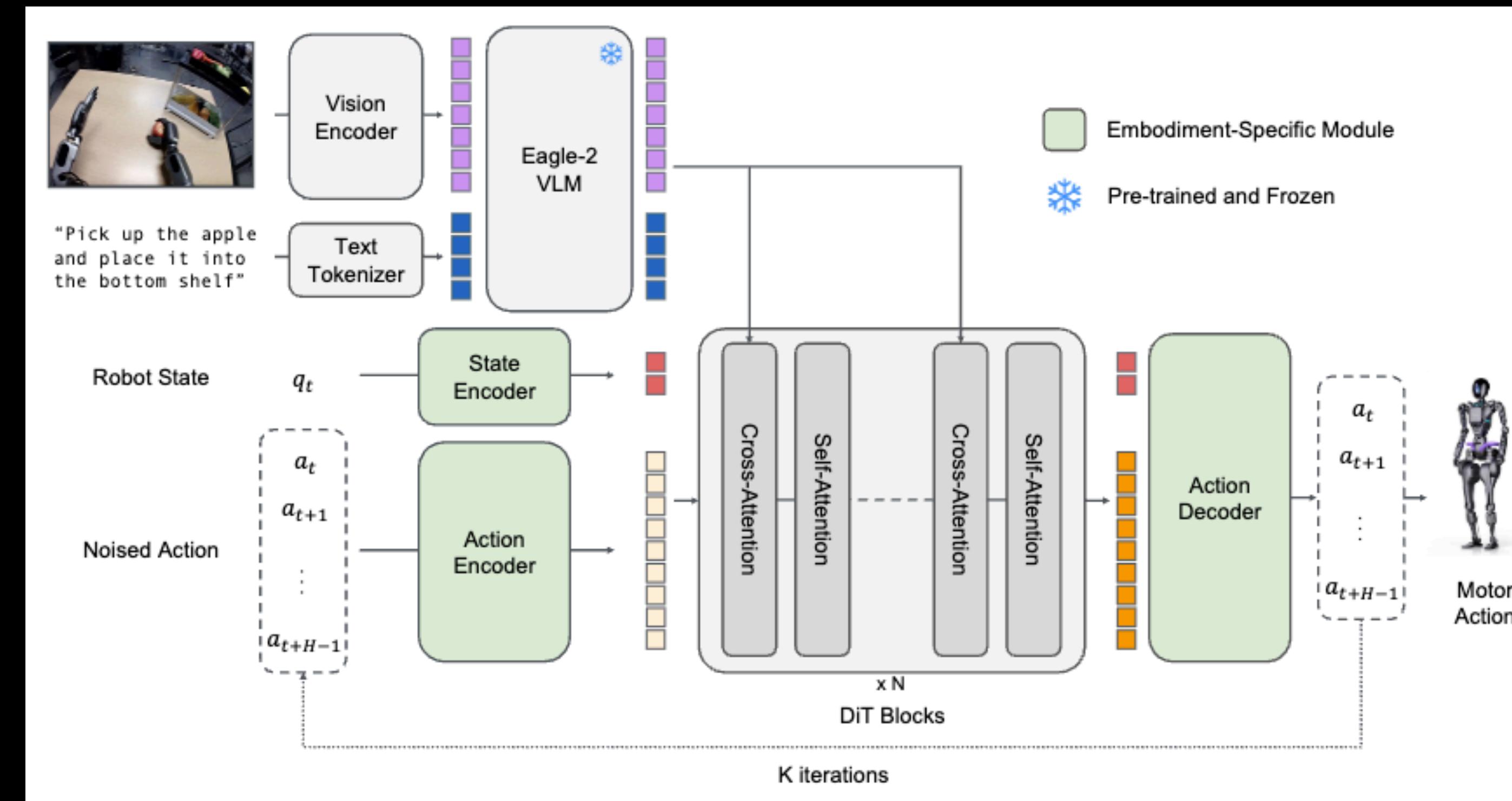
π_0

- π_0 : A Vision-Language-Action Flow Model for General Robot Control
- Architecture: separated weight params from VLMs and action expert



GROOT N1

- Separated params for VLMs and action models
- VLM's latent embedding → Cross-attention on action model



E-CoT

- Robotic Control via Embodied Chain-of-Thought Reasoning

Regular VLA Policy

User: Clean the sink.

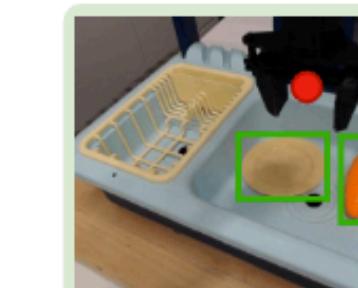
OpenVLA:
[Δx , $\Delta \theta$, ΔGrip] = ...



VLA w/ Embodied Chain-of-Thought

User: Clean the sink.

CoT OpenVLA:
Object bounding boxes: Gripper position:

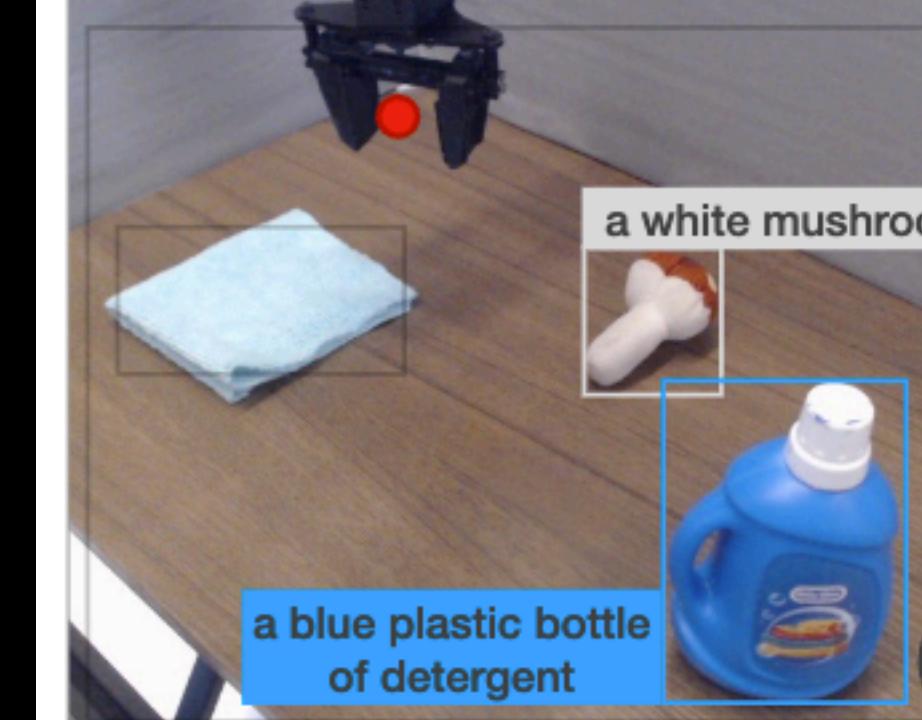


CoT OpenVLA:
Task Plan: 1. ..., 2. ...
Next Task: pick up carrot
Next Move: move right

CoT OpenVLA:
[Δx , $\Delta \theta$, ΔGrip] = ...

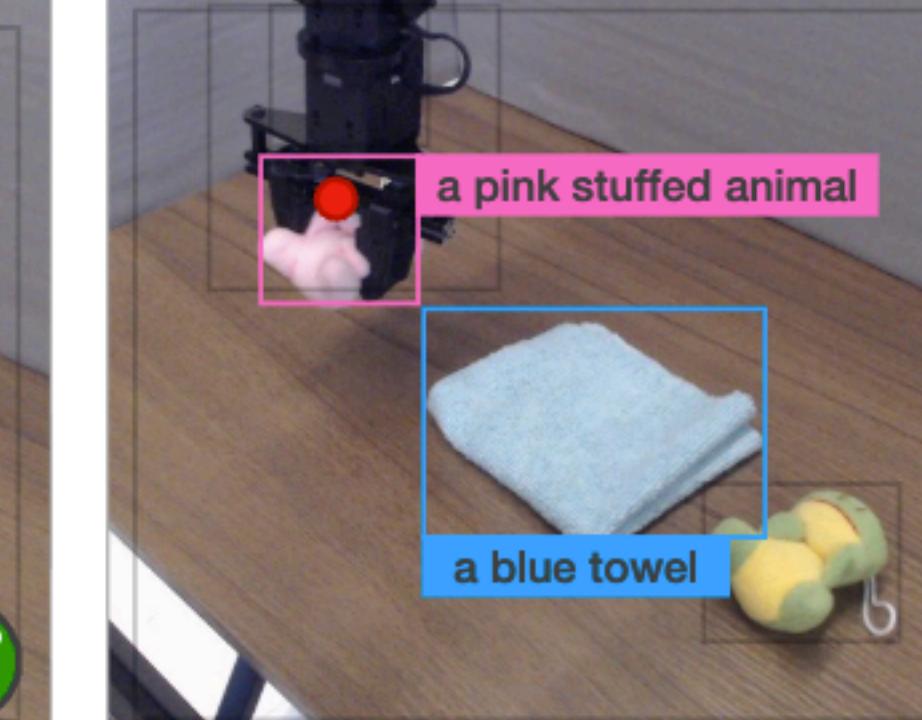


"Move the mushroom to the detergent"



a white mushroom
a blue plastic bottle of detergent

"Put the leftmost object on the middle object"



a pink stuffed animal
a blue towel

TASK: Move the mushroom to the detergent

PLAN: Move to the mushroom, grasp it, move to the detergent, release the mushroom

SUBTASK: The mushroom is the closest object to the robot, and needs to be grasped first → Move to the mushroom

MOVE: The mushroom is to the right of the robot, so the robot needs to move right → Move right

TASK: Put the pink stuffed animal on the blue towel

PLAN: Move to the pink stuffed animal. Pick up the pink stuffed animal. Move to the blue towel. Put the pink stuffed animal on the blue towel.

SUBTASK: The pink stuffed animal is held by the robotic arm and the blue towel is the middle object → Put the pink stuffed animal on the blue towel

MOVE: The pink stuffed animal is on the left of the blue towel and needs to be moved right → Move right

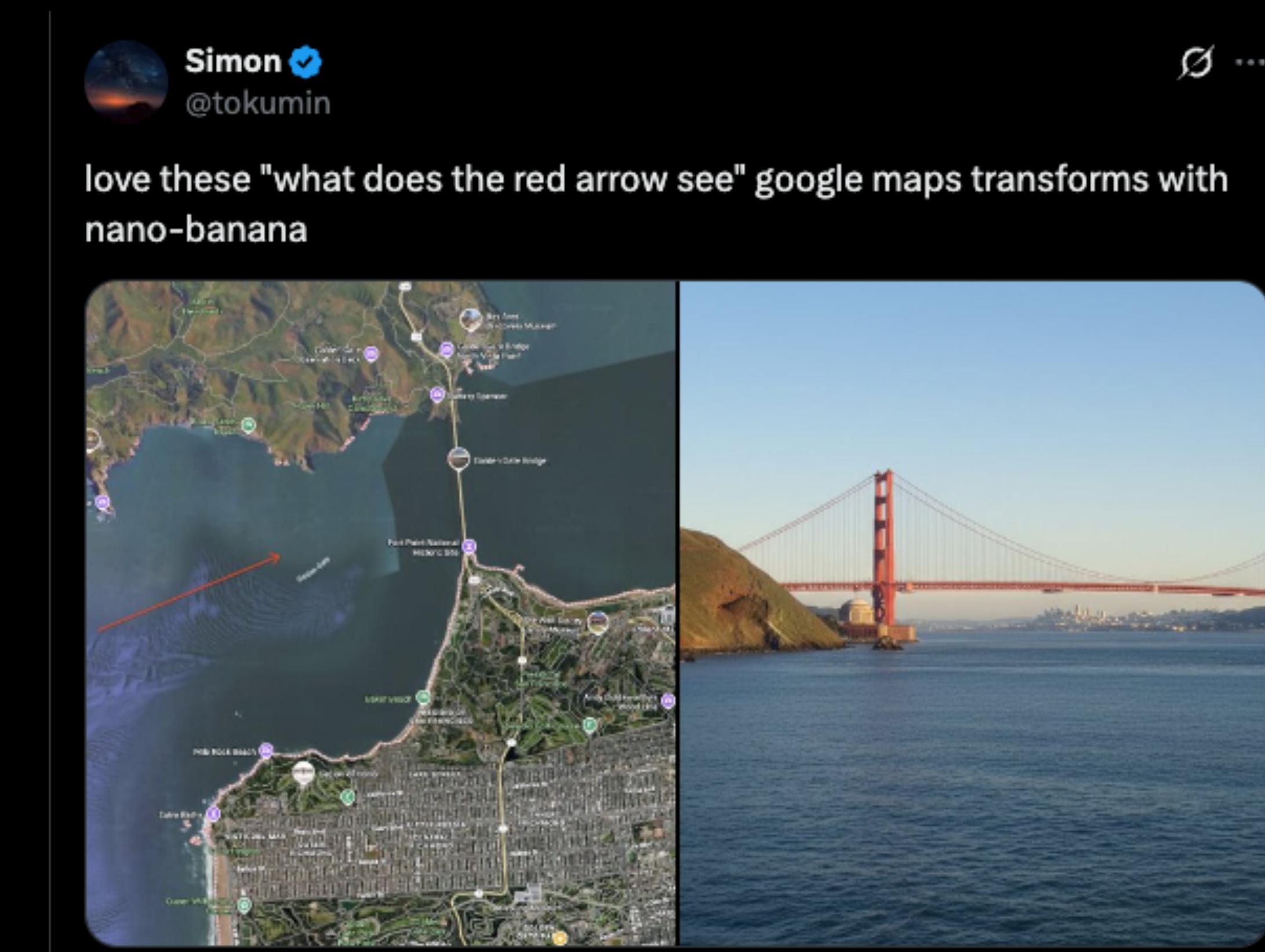
Next steps?

- Enhance perception module (e.g., vision encoder)
- Robotics data (robot's perception and action) are scarce
- VLAs are still slow, low-latency is necessary to interact with human

Appendix

Generation models

- Image generation (nano-banana)



<https://x.com/tokumin/status/1960583251460022626>

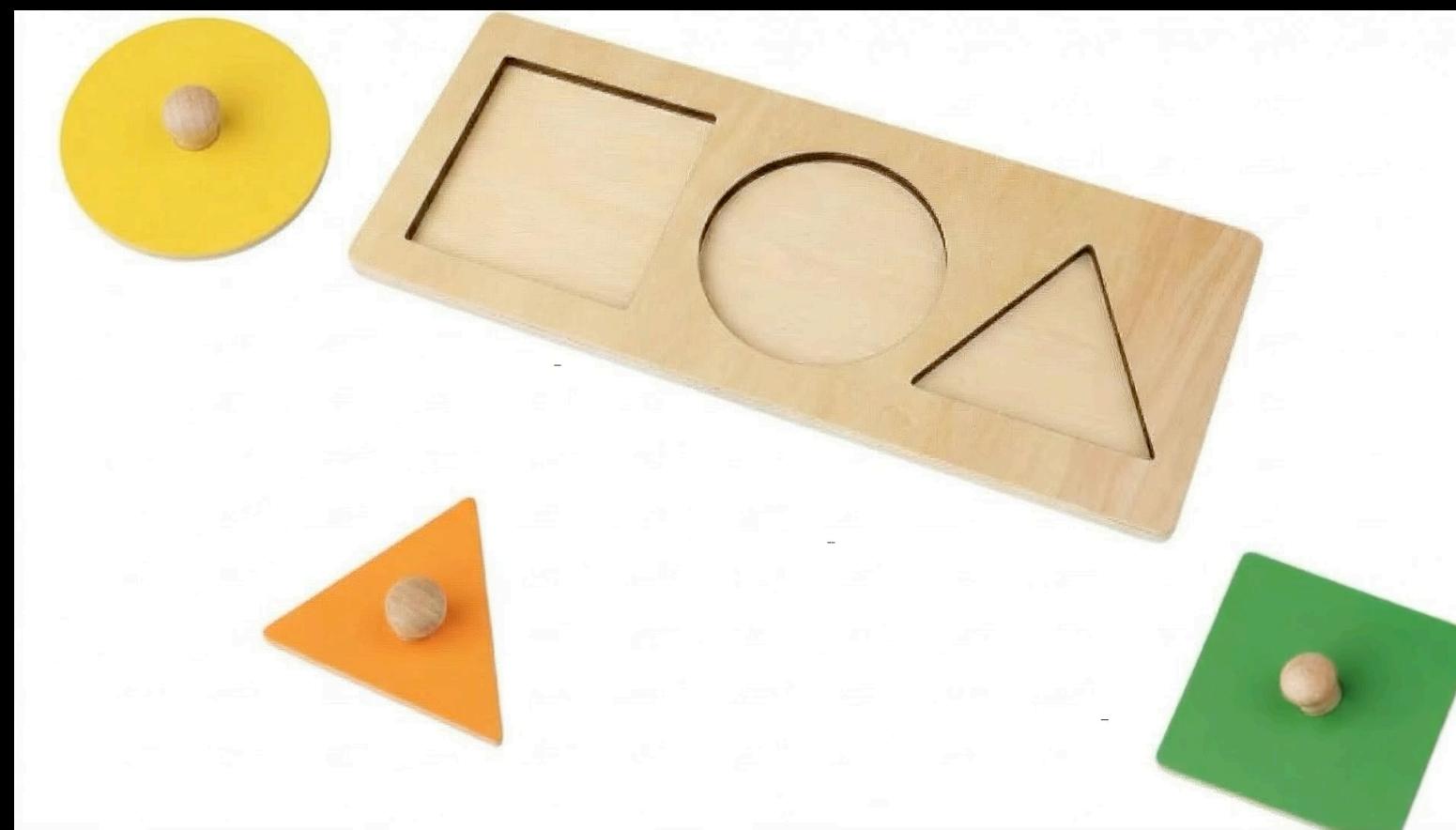
Generation models

- Video generation (Sora 2)



Generation models

- Visual reasoning with Veo3 (by Google)
- Prompt: describe the status and give an action



Generation models

- World models – Controllable video generation (Genie-3)



Conclusion

- Today, we study:
 - (Multimodal) Reasoning models
 - Vision-language-action models
 - Omni (any-to-any) models – Unified understanding and generation beyond modality boundaries

Thank You!