From Language Models to Native Multimodal Intelligence

Victoria Lin

CS25 Transformers United - Stanford University

victoria@thinkingmachines.ai @VictoriaLinML

Opinions are my own, and not of my employer

May 21, 2026

Language Models

ChatGPT

ChatGPT 4o

You 10:30 AM

I'm planning a 6-month sabbatical to learn, build, and reset. Help me brainstorm a high-level plan with goals, a learning path, and a few project ideas.

ChatGPT 10:30 AM

Great idea! Here's a high-level plan to structure your sabbatical around growth, building, and reflection.

1. Goals
   - Deepen technical skills in areas you enjoy
   - Build and ship meaningful projects
   - Develop a sustainable routine and reflect
   - Explore new ideas and expand your network

...

You 10:33 AM

Love this. Can you suggest a learning path for AI engineering and system design?

ChatGPT 10:34 AM

Absolutely. Here's a suggested path:
1. Foundations: Python, data structures, algorithms
2. System Design: Scalability, databases, caching, APIs
3. AI Engineering: LLMs, RAG, fine-tuning, evals
4. Build: Clone projects, then build your own
5. Share: Write, document, and engage with the community

...

Message ChatGPT...

Claude Code

```
Claude Code                                                    v1.0.0

> Claude Code is a coding assistant that lives in your terminal.
> Ask questions, edit files, run commands, and ship faster.

Type /help for available commands.

You:
Add a function to paginate results in api/src/handlers/projects.ts
using cursor-based pagination.

Claude Code:
I'll open the file and implement cursor-based pagination.

$ sed -n '1,200p' api/src/handlers/projects.ts
(file contents)
...

Claude Code:
I'll add a getProjects function with cursor-based pagination.

$ apply_patch
*** Begin Patch
*** Update File: api/src/handlers/projects.ts
@@
 export async function getProjects(req, res) {
-  const projects = await db.projects.findMany({ limit: 20 });
-  res.json(projects);
+   const { cursor, limit = 20 } = req.query;
+   const projects = await db.projects.findMany({
+     take: Number(limit) + 1,
+     cursor: cursor ? { id: cursor as string } : undefined,
+     orderBy: { id: 'asc' },
+   });
+   const nextCursor = projects.length > Number(limit)
+     ? projects[projects.length - 1].id
+     : null;
+   const items = projects.slice(0, Number(limit));
+   res.json({ items, nextCursor });
 }
*** End Patch

Claude Code:
Patch applied successfully.
Anything else you'd like to build?

You:
```

Codex

Codex Agent — Task: Add search + filters to projects list

Instructions

Add a text search and status filter to the projects list API and UI.
Update the query params, backend handler, and the table view.
Include tests.

Plan

1. Update API: accept q and status params
2. Update DB query with filters
3. Update UI: add search input and status dropdown
4. Add tests for API and UI
5. Run tests and ensure lint passes

Working

```
$ git status
On branch main
Changes not staged for commit:
   modified:  api/src/handlers/projects.ts
   modified:  web/src/pages/projects.tsx
   new file:  api/src/tests/projects.test.ts
...
```

Changes

api/src/handlers/projects.ts +18 −6

```
@@ export async function getProjects(req, res) {
-  const { cursor, limit = 20 } = req.query;
+  const { cursor, limit = 20, q, status } = req.query;
   const where: any = {};
+  if (q) where.name = { contains: q, mode: 'insensitive' };
+  if (status) where.status = status;
+  const projects = await db.projects.findMany({
   where,
   take: Number(limit) + 1,
...
```

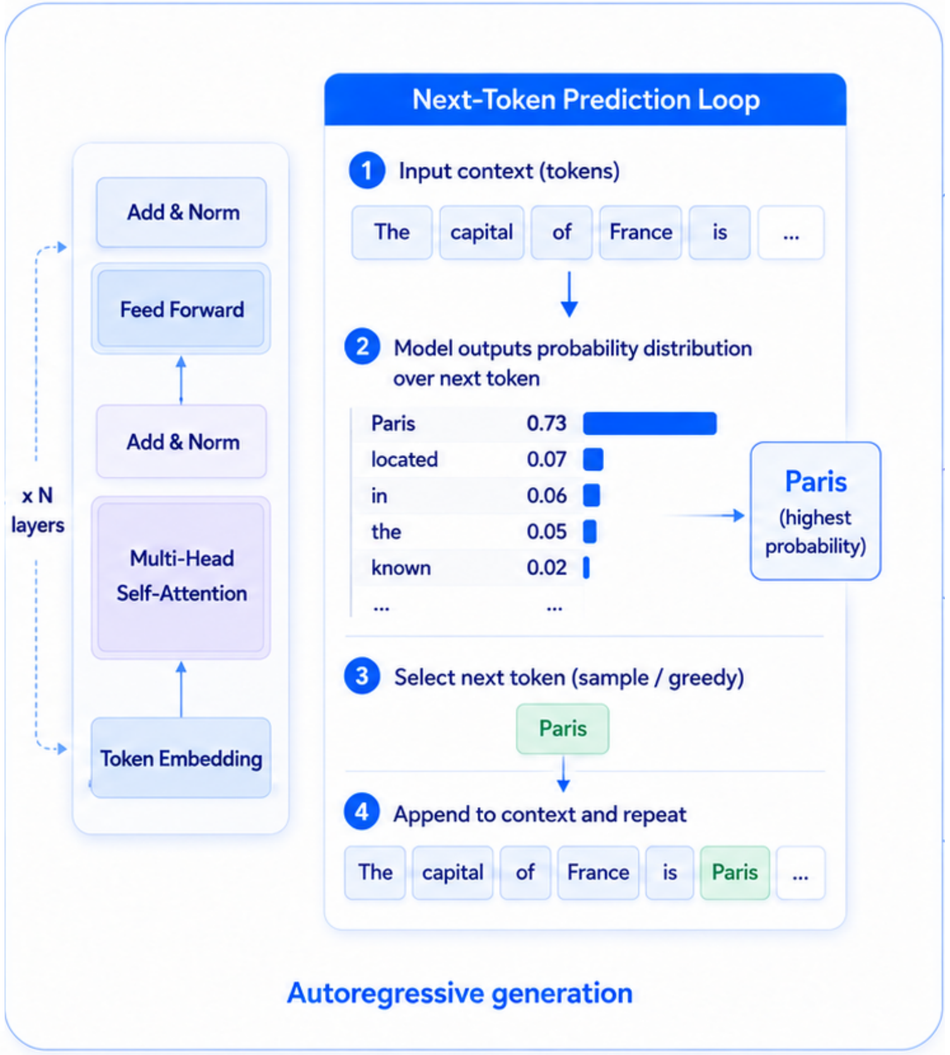
web/src/pages/projects.tsx +42 −15

api/src/tests/projects.test.ts +76 −0

Output

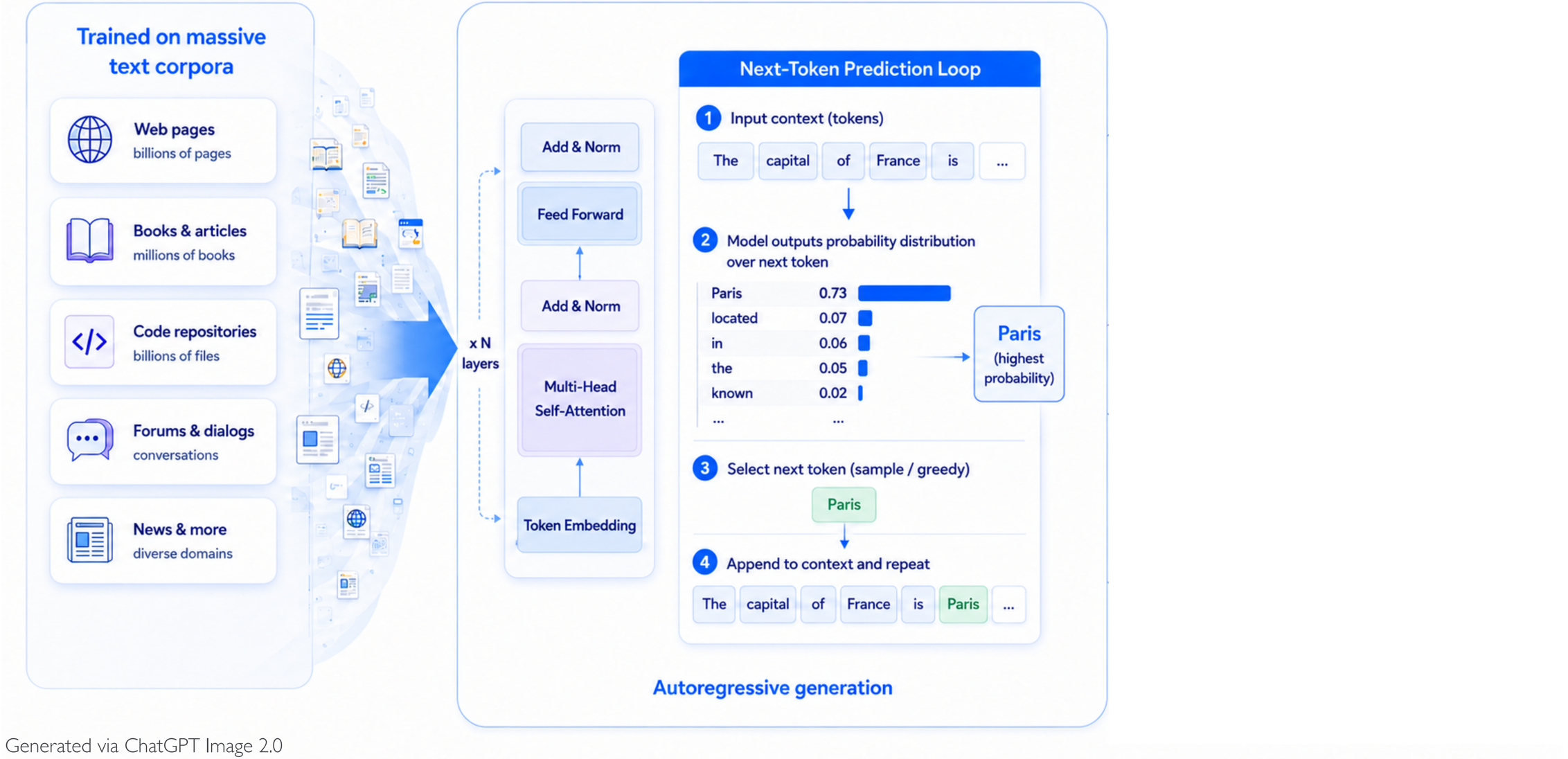
```
$ npm test
PASS  api/src/tests/projects.test.ts
Test Suites: 1 passed, 1 total
Tests:       12 passed, 12 total
Time:        1.23s
Done.
```

Language Models



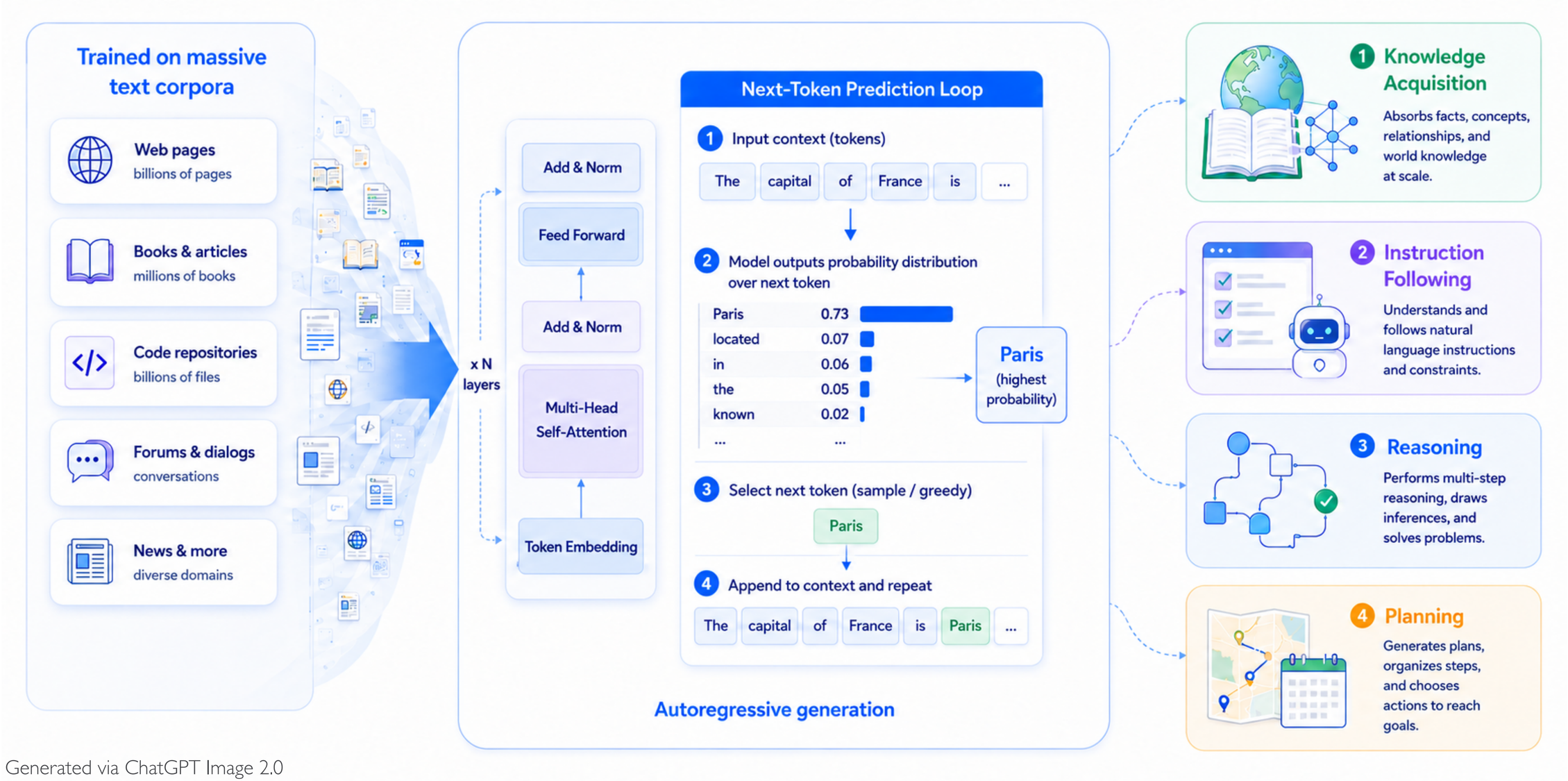
Generated via ChatGPT Image 2.0

Language Models



Generated via ChatGPT Image 2.0

Language Models



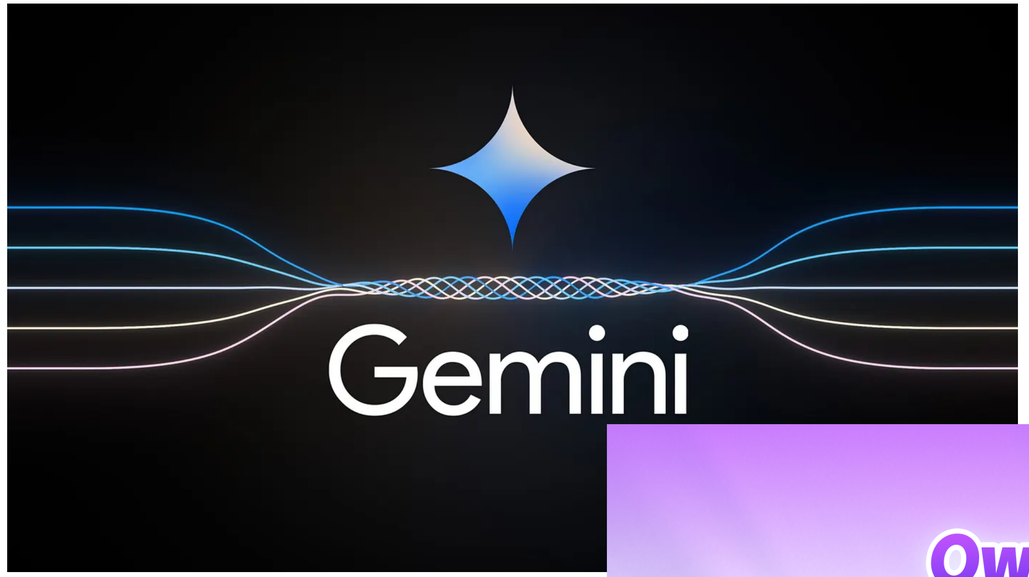
Generated via ChatGPT Image 2.0

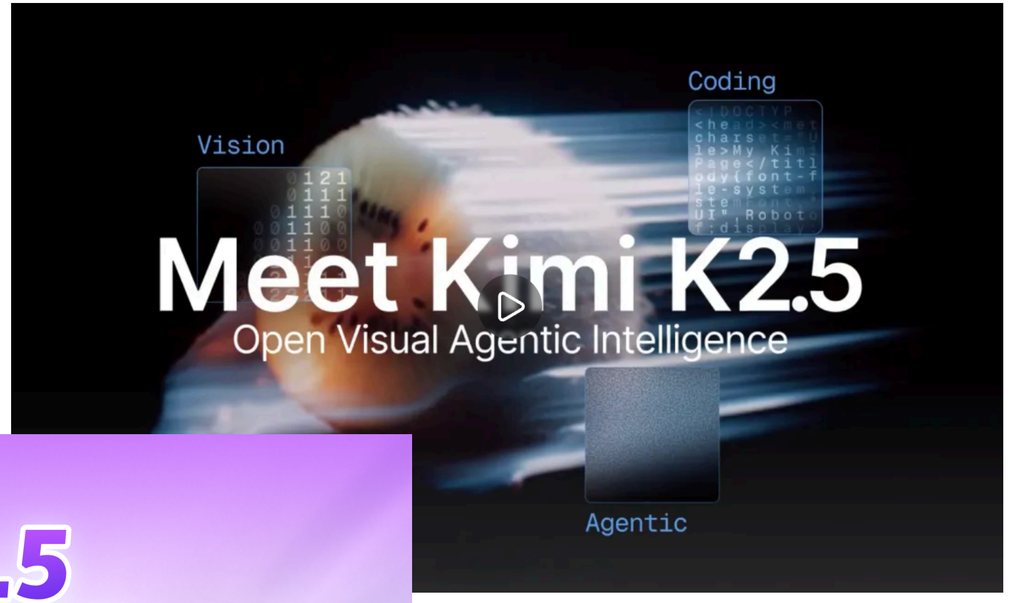
Language modeling alone is not enough…



Generated via ChatGPT Image 2.0

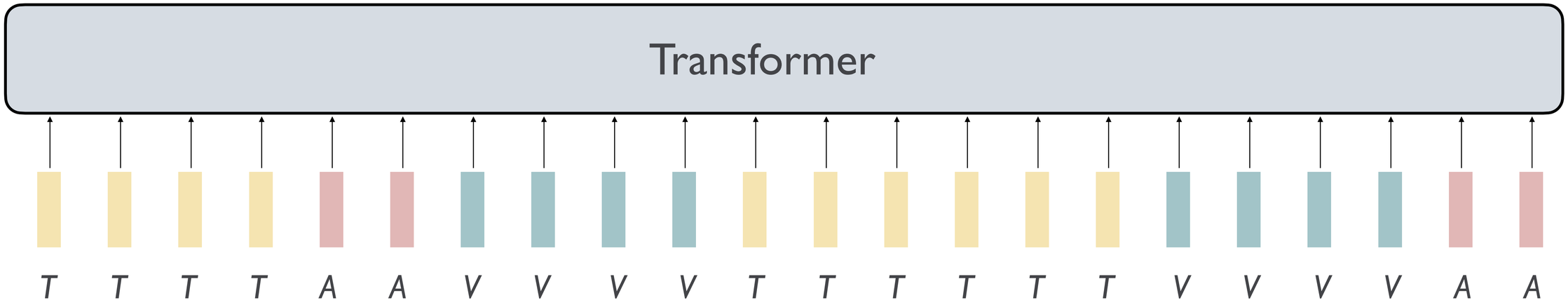
Native Multimodal LLMs







Native Multimodal LLMs

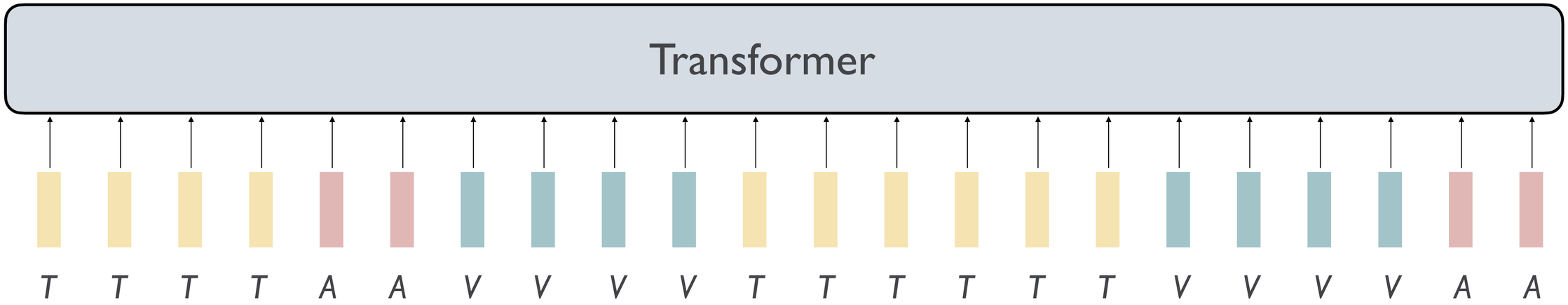


T - text token *V - image patch* *A - audio frame*

Native Multimodal LLMs

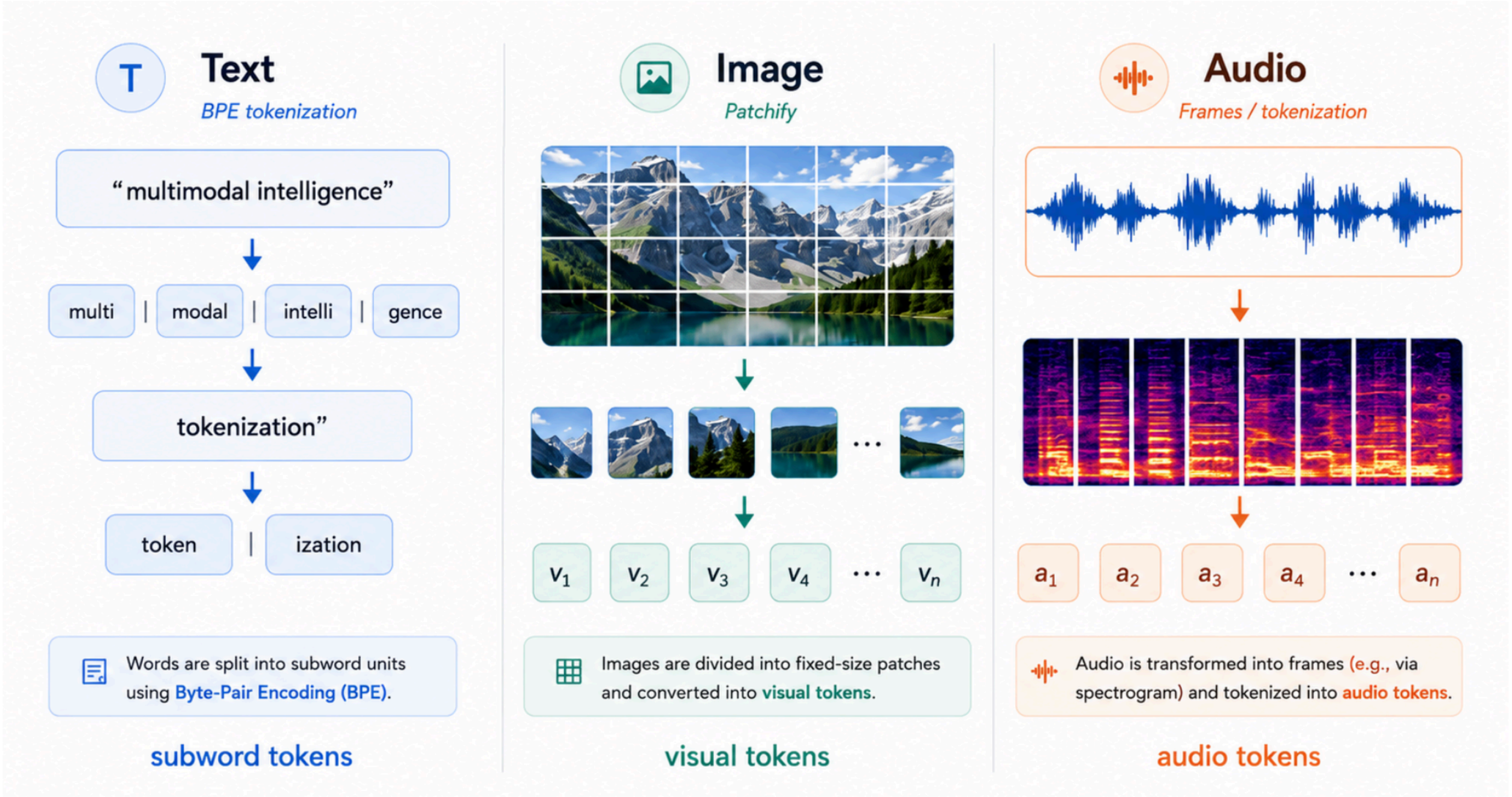
1. "Tokenization" across the board

2. Globally autoregressive generation



T - text token *V - image patch* *A - audio frame*

Multimodal Tokenization



Generated via ChatGPT Image 2.0

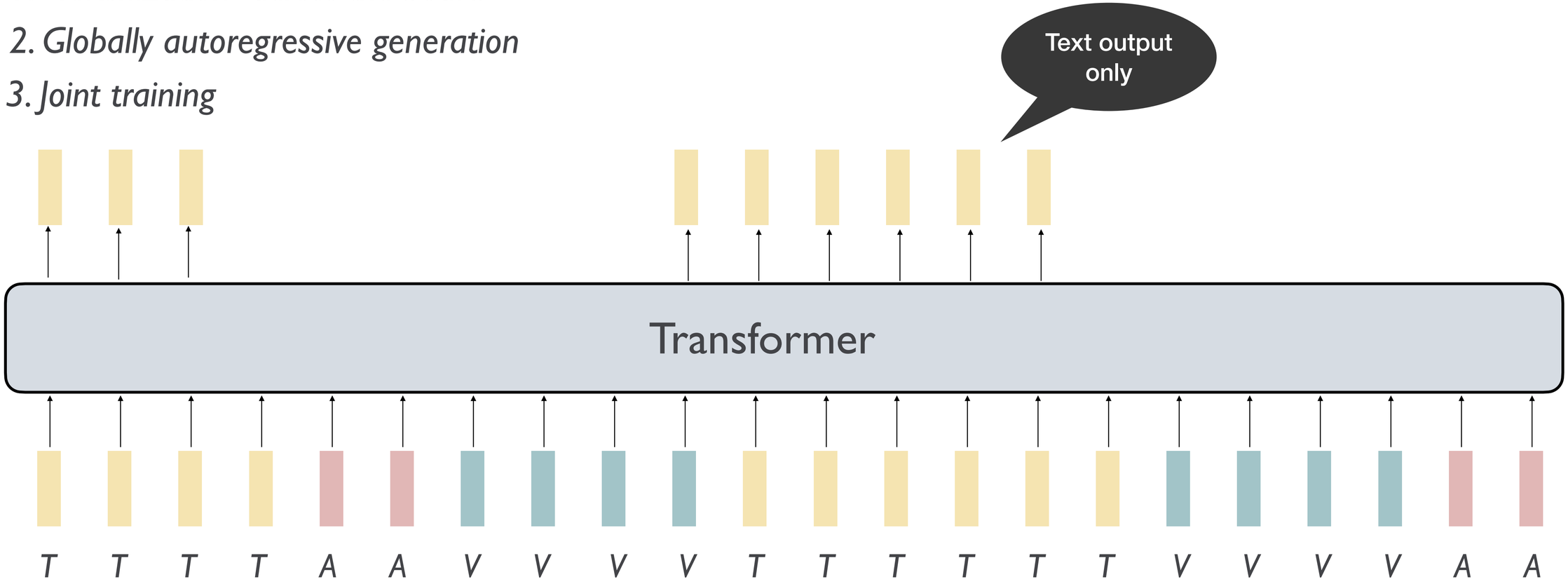
Multimodal Tokenization



Generated via ChatGPT Image 2.0

Training Objective

1. *"Tokenization" across the board*
2. *Globally autoregressive generation*
3. *Joint training*



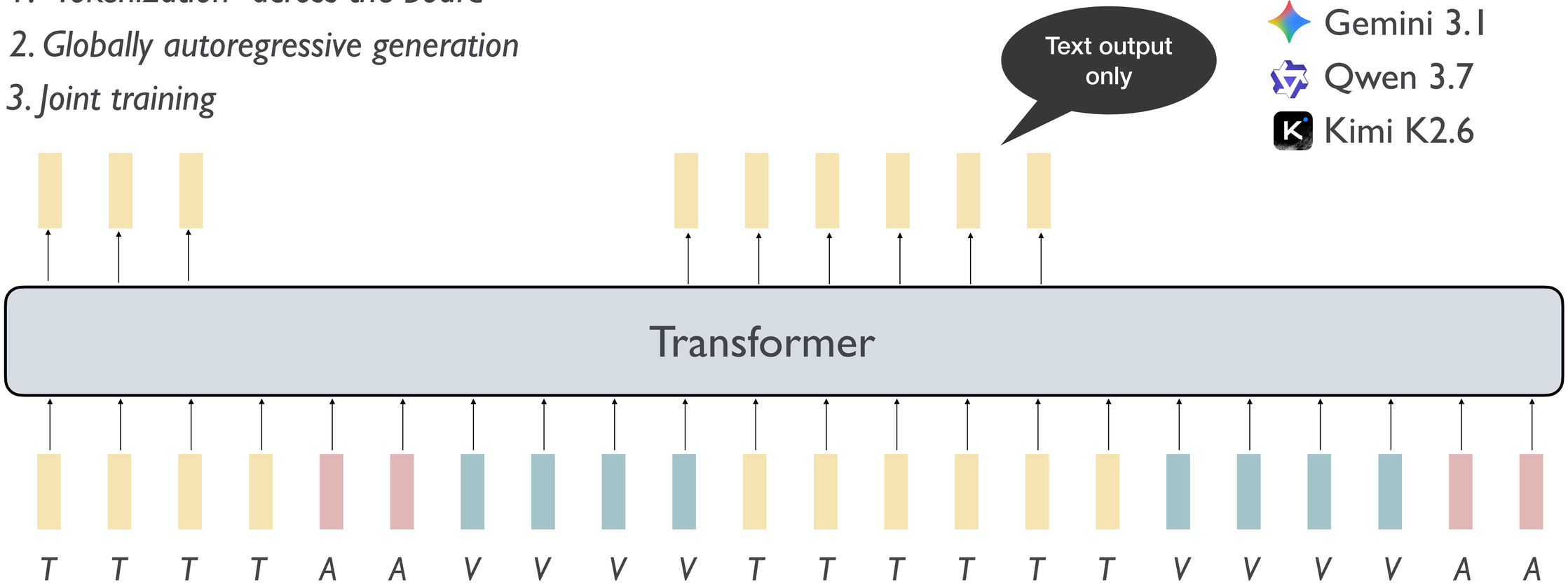
T - text token *V - image patch* *A - audio frame*

Training Objective

1. "Tokenization" across the board

2. Globally autoregressive generation

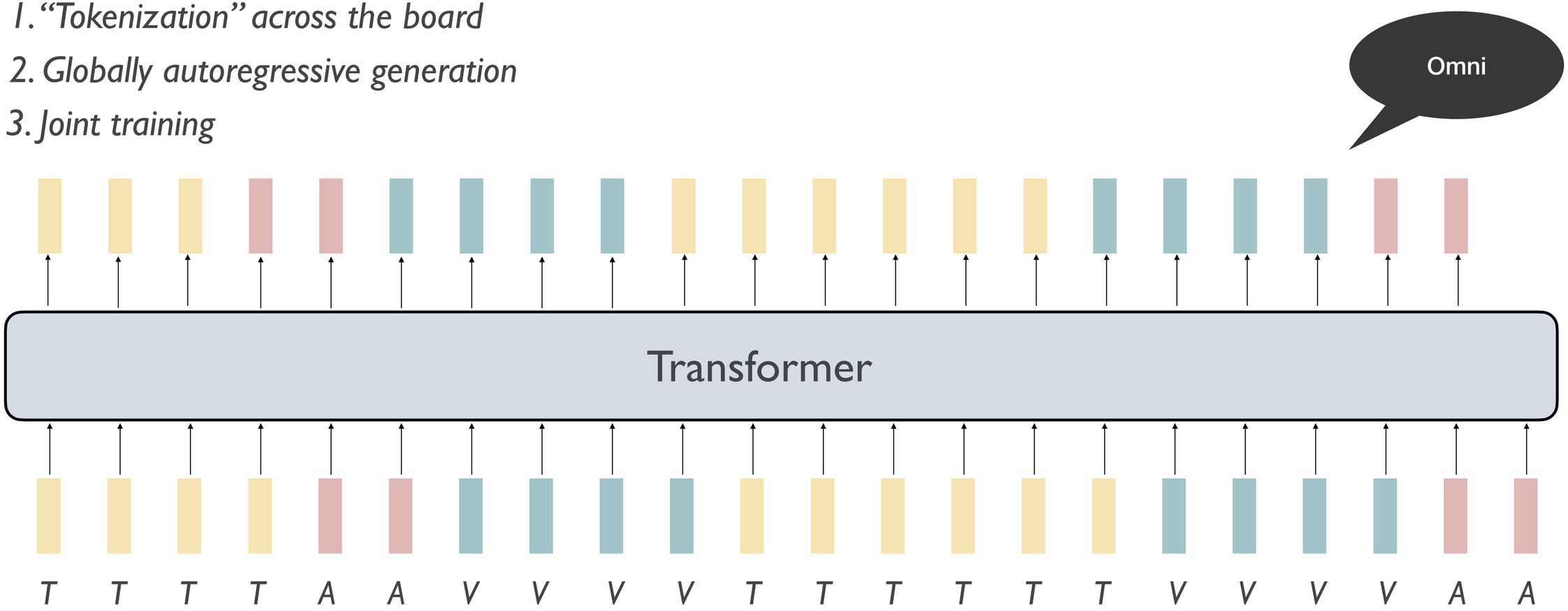
3. Joint training



T - text token *V - image patch* *A - audio frame*

Training Objective

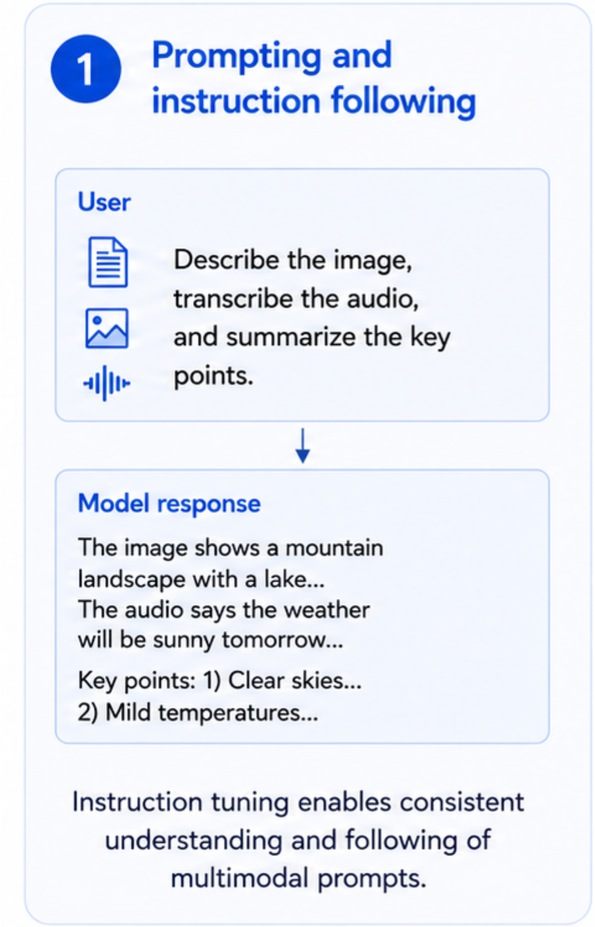
1. *"Tokenization" across the board*
2. *Globally autoregressive generation*
3. *Joint training*



T - text token *V - image patch* *A - audio frame*

A Scaling Approach to Multimodality

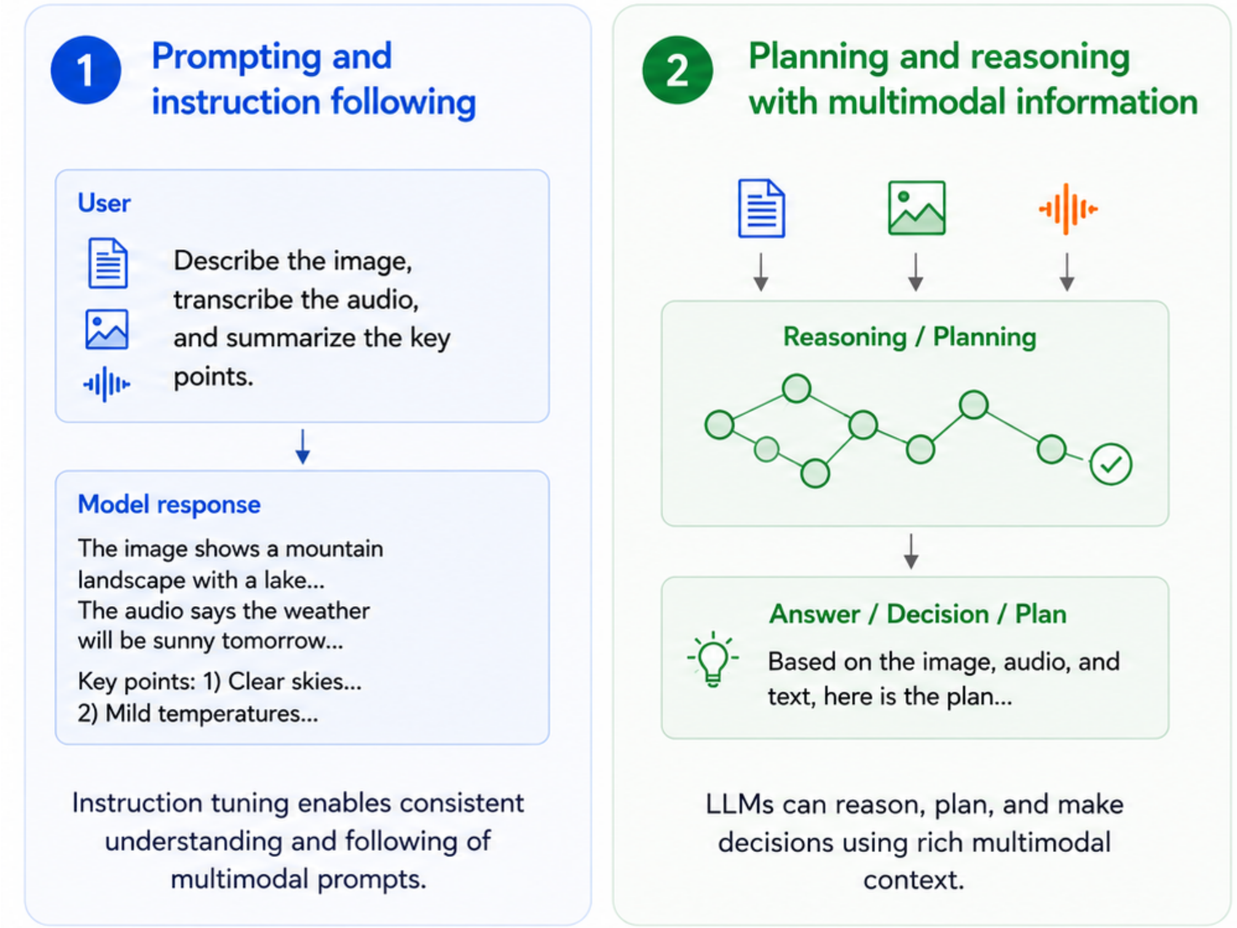
Architectural and training principles transferred from the LLM paradigm



Generated via ChatGPT Image 2.0

A Scaling Approach to Multimodality

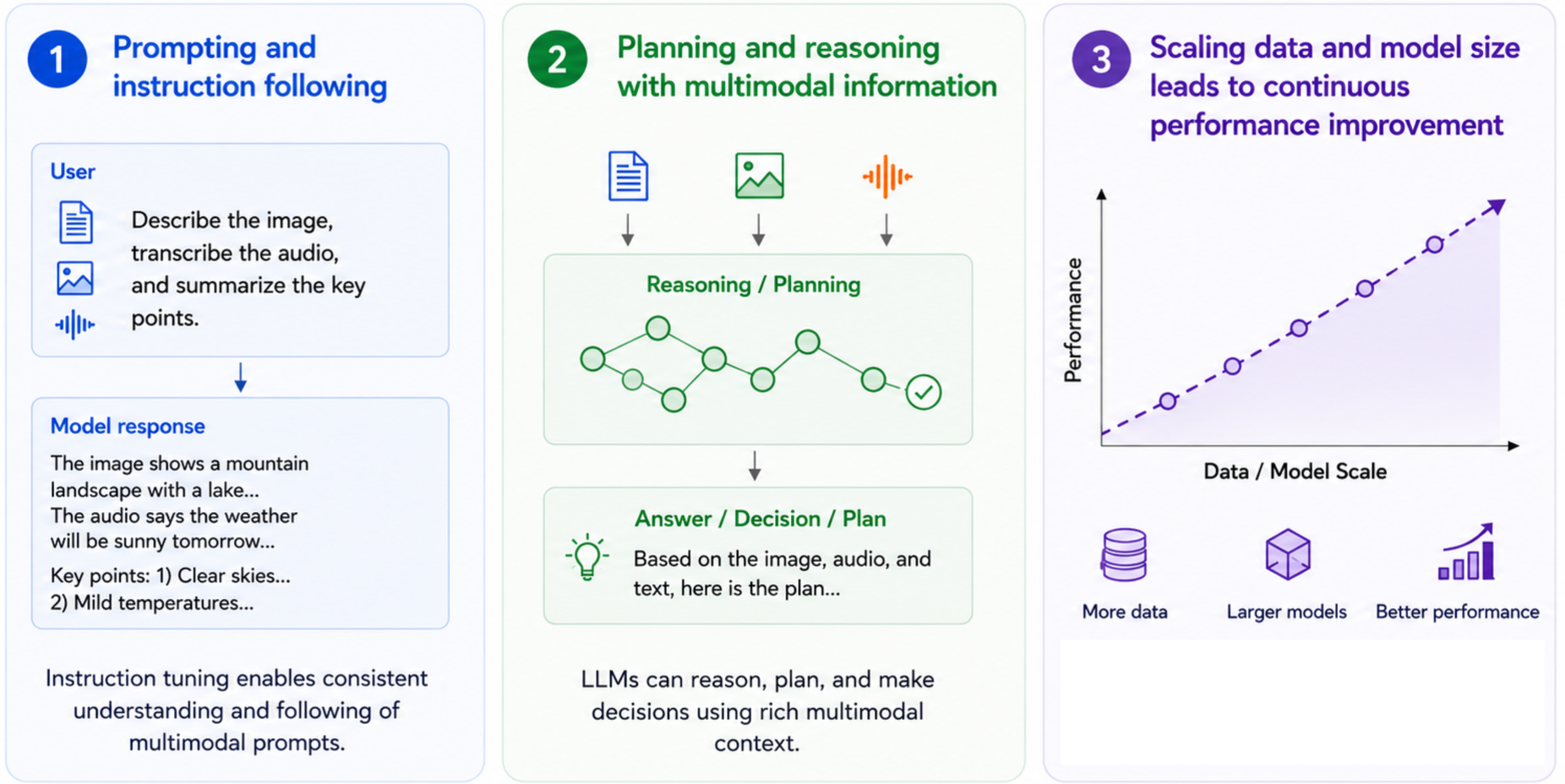
Architectural and training principles transferred from the LLM paradigm



Generated via ChatGPT Image 2.0

A Scaling Approach to Multimodality

Architectural and training principles transferred from the LLM paradigm



Generated via ChatGPT Image 2.0

A Scaling Approach to Multimodality

Architectural and training principles transferred from the LLM paradigm

1 Prompting and instruction following

User

Describe the image, transcribe the audio, and summarize the key points.

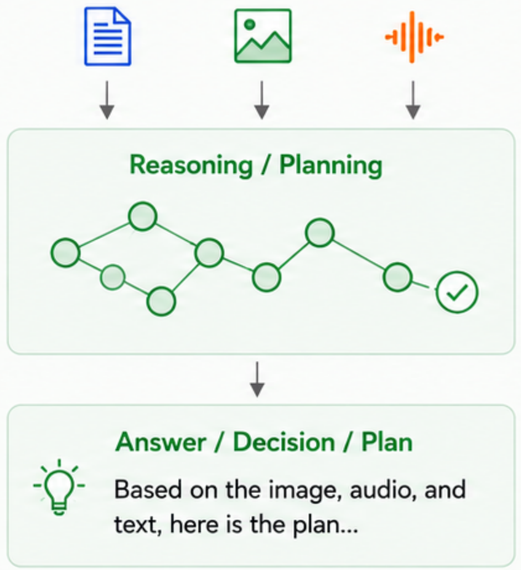
Model response

The image shows a mountain landscape with a lake...
The audio says the weather will be sunny tomorrow...

Key points: 1) Clear skies...
2) Mild temperatures...

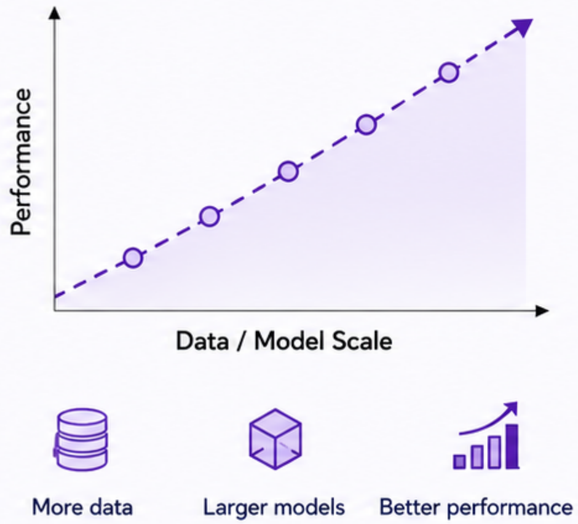
Instruction tuning enables consistent understanding and following of multimodal prompts.

2 Planning and reasoning with multimodal information

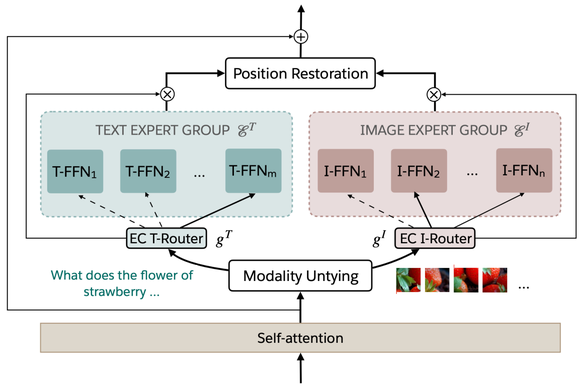


LLMs can reason, plan, and make decisions using rich multimodal context.

3 Scaling data and model size leads to continuous performance improvement



4 Architectural scaling techniques (e.g., MoE) and its variations



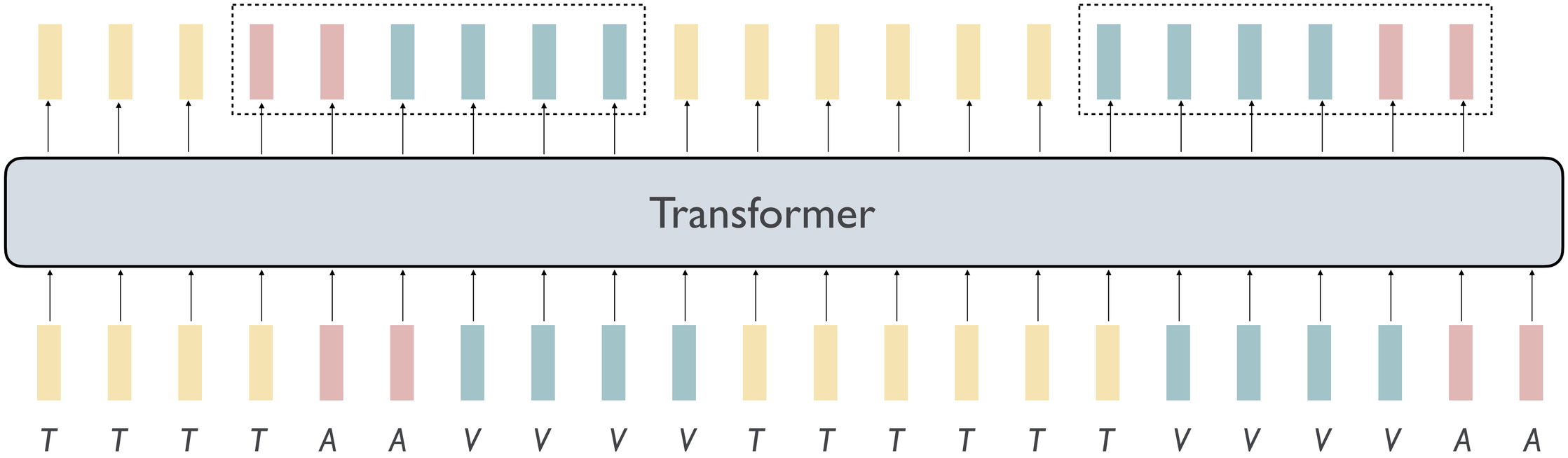
Techniques like Mixture-of-Experts, attention sinks, and sparse routing enable efficient scaling.

Generated via ChatGPT Image 2.0

Omni Models

Key research questions:

1. How to generate the non-text modalities?



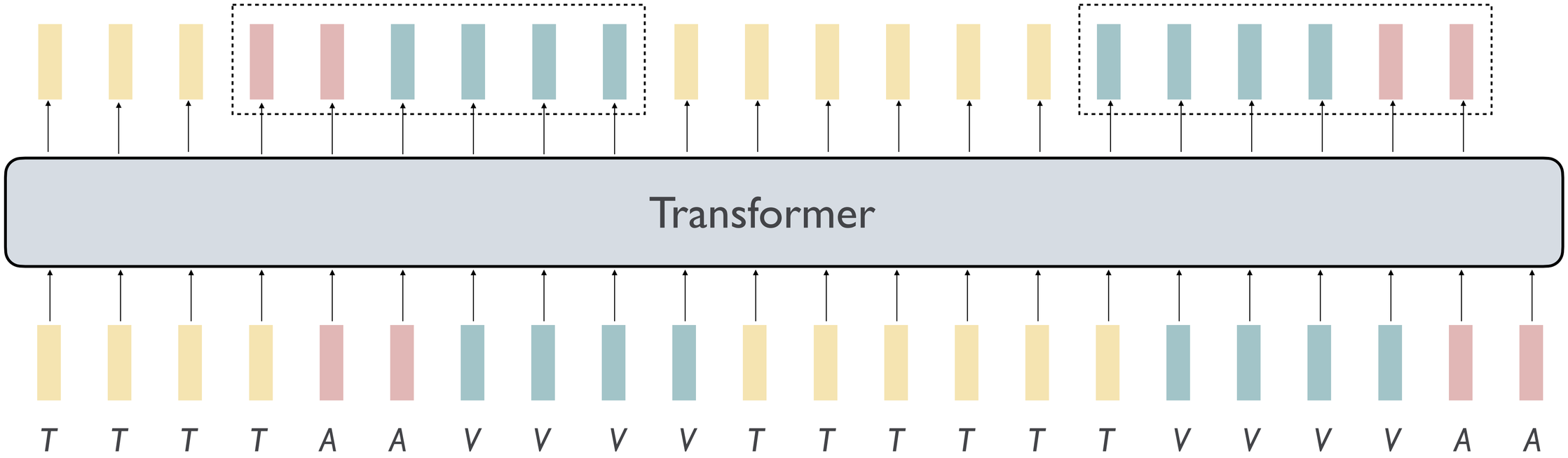
T - text token *V - image patch* *A - audio frame*

Omni Models

Key research questions:

1. How to generate the non-text modalities?

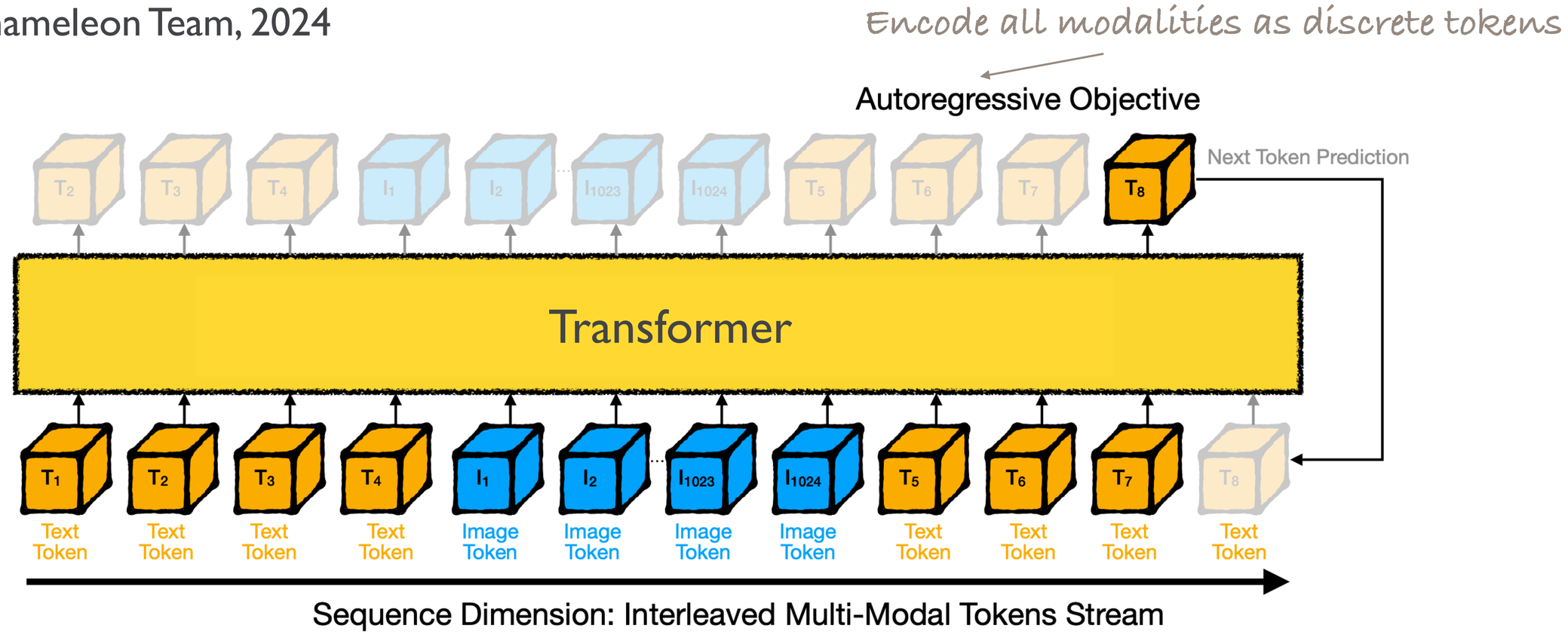
2. Is there positive transfer between understanding and generation of the non-text modalities?



T - text token *V - image patch* *A - audio frame*

Chameleon

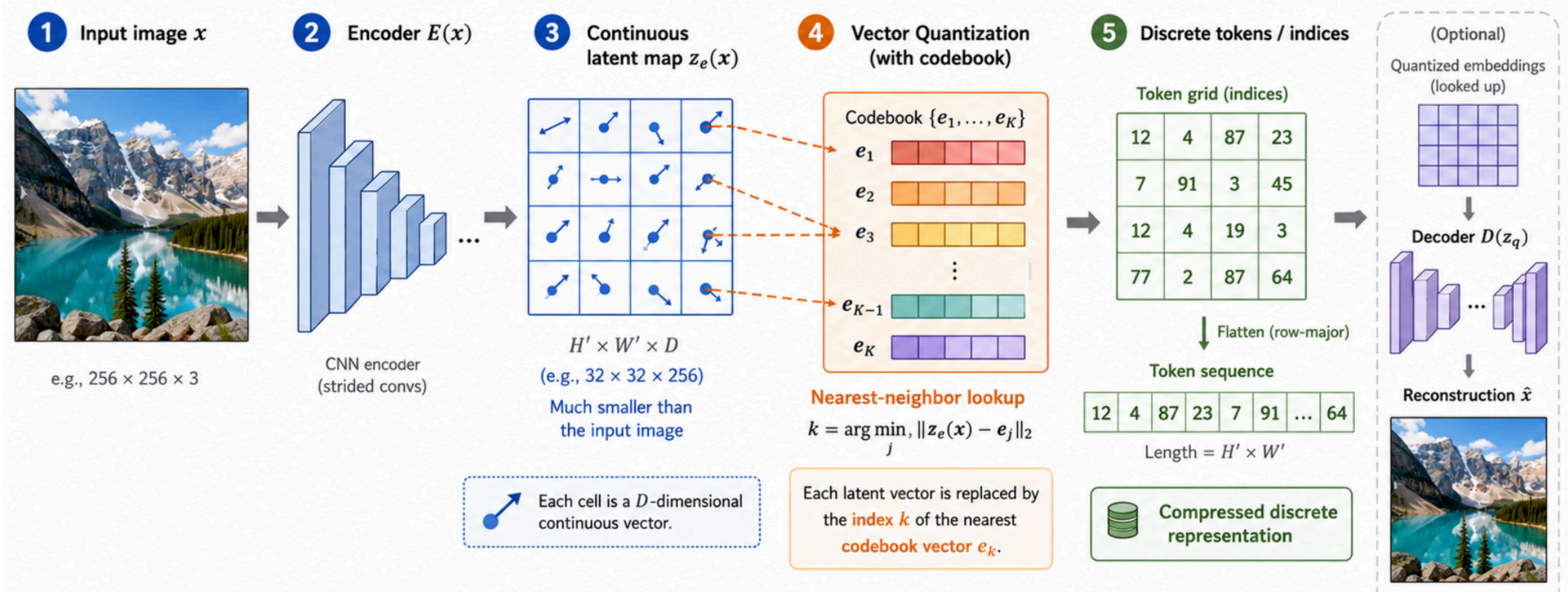
Chameleon Team, 2024



Chameleon: Mixed-Modal Early-Fusion Foundation Models. Chameleon Team. 2024

Chameleon

Vector Quantized Variational AutoEncoder (VQ-VAE)



Generating Diverse High-Fidelity Images with VQ-VAE-2. Razavi et al. 2019.

Generated via ChatGPT Image 2.0

Chameleon

Interleaved text-image generation in arbitrary order

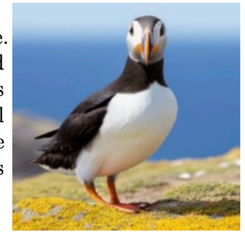
Prompt: I'm bored. Could you show me some cool, quirky-looking birds? Also tell me about them in 2–3 sentences.

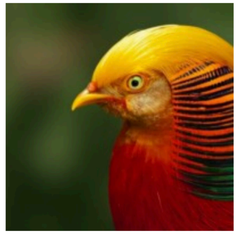


Response: Absolutely! Here are three birds with some unique features that you might find interesting:

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.

Chameleon: Mixed-Modal Early-Fusion Foundation Models. Chameleon Team. 2024

Chameleon

Multi-tasking

Advice: 10.2%

What does a meningitis rash look like? What are the other symptoms I should be on the lookout for?

How-to: 12.5%

How do I properly clean my TV screen? I used Windex and now there are towel fibers and wipe marks all over. Show me some reference photos.

Explanation: 14.4%

I've been studying classical French art, and my favorite so far is his painting seen here: **<img>** Could you please give me a few images of other contemporary artworks that have this same aesthetic?



Hypothetical: 5.6%

What would the modern-day vehicle look like if oil had never been discovered?

Brainstorming: 18.6%

Show me a Middle Eastern alternative to these dishes. **<img1> <img2>**





Article: 3.1%

Write me an introduction to a story about knick-knacks, and finish the story by shifting the focus with an image.

Story: 3.9%

Can you create and illustrate a short story for children about an octopus that can't stop eating pizza?

Identification: 9.3 %



Is the below image a Shetland Pony? If not, what is it, and can you show me a Shetland Pony? **<img>**

Comparison: 9.6%

Please tell me what the difference between these two creatures is, and show me some more examples. **<img1> <img2>**





Report: 5.4%

Who designed the church in the image below, and what's the name of the Church? **<img>** Can you please provide me with additional photos of famous landmarks designed by the same architect?



Other: 5.2%

Create a decal for my truck that features running horses as well as the TRD insignia. Use black to gray gradients.

Reasoning: 2.1%

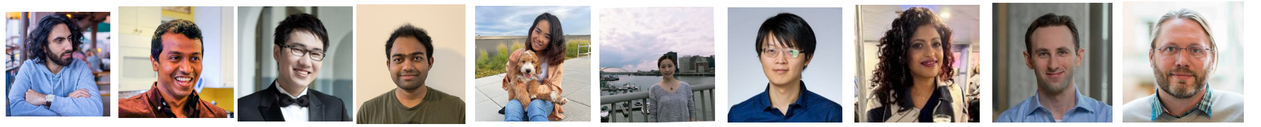
What is typically found at a construction site? Show me a construction site that has a crane.

Chameleon: Mixed-Modal Early-Fusion Foundation Models. Chameleon Team. 2024

Chameleon

Summary

1. **Early-fusion:** One of the first models to show that training from scratch on interleaved text–image sequences can induce multimodal capabilities while preserving strong text-only performance.
2. **Representation limitations:**
    1. VQ-VAE image tokens were later found to be less effective than continuous image embeddings, such as SigLIP, for capturing fine-grained visual semantics. Today's state-of-the-art multimodal LLMs typically use continuous image embeddings as input.
    2. By 2026, diffusion- and flow-matching-based approaches were proven to be more scalable and efficient formulations for image generation.

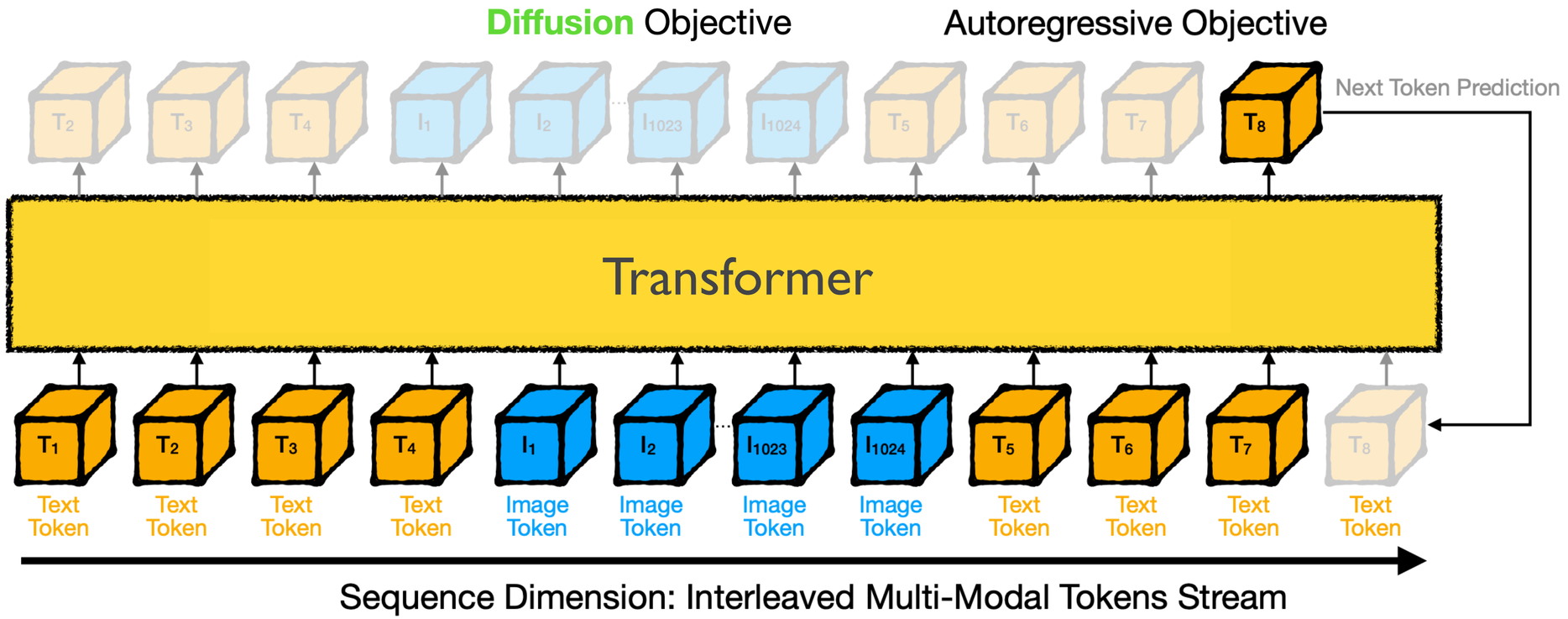


and many others …

Chameleon: Mixed-Modal Early-Fusion Foundation Models. Chameleon Team. 2024

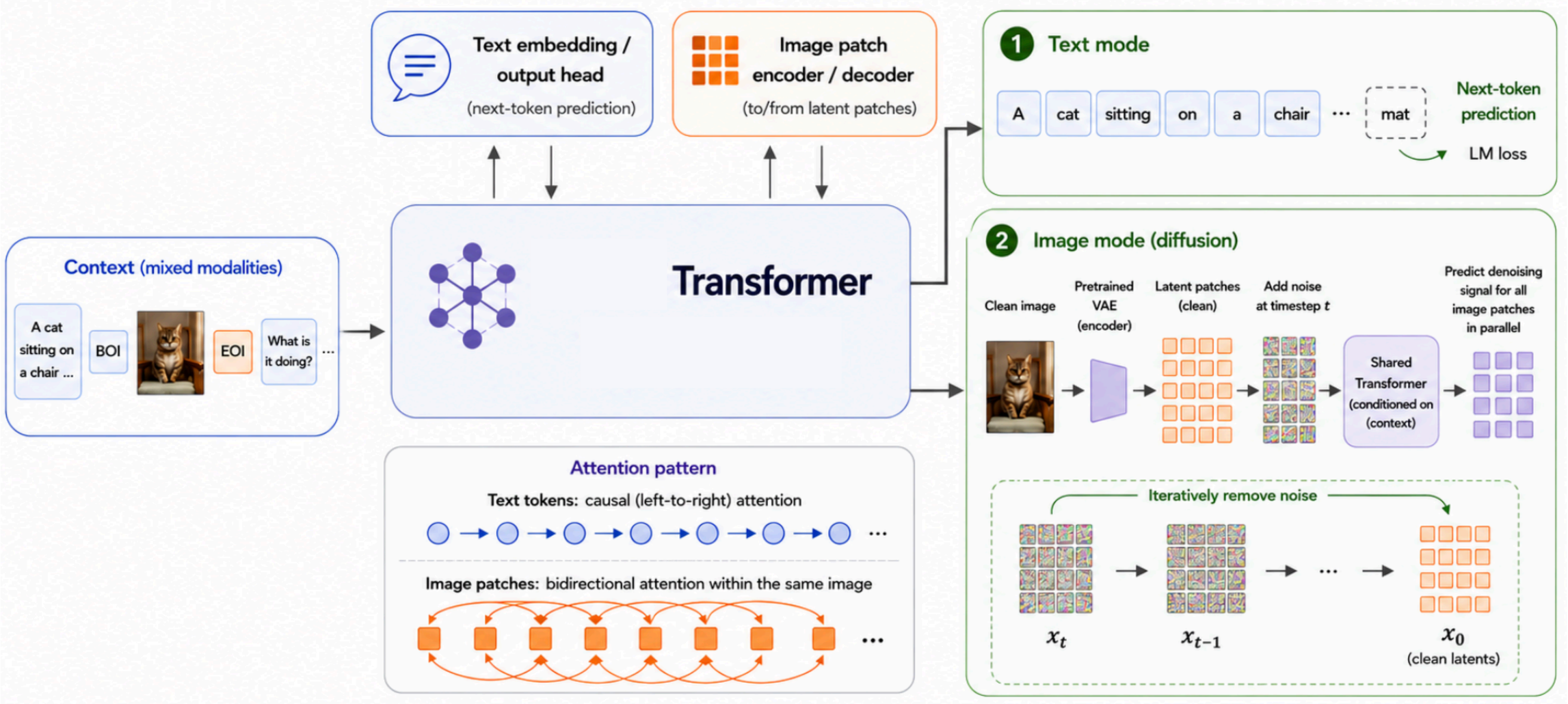
Transfusion

Transfusion (Zhou et al. 2024)



Transfusion: Predict the Next Token and Diffuse Images with One Multi-Modal Model. Zhou et al. 2024.

Transfusion



Transfusion: Predict the Next Token and Diffuse Images with One Multi-Modal Model. Zhou et al. 2024.

Generated via ChatGPT Image 2.0

Transfusion

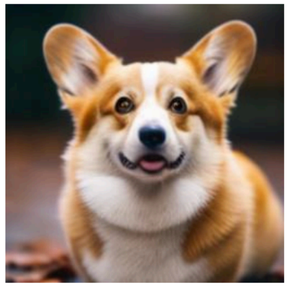
Transfusion (Zhou et al. 2024)



An armchair in the shape of an avocado



A bread, an apple, and a knife on a table



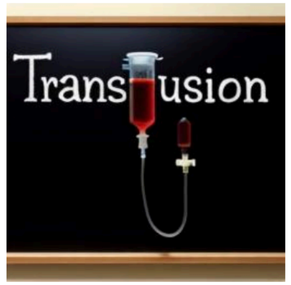
A corgi.



human life depicted entirely out of fractals



A blue jay standing on a large basket of rainbow macarons.



“Transfusion" is written on the blackboard.



A close up photo of a human hand, hand model. High quality



A cloud in the shape of two bunnies playing with a ball. The ball is made of clouds too.

Transfusion: Predict the Next Token and Diffuse Images with One Multi-Modal Model. Zhou et al. 2024.

Transfusion

Summary

1. First approach to seamlessly **unify auto-regressive next-token prediction and diffusion objective** in a single architecture.

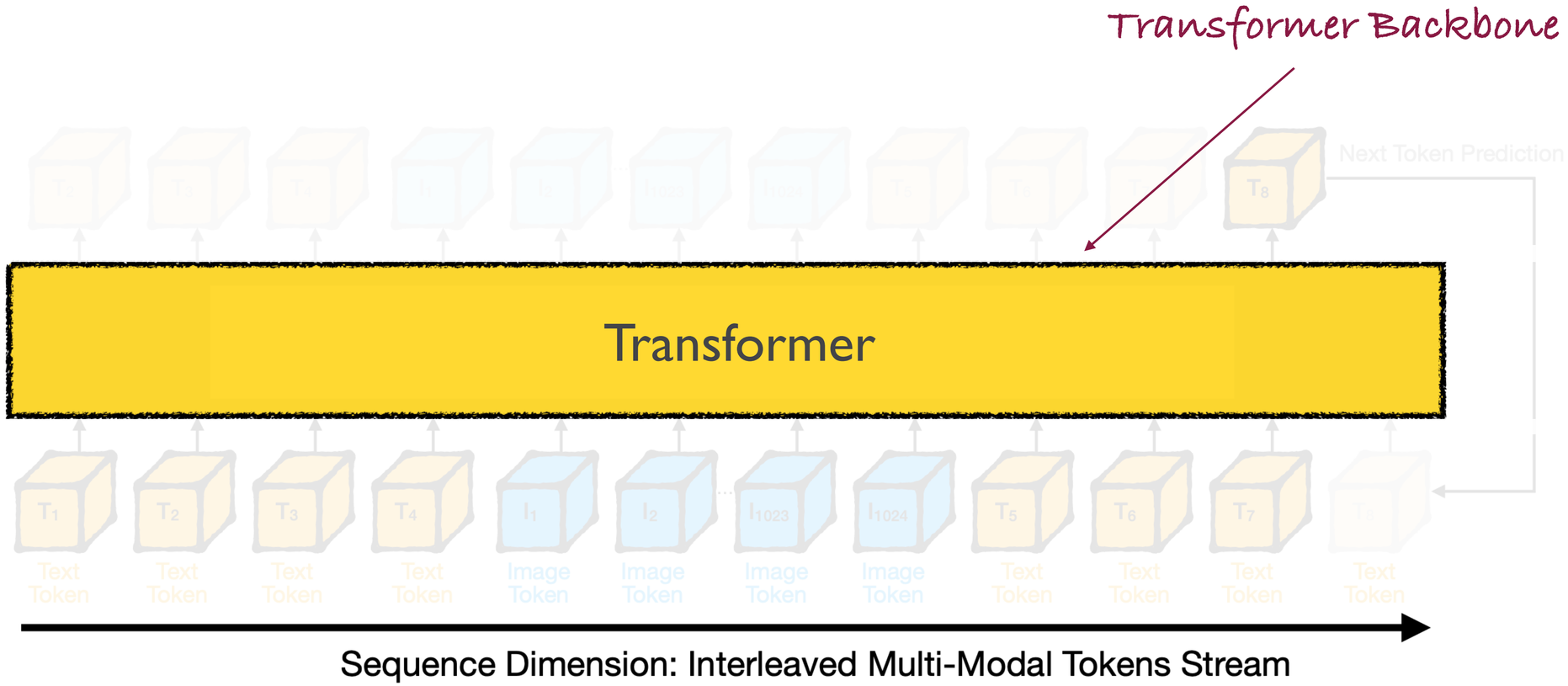
2. Significantly better image generation quality and training efficiency compared to discrete VQ-VAE based generation.

3. **Open research problem:** VAE representations are not effective for visual understanding. State-of-the-art omni models today often adopt separate representations for image understanding and generation respectively while there is active ongoing research in the field to unify them.

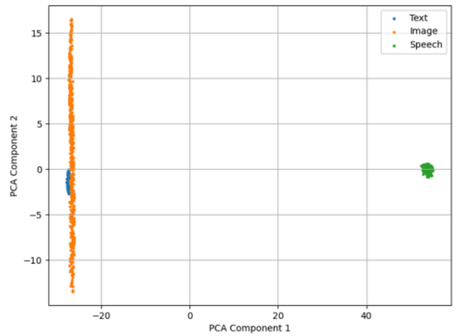


Transfusion: Predict the Next Token and Diffuse Images with One Multi-Modal Model

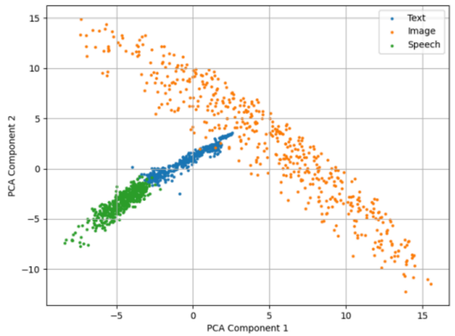
Multimodal Architecture Design



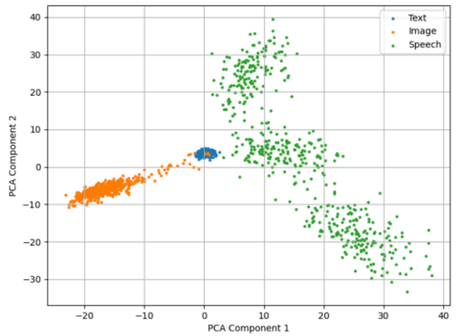
Modality-Aware Sparsity



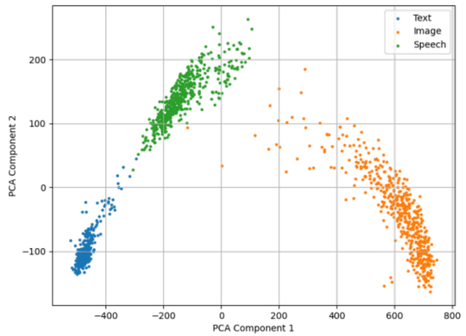
(b) Layer 1



(c) Layer 5



(d) Layer 17



(e) Layer 32

Mind the Gap: Understanding the Modality Gap in Multi-modal Contrastive Representation Learning. Liang et al. 2022.

Mixture-of-Transformers

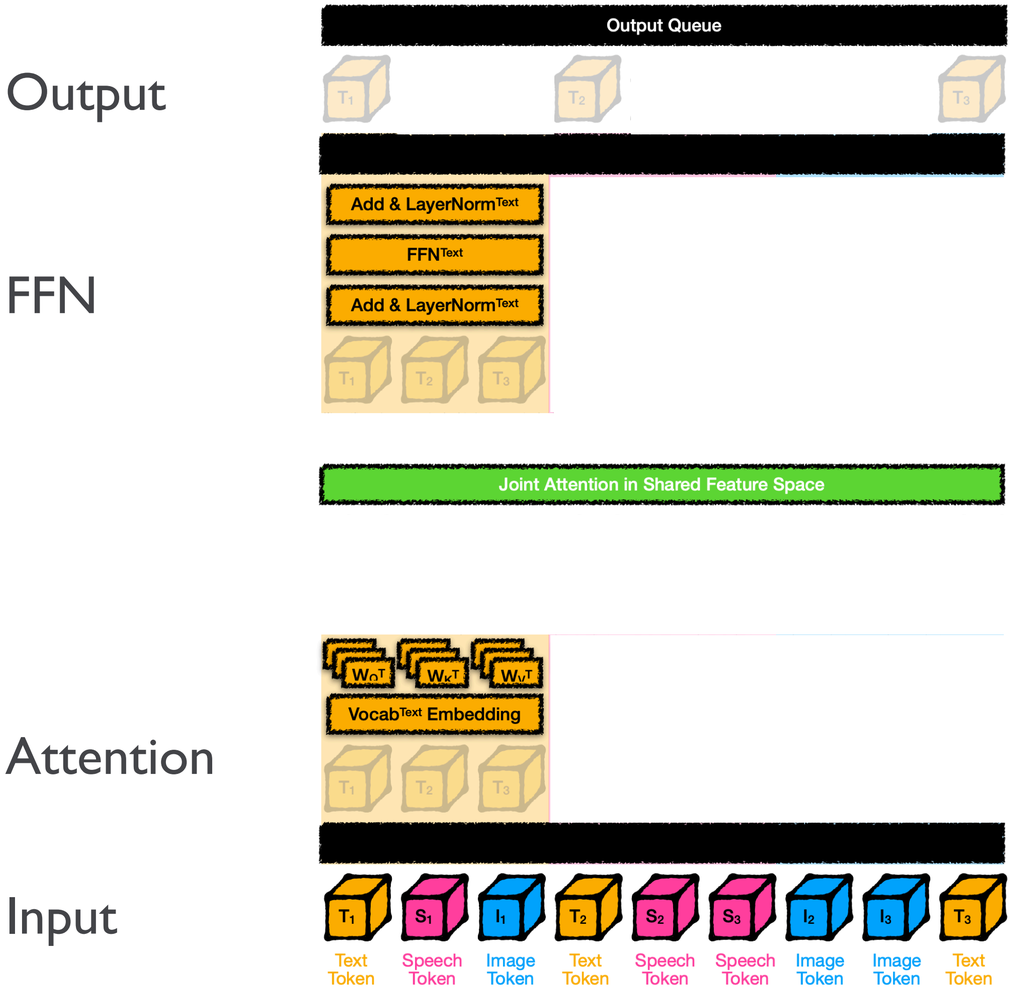


Independent transformer parameters for each modality

Deterministic routing

Mixture-of-Transformers: A Sparse and Scalable Architecture for Multi-Modal Foundation Models. Liang et al. 2024.

Mixture-of-Transformers



Independent transformer parameters for each modality

Deterministic routing

Mixture-of-Transformers: A Sparse and Scalable Architecture for Multi-Modal Foundation Models. Liang et al. 2024.

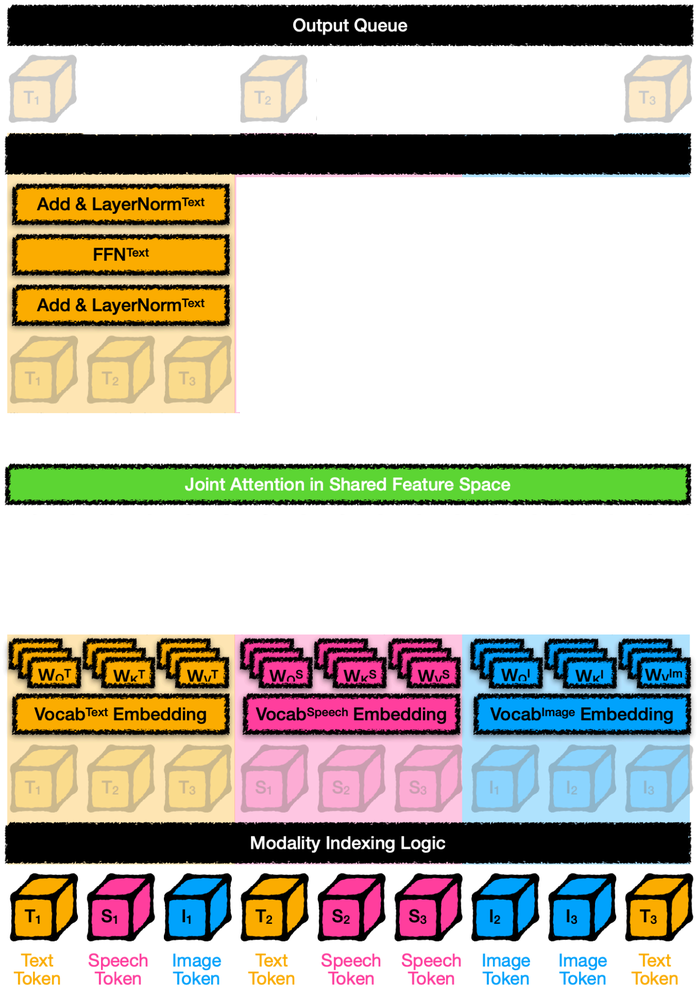
Mixture-of-Transformers

Output

FFN

Attention

Input



Independent transformer parameters for each modality

Deterministic routing

Mixture-of-Transformers: A Sparse and Scalable Architecture for Multi-Modal Foundation Models. Liang et al. 2024.

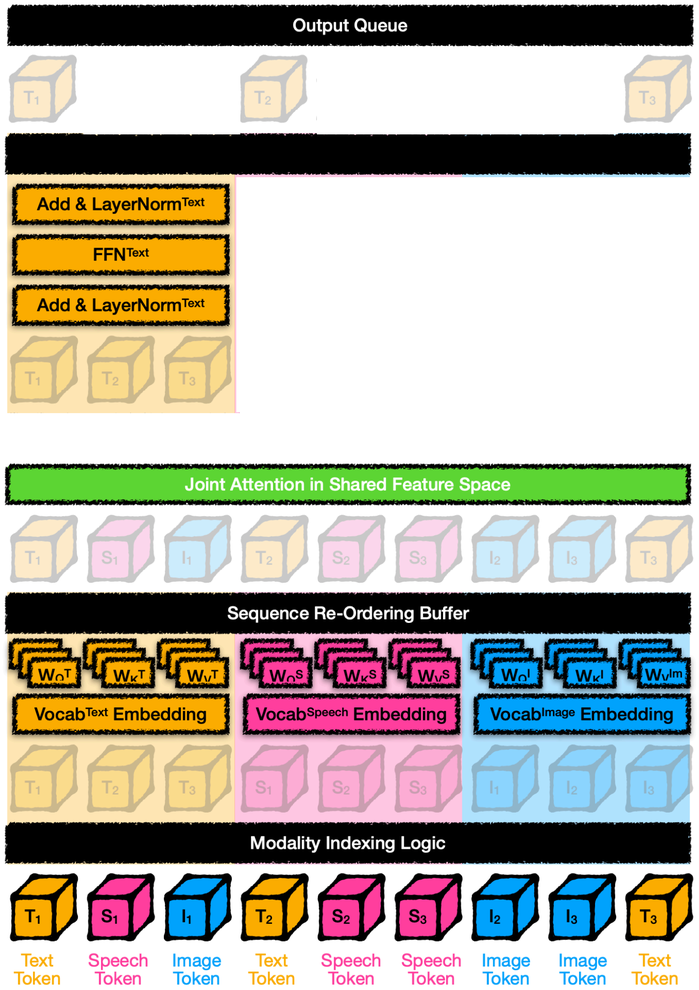
Mixture-of-Transformers

Output

FFN

Attention

Input



Independent transformer parameters for each modality

Deterministic routing

Mixture-of-Transformers: A Sparse and Scalable Architecture for Multi-Modal Foundation Models. Liang et al. 2024.

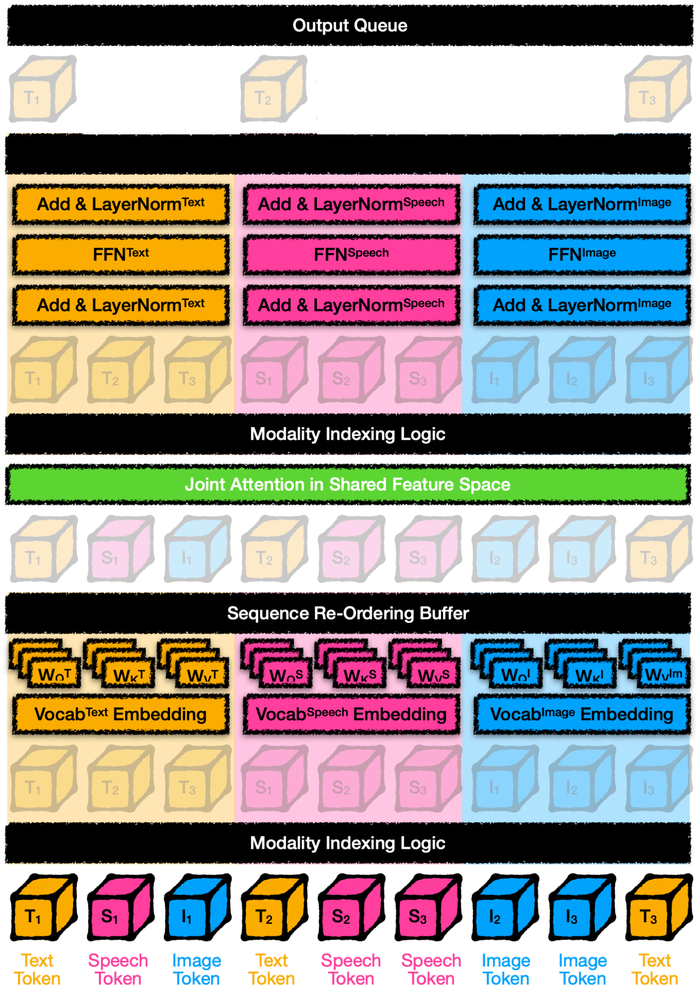
Mixture-of-Transformers

Output

FFN

Attention

Input



Independent transformer parameters for each modality

Deterministic routing

Mixture-of-Transformers: A Sparse and Scalable Architecture for Multi-Modal Foundation Models. Liang et al. 2024.

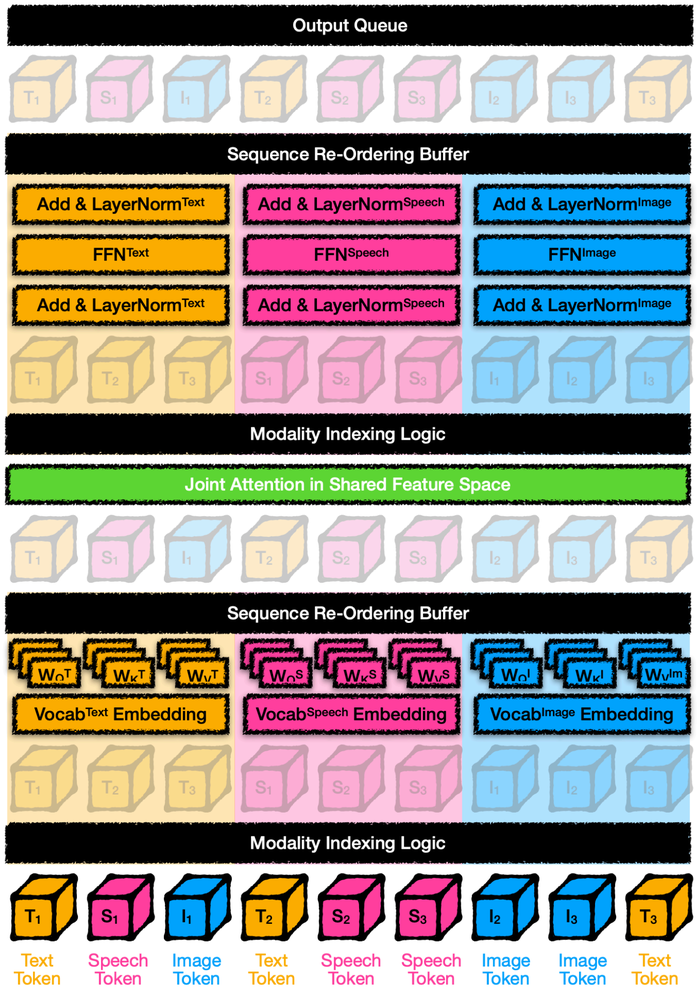
Mixture-of-Transformers

Output

FFN

Attention

Input



Independent transformer parameters for each modality

Deterministic routing

Mixture-of-Transformers: A Sparse and Scalable Architecture for Multi-Modal Foundation Models. Liang et al. 2024.

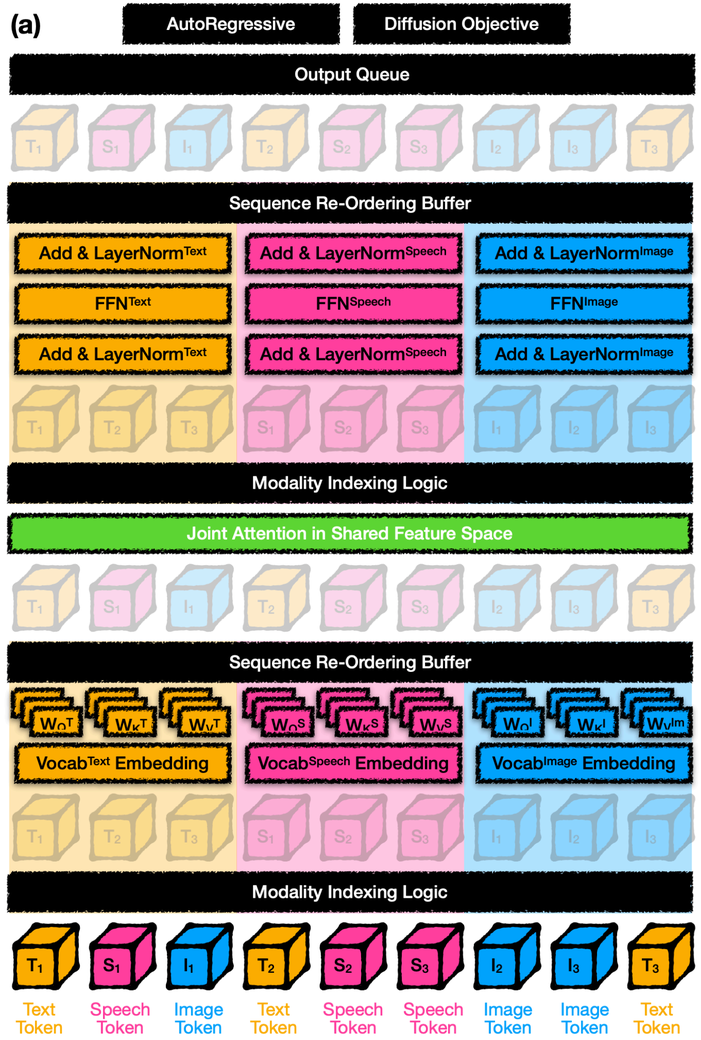
Mixture-of-Transformers

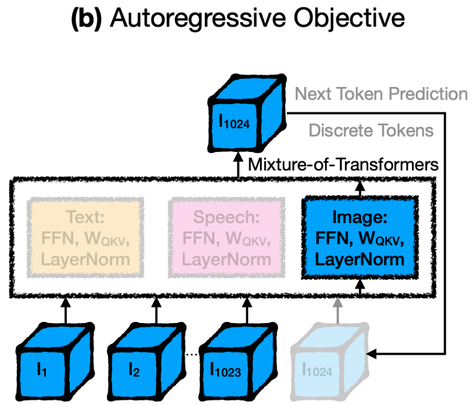
Output

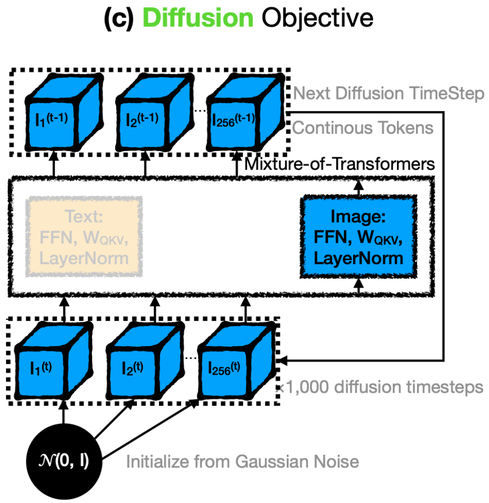
FFN

Attention

Input







Independent transformer parameters for each modality

Deterministic routing

Mixture-of-Transformers: A Sparse and Scalable Architecture for Multi-Modal Foundation Models. Liang et al. 2024.

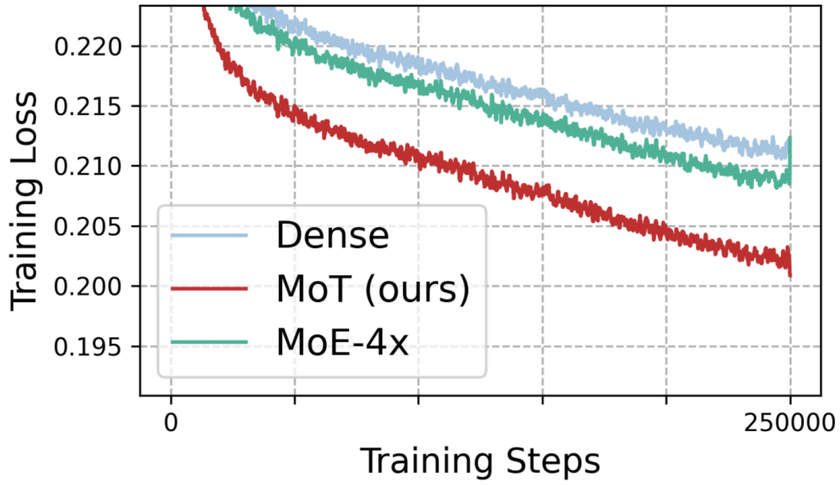
Experiments

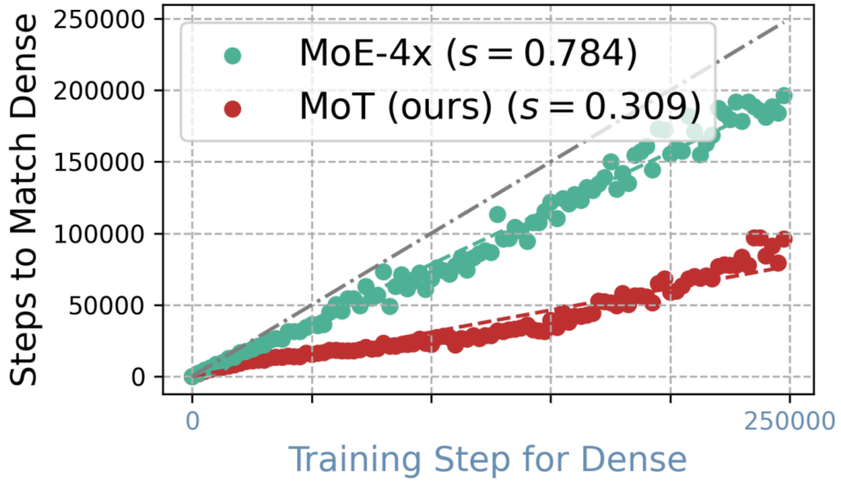
Model Size	Hidden Dim.	Layers	Heads	Seq. Length	Batch Size/GPU	GPUs	Tokens/Batch	Steps	Tokens (T)
163M	768	16	12	4096	4	128	2.10M	250k	0.524
760M	1536	24	24	4096	4	128	2.10M	250k	0.524
1.4B	2048	24	16	4096	2	256	2.10M	250k	0.524
7B	4096	32	32	4096	2	256	2.10N	250k	0.524

Table 4: **Architectural specifications and training configurations of models across different parameter scales (Transfusion setting).** The table lists the hidden dimension, number of transformer layers, attention heads, and sequence length for each model size. Additionally, we provide the batch size used per GPU, the total number of GPUs, training steps, and the corresponding total number of training tokens (in trillions).

Mixture-of-Transformers: A Sparse and Scalable Architecture for Multi-Modal Foundation Models. Liang et al. 2024.

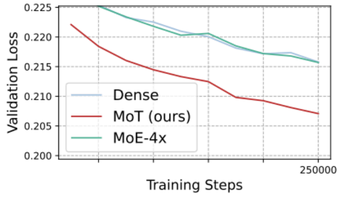
MoT + Transfusion 7B

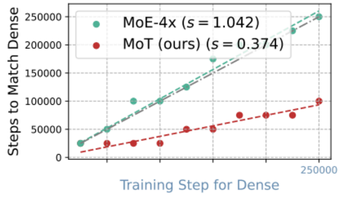


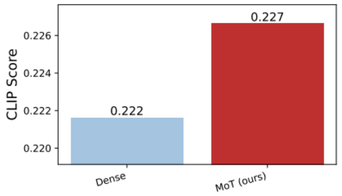


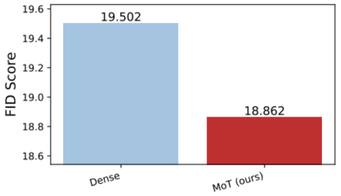
(b) 7B Image Modality Training Loss

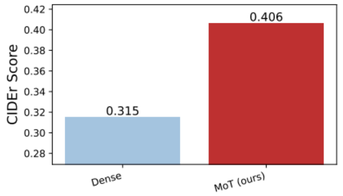
(c) Image Modality Training Loss vs dense steps











(d) Image Modality validation Loss

(e) Image validation Loss vs dense steps

(f) Image Eval: CLIP Score(↑)

(g) Image Eval: FID Score(↓)

(h) Captioning Eval: CIDEr Score(↑)

Mixture-of-Transformers: A Sparse and Scalable Architecture for Multi-Modal Foundation Models. Liang et al. 2024.

MoT + Transfusion 7B



(a) Photo of a lychee-inspired spherical chair, with a bumpy white exterior and plush interior, set against a tropical wallpaper.

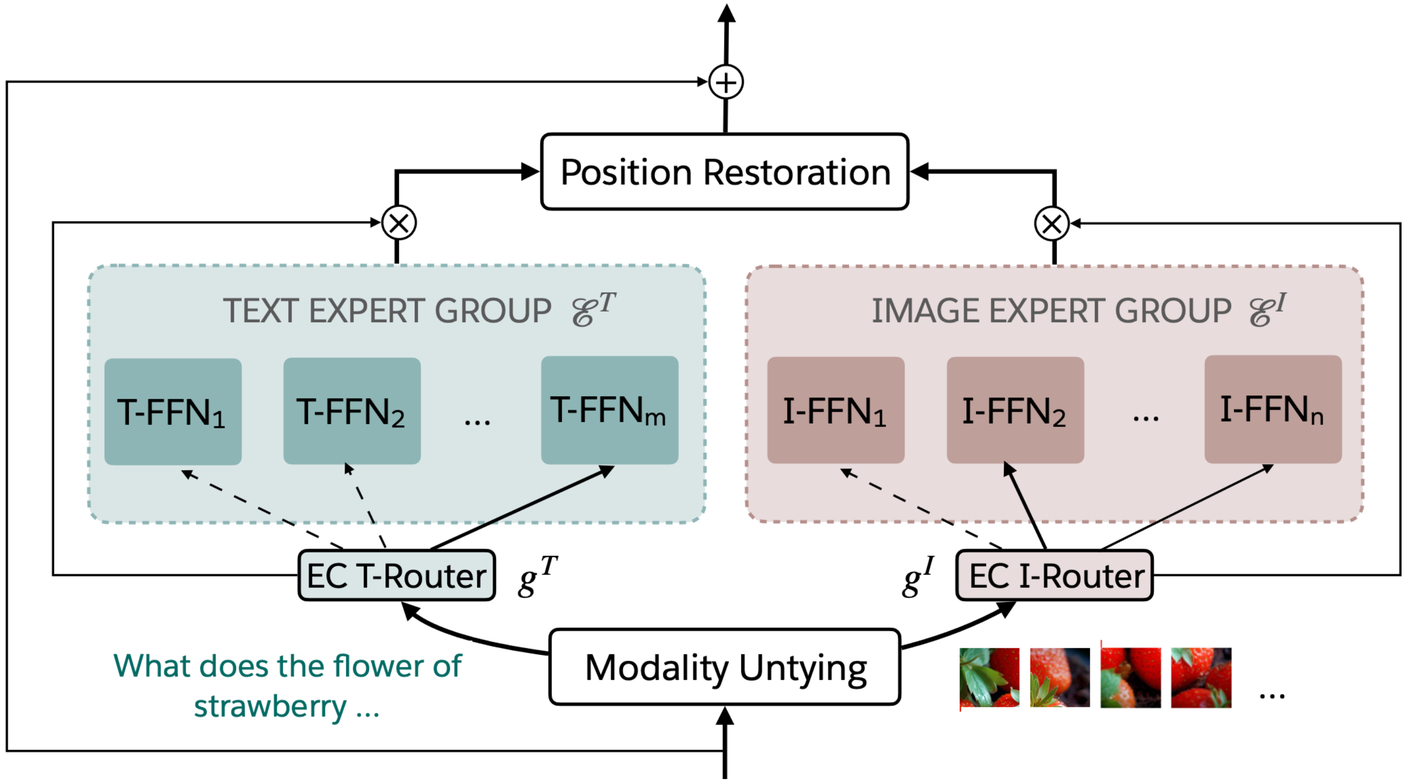
(b) tilt shift aerial photo of a cute city made of sushi on a wooden table in the evening.

(c) Film still of a long-legged cute big-eye anthropomorphic cheeseburger wearing sneakers relaxing on the couch in a sparsely decorated living room.

(d) A chrome-plated duck with a golden beak arguing with an angry turtle in a forest.

Mixture-of-Transformers: A Sparse and Scalable Architecture for Multi-Modal Foundation Models. Liang et al. 2024.

Combination w/ Mixture-of-Experts



Mixture-of-Transformers: A Sparse and Scalable Architecture for Multi-Modal Foundation Models. Liang et al. 2024.

Mixture-of-Transformers (MoT)

Summary

1. MoT uses **modality-specific parameters**, with **deterministic token routing** based on modality.
2. It substantially improves generation quality for non-text modalities.
3. It improves the **efficiency, stability, and controllability of mixed-modal training**, and enables asynchronous training of different modalities across stages *(e.g., LMFusion, Shi et al., 2025).*
4. Since publication, MoT has inspired a growing line of follow-up work on modality-aware architectures.



Mixture-of-Transformers: A Sparse and Scalable Architecture for Multi-Modal Foundation Models. Liang et al. 2024.

MoT in the Wild

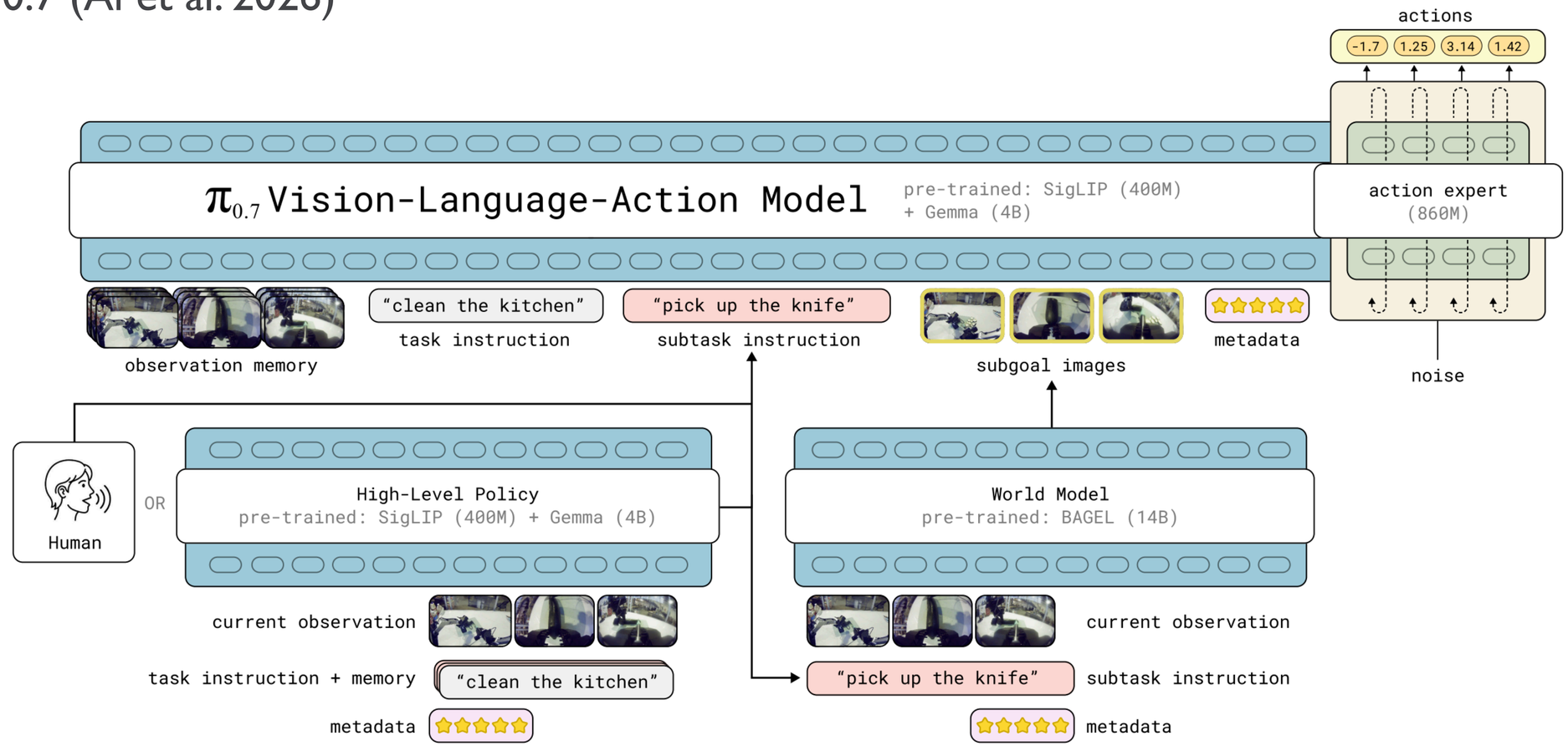
BAGEL (Deng et al. 2025)

Prompts	Results	Results *w/* Thinking	
A car made of small cars.		**<think>** The model should generate an image where numerous small cars are combined to form the shape and structure of a larger car. The refined prompt is: A highly detailed and creative image showcasing a large car composed entirely of numerous small cars, meticulously arranged to replicate the shape and structure of a full-sized vehicle. The small cars are tightly interlocked, creating a seamless and realistic exterior with accurate proportions and details. The setting is neutral and clean, with soft, diffused lighting to emphasize the intricacy of the composition without any distractions. The image captures the concept in a photorealistic style, with a focus on precision and craftsmanship, highlighting the innovative use of scale and form. **</think>**	
A cat is twice as large as the dog next to it.		**<think>** The model should generate an image where the cat appears significantly larger than the dog to emphasize the difference in size. The refined prompt is: A highly detailed and realistic image of a large, majestic cat standing next to a smaller dog, with the cat's size being visually emphasized by its towering presence compared to the dog. The cat should have sleek fur, vibrant colors, and a confident, regal posture, while the dog appears smaller, with a friendly and slightly curious expression. The setting is an outdoor environment with soft natural lighting, warm tones, and a gentle atmosphere. The perspective is slightly low-angle, capturing the contrast in size between the two animals in a visually striking composition, with sharp focus on the cat and a subtle depth of field for the background. **</think>**	

Emerging Properties in Unified Multimodal Pretraining, Deng et al. 2025.

MoT in the Wild

Pi 0.7 (Ai et al. 2026)



Pi0.7: A Steerable Generalist Robotic Foundation Model with Emergent Capabilities. Ai et al. 2026

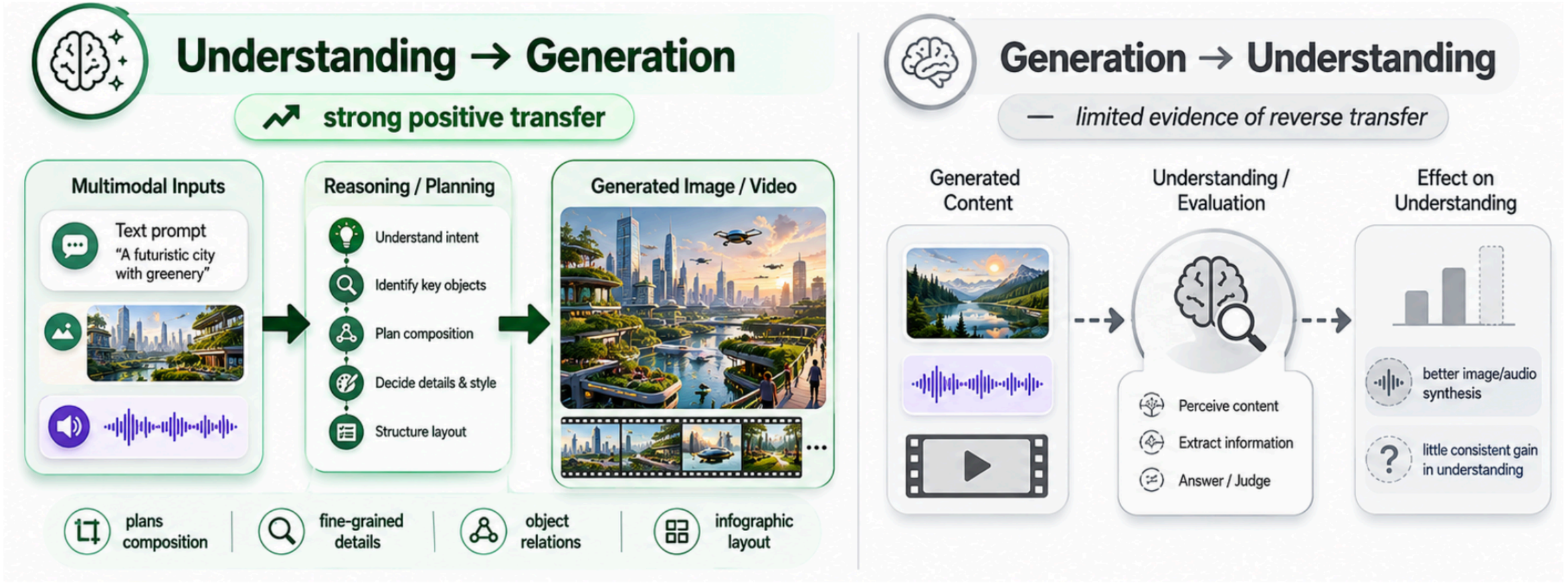
Transfer Between Multimodal Understanding and Generation

Key research questions:

1. How to generate the non-text modalities?

2. Is there positive transfer between understanding and generation of the non-text modalities?

Transfer Between Multimodal Understanding and Generation



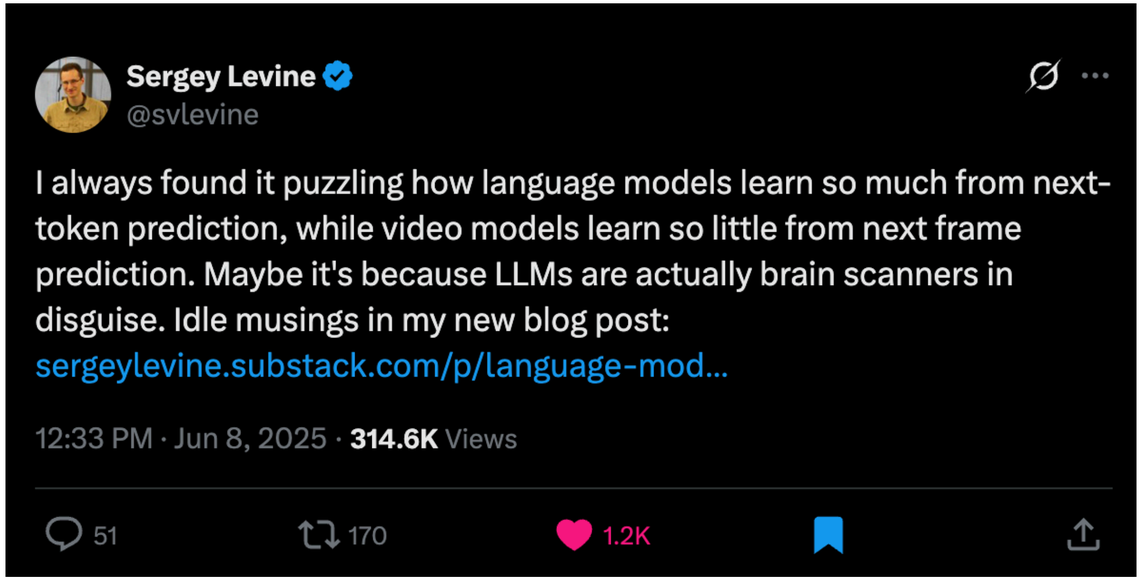
Generated via ChatGPT Image 2.0

Transfer Between Multimodal Understanding and Generation

Key research questions:

1. How to generate the non-text modalities?

2. Is there positive transfer between understanding and generation of the non-text modalities?



Generated via ChatGPT Image 2.0

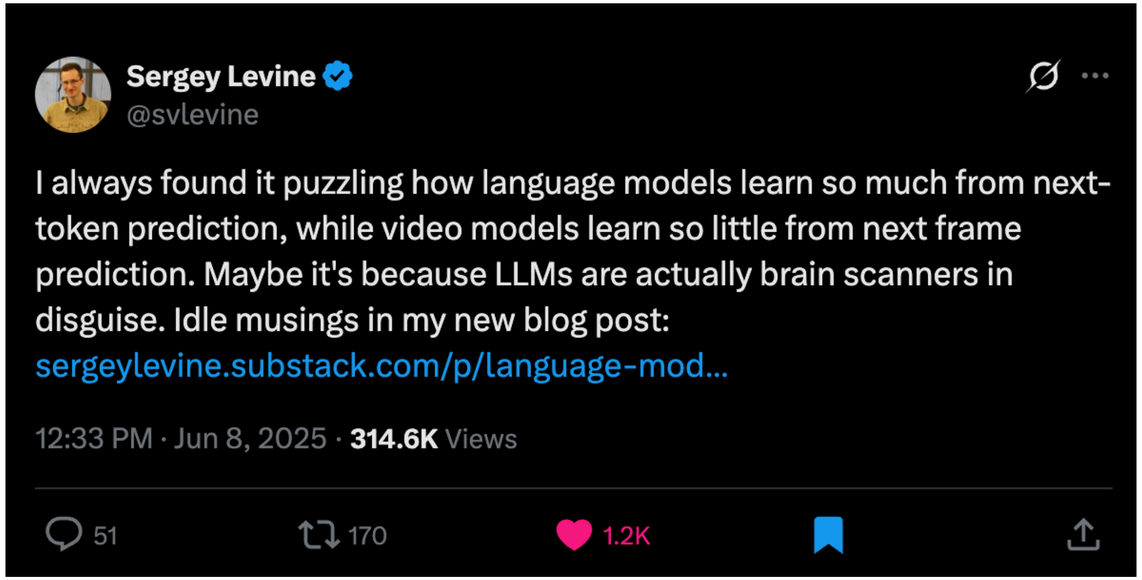
Transfer Between Multimodal Understanding and Generation

Key research questions:

1. How to generate the non-text modalities?

2. Is there positive transfer between understanding and generation of the non-text modalities?

Language is highly compressed abstraction of human cognition



Generated via ChatGPT Image 2.0

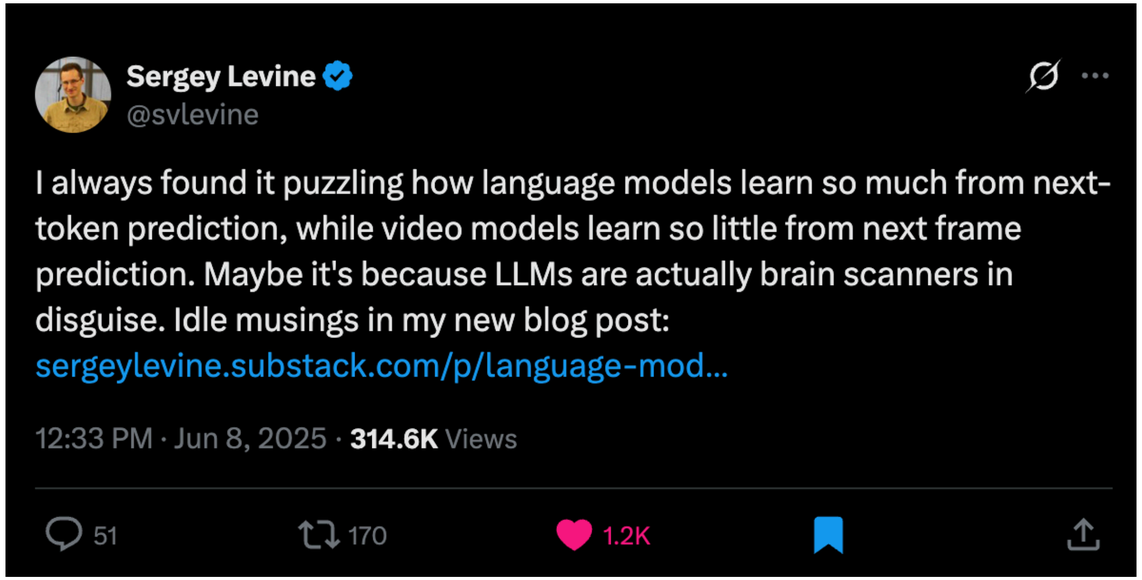
Transfer Between Multimodal Understanding and Generation

Key research questions:

1. How to generate the non-text modalities?

2. Is there positive transfer between understanding and generation of the non-text modalities?

Language is highly compressed abstraction of human cognition



Images/videos are sensory data, passive observation

Poor loss landscape

High redundancy / frame correlation

Generated via ChatGPT Image 2.0

Broader Multimodal Intelligence

The full spectrum of multimodal capabilities spanning digital information to the physical world

Digital / Information World

Well represented in current models

- Text
- Images
- Audio

What current multimodal LLMs do today

A small but important subset

- Chameleon
- Transfusion
- MoT

Cross-modal understanding

Interpret and reason across multiple modalities

Cross-modal generation

Generate content across multiple modalities

What remains largely open (still under-captured)

A much larger frontier of multimodal intelligence

Spatial-temporal understanding

Understand 3D space, movement, and dynamics over time

Real-world grounding

Connect perceptions and concepts to the real world and its constraints

Direct interaction with the physical world

Act on and manipulate objects and environments in the real world

Real-time perception-action

Closed-loop perception and action in real time

Navigation / control

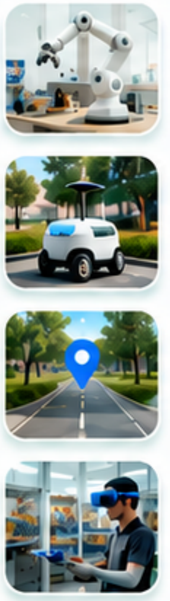
Plan, navigate, and control agents in complex environments

Long-horizon embodied feedback

Learn from long-horizon interactions and outcomes in the physical world

Physical World

Largely under-captured by current models



Today's multimodal LLMs excel at multimodal information processing, but fall far short of full physical-world multimodal intelligence.

Generated via ChatGPT Image 2.0

Interaction Models

https://thinkingmachines.ai/blog/interaction-models/





Interaction Models

https://thinkingmachines.ai/blog/interaction-models/





Interaction Models

Summary

1. **Natively multi-modal:** perceives visual, auditorial and textual information jointly; generates text and audio
2. **Multi-stream, micro-turn design** unlocks capabilities that current turn-based models struggle to express, such as continuously seeing and hearing and response in real-time.
3. **System I & II design** which delegates challenging, long-horizon work to a background model.
4. **Interactivity Research Grant:** https://thinkingmachines.ai/news/interactivity-research-grants/



Conclusion

1. Building models with **native multimodal capabilities** is a very active subfield with lots of open problems.
2. Today's multimodal LLMs excel at (digital) multimodal information processing, but still **fall far short of full physical-world multimodal intelligence**.
3. Multimodal models are computational heavy, raising additional challenges for training and inference infrastructure.
4. We will likely see an increasing number of **multimodal models specialized for certain capabilities** (e.g. embodiment vs. infographics understanding), given the complexity of the problem space.