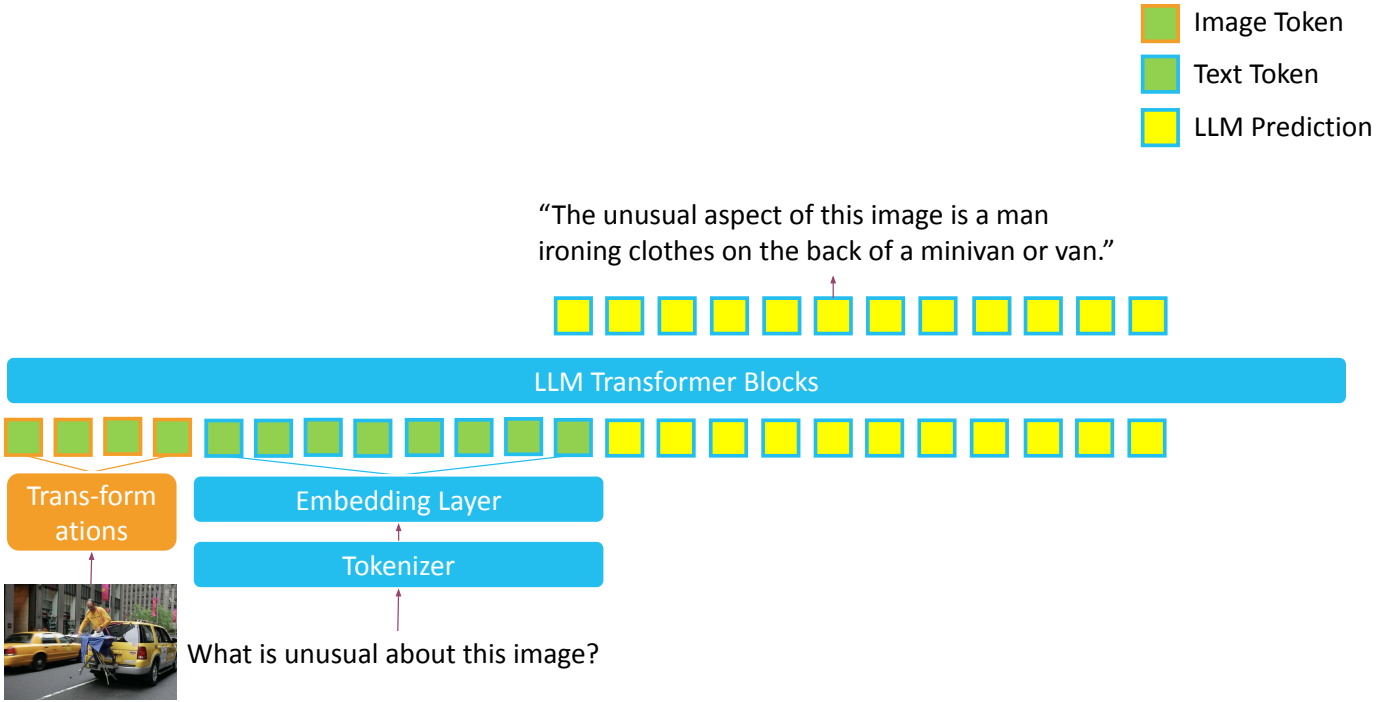
Vision-Language Models

Junyi Zhang & Colin Wang

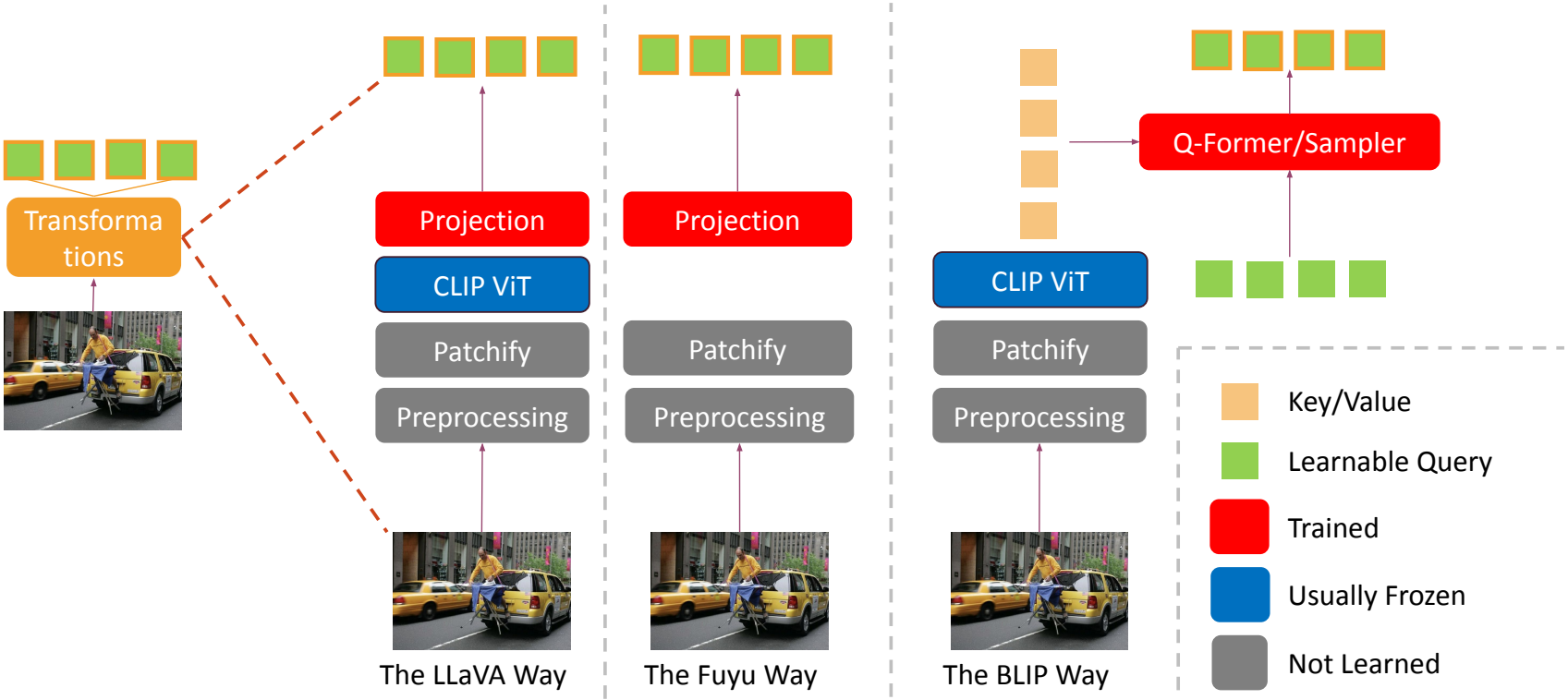
Nov 13, 2025

What is VLM (from my old presentation 1.5 yrs ago)

- Large Language Models that incorporate visual information, which can be natural images, printed documents, UI/Webpage screenshots, figures, etc.



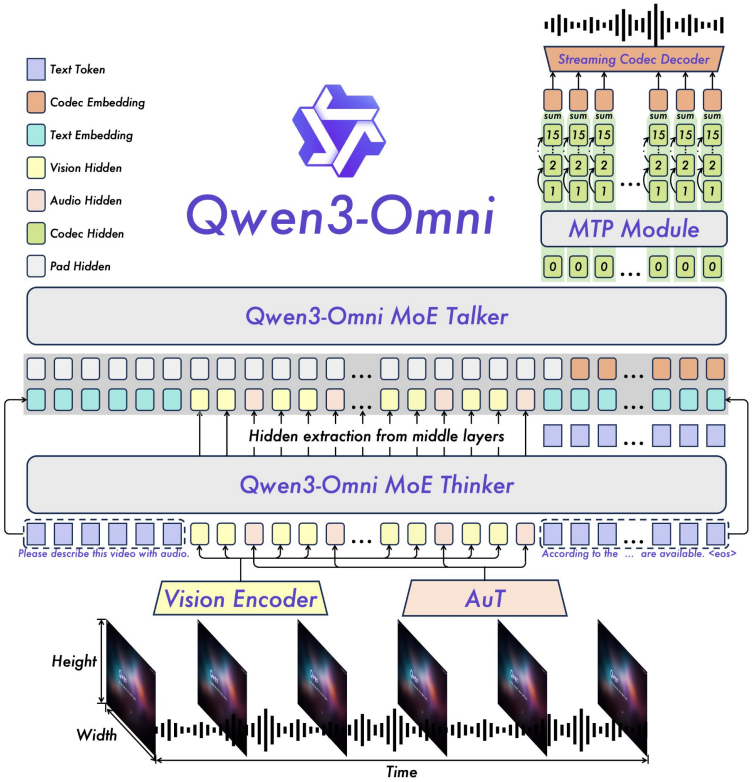
What is VLM (from my old presentation 1.5 yrs ago)



What has been changed?

- Input modalities can be more than images and text
- Output modalities can be more than text
- Test-time compute scaling (tho modality agnostic)
- Multi-image conversations
- Multi-turn conversations
- Computer-Use Agent
- Gaming Agent
- Vision-Language Action Models

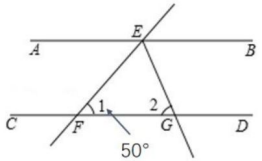
VLM Progression



- Input modalities can be more than images and text
- Output modalities can be more than text

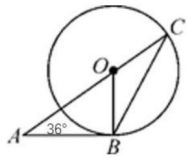
VLM Progression

- Test-time compute scaling (tho modality agnostic)



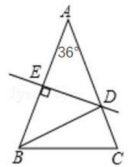
Q: As shown in the figure, AB parallel CD, EG bisects angle BEF, then angle 2 is equal to ()
Choices: A:50° B:60° C:65° D:90°

<think> Okay, let's see. The problem displays that AB is parallel to CD, and EG bisects angle BEF. We need to find the measure of angle 2. First, the image displays.....Wait, angle 1 is at point F, so.....be supplementary to angle 1 because they are same-side interior angles. So angle BEF would be 180 - 50 = 130 degrees. Then, EG bisects angle BEF, so it splits angle BEF into two equal parts. Therefore, each of the angles formed by the bisector EG would be 130 / 2 = 65 degrees. Since angle 2 is one of those bisected angles, angle 2 should be 65 degrees. Looking at the choices, option C is 65° . So the answer should be C. </think><answer> Final Answer:C </answer>



Q: As shown in the figure, AB is tangent to circle O at point B, then angle C is equal to ()
Choices: A:36° B:54° C:60° D:27°

<think> Okay, let's see. The problem displays.....I remember that the tangent to a circle is perpendicular to the radius at the point of contact.....so the remaining angle at O must be 180 - 90 - 36 = 54 degrees.....Now, the problem displays that angle C is an inscribed angle that subtends the same arc as the central angle O. Wait, the central angle subtended by arc OB would be angle O, which is 54 degrees. But angle C is an inscribed angle subtending the same arc. Remember that an inscribed angle is half the measure of the central angle subtending the same arc.....So, angle C is 27 degrees. Looking at the choices, D is 27° . So the answer should be D. </think><answer> Final Answer:D </answer>

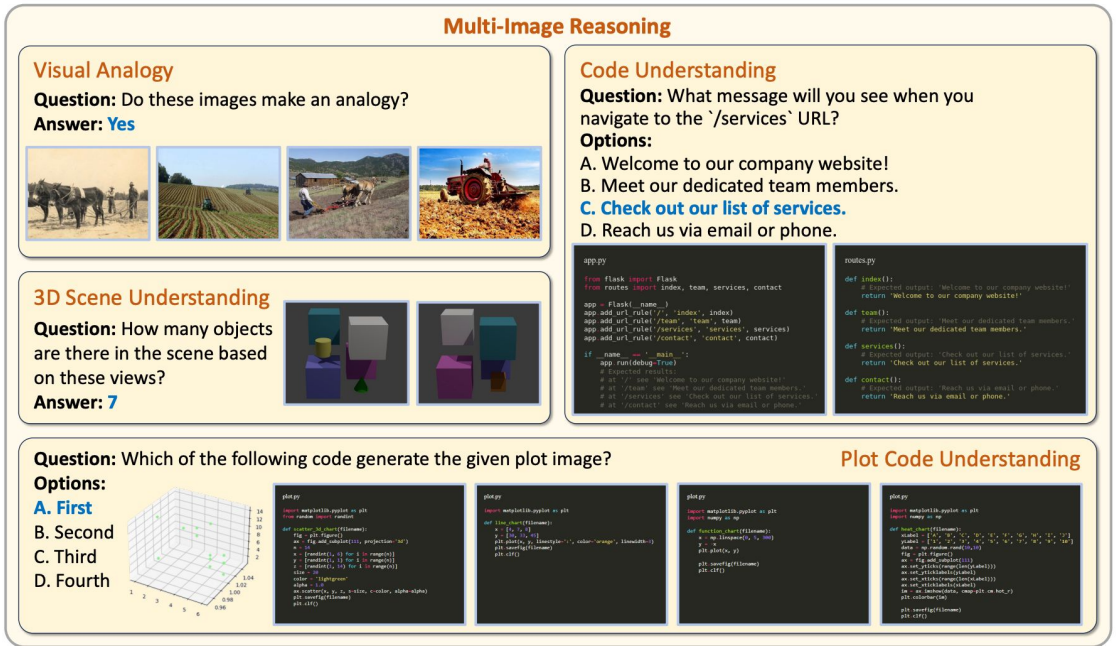


Q: As shown in the figure, AB = AC, angle A = 36.0, the perpendicular bisector of AB intersects AC at D, then the degree of angle BDC is ()
Choices: A:72° B:36° C:60° D:82°

<think> Okay, let's see. The problem displays..... First, since AB = AC.....Now, the perpendicular bisector of AB.....Let's call the midpoint of AB as point M. So BM = MA.....Wait, maybe I should consider the properties of the perpendicular bisector.....But wait, angle DBA is part of angle ABC.....But wait, maybe I'm getting confused here. Let me try a different approach..... Let's see. In triangle BDC.....But wait, that might not be directly applicable. Wait, maybe we can.....which is option A. Wait, but let me check again. Alternatively, maybe using.....So the answer is 72 degrees, which is option A. That seems correct. </think><answer> Final Answer:A </answer>

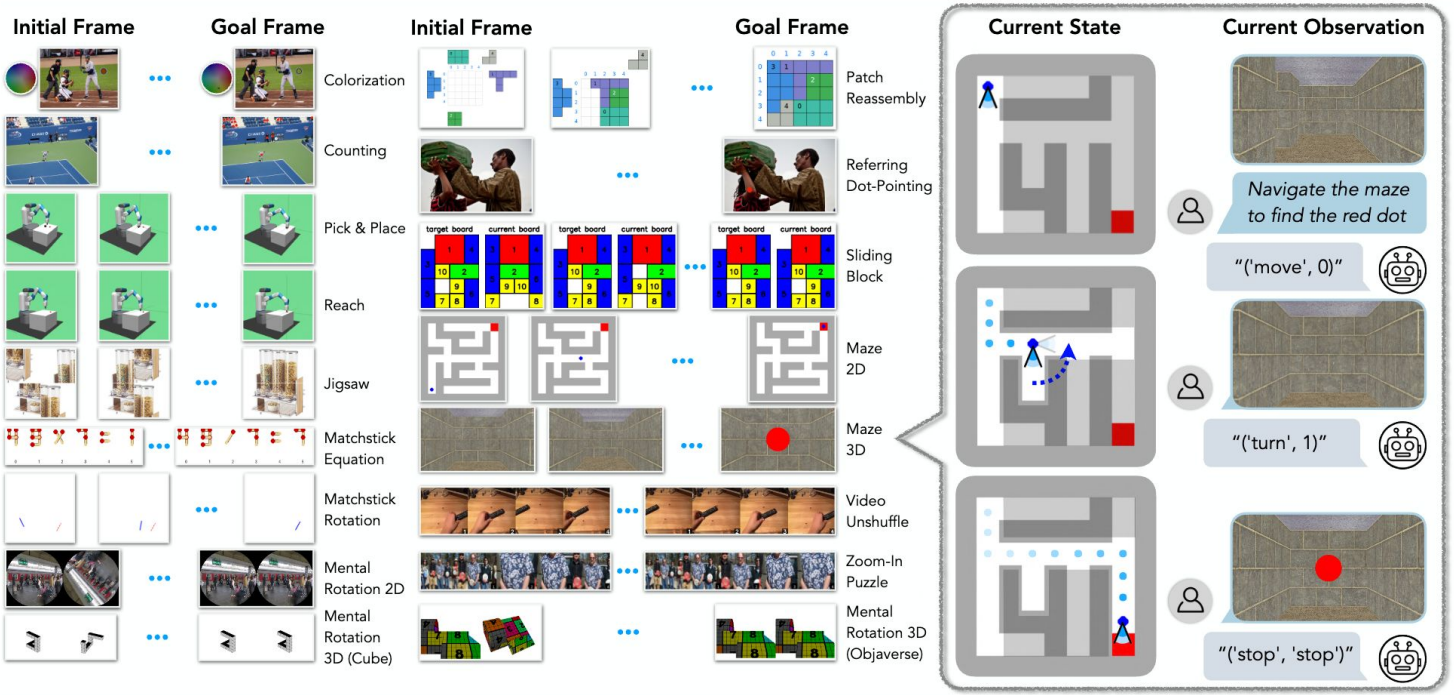
VLM Progression

- Multi-image conversations



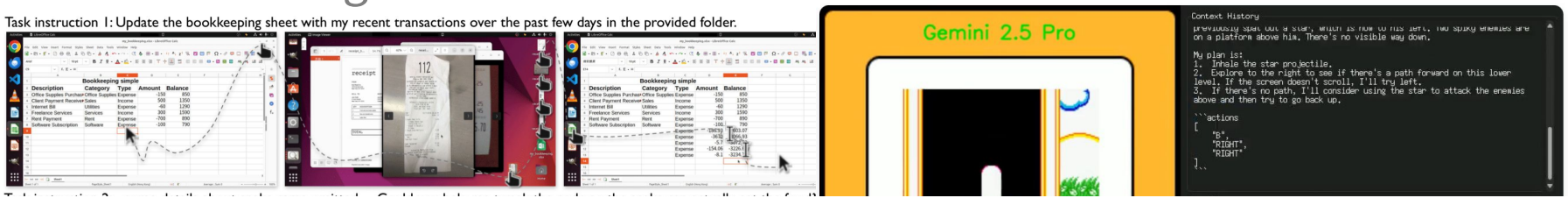
VLM Progression

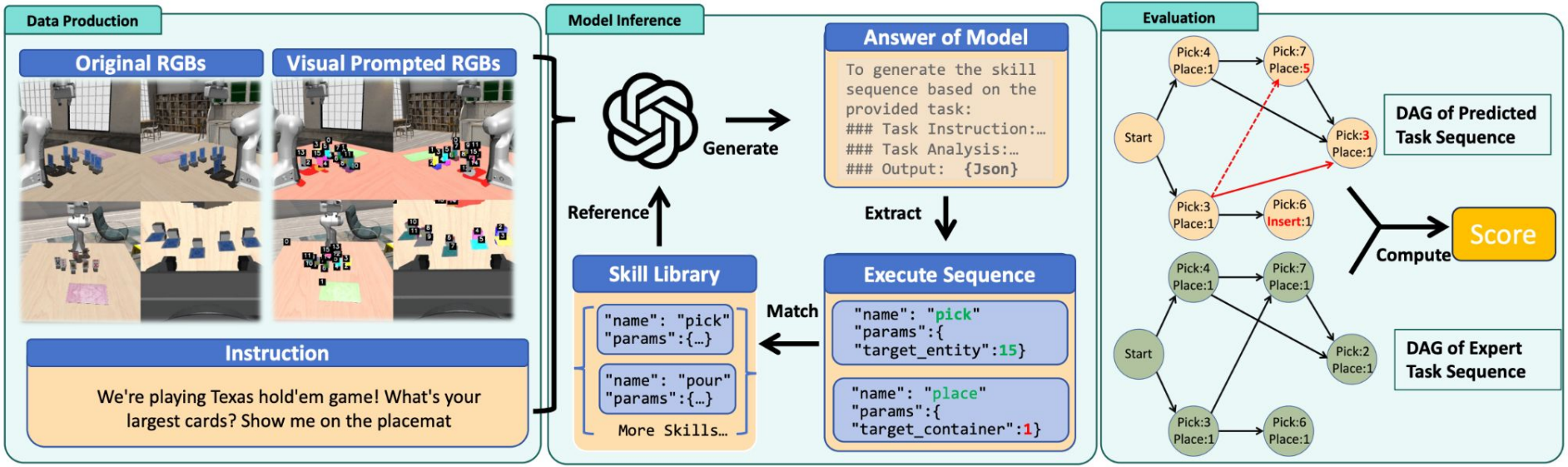
- Multi-turn conversations



VLM Progression

- Agents



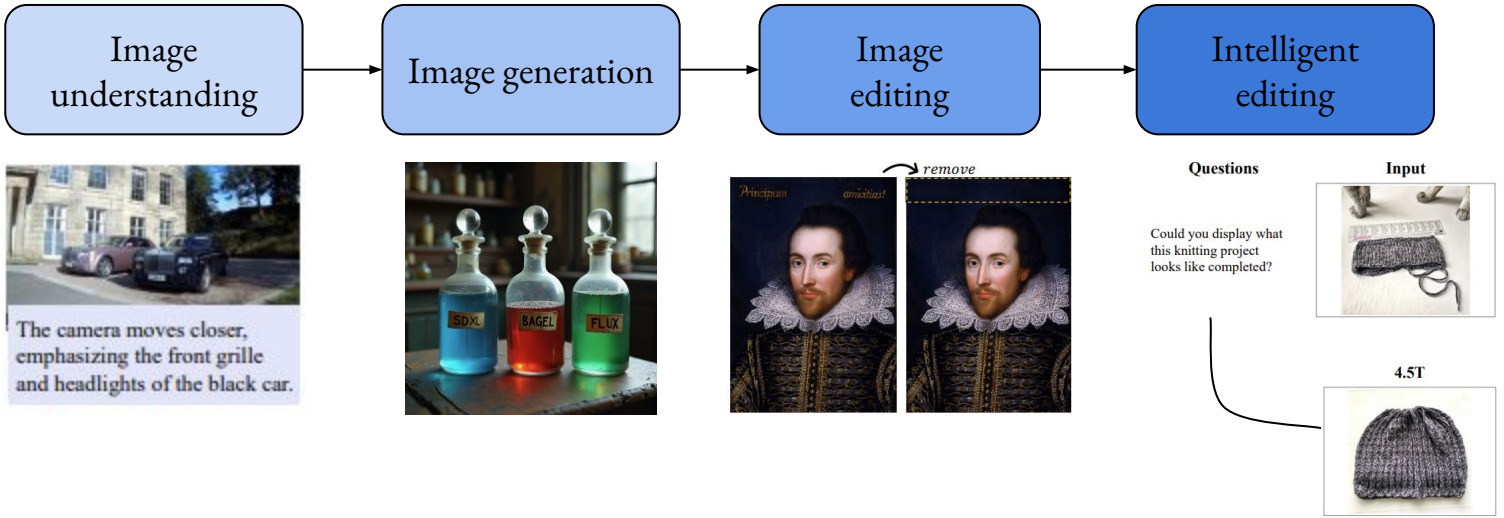


Emerging Properties in Unified Multimodal Pretraining

Junyi Zhang, CS 294
Nov 13, 2025

Paper summary

- Introduces BAGEL, a 14B multi-modal foundation model
- Highlights emerging properties of **the order of learned capabilities** in multimodal training

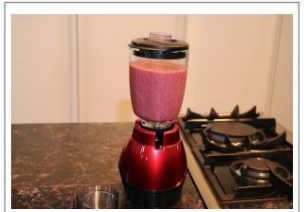


Results

Prompts	1.5T	2.5T	3.5T	4.5T
A clear image of a blackboard with a clean, dark green surface and the word 'Hello BAGEL' written precisely and legibly in the center with bold, white chalk letters.				

Prompts	Input	1.5T	2.5T	3.5T	4.5T
Replace the eagle with a parrot.					

Results

Questions	Input	1.5T	2.5T	3.5T	4.5T
Could you display what this knitting project looks like completed?					
Could you display the smoothie once it's blended?					

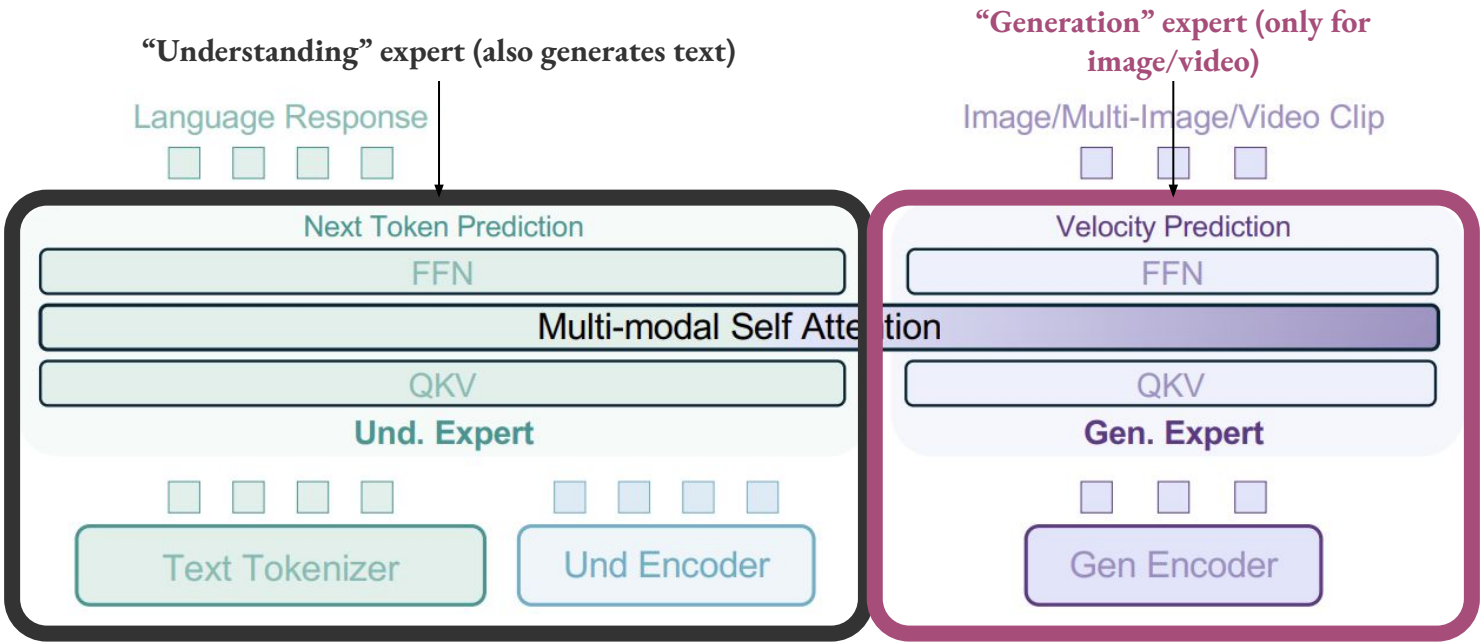
Results

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>** 

Move forward

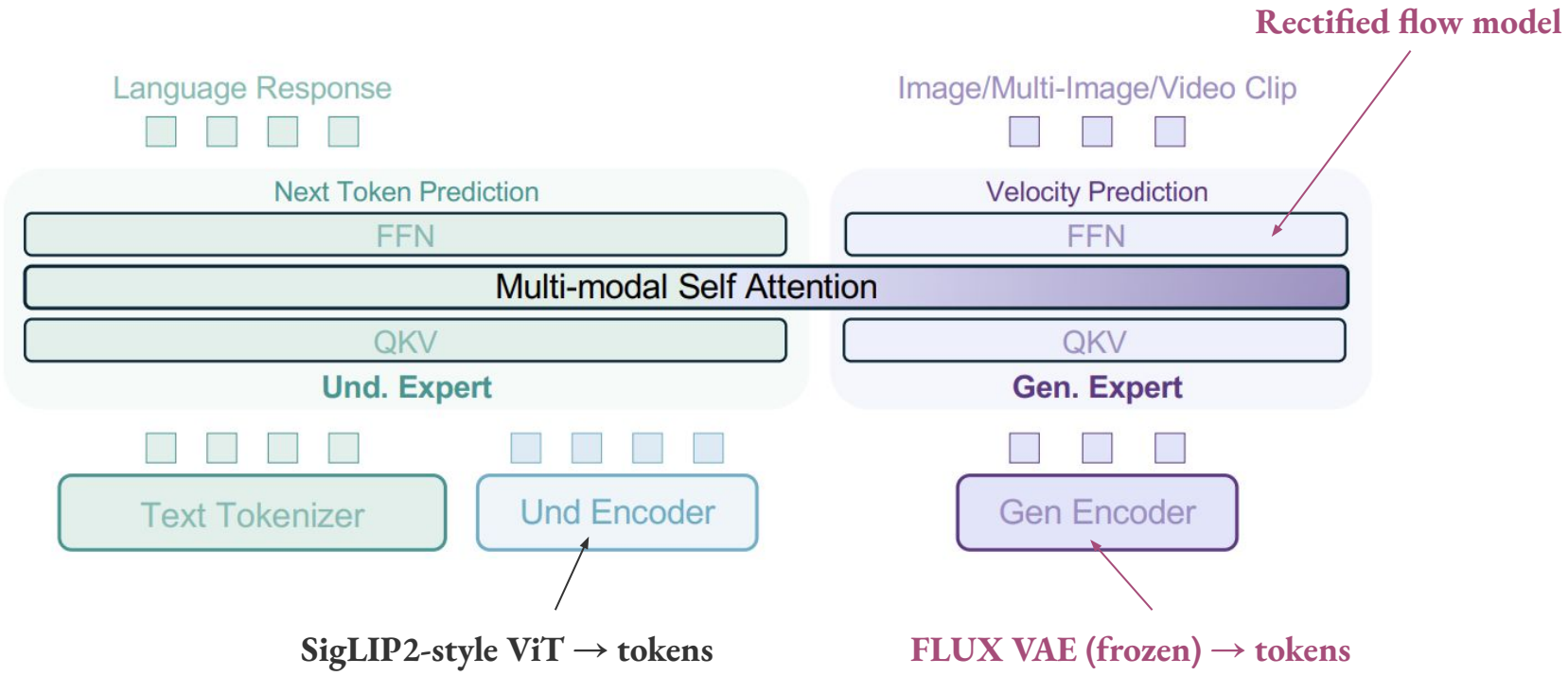
Look left

BAGEL: Scalable Generative Cognitive Model



"Mixture of Transformers" (MoT)

How is visual information represented?



How is visual information represented?

For multiple image inputs, create three sets of tokens:

ViT Tokens

Used for for understanding and generating

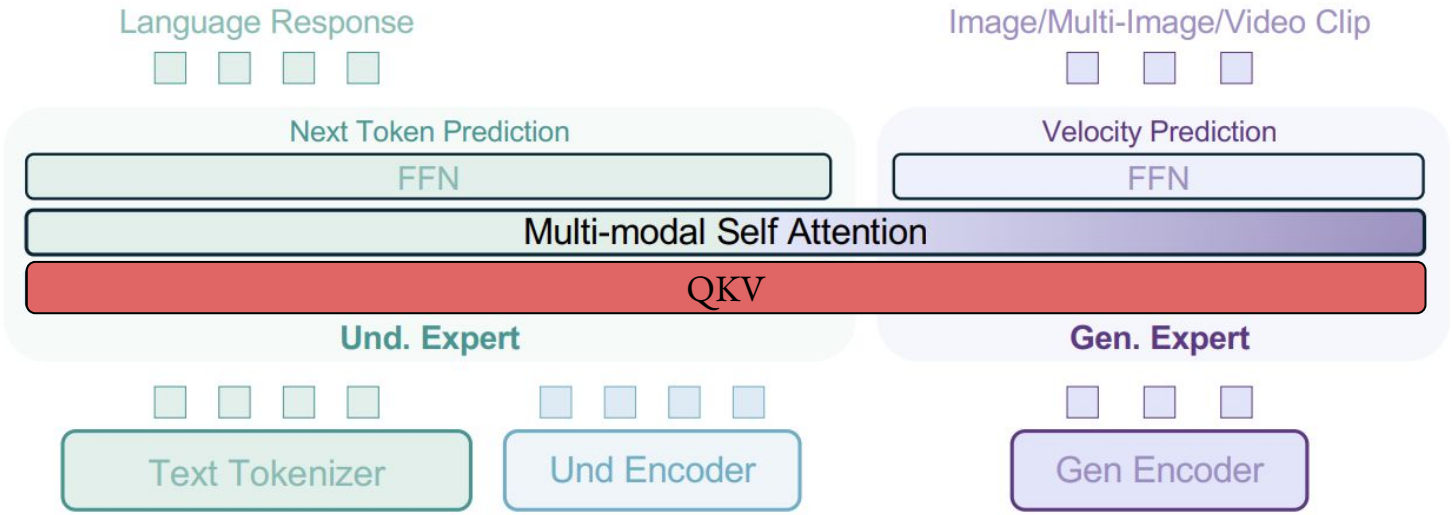
Clean VAE tokens

Used for conditioning for multi-image generation

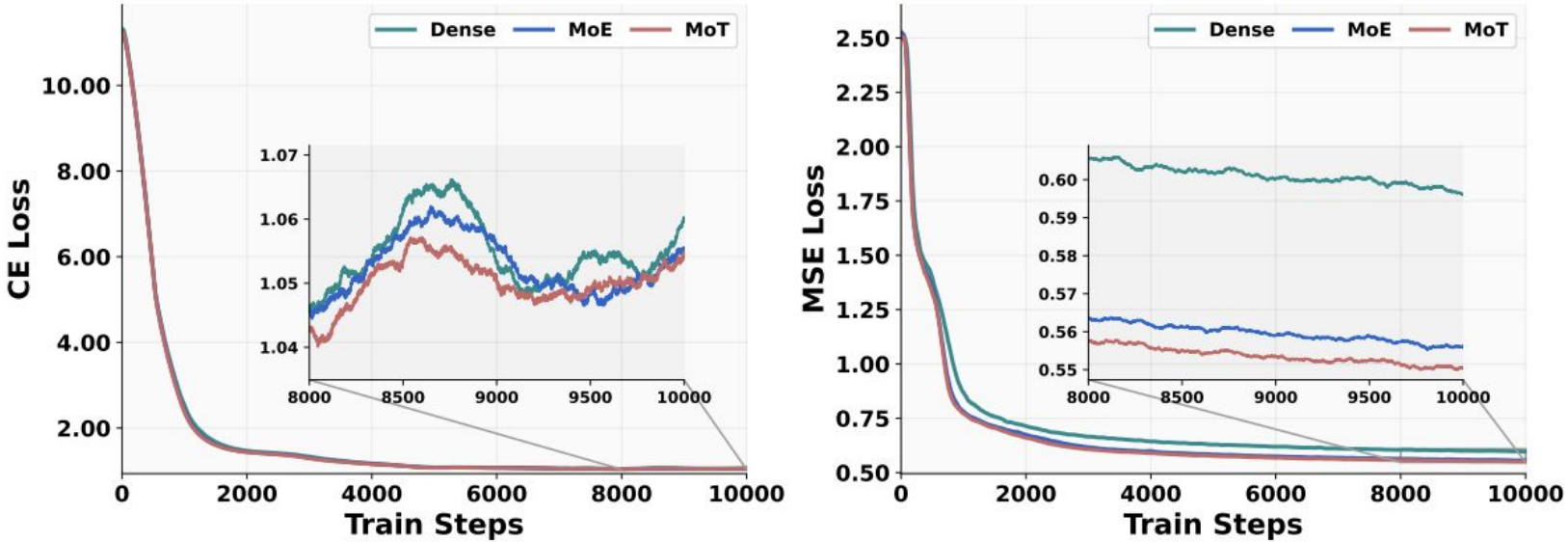
Noised VAE tokens

Used for rectified flow training

Mixture of Experts variant (MoE)



MoT outperforms MoE



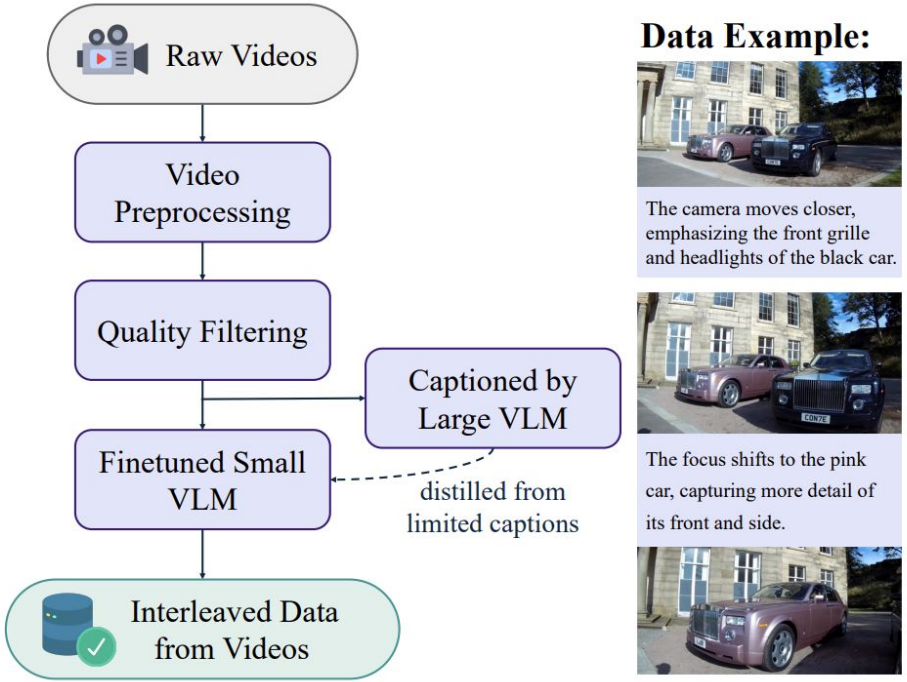
Advantage to decoupling
understanding and generation

Dataset curation

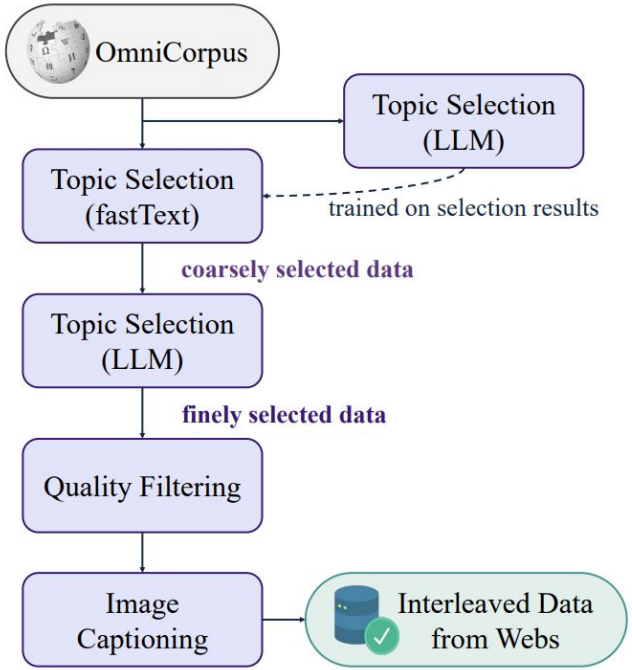
Data Source	# Data (M)	# Tokens (T)
Text Data	400	0.4
Image-Text-Pair Understanding Data	500	0.5
Image-Text-Pair Generation Data	1600	2.6
Interleaved Understanding Data	100	0.5
Interleaved Generation Data: Video	45	0.7
Interleaved Generation Data: Web	20	0.4

Uses VLM annotation

Interleaved dataset curation: internet videos



Interleaved dataset curation: web image captioning



Data Example:

[Orig. Text] Tips to Take Care of Your Car

[Img Cap.] A mechanic in a garage inspects a car engine, holding a tablet and a tool. They wear gloves and a cap, focusing on the task.

[Image]

[Orig. Text] Follow Maintenance Schedule
The easiest way to take care of your car is by following the maintenance schedule ……

[Img Cap.] A gloved hand holds a car's coolant reservoir cap, emphasizing routine maintenance. The yellow and blue glove contrasts with the dark engine bay

[Image]

[Orig. Text] Check Oil & Other Fluids
Your car depends on several oils and fluids such as engine oil, coolant, and brake fluid in order to ……

Interleaved dataset curation: reasoning augmented data

Text to image generation

Query

A car made of small cars.

Reasoning trace

<think> The model should generate 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...
</think>

Image

Free-form image editing

Query + reasoning

Source

remove

Target

Sources: OmniEdit, videos

Conceptual editing

Questions

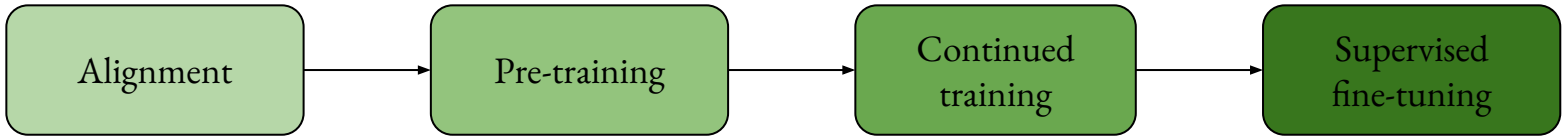
Could you display what this knitting project looks like completed?

Input

Using VLM:

1. Generate I/O pairs from image sequence
2. Generate questions aligned with pair

Training



Align SigLIP2 encoder by training only MLP connector to und. expert

Train all parameters except VAE gen. encoder

Increase visual input resolution

Increase sampling ratio of interleaved data → cross-modal reasoning

Train on high-quality subset from image-text pair and interleaved data

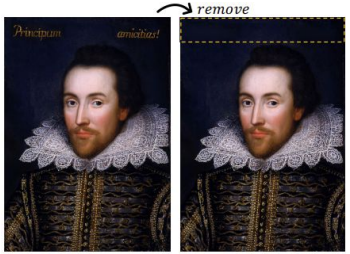
Data sampling ratio				
Text	0.0	0.05	0.05	0.05
Image-Text pair (T2I)	0.0	0.6	0.4	0.3
Image-Text pair (I2T)	1.0	0.1	0.1	0.05
Interleaved understanding	0.0	0.1	0.15	0.2
Interleaved generation: video	0.0	0.1	0.15	0.2
Interleaved generation: web	0.0	0.05	0.15	0.2

Emerging properties

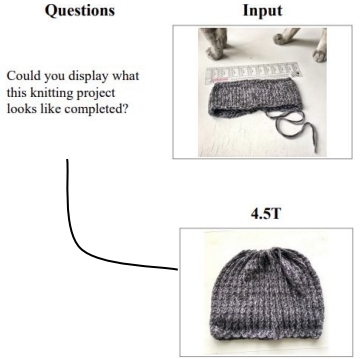
Image understanding

Image generation

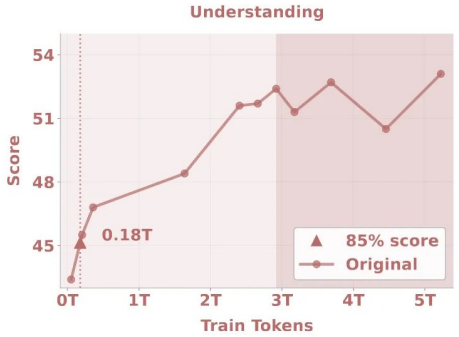
Image editing



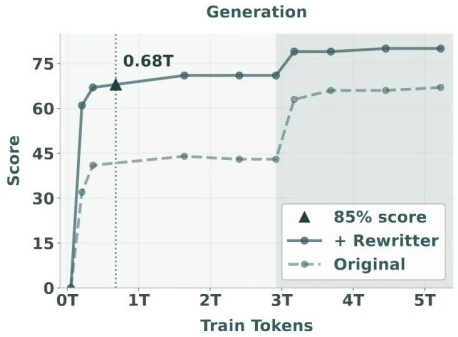
Intelligent editing



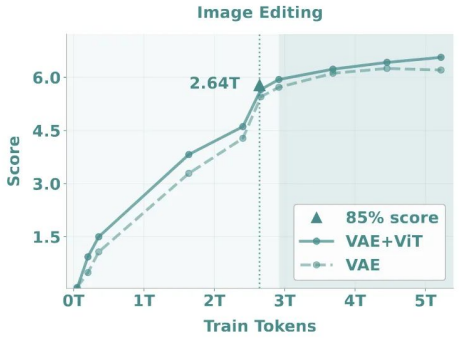
Emerging properties



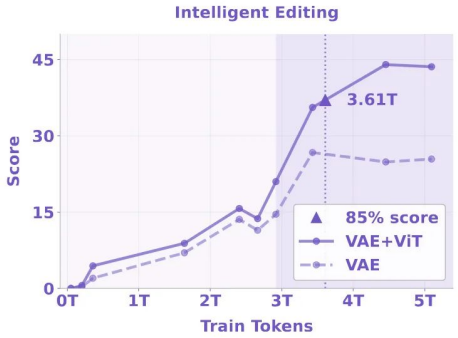
(a) Average score on Image Understanding tasks.



(b) GenEval score on Image Generation task.



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



(d) IntelligentBench Score on Intelligent Editing task.

Are the emerging properties a result of the training schedule?

- Common training pipeline: pre-training → train generative model → train on high-quality subset of data
- Emphasizes the “understanding → advanced generation” emergent property
- Do all “well-trained” models follow this emergent pattern?

Holistic Evaluation for Interleaved Text-and-Image Generation

Minqian Liu[♠] **Zhiyang Xu**[♠] **Zihao Lin**[♠] **Trevor Ashby**[♠]
Joy Rimchala[♡] **Jiaxin Zhang**[♡] **Lifu Huang**[♠,♣]
[♠]Virginia Tech [♡]Intuit AI Research [♣]University of California, Davis
{minqianliu, zhiyangx, zihaol, trevorashby, lifuh}@vt.edu
{joy_rimchala, jiaxin_zhang}@intuit.com

https://vt-nlp.github.io/InterleavedEval/

Colin Wang, CS 294
Nov 13, 2025

Motivation

- Interleaved text-and-image generation has been trending, but evals lag behind
  - **Limited output**: existing works for interleaved generation checks 1 image output
  - **Outdated metrics**: BLEU, FID; on the other extreme human eval is costly
  - **Distinct aspect important for the task**: perceptual quality, coherence and helpfulness
- They proposed:
  - **InterleavedBench**, a dataset
  - **InterleavedEval**, an eval framework based on GPT-4o

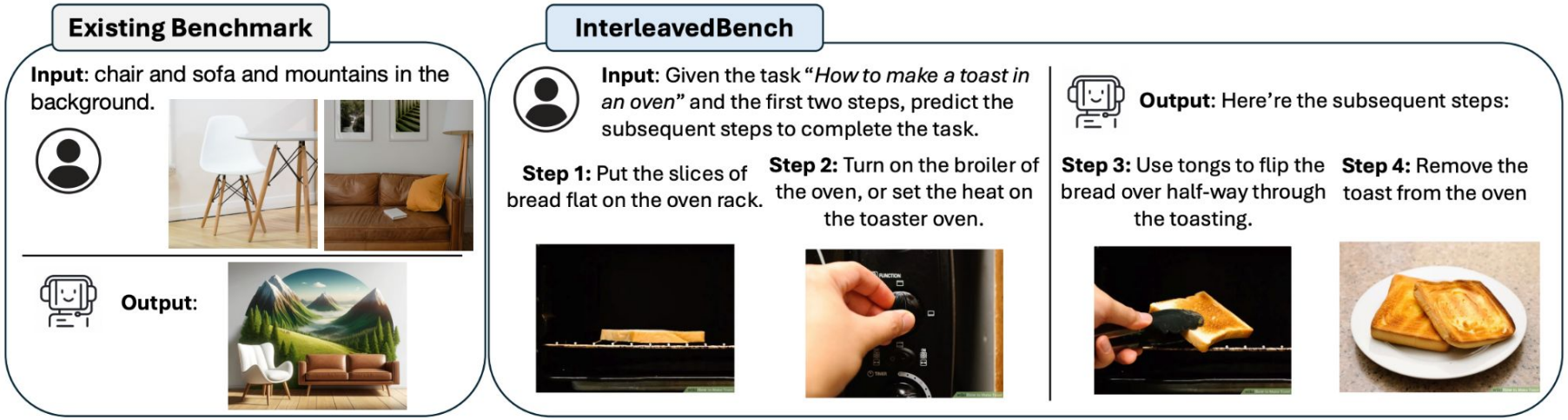
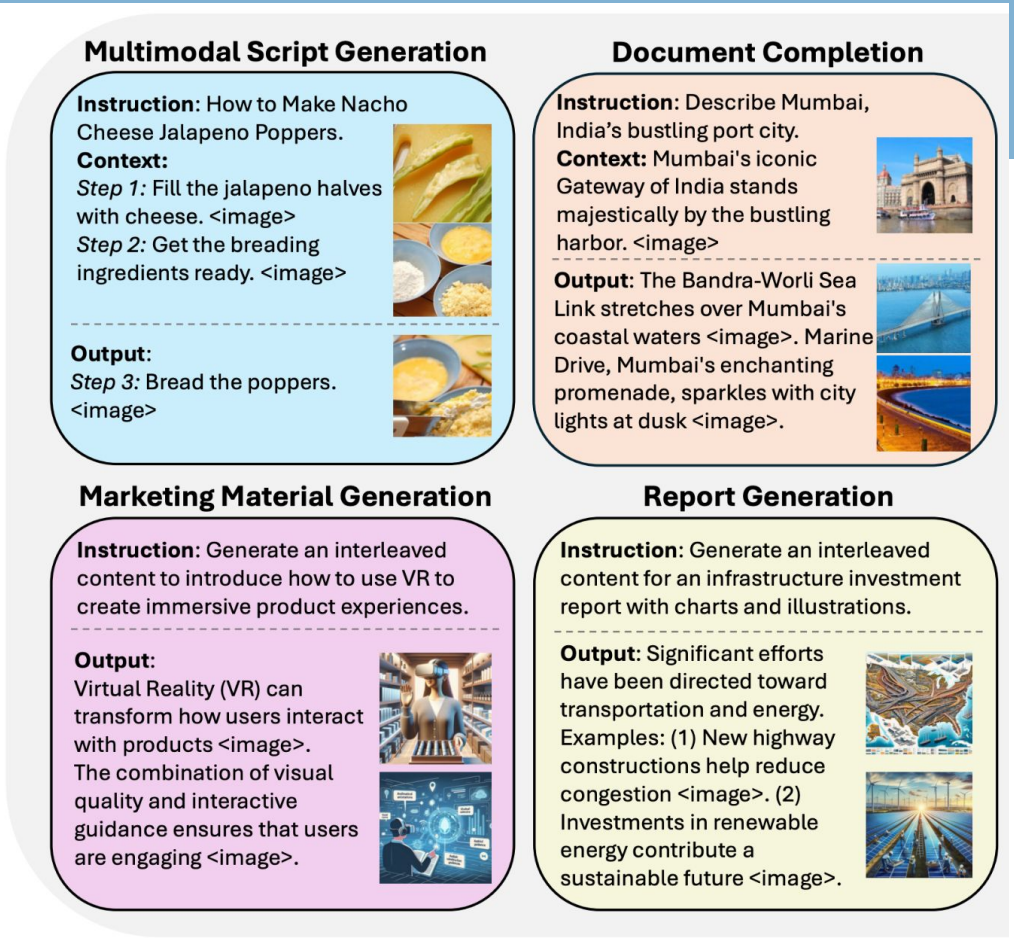


Figure 2: Comparison between the existing benchmark (multi-concept image composition (Kumari et al., 2023a)) and our INTERLEAVEDBENCH. Compared with the existing benchmark, INTERLEAVEDBENCH has the following features: (1) both input and output can have arbitrarily interleaved text and images, and (2) each instance has a detailed instruction to benchmark models' instruction-following capability.

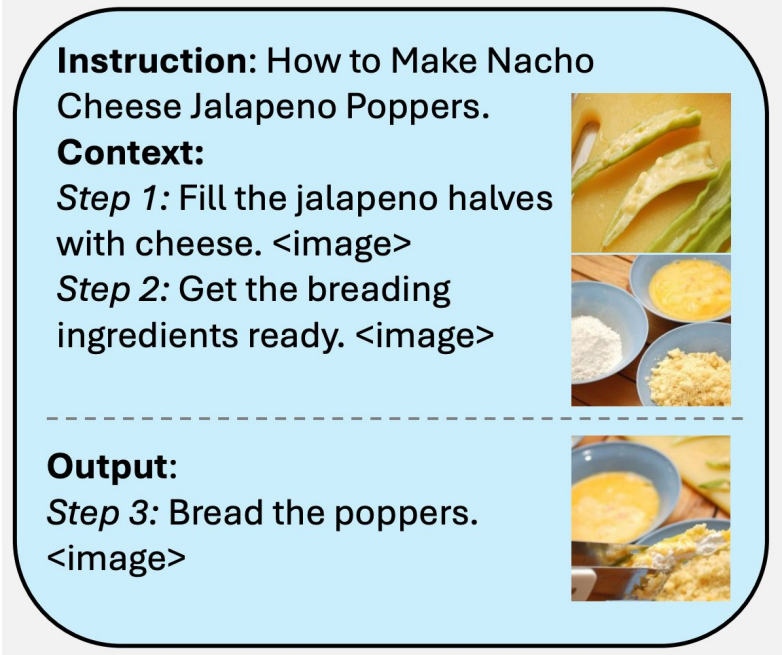
InterleavedBench

- 2 subsets: context-based and context-free
  - Context-based: img+text generation *continuation with img+text context*
  - Context-free: img+text generation
- 815 instances across 10 use cases.



Dataset Curation (context-based)

- Sources: wikihow, vist, activitynet (captions), etc
- Then, human selection based on quality and diversity
- Finally, humans write instructions
  - Article split: pick first k images and contexts associated with them;
    1 <= k <= max_num_imgs - 1
  - The remaining used as the gold reference

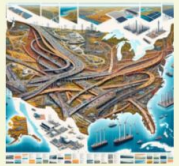


Dataset Curation (context-free)

- Sources: GPT-4o generates synthetic instances and instructions for marketing materials, report, education content, etc.
- Then, human selection based on quality and reasonableness
- No gold references

Instruction: Generate an interleaved content for an infrastructure investment report with charts and illustrations.

Output: Significant efforts have been directed toward transportation and energy. Examples: (1) New highway constructions help reduce congestion <image>. (2) Investments in renewable energy contribute a sustainable future <image>.



Interleaved Eval

- Traditionally, people’ve been using BLUE and FID to compare interleaved generation for text and images against gold reference
- But this can be unrealistic because generation can be open-ended
  - For example, there can be 100000000000000 different ways to date
- So reference-free based metrics are preferred
- GPT-4o is used as a judge
  - Text quality, perceptual quality, image coherence, text-image coherence and helpfulness

Interleaved Eval

- During eval, judge takes in instruction, model output, eval criteria, and optionally context. Each aspect rated 0-5 w/ explanations

Aspect	Definition
Text Quality	Text quality measures how clear, coherent, and error-free the output text is. It considers grammar, spelling, readability, coherence with the instruction and context, and whether it contains duplicate content.
Perceptual Quality	Perceptual quality measures how visually convincing, natural, and free from distortions or artifacts a generated image appears. It considers how accurately the image mimics reality without unnatural disruptions in structure, colors, or composition.
Image Coherence	Image coherence measures the consistency in style and subject representation across images. This includes textures, color palette, lighting, rendering styles, and maintaining consistent physical attributes, clothing, and behavioral traits. Image coherence also penalizes image duplication, where the output images are too similar, or within the output images themselves.
Text-Image Coherence	Text-to-image coherence measure the alignment and integration between textual and visual elements in a pairwise manner, ensuring they work together to convey a unified and cohesive narrative.
Helpfulness	Helpfulness measures how well the output text and images follow the task instructions and provide complete information to achieve the task. It also considers whether the outputs follow a reasonable logic flow.

Experiments

- 2 types of paradigms
  - Integrated text-and-image generation models
    - MiniGPT5, GILL, Emu2: models learn to generate hidden states that can be mapped to the embedding space of diffusion models
  - Agentic text and image generation models
    - Emu-2 Gen + Gold Text: combine GT text and let Emu-2 Gen to generate images
    - GPT-4o + DALL E3; Gemini 1.5 + SDXL: VLMs generate text and captions for T2I models to generate images
  - (BAGEL does not exist at the time of eval)

Experiments

- Baseline metrics: BERTScore, CLIPScore and DreamSim
  - They validate that their LLM judge setup correlates strongest with human eval.

Metric	Ref-free?	Text Quality	Perceptual Quality	Image Coherence	TIC	Helpfulness
BERTScore	✗	0.21	-	-	-	0.37
DreamSim	✗	-	0.02	0.1	-	0.06
Image-Image CLIPScore	✗	-	0.08	0.2	-	-0.01
Text-Image CLIPScore	✓	-	-	-	0.2	0.09
INTERLEAVEDEVAL-LLaVA	✓	0.06	**0.32**	0.24	0.23	0.3
INTERLEAVEDEVAL-GPT-4o	✓	**0.72**	0.30	**0.43**	**0.4**	**0.57**

Table 4: **Mete-evaluation on evaluation metrics** in terms of Spearman correlation between automatic evaluation results with human judgments. For baseline metrics, we only report the correlation on the corresponding aspects (e.g., BERTScore can correspond to *text quality*) as well as *helpfulness*.

Experiment

- Results

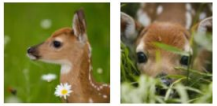
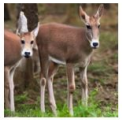
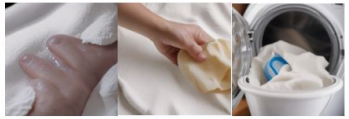
Model	Text Quality	Perceptual Quality	Image Coherence	TIC	Helpfulness	AVG
MiniGPT-5	1.22	2.45	1.62	2.03	1.77	1.82
GILL	0.75	3.21	2.25	1.53	1.48	1.84
EMU-2	1.26	2.28	1.89	1.34	1.64	1.68
EMU-2 (Gold Text)	1.56	3.35	2.89	1.43	2.10	2.27
Gemini1.5 + SDXL	**4.40**	3.99	**3.64**	4.13	3.62	3.96
GPT-4o + DALL·E 3	4.37	**4.36**	3.51	**4.55**	**3.88**	**4.13**

Table 2: **Automatic evaluation** results of existing interleaved generation models on INTERLEAVEDBENCH using INTERLEAVEDEVAL based on GPT-4o. TIC is the abbreviation for 'Text-Image Coherence'. The best results are highlighted in **bold**.

Experiment

- Interpretation
  - Agentic models better than integrated models (?)
  - Agentic models are much better in text quality
  - Issues w/ integrated models
    - Empty text or image output
    - Poor quality
    - Duplications
  - Image coherence is the biggest challenge for all models
- (I don’t know how many of these interpretations are still true…)

Case Study

Input	GILL	EMU-2	Gemini+SDXL	GPT-4+DALLE3
Instruction: Describe a doe's protective behavior towards her fawn. **Context**: Fawns usually don't bleat unless distressed <image>. A distressed fawn call prompts the mother doe <image>. 	**Output**: mate and produce offspring <image>. 	**Output**: Fawns are born with a spotted coat that provides camouflage <image>. 	**Output**: This behavior is rooted in the harsh realities of environment <image>. She keeps her fawn hidden in thick vegetation and only visits it for feeding <image>. 	**Output**: The bond between a doe and fawn creates security <image>. This is vital as fawns learn survival skills from interaction with their mothers <image>. 
Instruction: You need to help user to complete the task 'How to Remove Banana Stains from Fabric'. **Context**: Scrape off excess banana pieces <image>. 	**Output**: the banana pieces are stuck to the fabric <image>. 	**Output**: Soak the fabric in cold water. Soak the fabric in warm water. <image>. 	**Output**: Rinse the stain with cold water <image>. Apply a stain remover to the stain <image>. Wash the garment <image>. 	**Output**: Rinse the Stained Area <image>. Apply Liquid Laundry Detergent <image>. Gently Rub the Detergent into the Fabric <image>. 
Instruction: Generate a visual story with the following description: The Hidden Library: A library that holds all the knowledge of the world appears only to those who seek it, and a young scholar named Elara must find it to save her kingdom.	**Output:** 	**Output:** 	**Output:** The king, Elara's father, was seeking solace in the royal library <image>. She traversed treacherous mountains <image>. This was the Hidden Library <image>. 	**Output:** Elara ventures into the ancient forest <image>. Elara stumbles upon a hidden cave <image>. She speaks the incantations <image>. 

GILL is not good because both text and image are not coherent

EMU is more coherent but it's not very helpful; image qualities are poor

Agentic setups are better, but DALLE 3 has style bias

Maintaining image coherence seems hard

Context-based vs context-free

Model	Text Quality	Perceptual Quality	Image Coherence	TIC	Helpfulness	AVG
MiniGPT5	1.29	3.47	2.04	2.64	1.76	2.24
GILL	1.37	3.96	2.01	2.61	1.51	2.29
EMU-2	1.29	2.22	1.65	1.18	1.84	1.64
Gemini1.5+SDXL	**3.29**	4.24	**3.26**	3.94	3.25	3.60
GPT-4o+DALLE3	3.12	**4.39**	3.08	**4.36**	**3.48**	**3.69**

Table 5: **Automatic evaluation** results of the *context-based* subset on INTERLEAVEDBENCH. TIC is the abbreviation for 'Text-Image Coherence'. The best results are highlighted in **bold**.

Model	Text Quality	Perceptual Quality	Image Coherence	TIC	Helpfulness	AVG
MiniGPT5	1.00	1.09	1.07	1.06	1.78	1.20
GILL	0.12	2.23	2.58	0.23	1.45	1.32
EMU-2	0.77	2.35	2.20	1.05	1.38	1.55
Gemini1.5+SDXL	4.50	3.66	**4.13**	3.98	4.10	4.07
GPT-4o+DALLE3	**4.60**	**4.31**	4.05	**4.52**	**4.41**	**4.38**

Table 6: **Automatic evaluation** results of the *context-free* subset on INTERLEAVEDBENCH. TIC is the abbreviation for 'Text-Image Coherence'. The best results are highlighted in **bold**.

Number of steps

- How number of output steps affect scores compared to gold reference w/ gpt4o + dalle3

Output Steps	Text Quality	Perceptual Quality	Image Coherence	TIC	Helpfulness	AVG
Less	1.8	1.1	1.2	1.3	2.1	1.5
Equal	**2.7**	**3.8**	**4.0**	**4.0**	**3.0**	**3.5**
More	1.7	3.5	2.4	3.3	2.0	2.6

Table 7: Analysis of the number of output steps compared with ground truths.

QA?

Small-group discussion

Group	Leads				
1	Yichuan Wang & Junyi Zhang	Prasann Singhal	Ryan Wang	Nathan Ju	Yuezhou Hu
2	Jongho Park & Colin Wang	Huanzhi Mao	Sidhika Balachandar	Shangyin Tan	Hanchen Li
3	Sangdae Nam & Dongwei Lyu	Donghyun Lee	Bhavya Chopra	Dennis Jacob	Charlie Ruan
4	Téa Wright & Harman Singh	Sanjay Adhikesaven	Xutao Ma	Kaiwen Hu	Juno Kim

Write a short review at #class-discussion

Discussion questions

1. Why would you want a unified (omni model), instead of separate models for image understanding and image generation?
2. A Unified Transformers vs. A Mixture-of-Transformers approach?
3. Why would we want interleaved text/image generation?
4. What's the best way to evaluate interleaved generation (w.r.t. the limitations of the second paper)?
5. How could this be extended for other modalities, beyond text and vision? (e.g., Pi-0.5 for a Vision-Language-Action model)