International Conference On Learning Representations

Project Page

Matryoshka Multimodal Models

Mu Cai[1,2], Jianwei Yang[2], Jianfeng Gao[2], Yong Jae Lee[1]

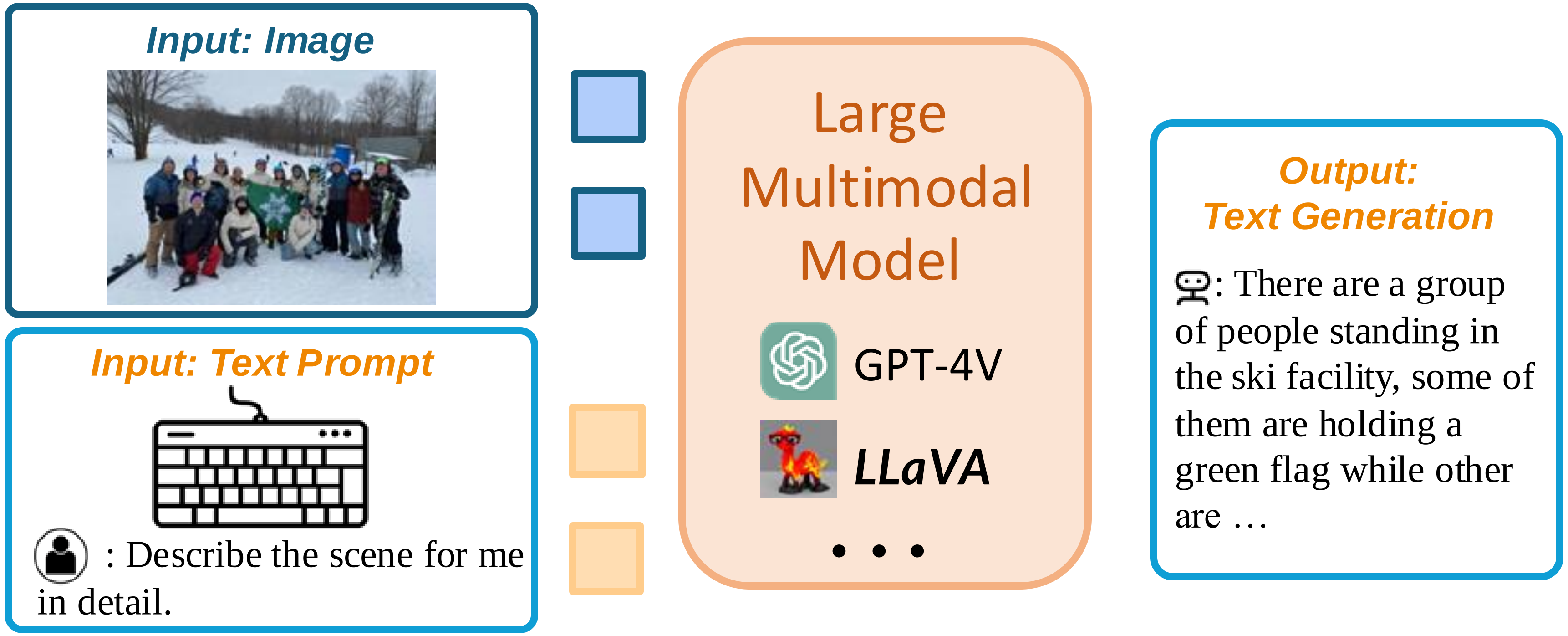
1 Department of Computer Sciences UNIVERSITY OF WISCONSIN–MADISON

2 Microsoft

ICLR 2025

Motivation

Most Large Multimodal Models are pretty good
at understanding a standard-resolution image:



But this is far from practice

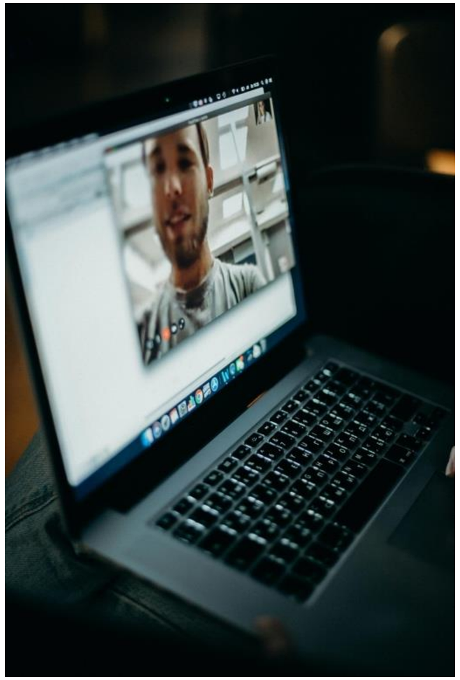
Motivation

How can such multimodal systems be real-world assistants?

Motivation

How can such multimodal systems be real-world assistants?

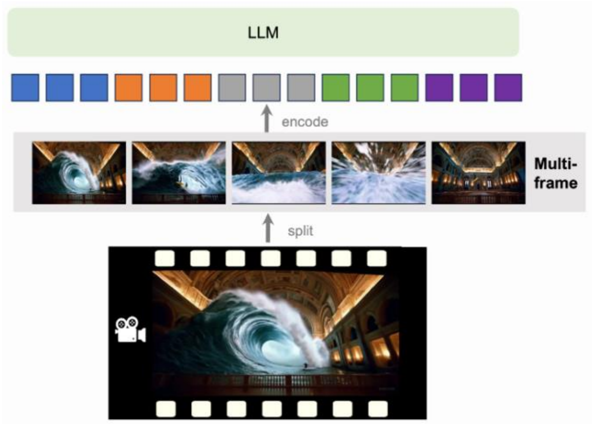
(a) High-resolution images
(**thousands of tokens**)



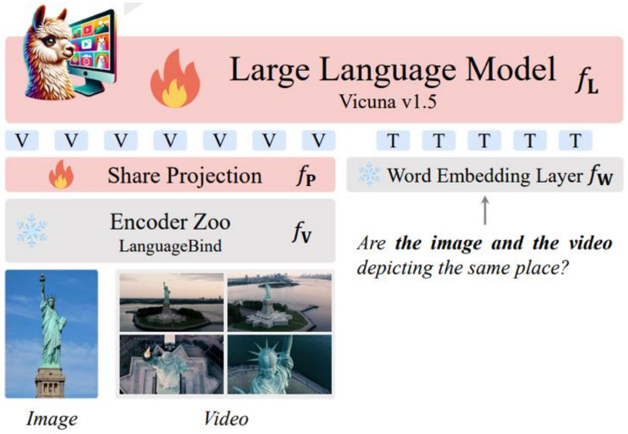
(b) Long videos
(**millions of tokens**)

Motivation

Bottleneck of Current Multimodal Models

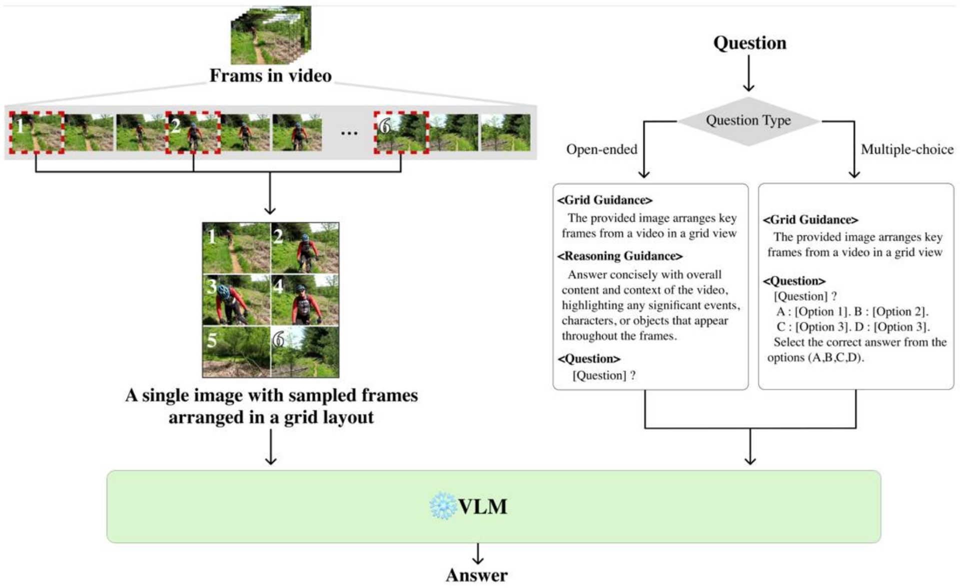


(a) LLaVA-OneVision: 6272 visual tokens for 32 frames



(a) Illustration of Video-LLaVA

(b) Video-LLaVA: 2048 visual tokens for 8 frames

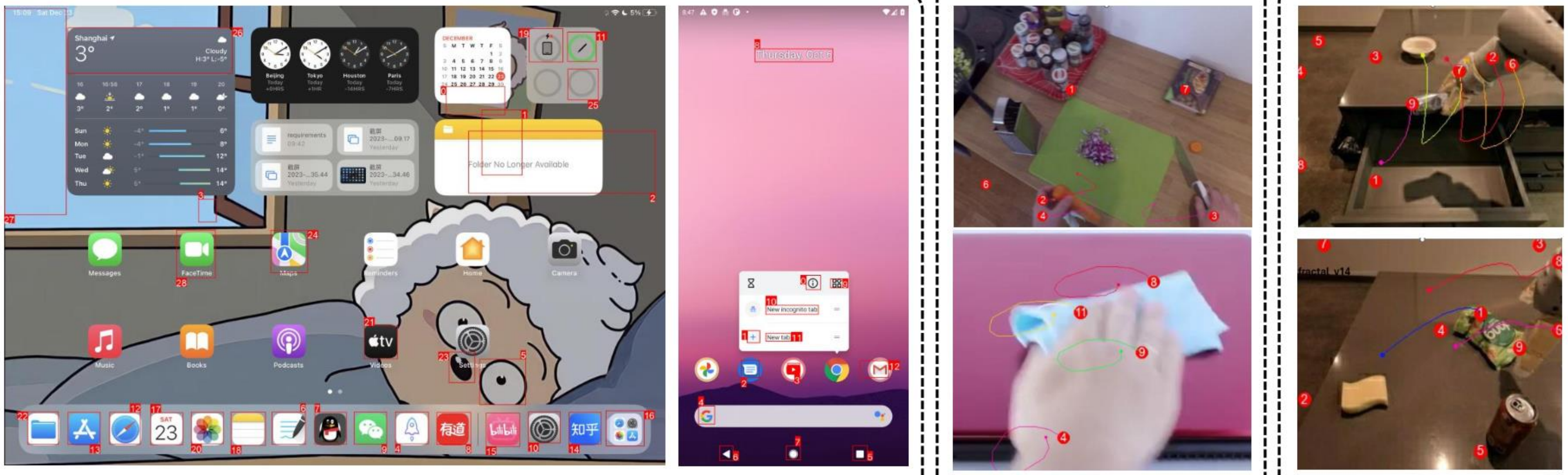


(c) IG-VLM: 2880 visual tokens for 6 frames

Motivation

Too many tokens is a serious problem, especially for agentic systems.

Because long video understanding is the first step for multimodal agents!

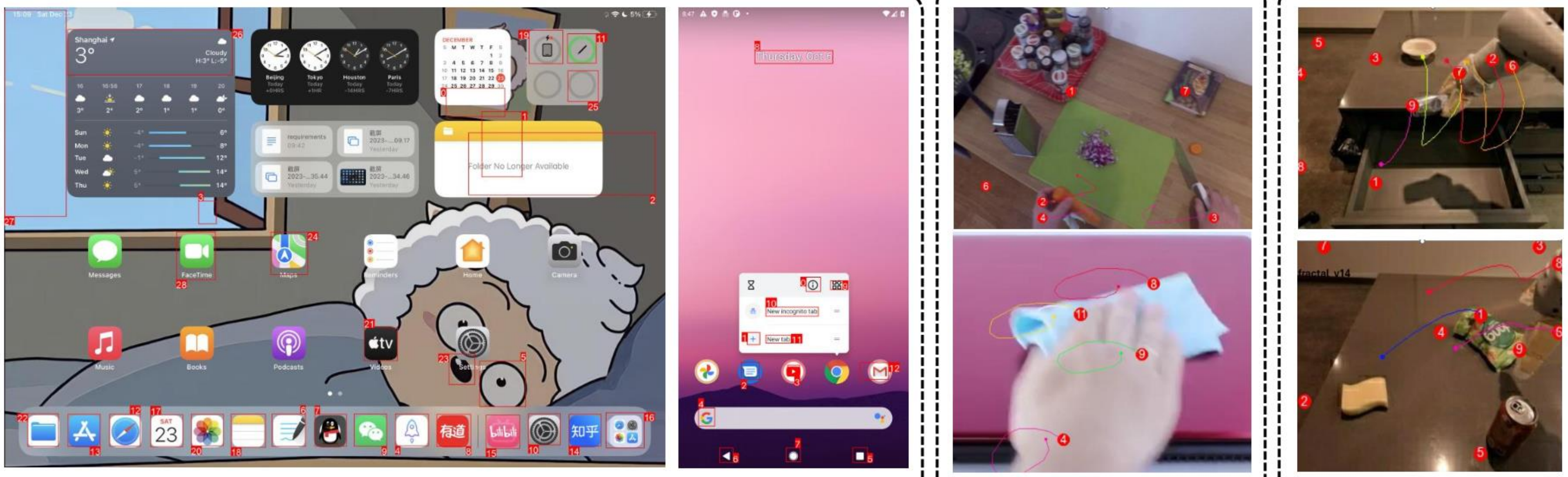


(a) UI Navigation (b) Action in Video (c) Robot Mani.

Motivation

Too many tokens is the bottleneck of current multimodal system!

- **Inefficiency**, especially multimodal agents!
- **Distract** LMMs from focusing on the key information.



(a) UI Navigation (b) Action in Video (c) Robot Mani.

Matryoshka Multimodal Models

Mu Cai[1], Jianwei Yang[2], Jianfeng Gao[2], Yong Jae Lee[1]

1 Department of Computer Sciences
UNIVERSITY OF WISCONSIN–MADISON

2 Microsoft

ICLR 2025

Motivation

Why content representation should be flexible?

Motivation

Why content representation should be flexible?

Image Complexity Varies

Motivation

What do we really want for content representation?

- A versatile, flexible system that can represent visual content in diverse granularities!

Motivation

What do we really want for content representation?

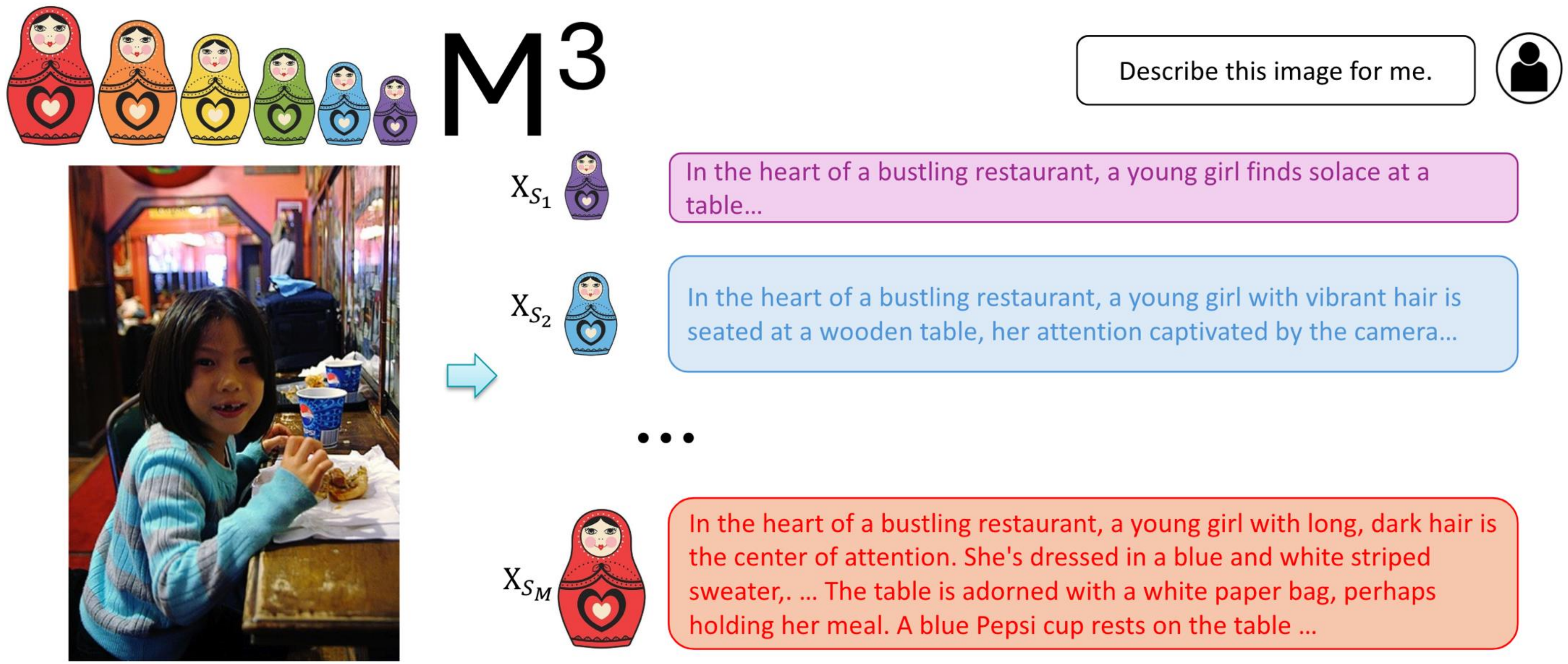
- A versatile, flexible system that can represent visual content in diverse granularities!

Merits:

- Users can manipulate how many tokens to use for a specific task.
- Increase the number of frames for video understanding.

Our Approach

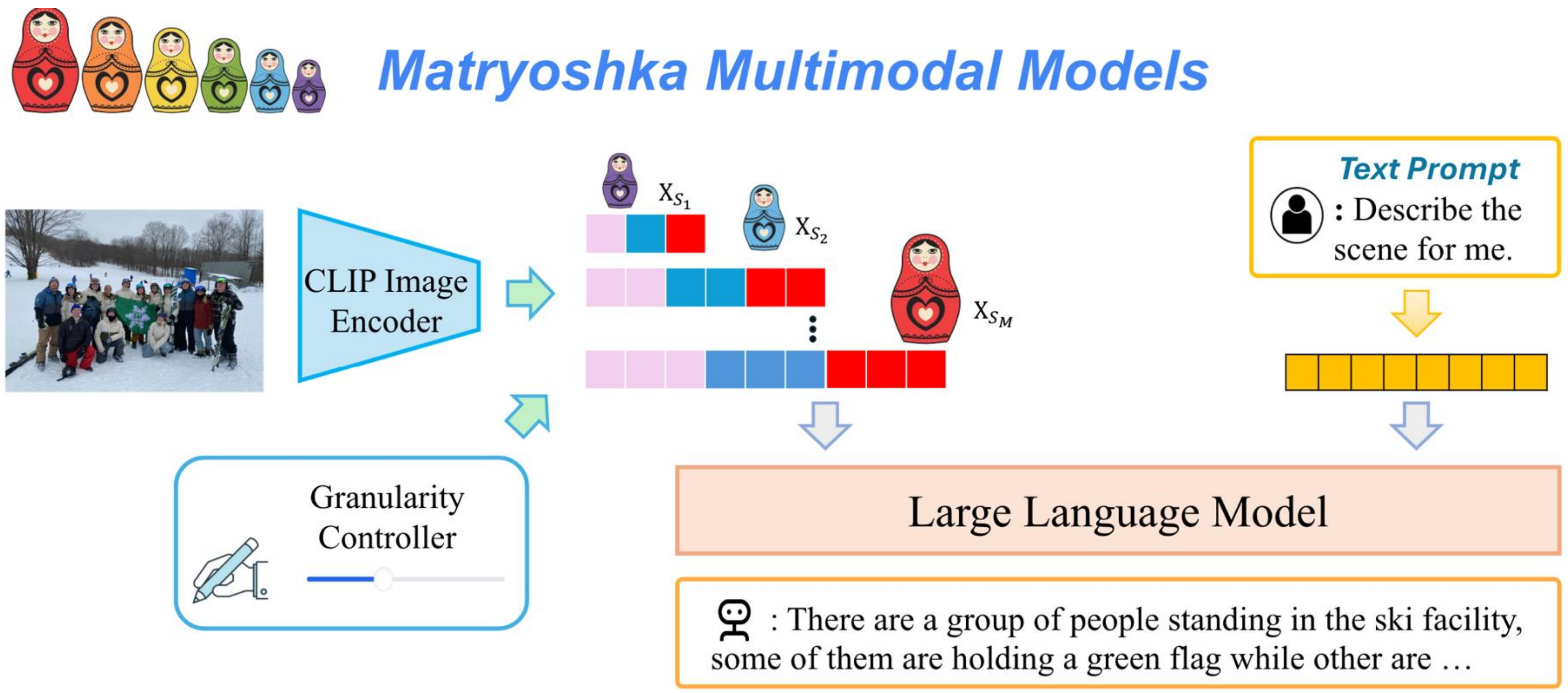
Represent visual features like Matryoshka dolls!



Ou Approach

Extremely simple

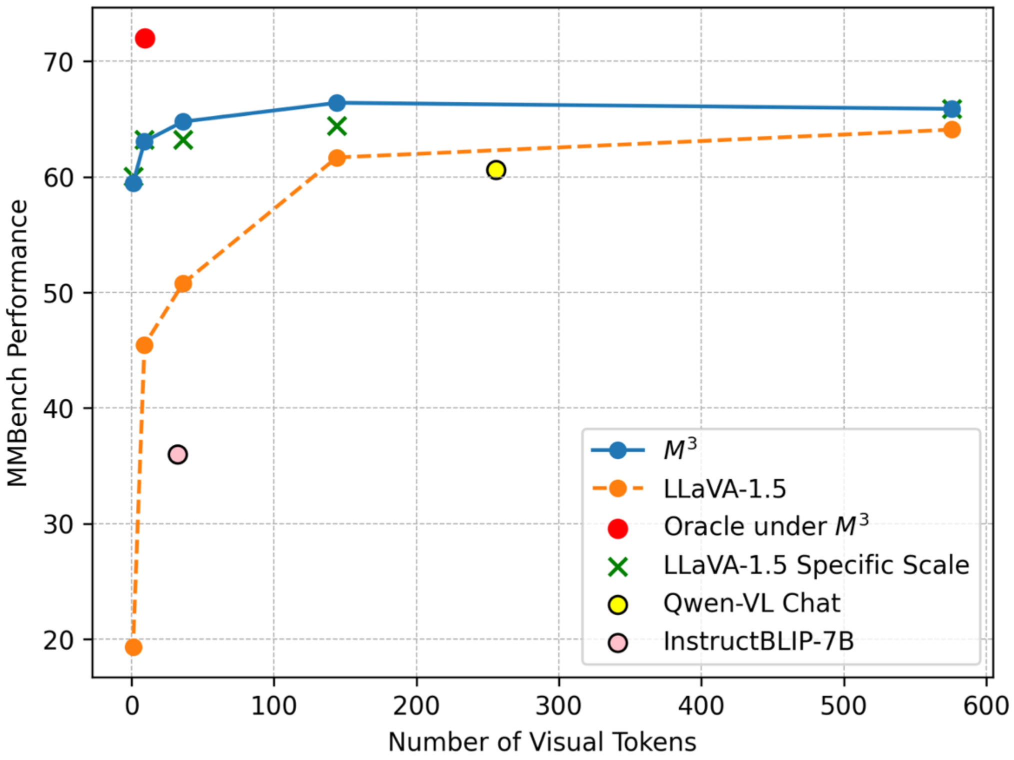
- Exact LLaVA architecture, but gradually apply **average pooling** to the $[H, W]$ visual features!
- Resulting in visual features with shape $[H, W]$, $\left[\frac{H}{2}, \frac{W}{2}\right]$, $\left[\frac{H}{4}, \frac{W}{4}\right]$, $\cdots, [1,1]$.
- Backpropagate the LLM loss upon all scales during training.



Applications # 1

Now users can control how many tokens they want!

- In MMBench, we achieve comparable performance using 9 or 36 tokens instead of full tokens.
- We are at least better than the vanilla model (LLaVA) trained at a specific scale.



Applications # 2

Different types of data prefer different number of visual tokens

- Documents need more tokens while COCO-style benchmarks need as few as 9~36 tokens.

Table: Performance of M3 on LLaVA-NeXT

# Tokens Per Grid	Approach	TextVQA	AI2D	ChartQA	DocVQA	MMBench	POPE	ScienceQA	MMMU
576	*SS*	64.53	64.83	59.28	75.40	66.58	87.02	72.29	34.3
	M^3	63.13	66.71	58.96	72.61	67.96	87.20	72.46	34.0
144	*SS*	62.16	65.77	55.28	67.69	67.78	87.66	72.15	36.4
	M^3	62.61	68.07	57.04	66.48	69.50	87.67	72.32	36.1
36	*SS*	58.15	65.90	45.40	56.89	67.01	86.75	71.87	36.2
	M^3	58.71	67.36	50.24	55.94	68.56	87.29	72.11	36.8
9	*SS*	50.95	65.06	37.76	44.21	65.29	85.62	72.37	36.8
	M^3	51.97	66.77	42.00	43.52	67.35	86.17	71.85	35.2
1	*SS*	38.39	63.76	28.96	33.11	61.43	82.83	72.32	35.3
	M^3	38.92	64.57	31.04	31.63	62.97	83.38	71.19	34.8
		-26%				-4%			

Applications # 2

Most video benchmarks achieve similar accuracies with 1.6% tokens…

- We can prune visual token more than we imagined
- Using full tokens does not always result in best performance

Approach	# Tokens	MSVD	MSRVTT	ActivityNet	NextQA	IntentQA	EgoSchema
Video-LLaMA [11]	32	51.6	29.6	12.4	-	-	-
LLaMA-Adapter [62]	-	54.9	43.8	34.2	-	-	-
Video-ChatGPT [63]	264+	64.9	49.3	35.2	-	-	-
Video-LLaVA [64]	2048	70.7	59.2	45.3	-	-	-
InternVideo [65]	-	-	-	-	59.1	-	32.1
LLaVA-NeXT-7B [4]	2880	78.8	63.7	54.3	**63.1**	**60.3**	35.8
LLaVA-NeXT-7B-M^3	2880	78.2	**64.5**	53.9	**63.1**	58.8	36.8
	720	**79.0**	**64.5**	**55.0**	62.6	59.6	37.2
	180	77.9	63.7	**55.0**	61.4	59.3	37.6
	45	75.8	63.0	53.2	59.5	58.7	**38.8**
	5	73.5	62.7	50.8	56.5	56.7	36.2

Applications # 2

Most video benchmarks achieve similar accuracies with 1.6% tokens…

Approach	# Tokens	MSVD	MSRVTT	ActivityNet	NextQA	IntentQA	EgoSchema
Video-LLaMA [11]	32	51.6	29.6	12.4	-	-	-
LLaMA-Adapter [62]	-	54.9	43.8	34.2	-	-	-
Video-ChatGPT [63]	264+	64.9	49.3	35.2	-	-	-
Video-LLaVA [64]	2048	70.7	59.2	45.3	-	-	-
InternVideo [65]	-	-	-	-	59.1	-	32.1
LLaVA-NeXT-7B [4]	2880	78.8	63.7	54.3	**63.1**	**60.3**	35.8
LLaVA-NeXT-7B-M^3	2880	78.2	**64.5**	53.9	**63.1**	58.8	36.8
	720	**79.0**	**64.5**	**55.0**	62.6	59.6	37.2
	180	77.9	63.7	**55.0**	61.4	59.3	37.6
	45	75.8	63.0	53.2	59.5	58.7	**38.8**
	5	73.5	62.7	50.8	56.5	56.7	36.2

Or such video benchmarks are not really evaluating video understanding?

We propose ***TemporalBench*** and ***Vinoground*** to solve this problem.

Applications # 3

Good side-effect:

- M3 serving as an image complexity evaluator.

Q: how much is a polos crazy bike?

Q: what directive is the sign giving?

# Tokens	576	144	36	9	1
Correct?	✓	✓	✓	✓	✓

Q: what number is on the black and white sign?

Q: what brand is the apricot brandy?

# Tokens	576	144	36	9	1
Correct?	✓	✓	✓	✗	✗

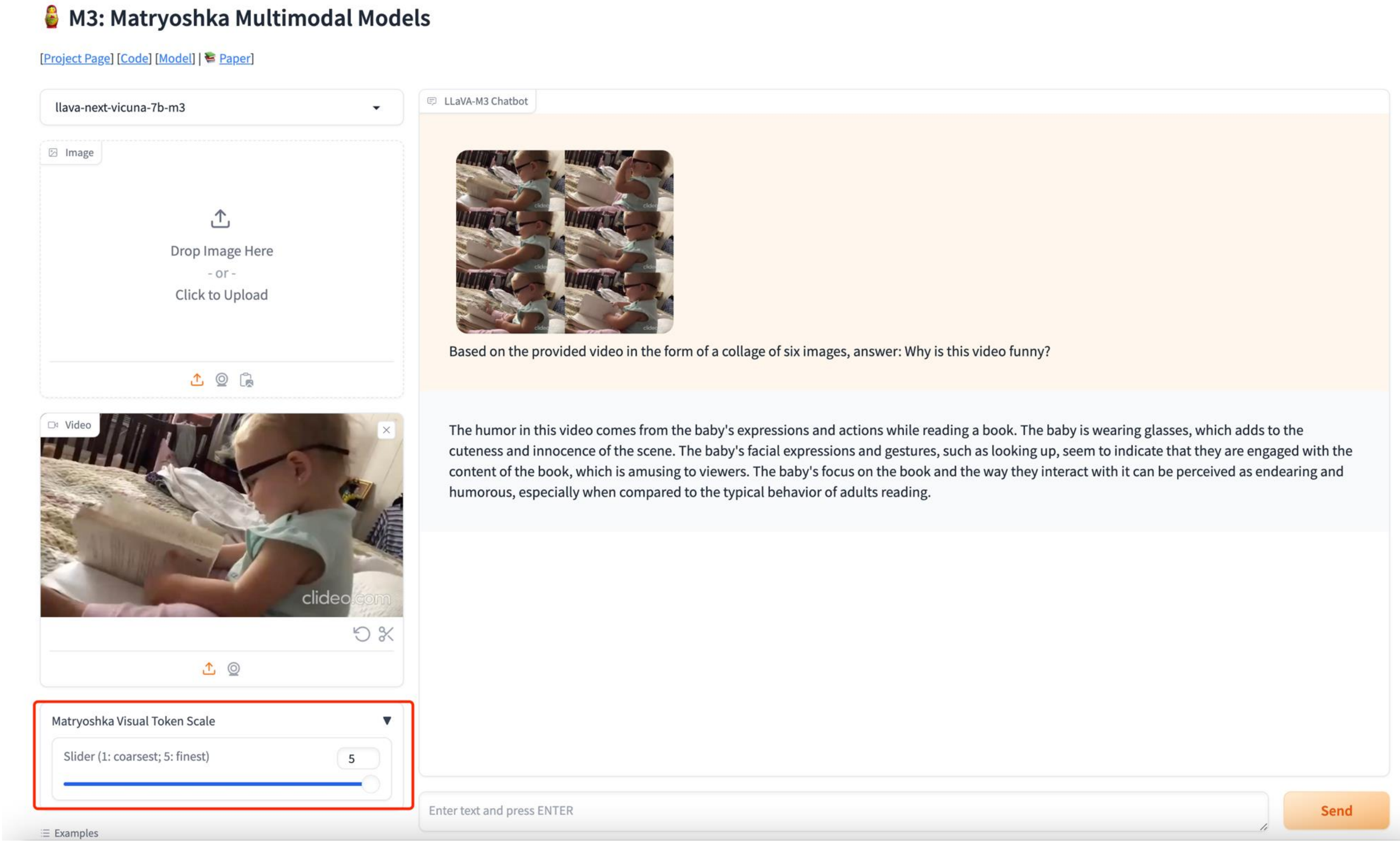
Q: what beer company is a sponsor on the score board?

Q: what is the telephone number of andrew yates?

# Tokens	576	144	36	9	1
Correct?	✗	✗	✗	✗	✗

Demo

https://pages.cs.wisc.edu/~mucai/matryoshka-mm.html



Impact

Matryoshka has already ignited many interesting tasks, such as (1) Matryoshka VQVAE.

ADAPTIVE LENGTH IMAGE TOKENIZATION VIA RECURRENT ALLOCATION

Shivam Duggal **Phillip Isola** **Antonio Torralba** **William T. Freeman**
MIT CSAIL

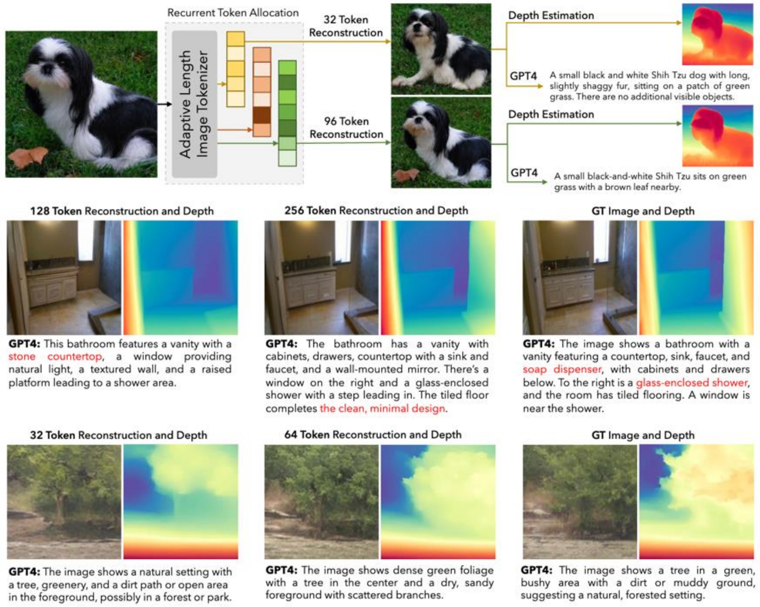


Figure 1: **Adaptive Length Image Tokenization** maps an image to multiple variable-length representations through a recurrent token allocation process, **enabling task-specific sampling**. We learn the tokenizer via image reconstruction as a self-supervised objective. While a compressed representation can be optimized for specific tasks (e.g., fewer tokens for "dog", "leaf", "grass" may suffice for a VLM task), reconstruction objective supports learning a universal, task-agnostic tokenizer.

CAT: Content-Adaptive Image Tokenization

Junhong Shen[1*] Kushal Tirumala[2] Michihiro Yasunaga[2]
Ishan Misra[2] Luke Zettlemoyer[2] Lili Yu[2†] Chunting Zhou[†‡]

[1] Carnegie Mellon University
[2] Meta

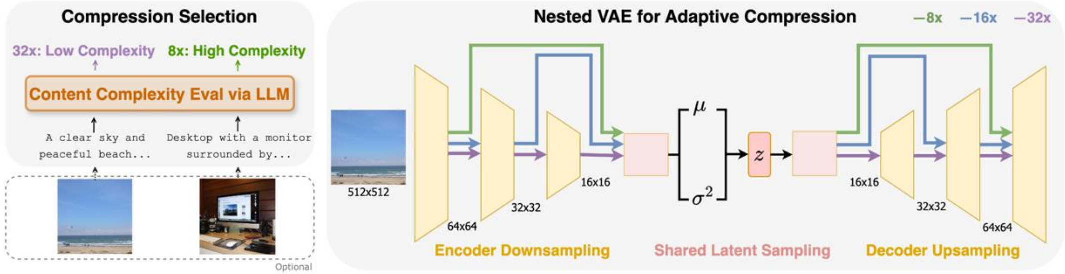


Figure 1. **Content-Adaptive Tokenization.** CAT uses an LLM to evaluate the content complexity and determine the optimal compression ratio based on the image's text description. The image is processed by a nested VAE architecture that dynamically routes the input according to the selected compression ratio. The resulting latent representations thus have varying spatial dimensions. Images shown in the figure are taken from COCO 2014 [9].

Thanks for Listening!

- Looking forward to any comment!

THANK YOU!

Mu CAI, UW-Madison CS

Project Page