

COMP3361 – Natural Language Processing: Multimodal Language Models

Chen Wu

Carnegie Mellon University

Language models are becoming multimodal

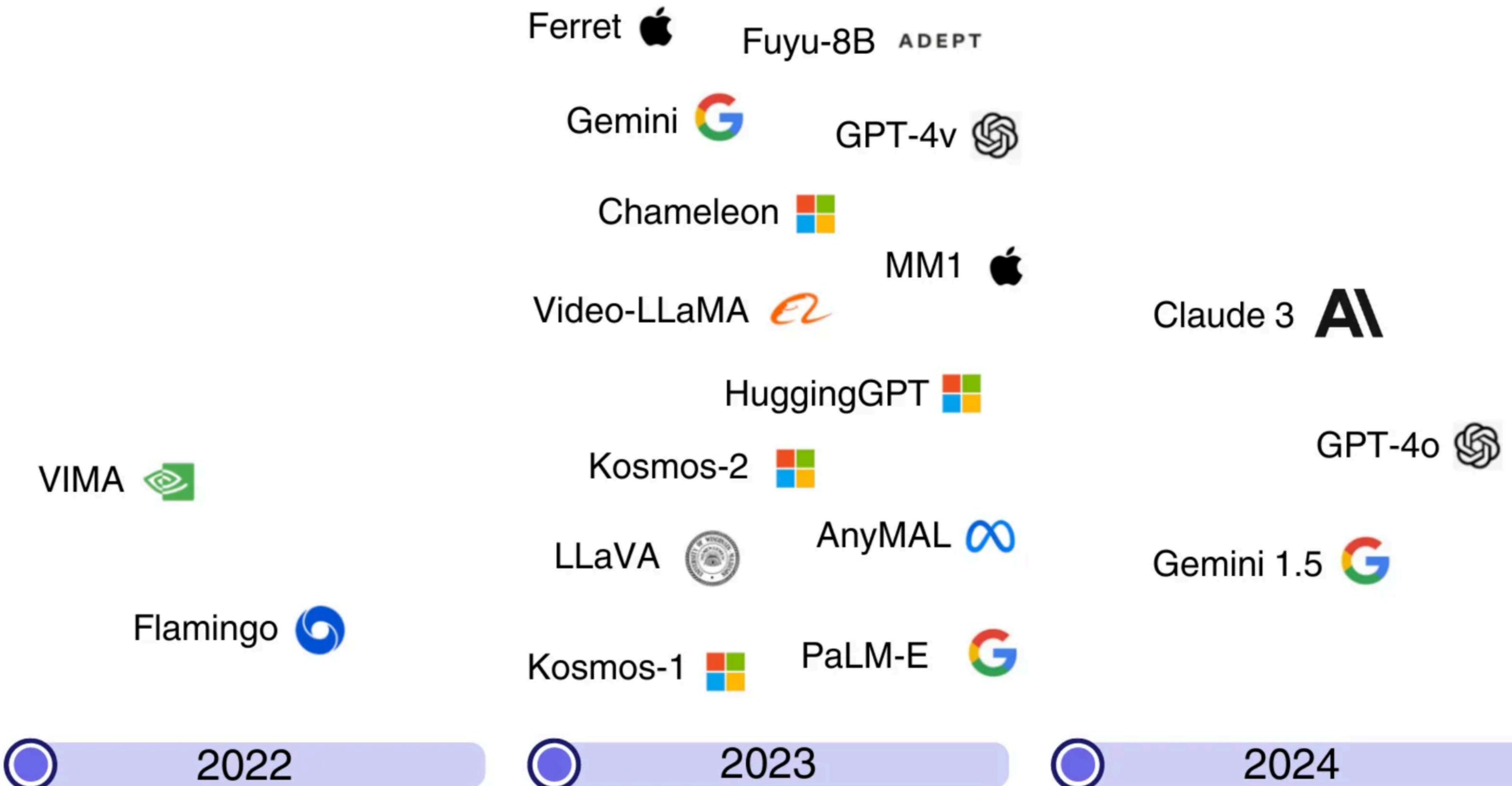


Figure source: Multimodal Large Language Models (MLLMs) transforming Computer Vision (Medium post)

Multimodal LMs good



swax

@swax



...

Wow, nailed it and not even a tree in sight.

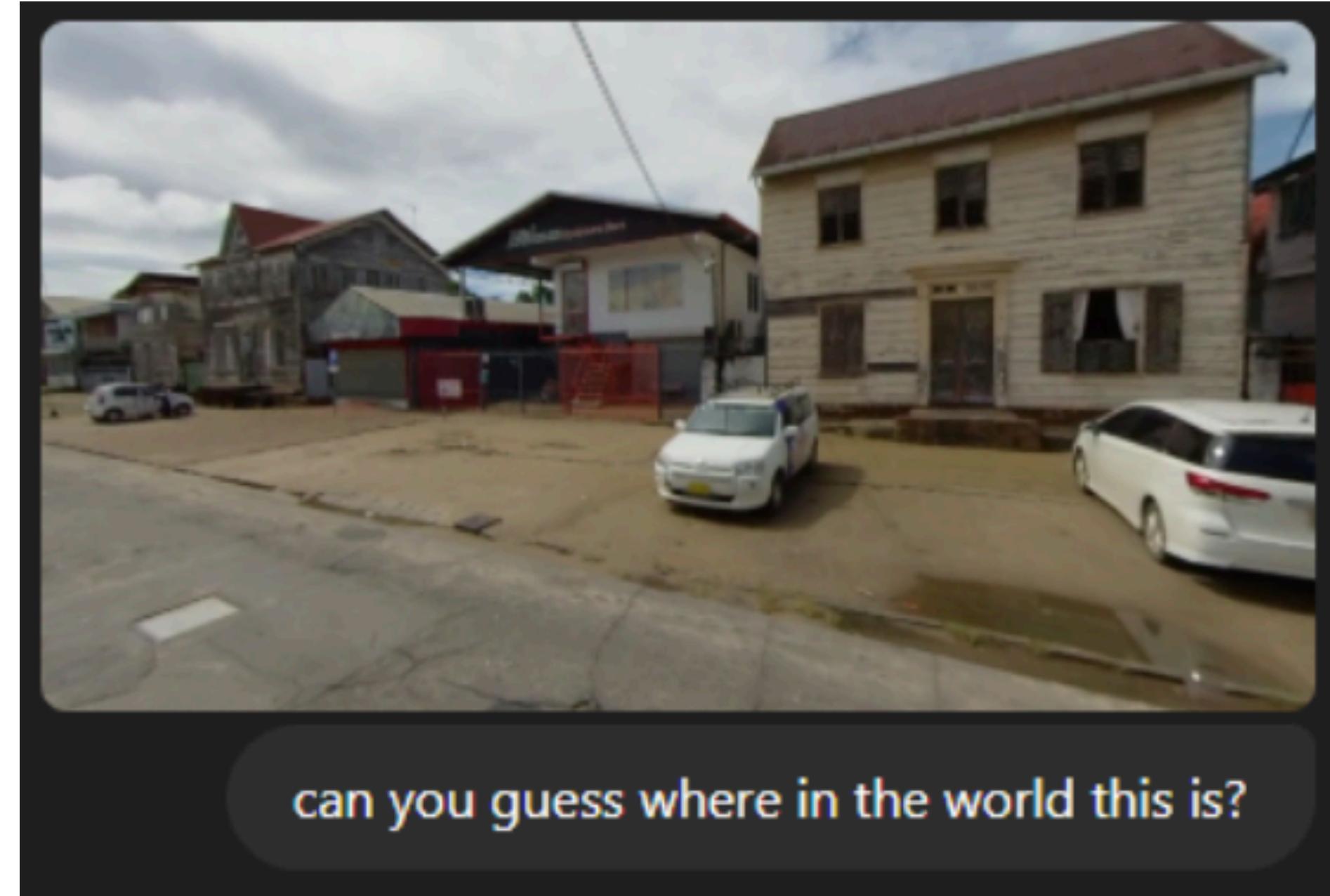
can you guess where in the world this is?

Thought for 1m 15s >

My best guess is Paramaribo, Suriname.

Why I think so

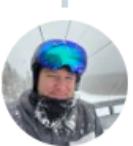
Clue	What it suggests
Yellow-on-black license plates	That color scheme is standard for private cars in Suriname.
Tall wooden colonial houses with louvered shutters	Very typical of the Dutch-colonial core of Paramaribo (a UNESCO World-Heritage area).
Left-hand-drive cars, but traffic keeps to the left	Both Suriname and neighboring Guyana drive on the left, unlike most of South America.
Language on the shop fascia looks like a Latin alphabet business name rather than Spanish or Portuguese	Fits Suriname, where Dutch is official but English signage is common.
Low-lying, sandy lot and drainage puddles	Characteristic of Paramaribo's coastal flood-plain setting.



Thought for 1m 15s >

My best guess is Paramaribo, Suriname.

Multimodal LMs good



swax
@swax



Wow, nailed it and not even a tree in sight.

Thought for 1m 15s >

My best guess is Paramaribo, Suriname.

Why I think so

Clue

Yellow-on-black license plates

Tall wooden colonial houses with louvers

Left-hand-drive cars, but traffic keeps to the left

Language on the shop fascia looks like a Latin alphabet business name rather than Spanish or Portuguese

Low-lying, sandy lot and drainage puddles

is standard for private cars in Suriname.

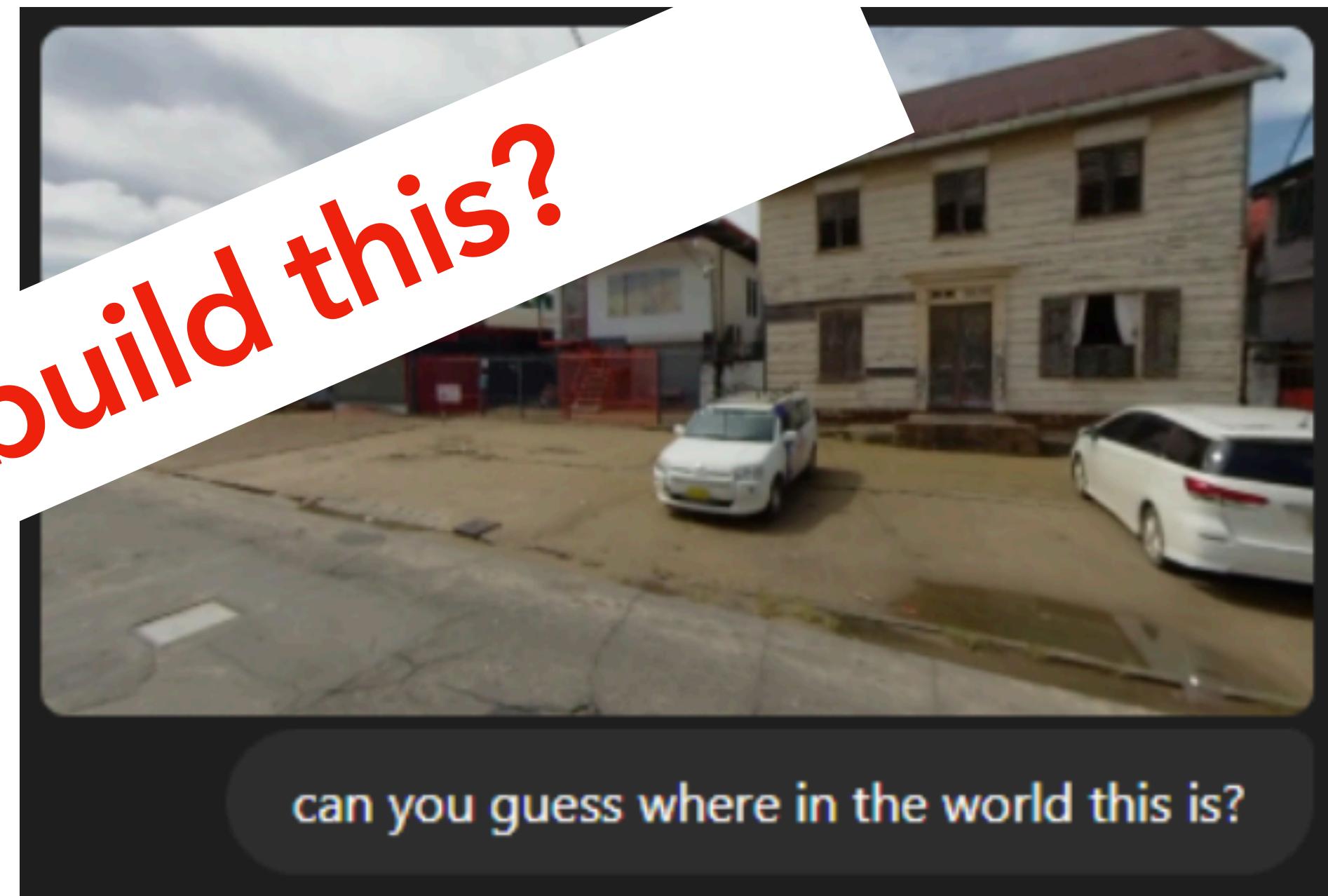
is typical of the Dutch-colonial core of Paramaribo (a UNESCO World-Heritage area).

Both Suriname and neighboring Guyana drive on the left, unlike most of South America.

Fits Suriname, where Dutch is official but English signage is common.

Characteristic of Paramaribo's coastal flood-plain setting.

How do we build this?



Thought for 1m 15s >

My best guess is Paramaribo, Suriname.

A lot of stuff going on...



The screenshot shows the arXiv search interface. At the top is a red header bar with the arXiv logo on the left, a search bar with a dropdown menu for 'All fields' and a 'Search' button on the right, and links for 'Help | Advanced Search' and 'Login' below it.

Showing 1–50 of 3,446 results

Search v0.5.6 released 2020-02-24

Query: `order: -announced_date_first; size: 50; hide_abstracts: True; date_range: from 2024-04-01 to 2025-04-24; include_cross_list: True; terms: AND title=multimodal`

[Simple Search](#)

[Refine query](#)

[New search](#)

50 results per page. Sort results by Announcement date (newest first) [Go](#)

[1](#) [2](#) [3](#) [4](#) [5](#) ...

[Next](#)

1. [arXiv:2504.16798](#) [pdf, other] [cs.MM](#) [cs.CV](#) [cs.LG](#)

4D Multimodal Co-attention Fusion Network with Latent Contrastive Alignment for Alzheimer's Diagnosis

Authors: [Yuxiang Wei](#), [Yanteng Zhang](#), [Xi Xiao](#), [Tianyang Wang](#), [Xiao Wang](#), [Vince D. Calhoun](#)

Submitted 23 April, 2025; originally announced April 2025.

2. [arXiv:2504.16656](#) [pdf, other] [cs.CV](#)

Skywork R1V2: Multimodal Hybrid Reinforcement Learning for Reasoning

Authors: [Chris](#), [Yichen Wei](#), [Yi Peng](#), [Xiaokun Wang](#), [Weijie Qiu](#), [Wei Shen](#), [Tianyidan Xie](#), [Jiangbo Pei](#), [Jianhao Zhang](#), [Yunzhuo Hao](#), [Xuchen Song](#), [Yang Liu](#), [Yahui Zhou](#)

Submitted 23 April, 2025; originally announced April 2025.

This lecture:

Focus on the Qwen2.5-VL tech report

Qwen2.5-VL good

		Qwen2.5-VL 72B	Gemini-2 Flash	GPT-4o	Claude3.5 Sonnet	Qwen2-VL 72B	Other Best Open LLM
College-level Problems	MMMU	70.2	70.7	70.3	70.4	64.5	70.1
	MMMU Pro	51.1	57.0	54.5	54.7	46.2	52.7
Document and Diagrams Reading	DocVQA	96.4	92.1	91.1	95.2	96.5	96.1
	InfoVQA	87.3	77.8	80.7	74.3	84.5	84.1
	CC-OCR	79.8	73.0	66.6	62.7	68.7	68.7
	OCRBenchV2	61.5	-	46.5	45.2	47.8	47.8
General Visual Question Answering	MegaBench	51.3	55.2	54.2	52.1	46.8	47.4
	MMStar	70.8	69.4	64.7	65.1	68.3	69.5
	MMBench1.1	88.0	83.0	82.1	83.4	86.6	87.4
Math	MathVista	74.8	73.1	63.8	65.4	70.5	72.3
	MathVision	38.1	41.3	30.4	38.3	25.9	32.2
Video Understanding	VideoMME	73.3	-	71.9	60.0	71.2	72.1
	MMBench-Video	2.0	-	1.7	1.4	1.7	1.9
	LVBench	47.3	-	30.8	-	-	43.6
	CharadesSTA	50.9	-	35.7	-	-	48.4
Visual Agent	AITZ	83.2	-	35.3	-	-	53.3
	Android Control	67.4	-	-	-	-	66.4
	ScreenSpot	87.1	84.0	18.1	83.0	-	89.5
	ScreenSpot Pro	43.6	-	-	17.1	-	38.1
	AndroidWorld	35.0	-	34.5	27.9	-	46.6
	OSWorld	8.8	-	5.0	14.9	-	22.7



March 5, 2025

Qwen2.5-VL Technical Report

Qwen Team, Alibaba Group

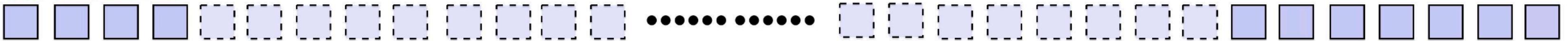
-  <https://chat.qwenlm.ai>
-  <https://huggingface.co/Qwen>
-  <https://modelscope.cn/organization/qwen>
-  <https://github.com/QwenLM/Qwen2.5-VL>

Outline

- ▶ Architecture (main focus)
- ▶ Pretraining
- ▶ Post-training
- ▶ What's next

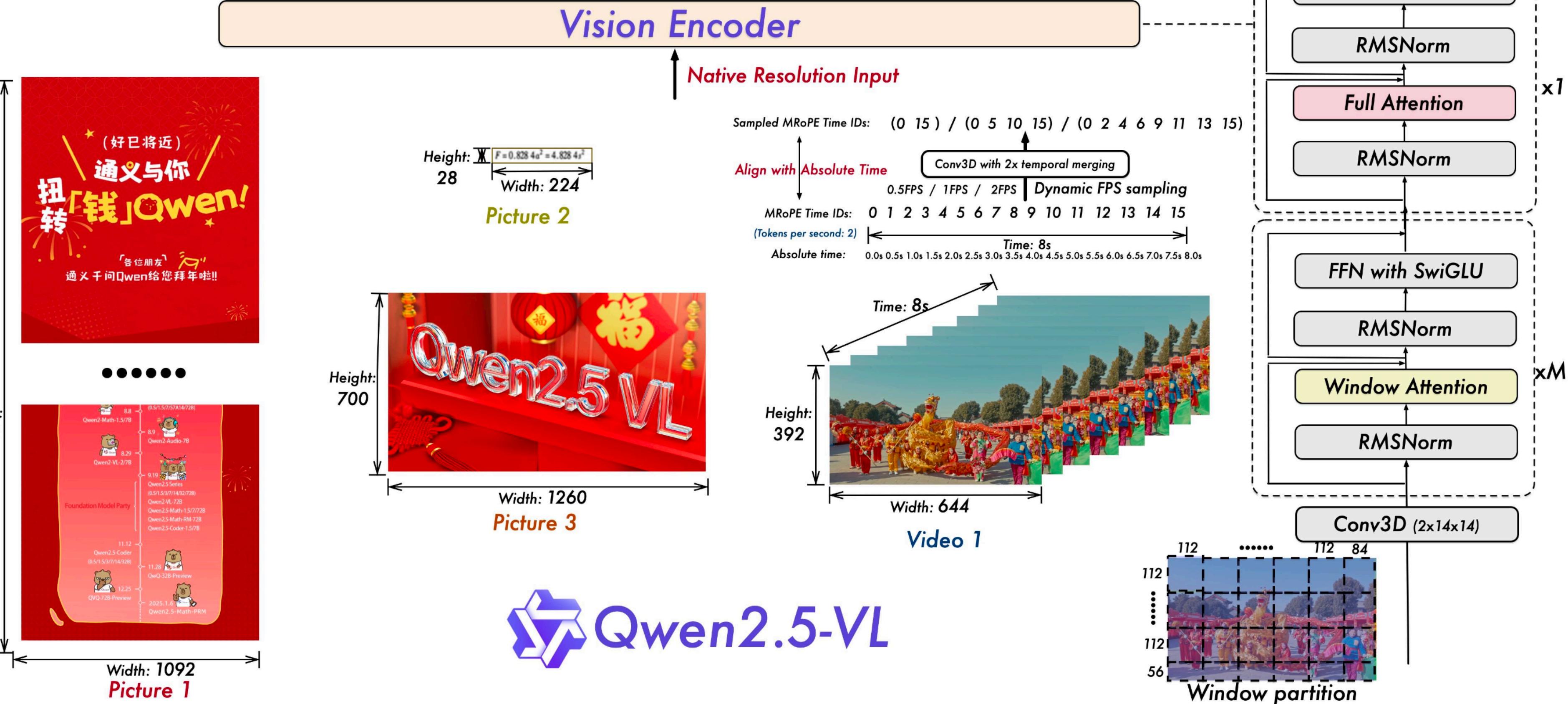
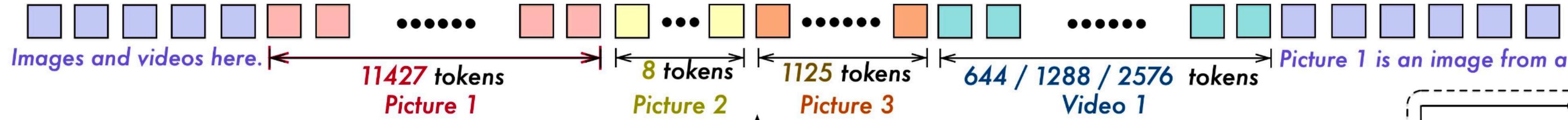
Architecture

and videos here.



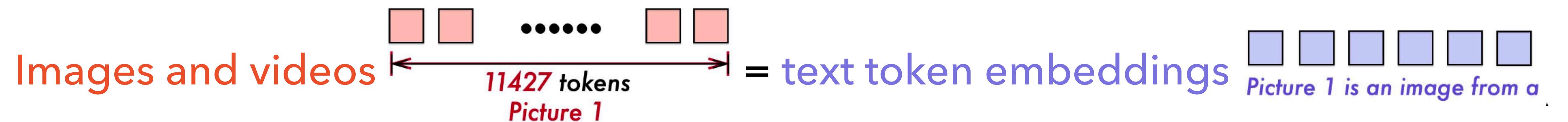
Picture 1 is an image from a blog

Qwen2.5 LM Decoder

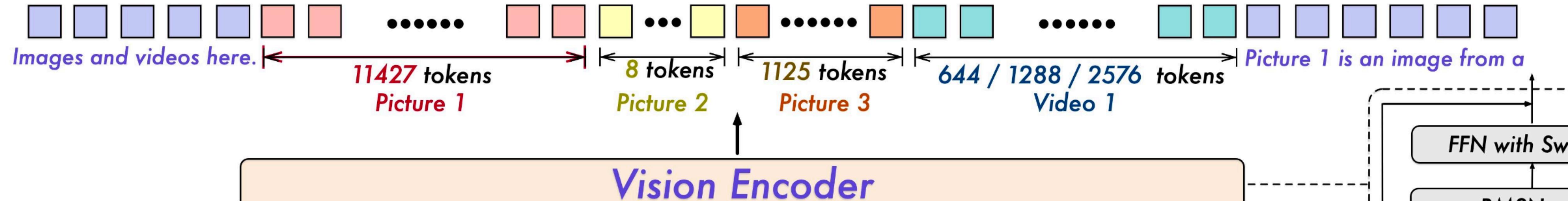


 **Qwen2.5-VL**

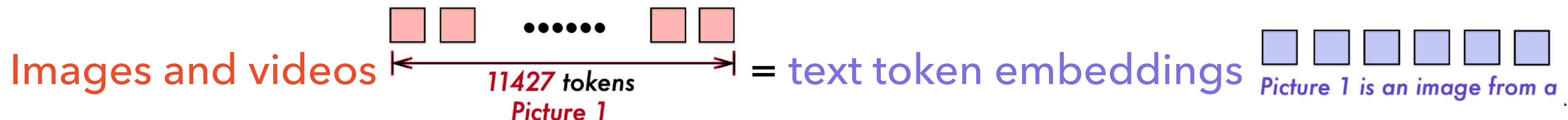
Visual tokens



Qwen2.5 LM Decoder



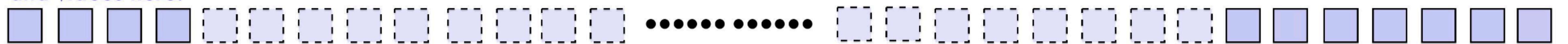
Visual tokens



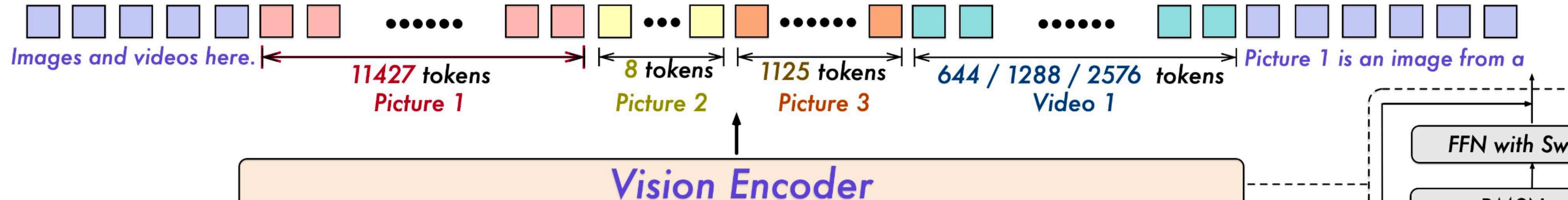
So everything else is just language modeling!

- Except a small change that we'll talked about later :)

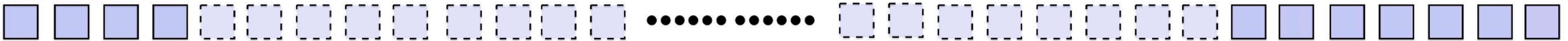
and videos here.



Qwen2.5 LM Decoder

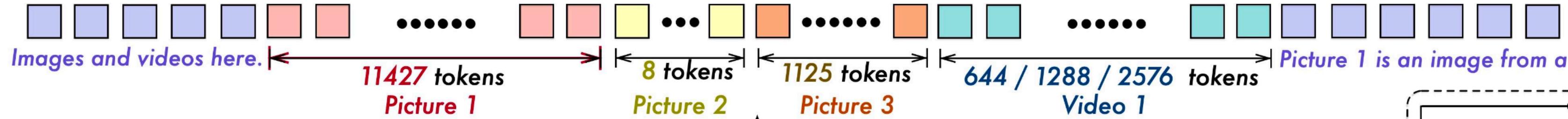


and videos here.



Picture 1 is an image from a blog

Qwen2.5 LM Decoder



Vision Encoder



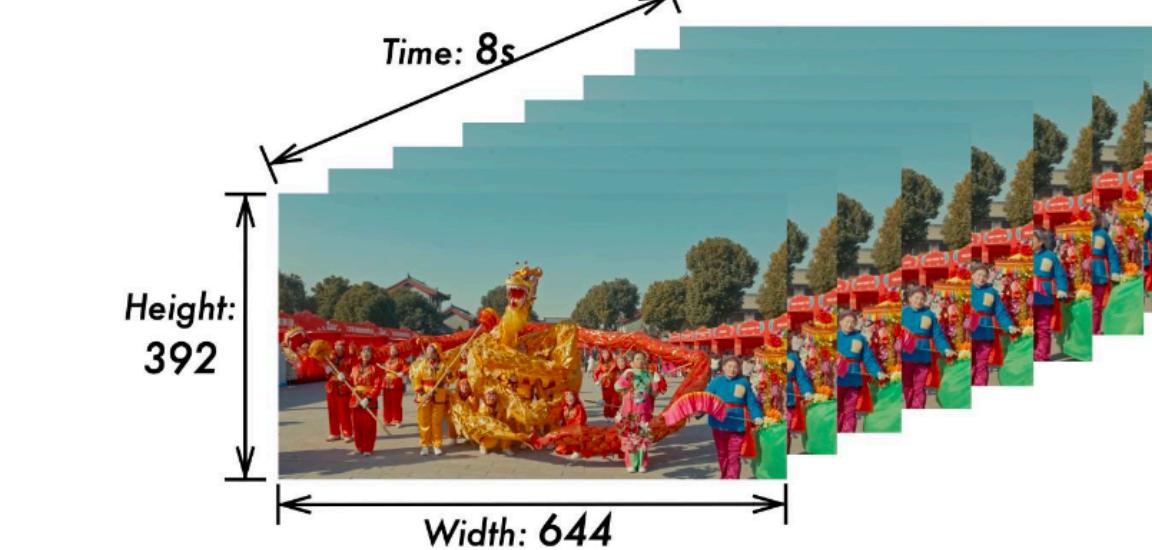
Height: 28
Width: 224

Picture 2

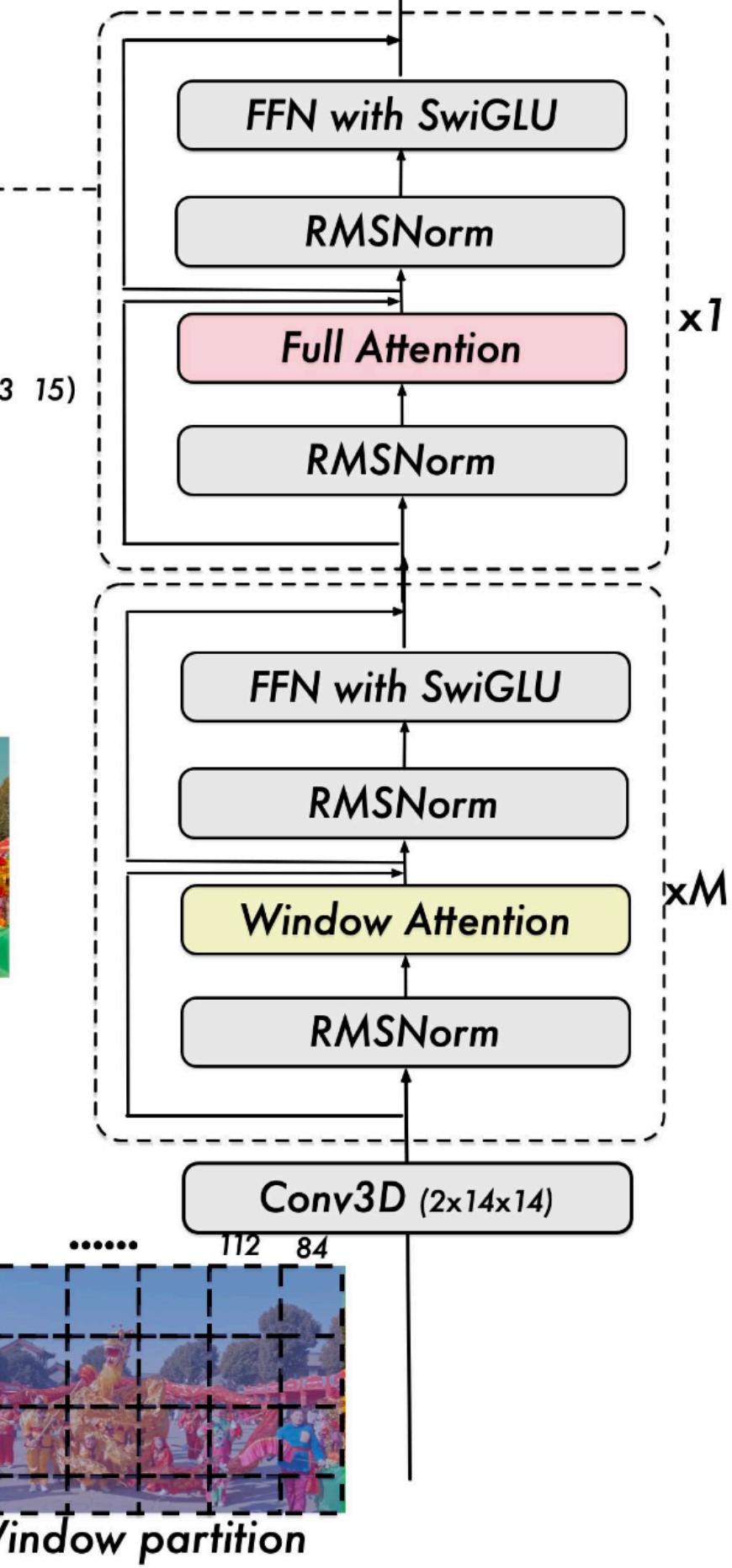


Picture 3

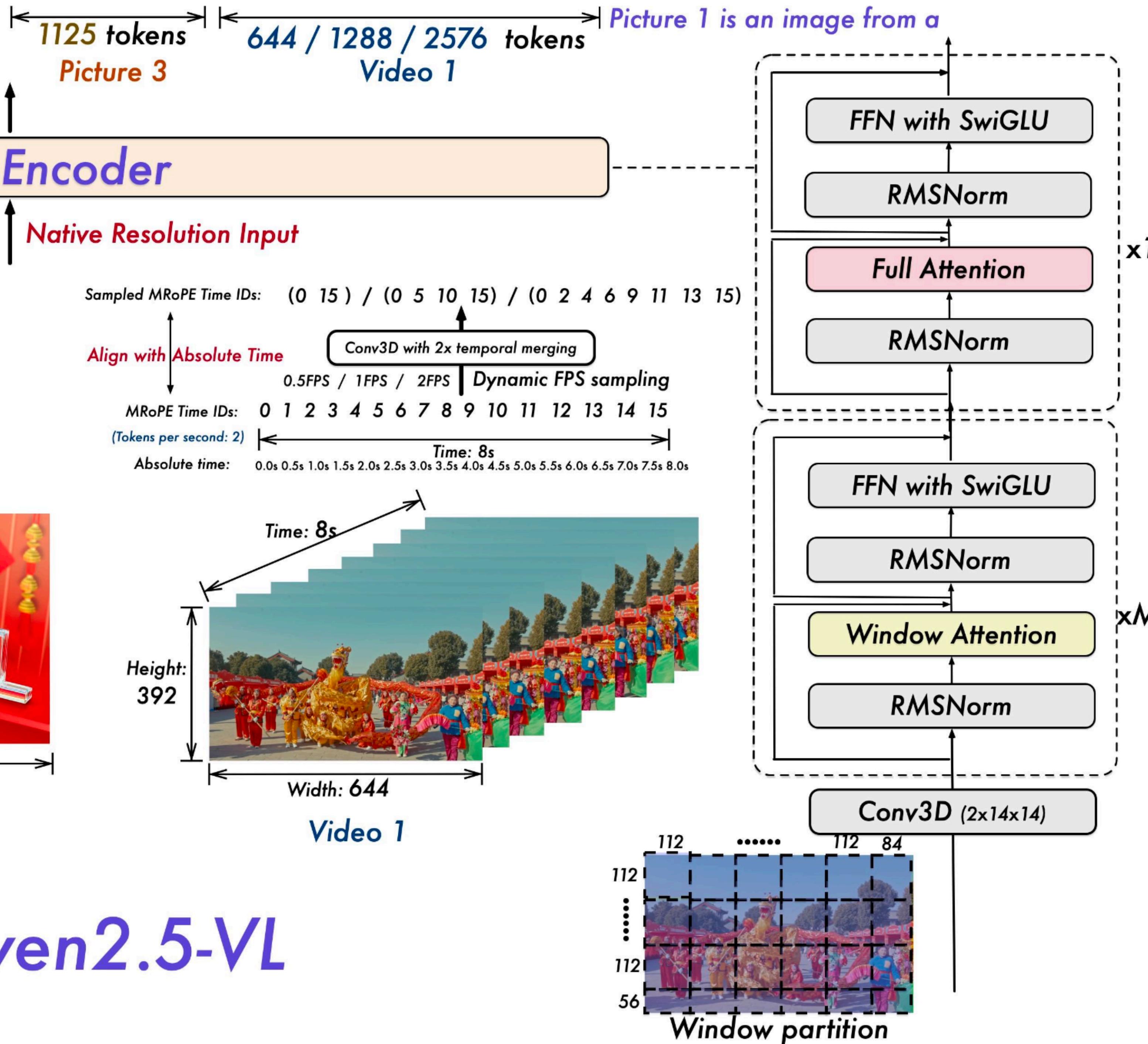
Sampled MRoPE Time IDs: (0 15) / (0 5 10 15) / (0 2 4 6 9 11 13 15)
Align with Absolute Time
MRoPE Time IDs: 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
(Tokens per second: 2) Time: 8s
0.0s 0.5s 1.0s 1.5s 2.0s 2.5s 3.0s 3.5s 4.0s 4.5s 5.0s 5.5s 6.0s 6.5s 7.0s 7.5s 8.0s
Conv3D with 2x temporal merging
0.5FPS / 1FPS / 2FPS Dynamic FPS sampling



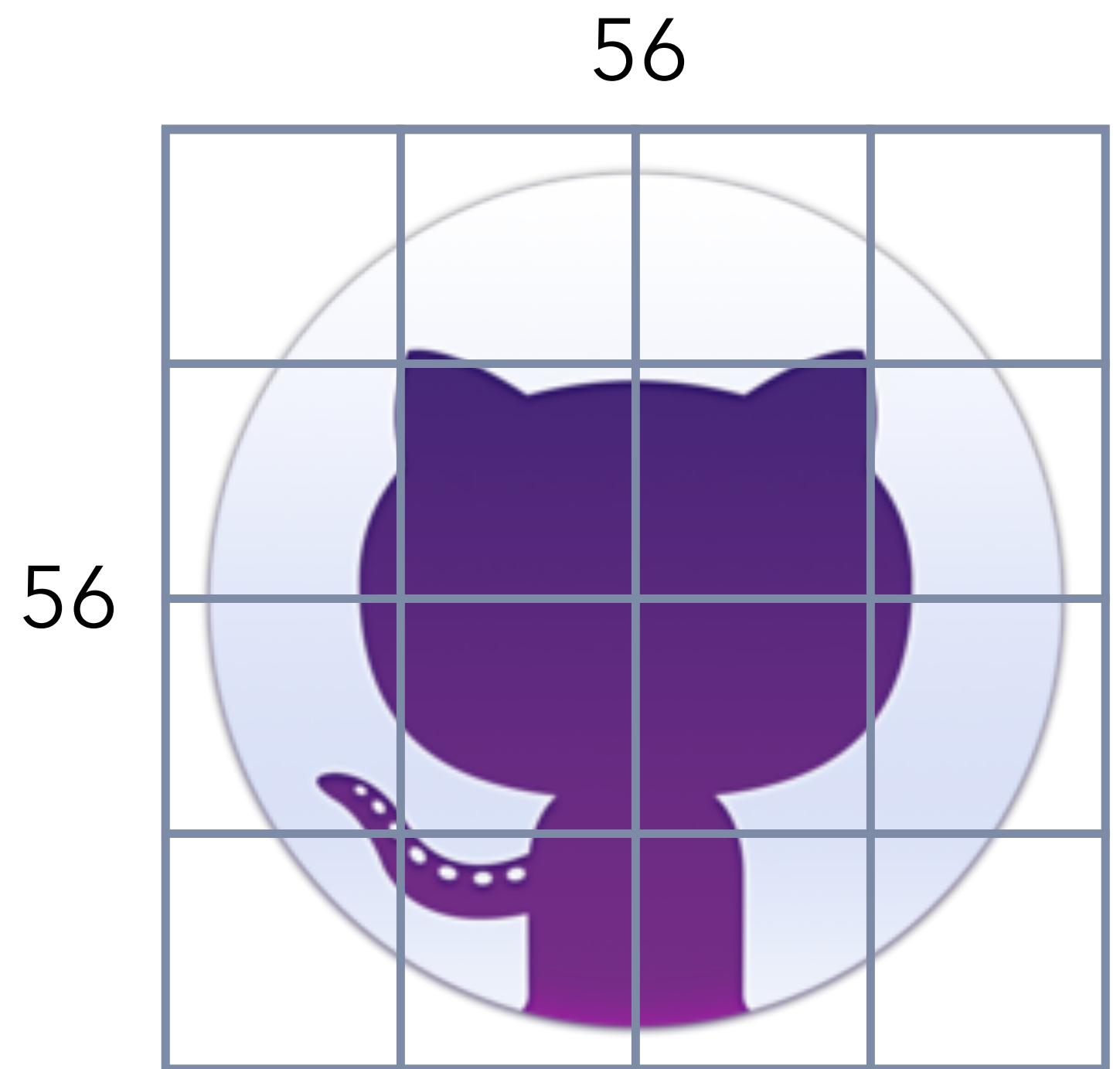
Video 1



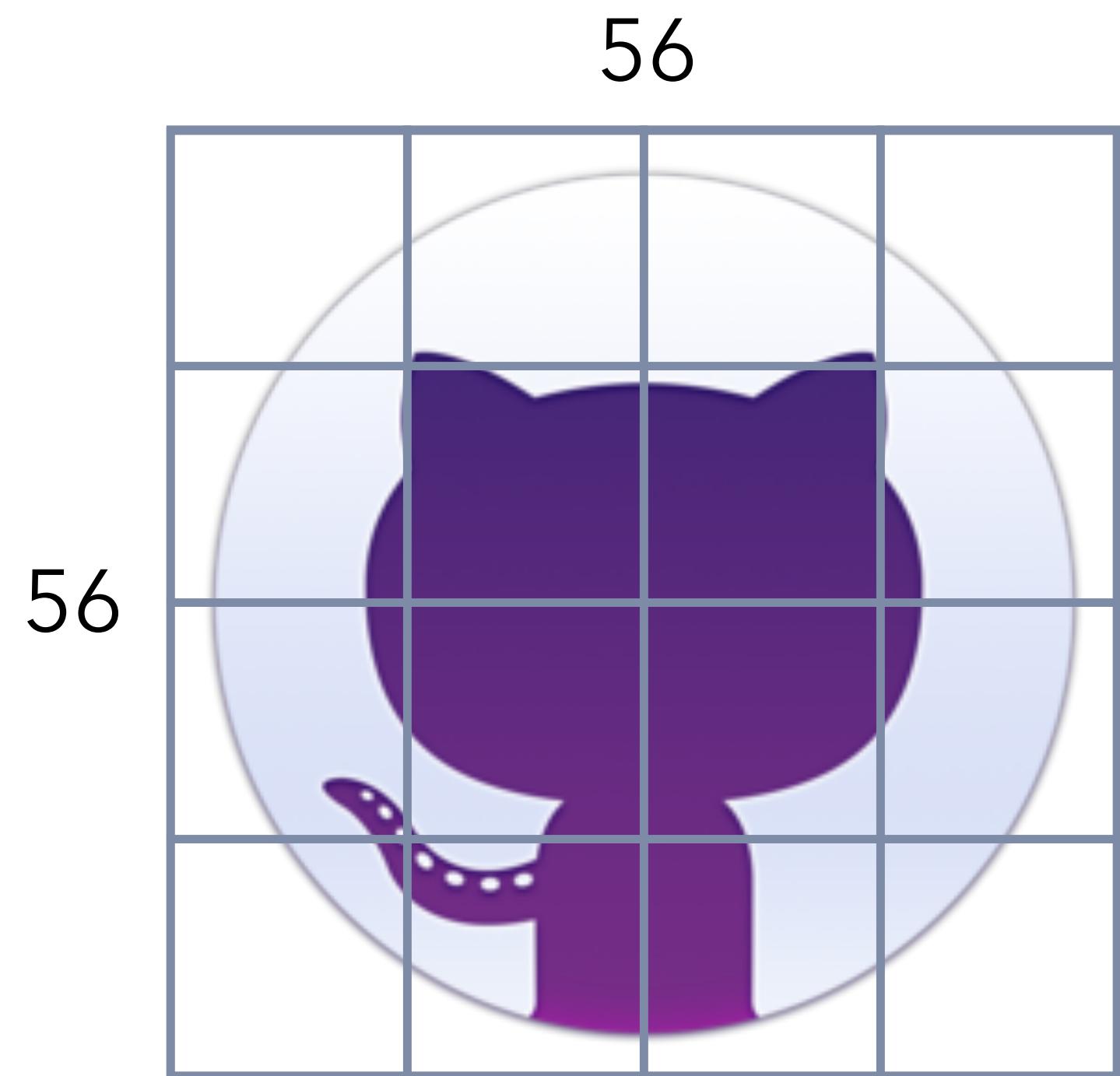
 **Qwen2.5-VL**



Vision encoder



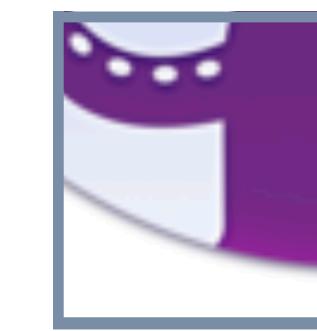
Vision encoder



Input: image patches



...

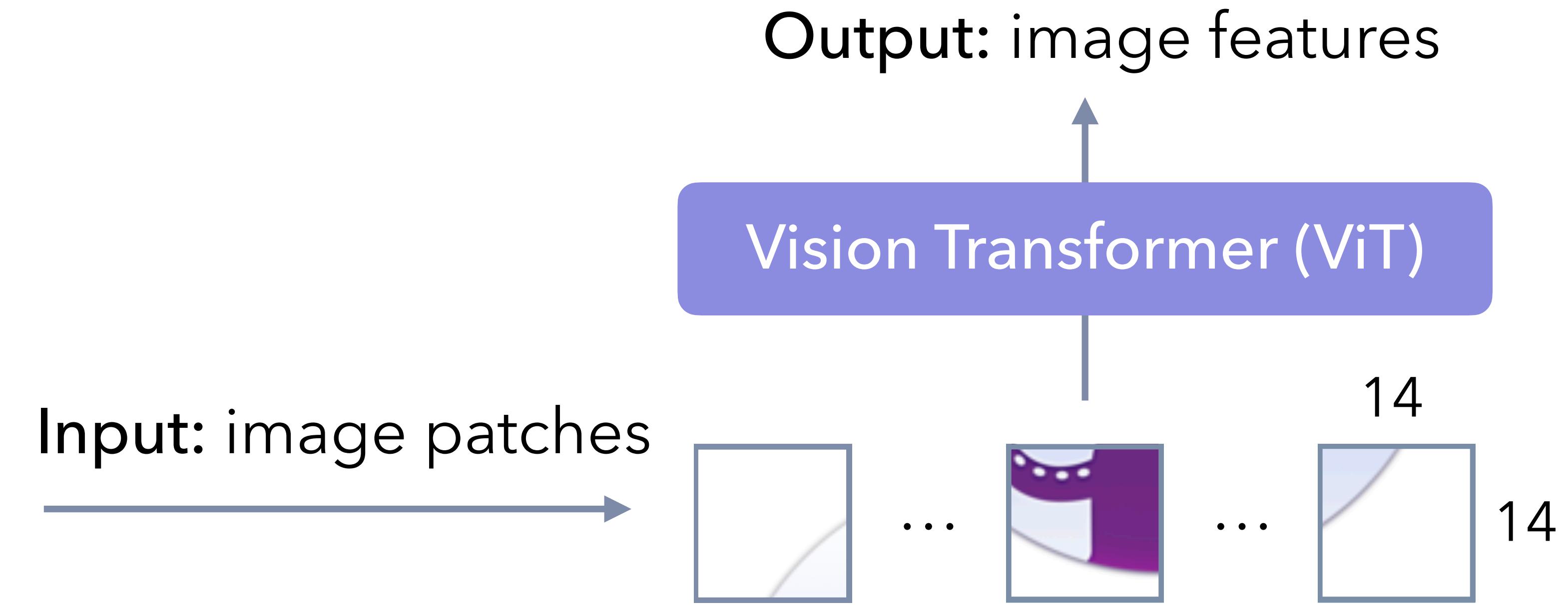
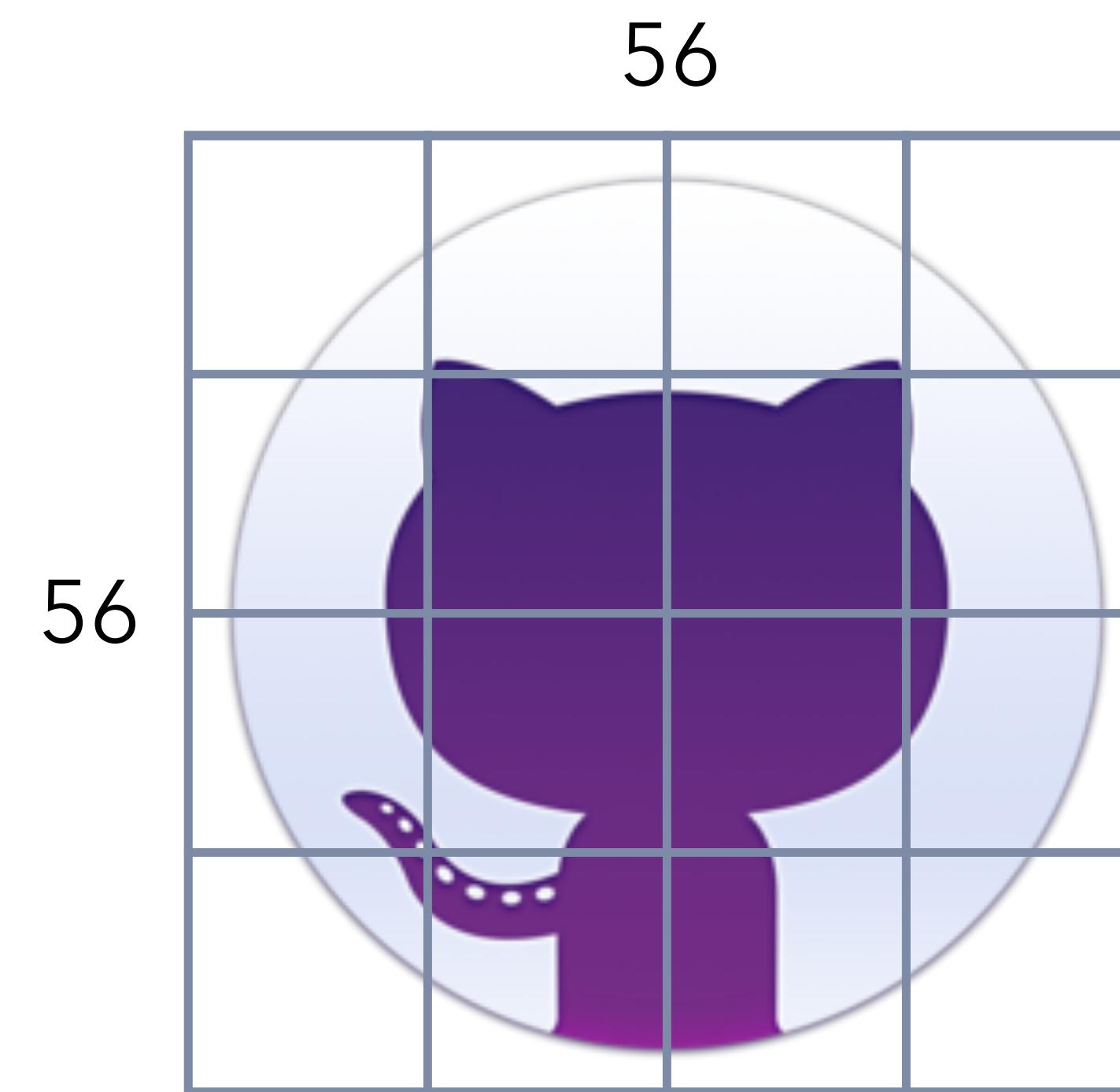


...

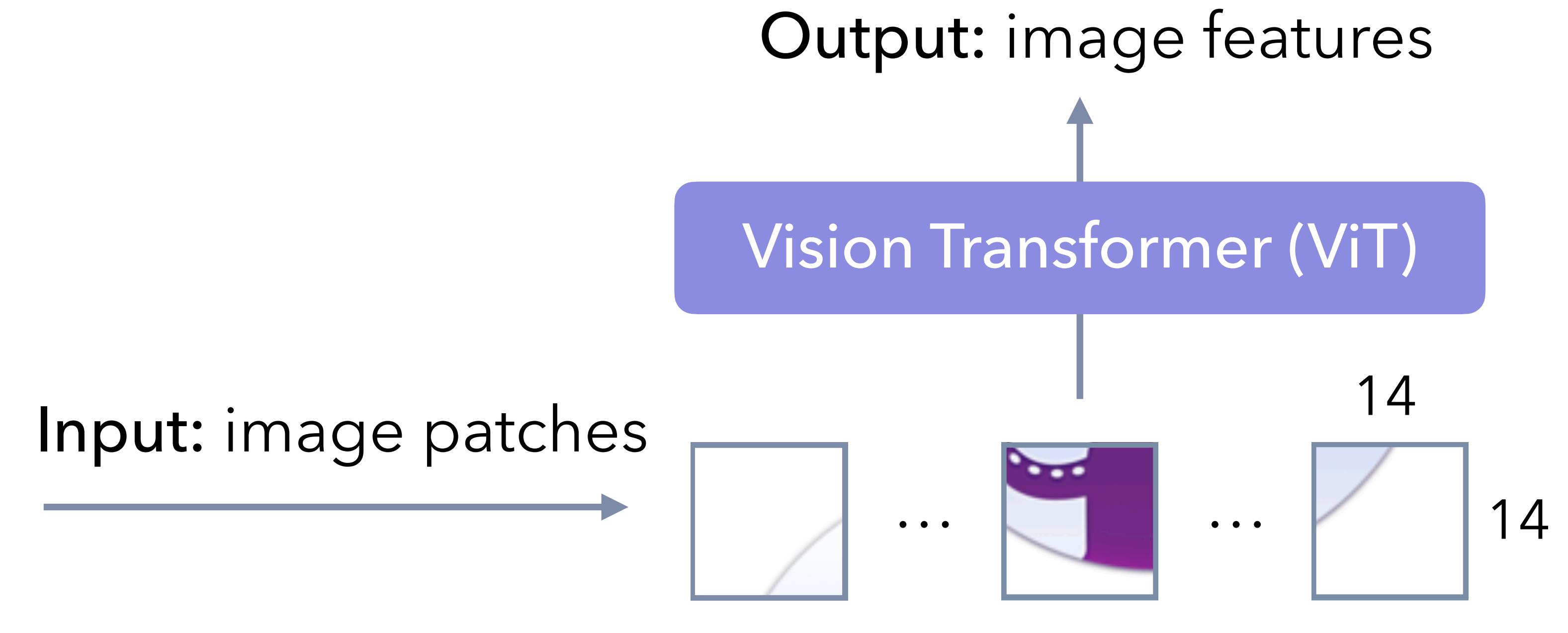
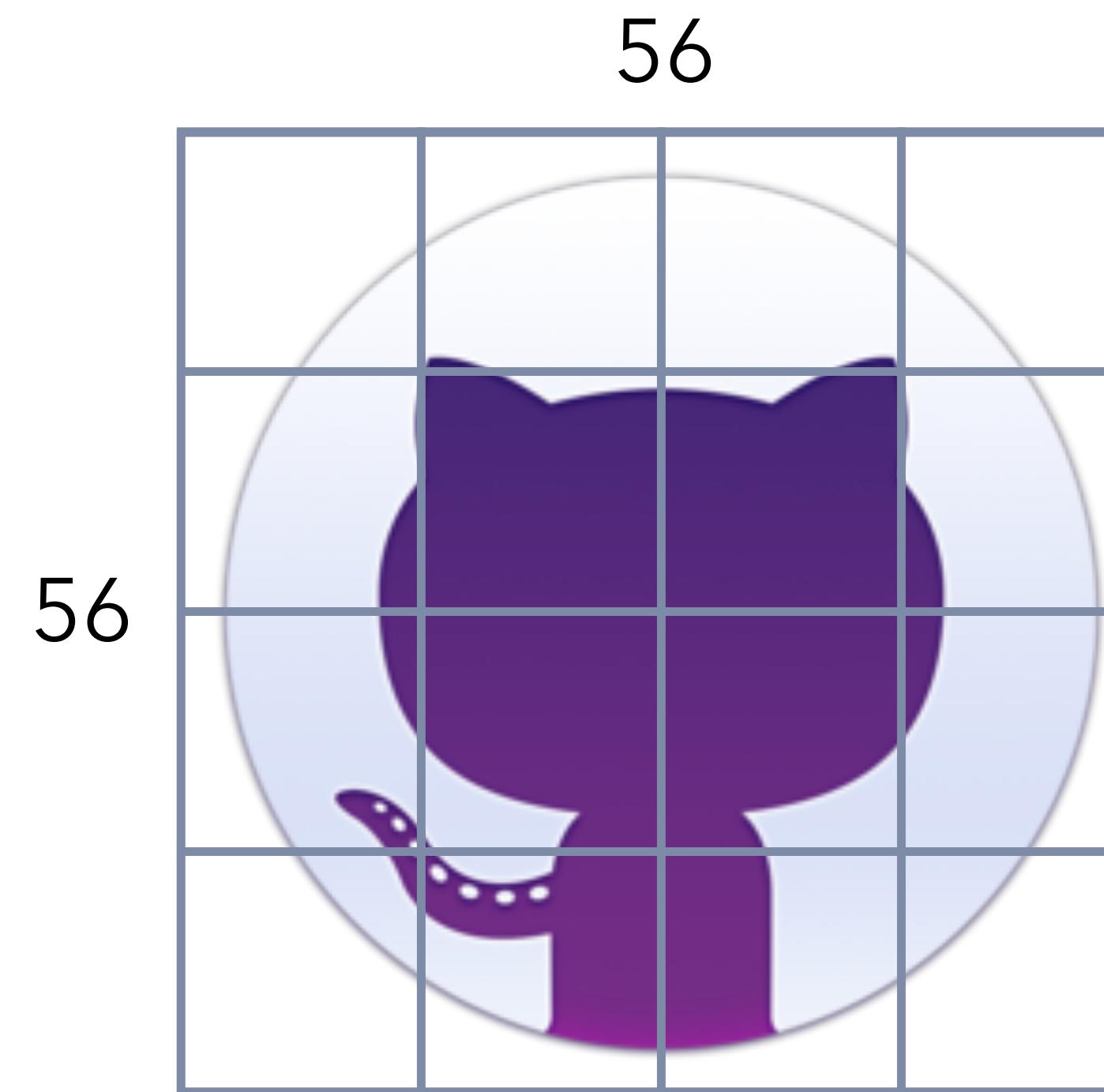


14
14

Vision encoder

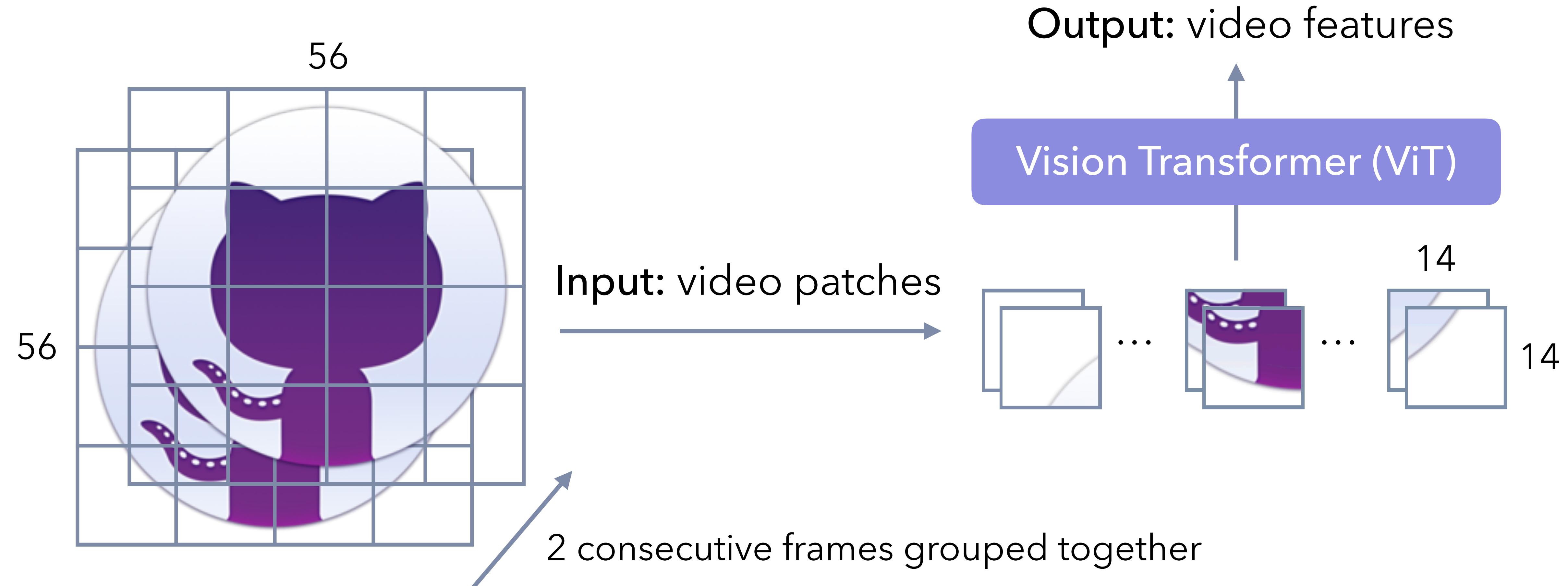


Vision encoder

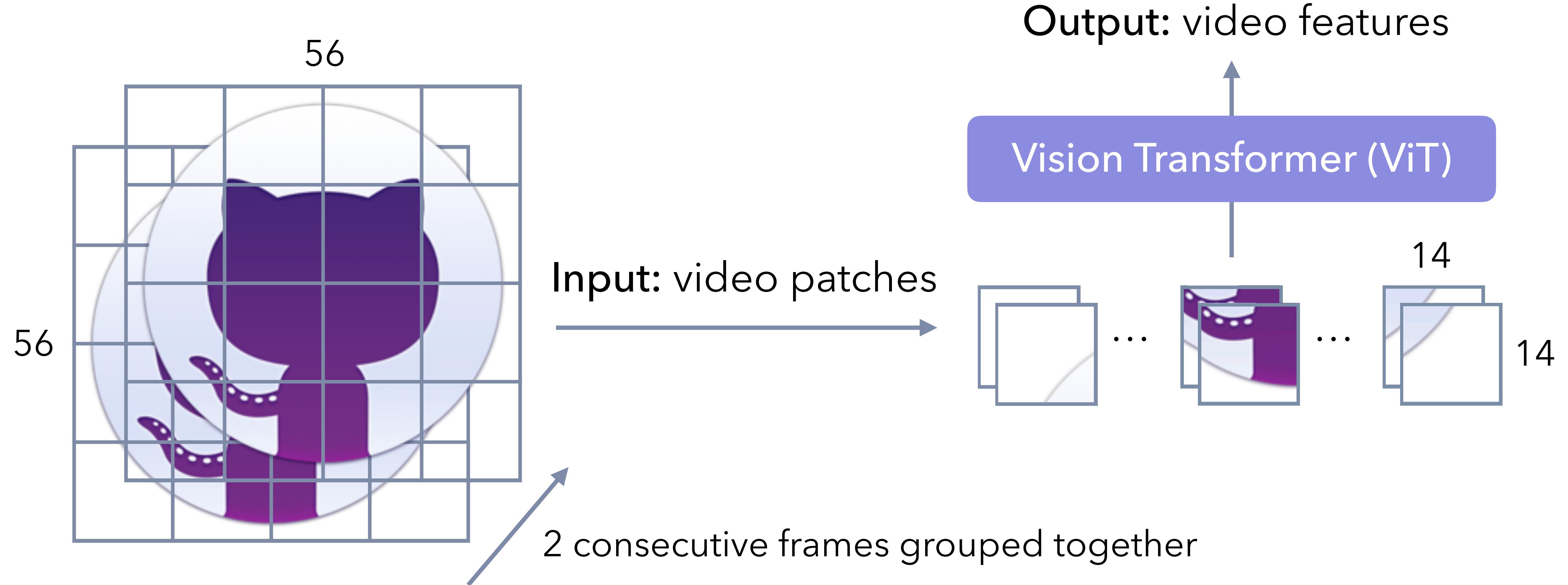


How about videos?

Vision encoder



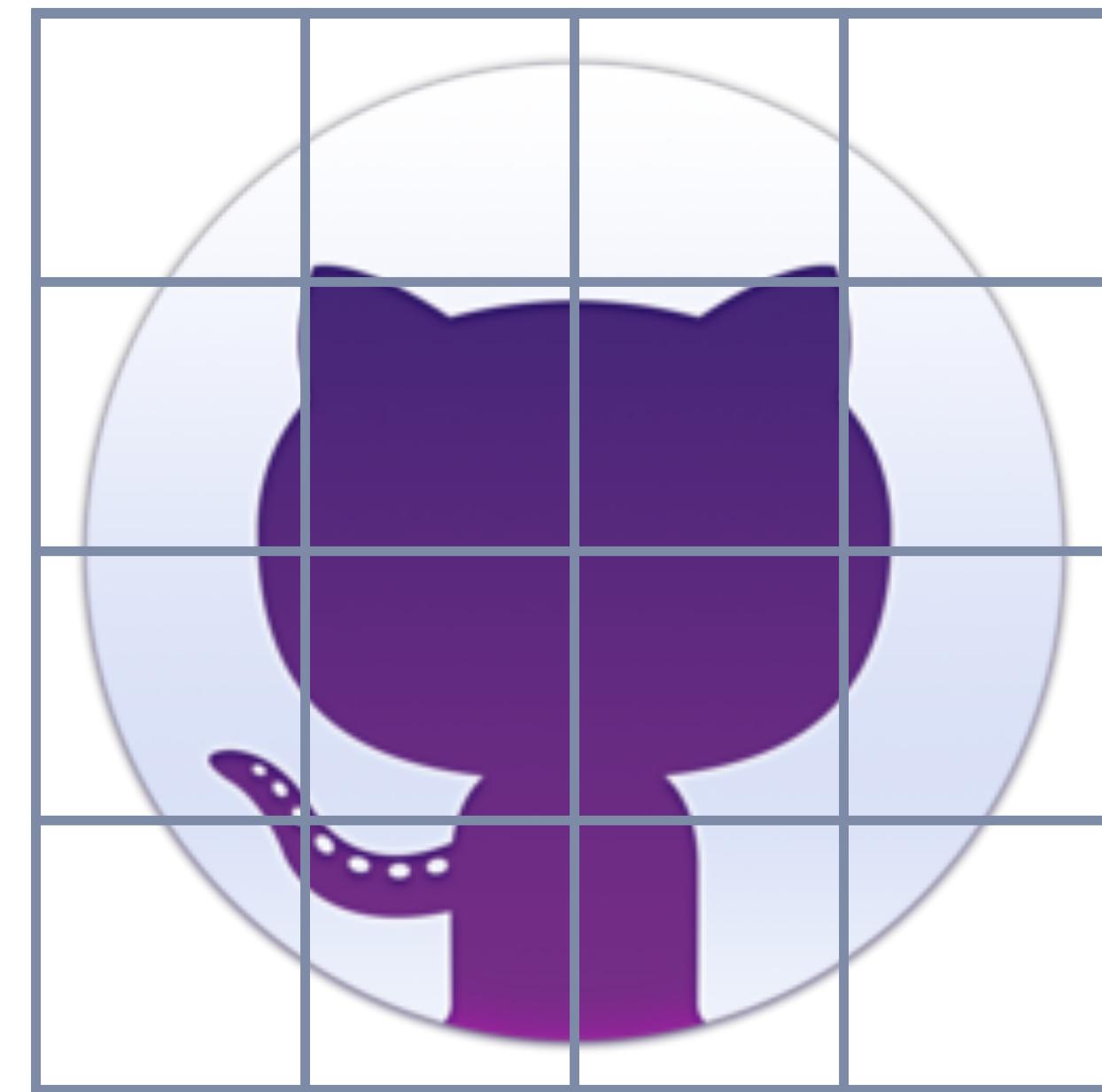
Vision encoder



Recall: time complexity of Transformers

Vision encoder

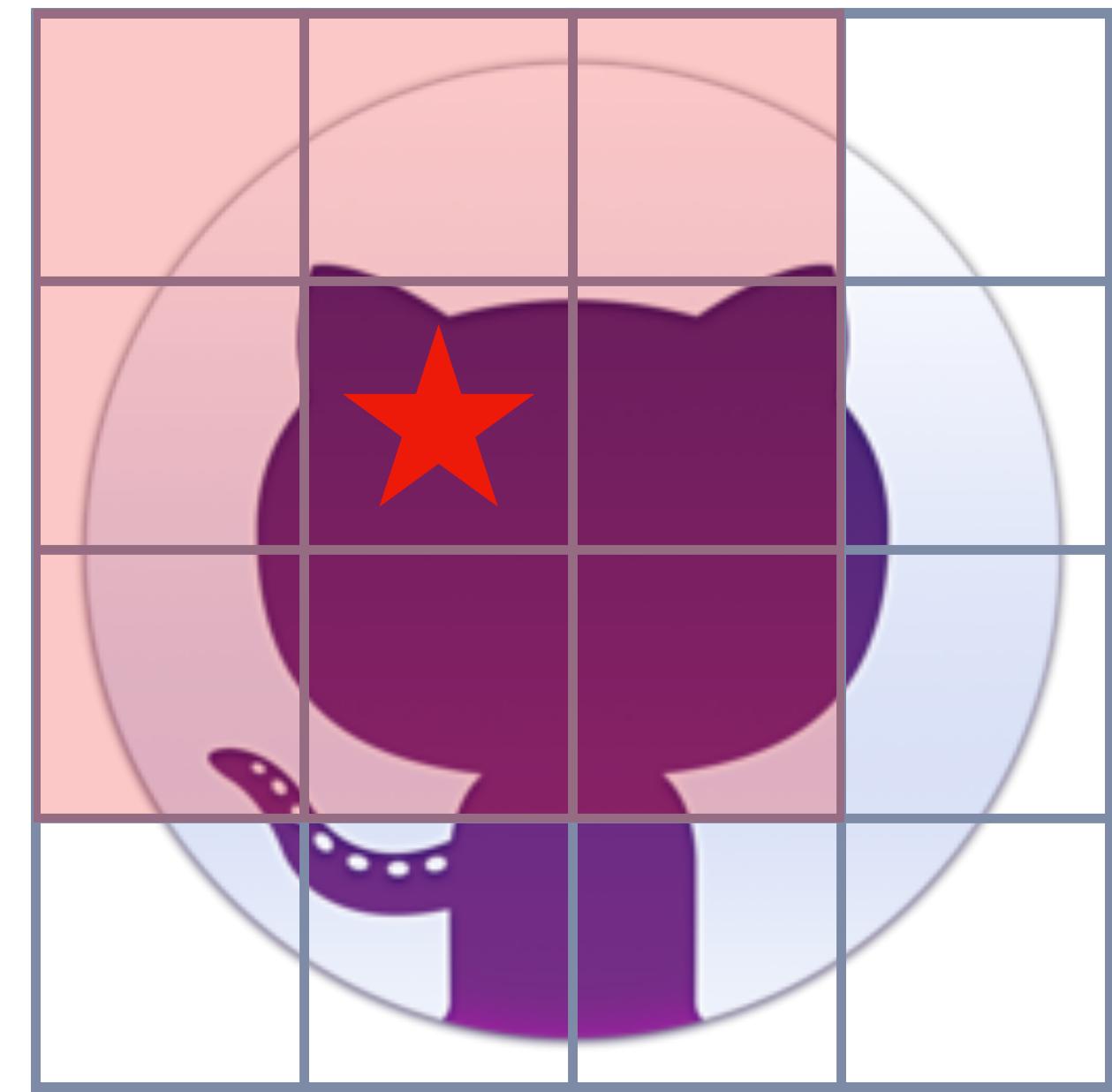
Efficiency matters for long visual inputs



Vision encoder

Efficiency matters for long visual inputs

Window attention: a visual patch only attends to its neighbors



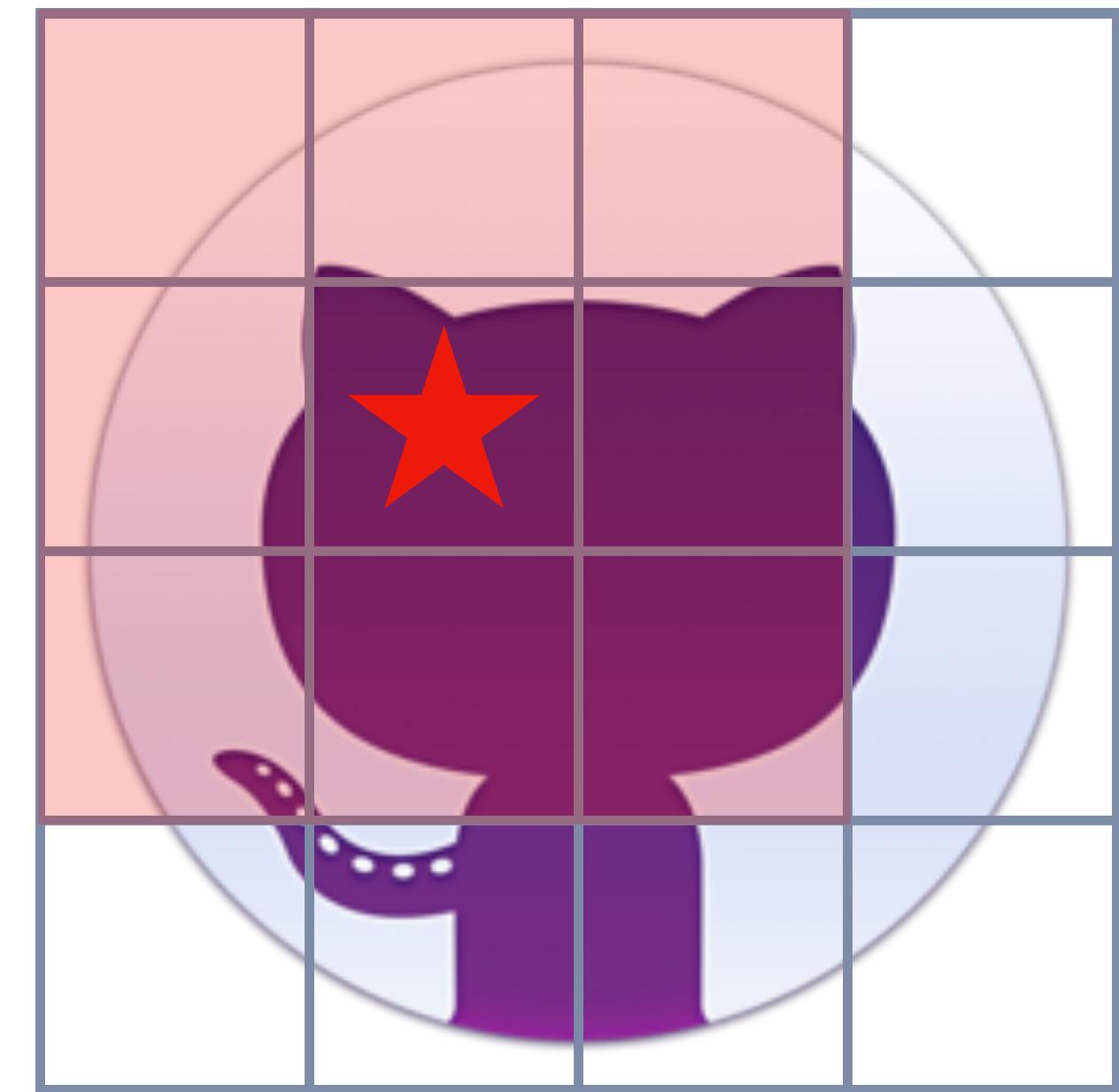
Vision encoder

Efficiency matters for long visual inputs

Window attention: a visual patch only attends to its neighbors

Now time complexity is linear :)

However, loss of global information :(



Vision encoder

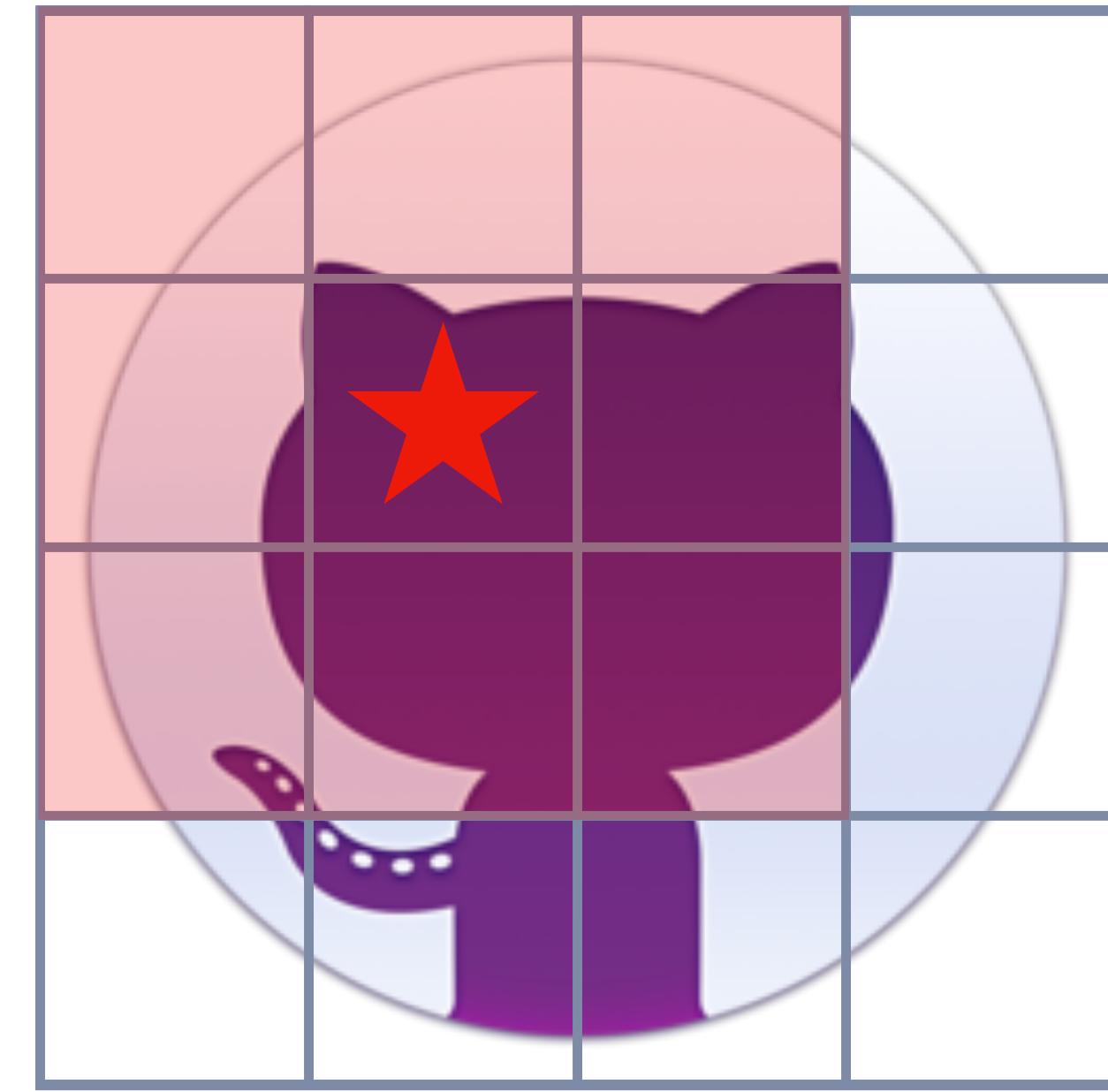
Efficiency matters for long visual inputs

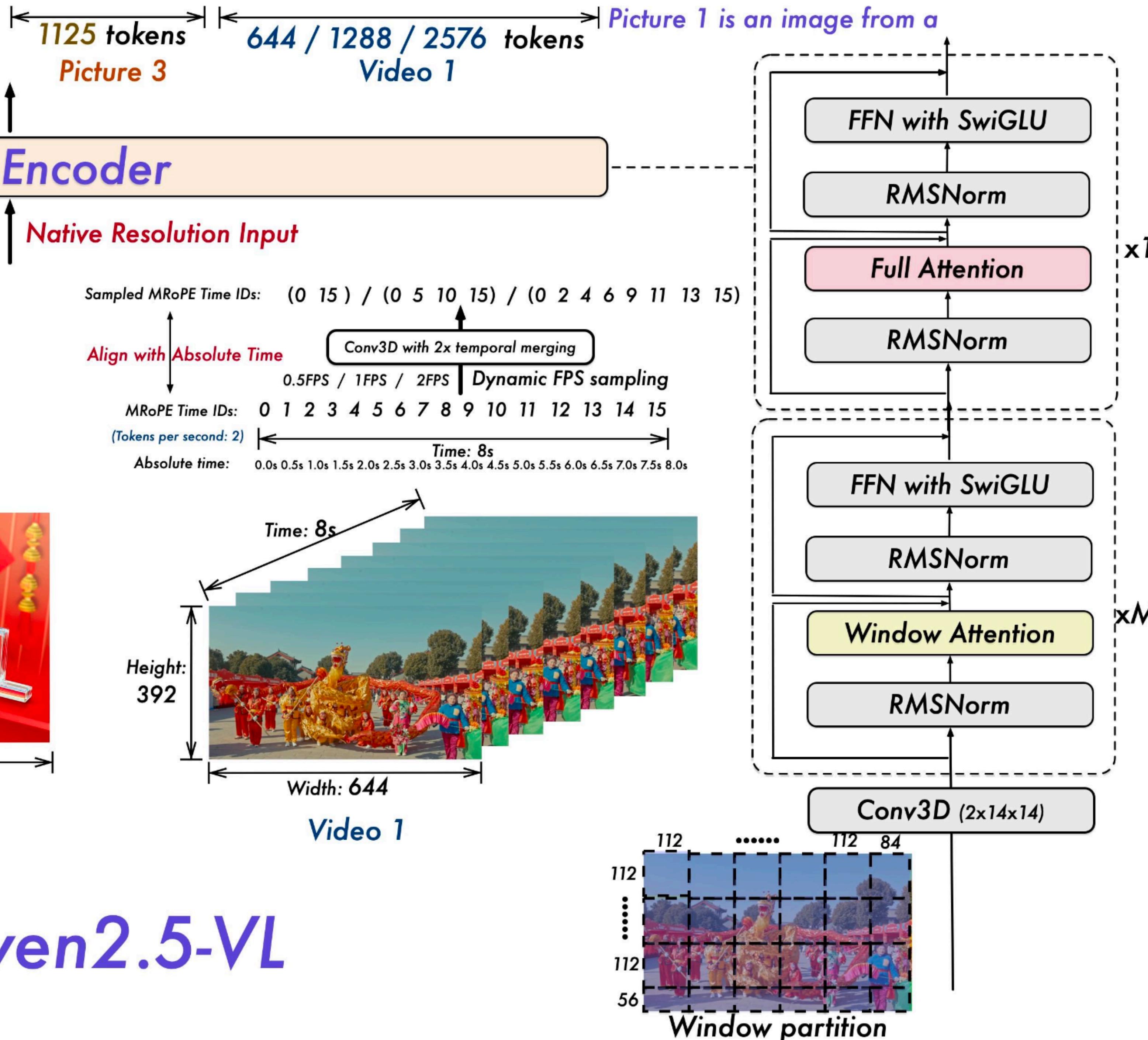
Window attention: a visual patch only attends to its neighbors

Now time complexity is linear :)

However, loss of global information :(

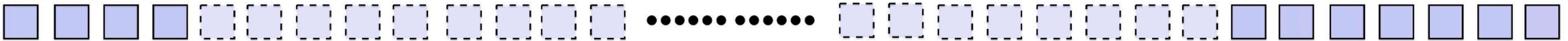
Solution: use full self-attention every M layers





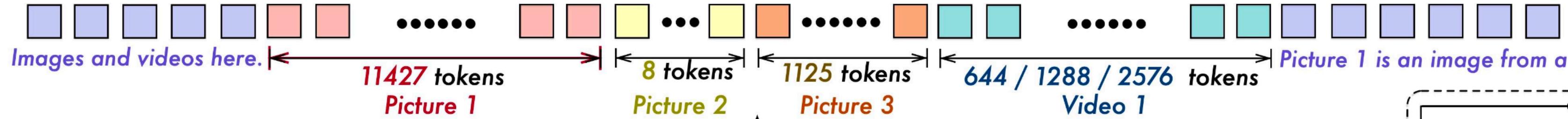
wen2.5-VL

and videos here.



Picture 1 is an image from a blog

Qwen2.5 LM Decoder



Vision Encoder



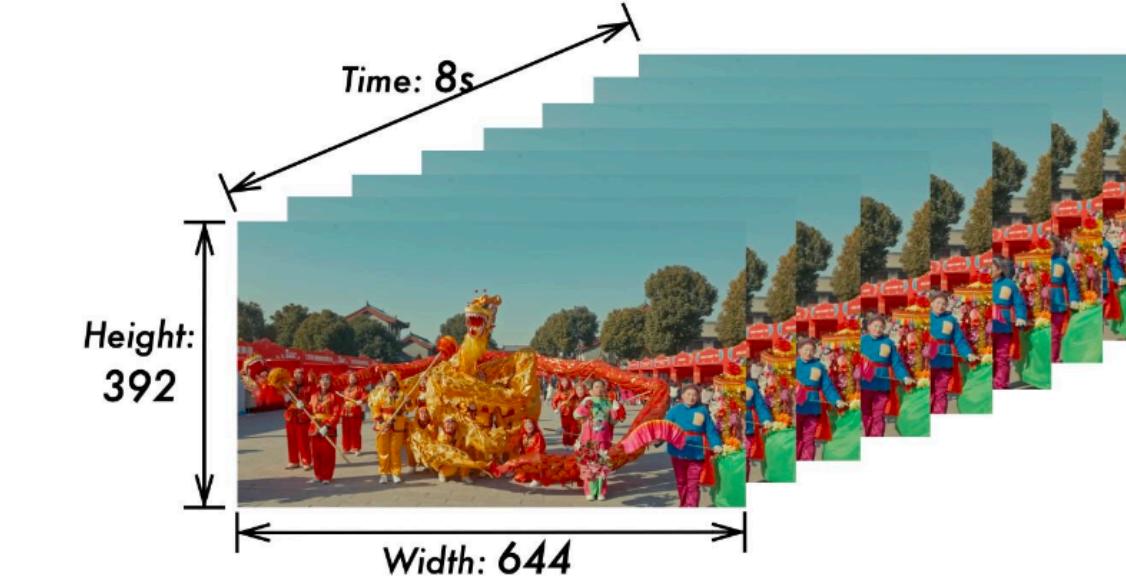
Height: 28
Width: 224

Picture 2

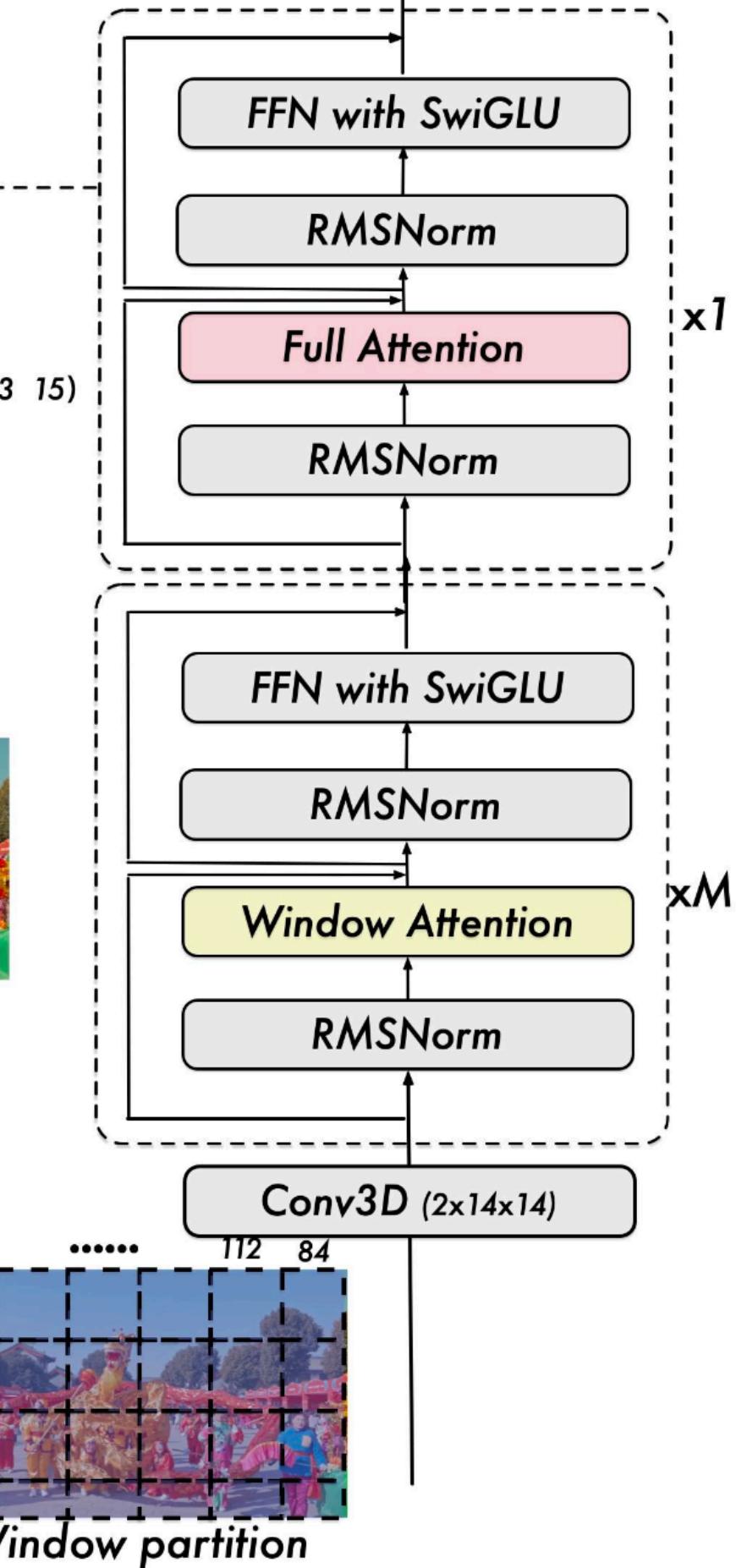


Picture 3

Sampled MRoPE Time IDs: (0 15) / (0 5 10 15) / (0 2 4 6 9 11 13 15)
Align with Absolute Time
MRoPE Time IDs: 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
(Tokens per second: 2) Time: 8s
0.0s 0.5s 1.0s 1.5s 2.0s 2.5s 3.0s 3.5s 4.0s 4.5s 5.0s 5.5s 6.0s 6.5s 7.0s 7.5s 8.0s
Conv3D with 2x temporal merging
0.5FPS / 1FPS / 2FPS Dynamic FPS sampling



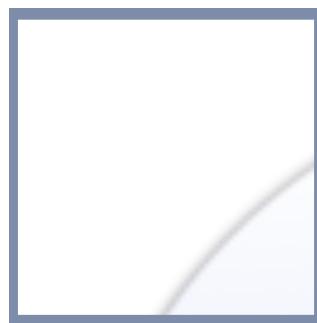
Video 1



 **Qwen2.5-VL**

MLP-based vision-language merger

Visual patches



...



...



...



Visual features



...



...



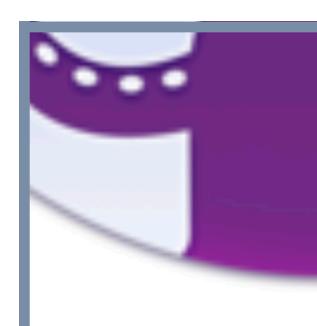
Native Resolution Input

MLP-based vision-language merger

Visual patches



...



...



Visual features



...



...



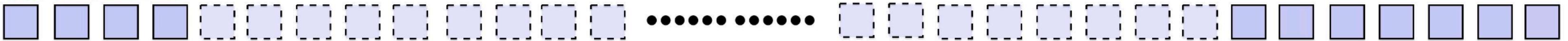
Visual tokens



Vision Encoder

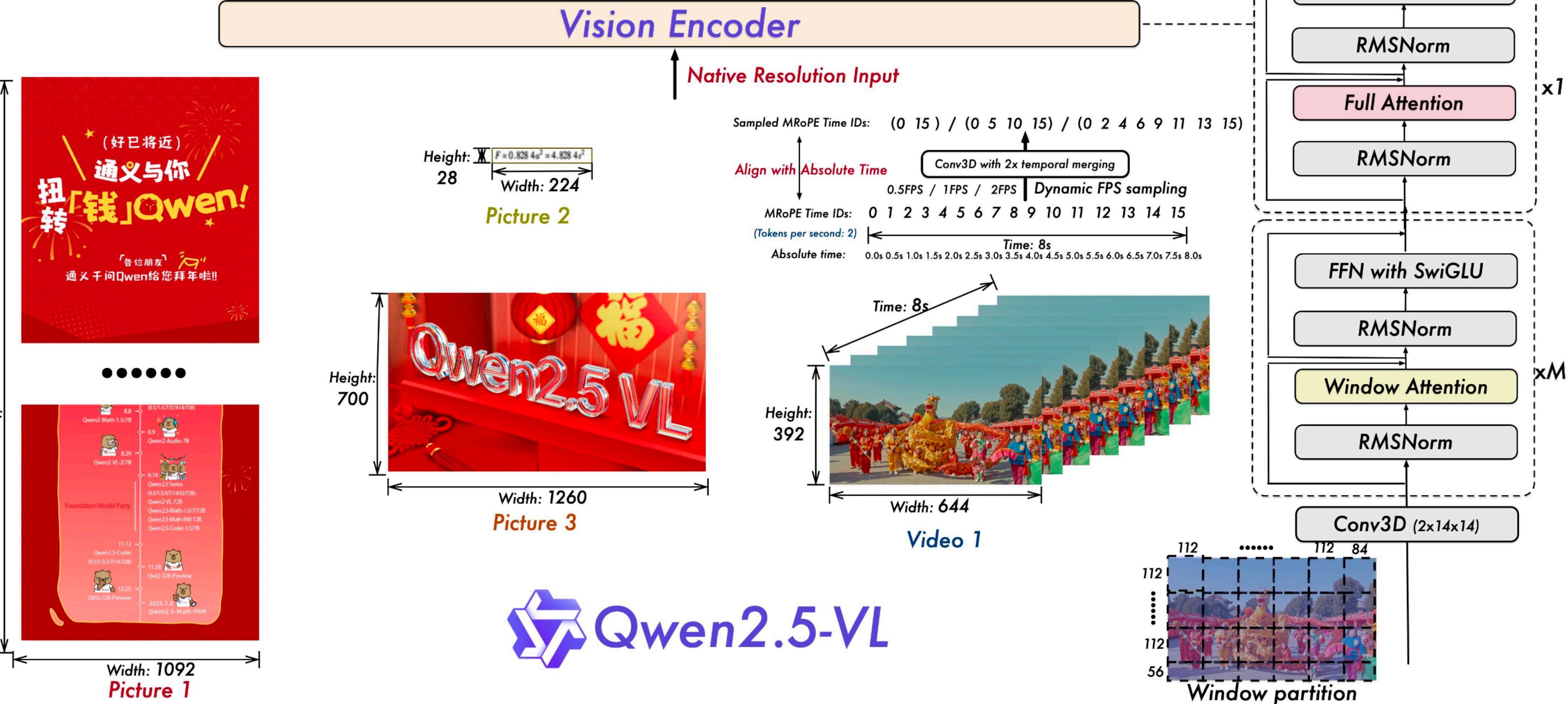
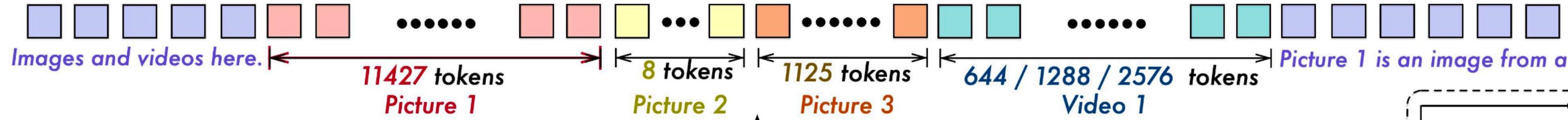
Native Resolution Input

and videos here.



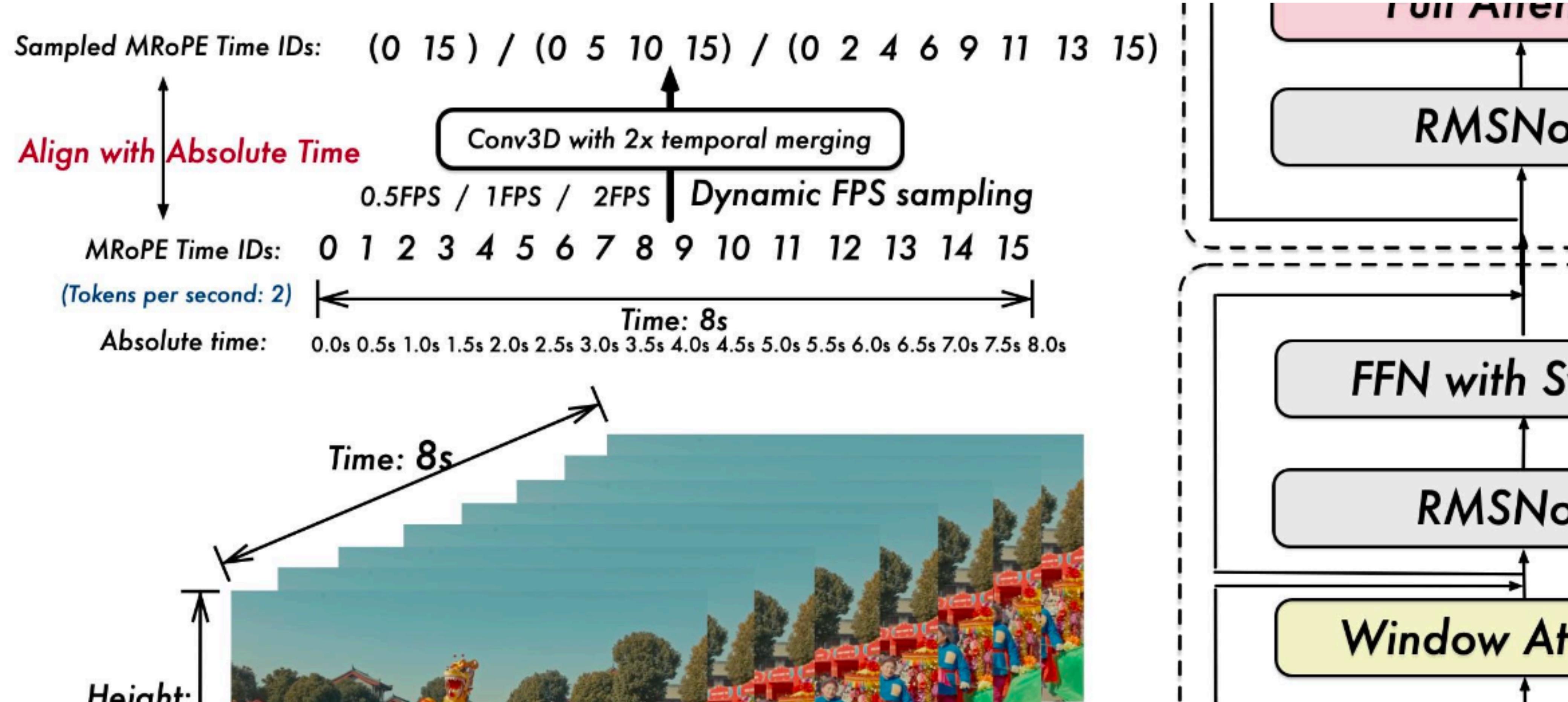
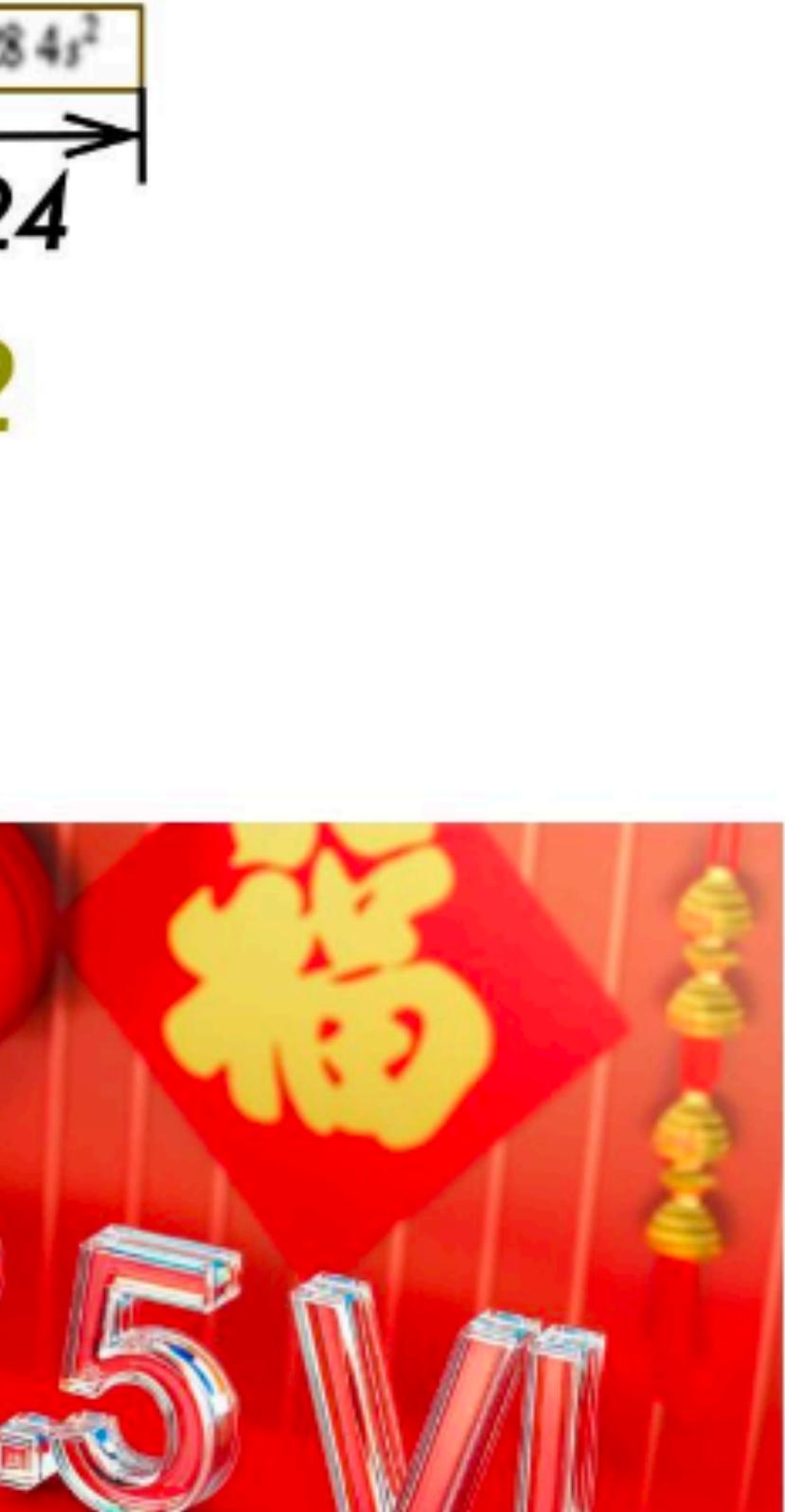
Picture 1 is an image from a blog

Qwen2.5 LM Decoder



 **Qwen2.5-VL**

Positional embeddings



Positional embeddings

Recall: self-attention is permutation invariant!

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

Positional embeddings

Recall: self-attention is permutation invariant!

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

Solution: positional embeddings

- ▶ Absolute positional embedding

$$\text{PE}(pos, 2i) = \sin\left(pos/10000^{2i/d_{model}}\right)$$

$$\text{PE}(pos, 2i + 1) = \cos\left(pos/10000^{2i/d_{model}}\right)$$

Positional embeddings

Recall: self-attention is permutation invariant!

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

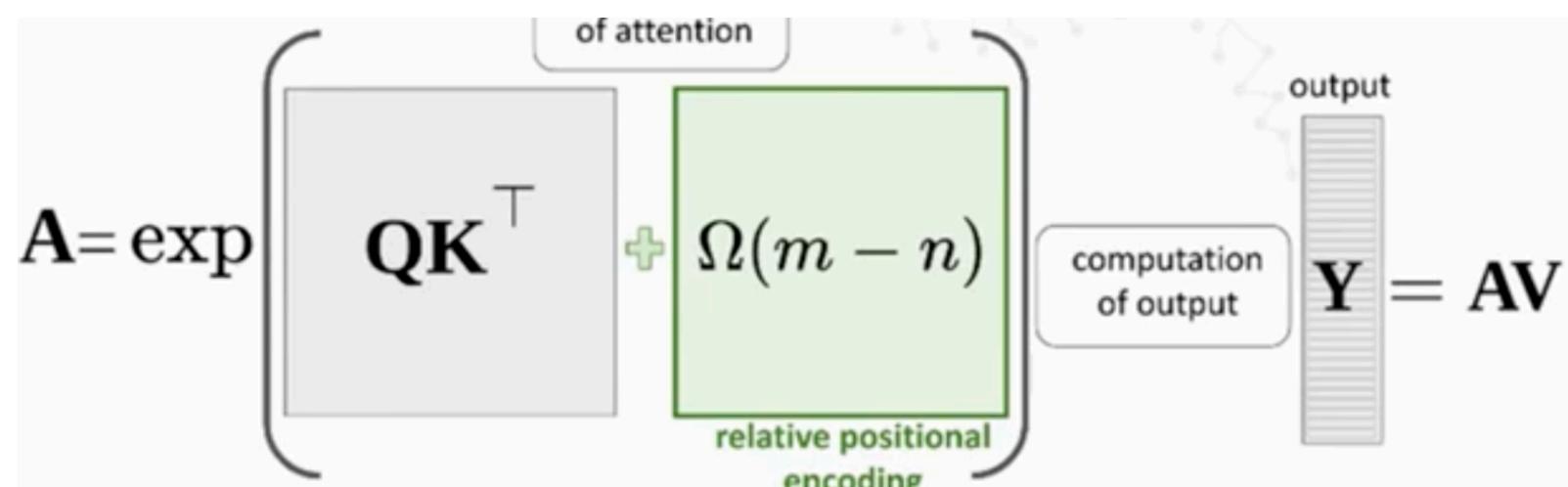
Solution: positional embeddings

- ▶ Absolute positional embedding

$$\text{PE}(pos, 2i) = \sin\left(pos/10000^{2i/d_{model}}\right)$$

$$\text{PE}(pos, 2i + 1) = \cos\left(pos/10000^{2i/d_{model}}\right)$$

- ▶ Relative positional embedding



Positional embeddings

Recall: self-attention is permutation invariant!

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

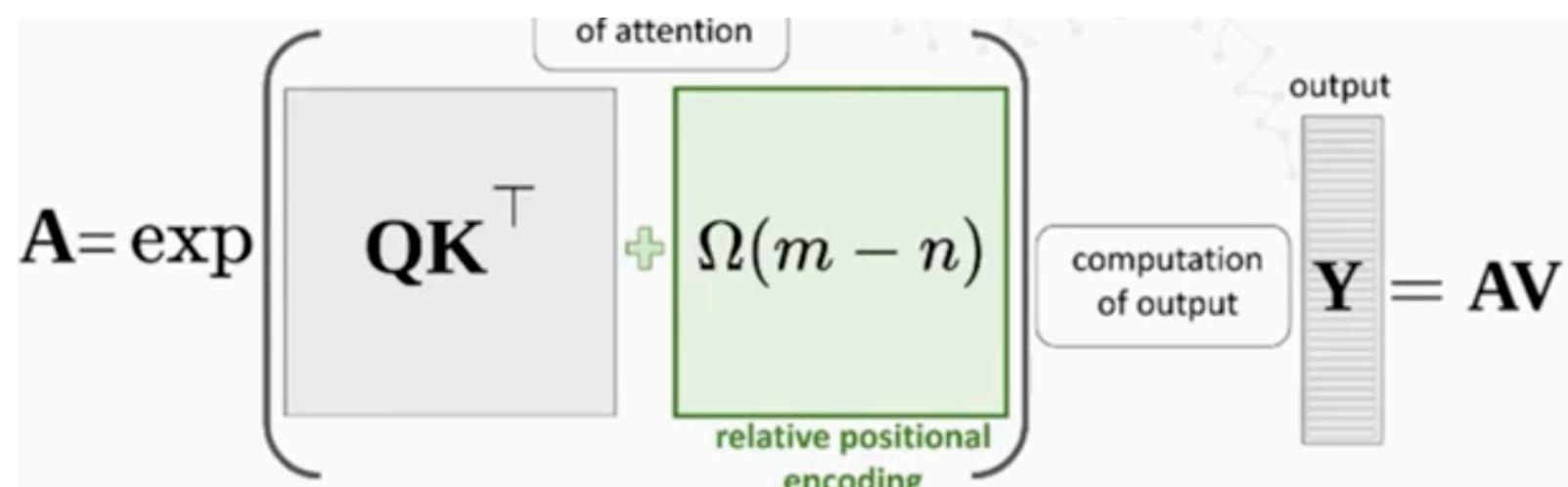
Solution: positional embeddings

- ▶ Absolute positional embedding

$$\text{PE}(pos, 2i) = \sin\left(pos/10000^{2i/d_{model}}\right)$$

$$\text{PE}(pos, 2i + 1) = \cos\left(pos/10000^{2i/d_{model}}\right)$$

- ▶ Relative positional embedding



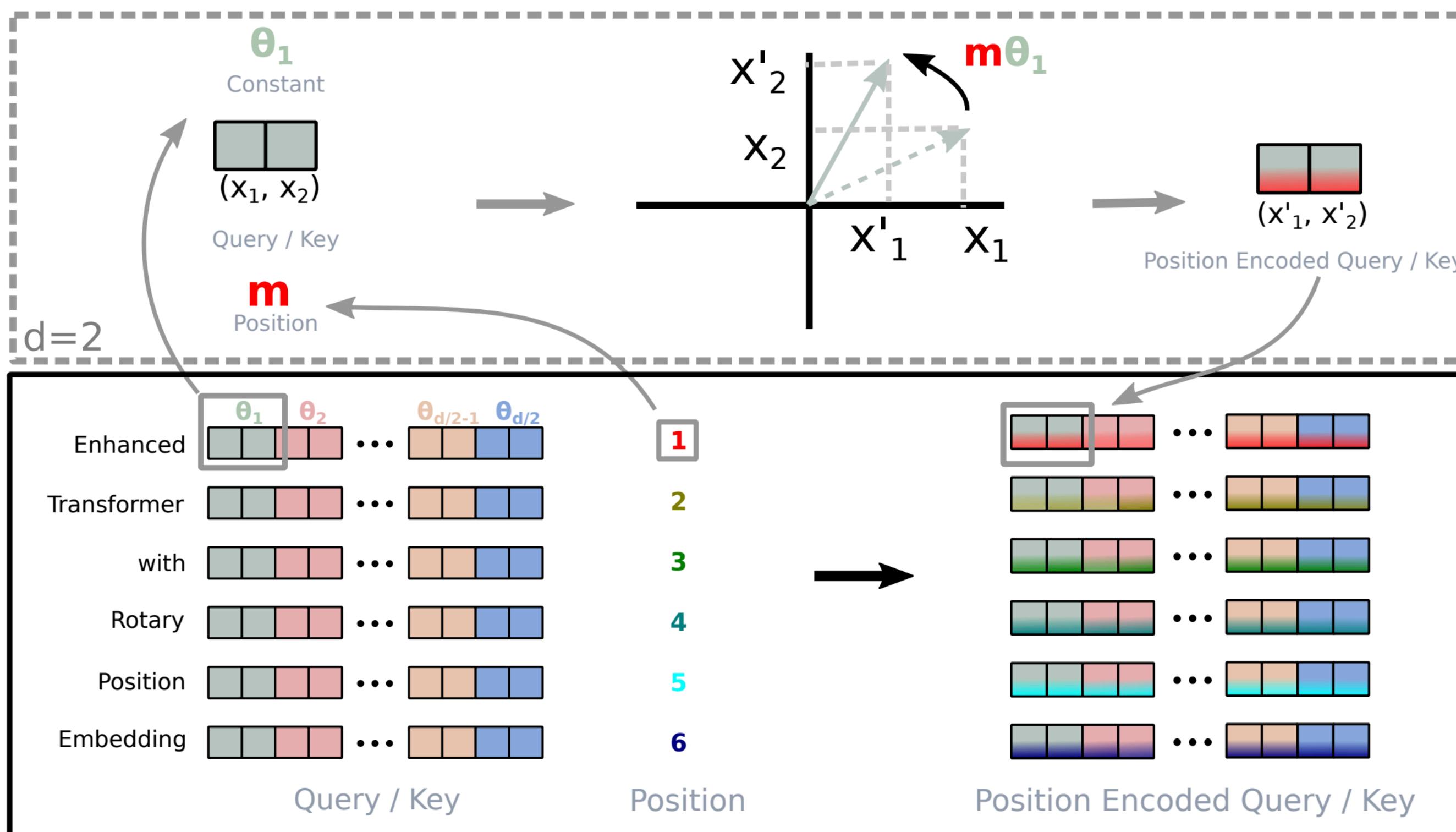
- ▶ Rotary Position Embedding (RoPE) [1]

[1] Su et al., 2024. Roformer: Enhanced Transformer with rotary position embedding

Positional embeddings

Rotary Position Embedding (RoPE) [1]

- The 2D case



$$f_{\{q,k\}}(\mathbf{x}_m, m) = \begin{pmatrix} \cos m\theta & -\sin m\theta \\ \sin m\theta & \cos m\theta \end{pmatrix} \begin{pmatrix} W_{\{q,k\}}^{(11)} & W_{\{q,k\}}^{(12)} \\ W_{\{q,k\}}^{(21)} & W_{\{q,k\}}^{(22)} \end{pmatrix} \begin{pmatrix} x_m^{(1)} \\ x_m^{(2)} \end{pmatrix}$$

Positional embeddings

Rotary Position Embedding (RoPE) [1]

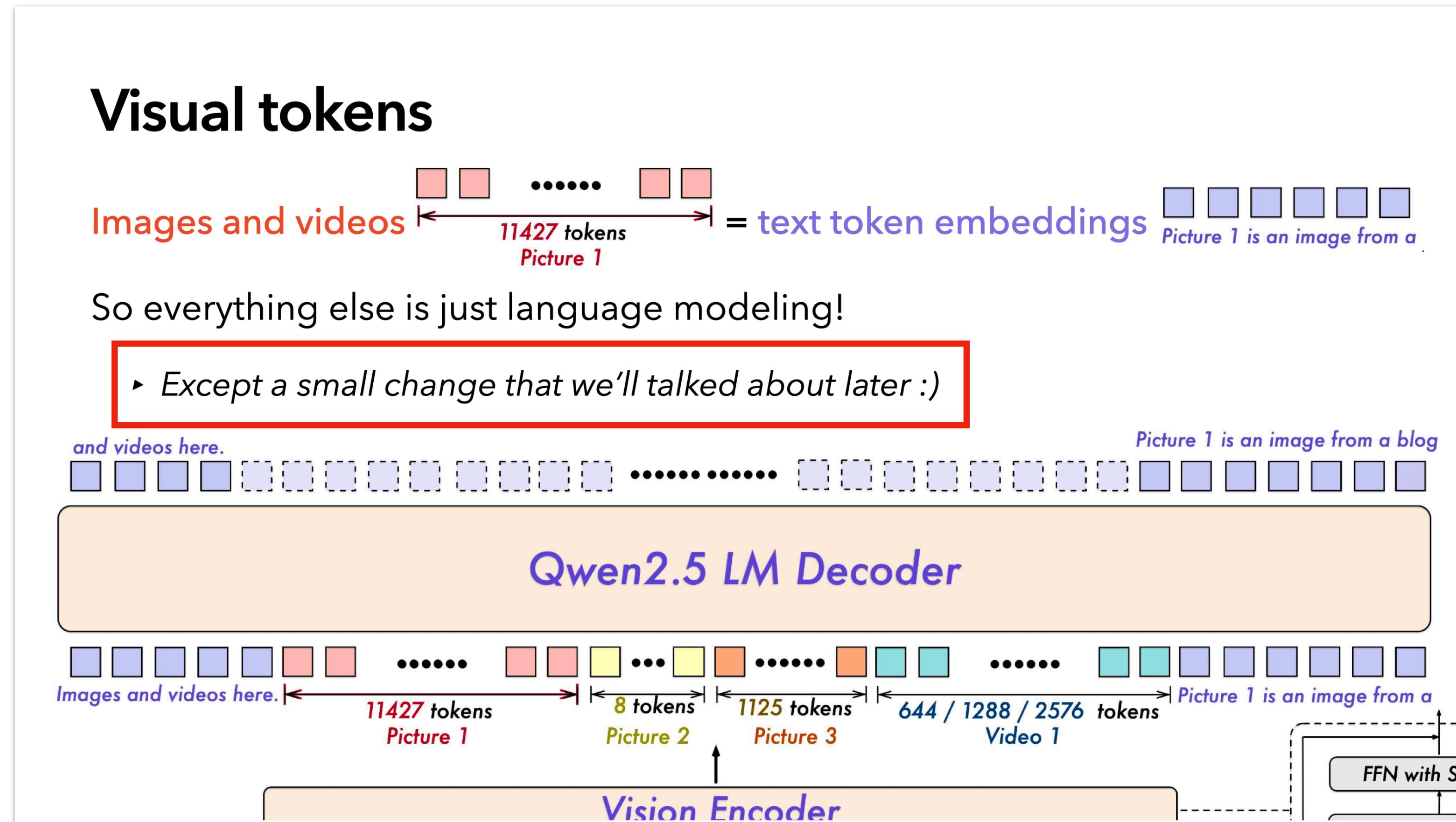
- ▶ The 2D case
- ▶ The general case: group the feature; 2 dimensions in each group

$$\mathbf{R}_{\Theta,m}^d = \begin{pmatrix} \cos m\theta_1 & -\sin m\theta_1 & 0 & 0 & \cdots & 0 & 0 \\ \sin m\theta_1 & \cos m\theta_1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \cos m\theta_2 & -\sin m\theta_2 & \cdots & 0 & 0 \\ 0 & 0 & \sin m\theta_2 & \cos m\theta_2 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & \cos m\theta_{d/2} & -\sin m\theta_{d/2} \\ 0 & 0 & 0 & 0 & \cdots & \sin m\theta_{d/2} & \cos m\theta_{d/2} \end{pmatrix}$$

$$\mathbf{q}_m^\top \mathbf{k}_n = (\mathbf{R}_{\Theta,m}^d \mathbf{W}_q \mathbf{x}_m)^\top (\mathbf{R}_{\Theta,n}^d \mathbf{W}_k \mathbf{x}_n) = \mathbf{x}^\top \mathbf{W}_q \mathbf{R}_{\Theta,n-m}^d \mathbf{W}_k \mathbf{x}_n$$

Positional embeddings

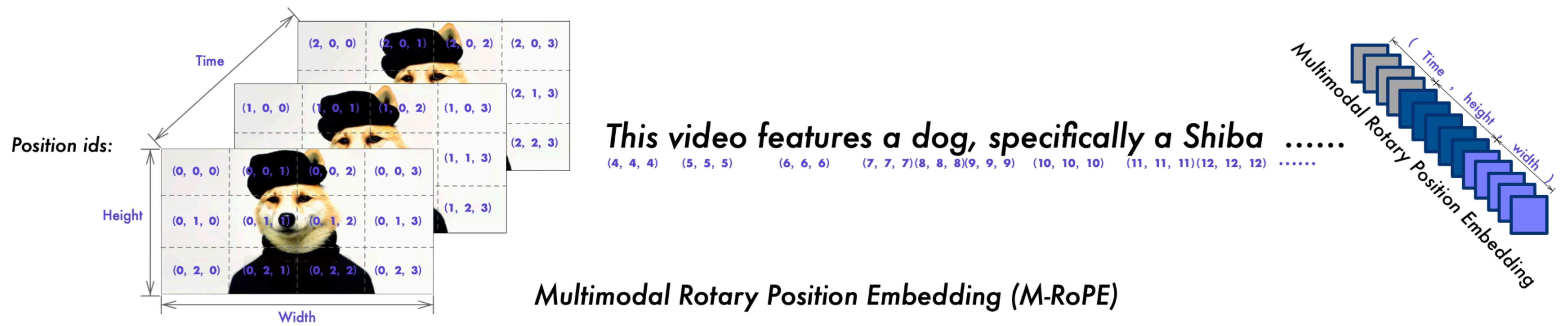
How about multimodal positions?



Positional embeddings

Multimodal Rotary Position Embedding (MRoPE)

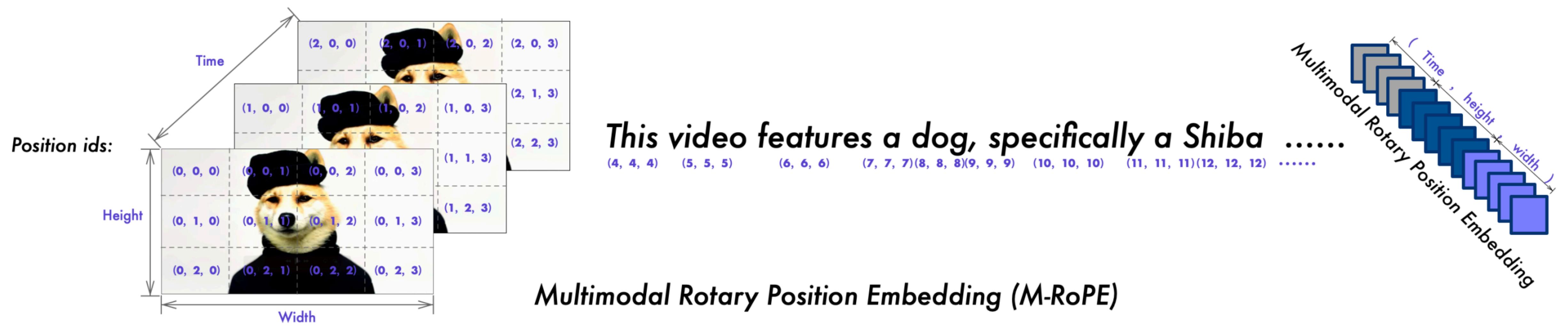
- ▶ Three sets of IDs (instead of just 1): width, height, time



Positional embeddings

Multimodal Rotary Position Embedding (MRoPE)

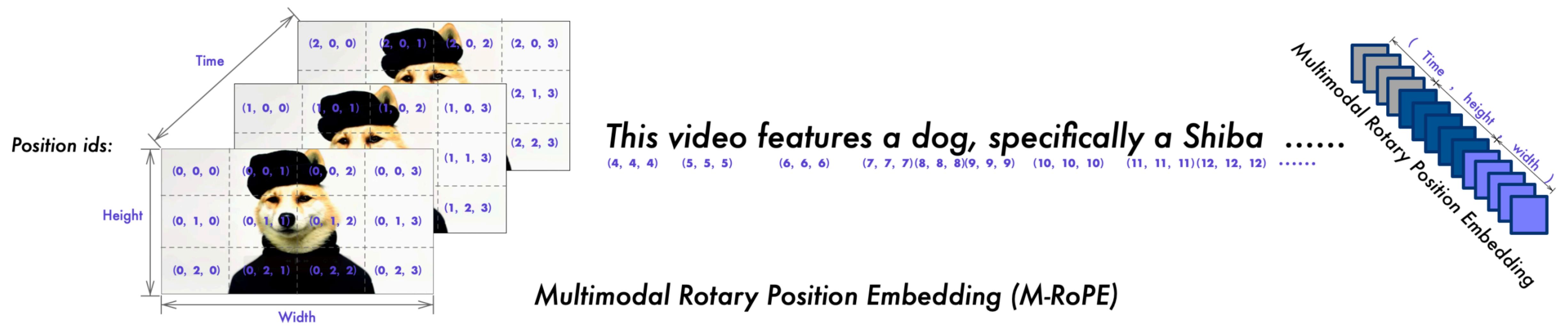
- ▶ Three sets of IDs (instead of just 1): width, height, time
- ▶ Text position ID is a special case of time ID; and its height ID = width ID = time ID



Positional embeddings

Multimodal Rotary Position Embedding (MRoPE)

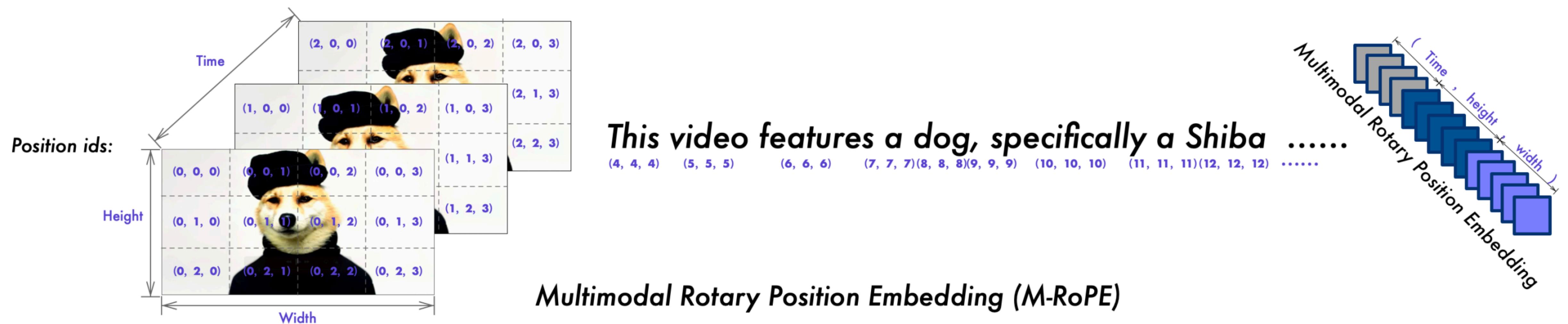
- ▶ Three sets of IDs (instead of just 1): width, height, time
- ▶ Text position ID is a special case of time ID; and its height ID = width ID = time ID
- ▶ The feature is split into three partitions; each for one ID



Positional embeddings

Multimodal Rotary Position Embedding (MRoPE)

- ▶ Three sets of IDs (instead of just 1): width, height, time
- ▶ Text position ID is a special case of time ID; and its height ID = width ID = time ID
- ▶ The feature is split into three partitions; each for one ID



Challenge: the time ID can be tricky

Positional embeddings

MRoPE in Qwen2.5-VL

- ▶ Videos can be subsampled for efficiency considerations

Positional embeddings

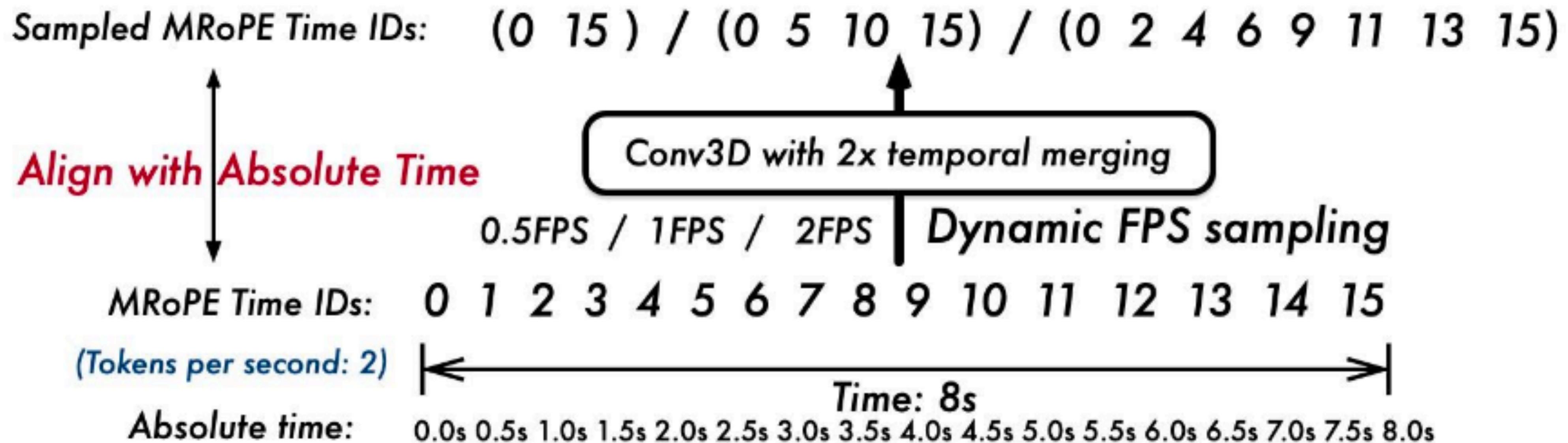
MRoPE in Qwen2.5-VL

- ▶ Videos can be subsampled for efficiency considerations
- ▶ How does the model distinguish between a slow move vs a subsampled fast move?

Positional embeddings

MRoPE in Qwen2.5-VL

- ▶ Videos can be subsampled for efficiency considerations
- ▶ How does the model distinguish between a slow move vs a subsampled fast move?
- ▶ Solution: Align the time ID to absolute time (instead of frames)



Pretraining

Pretraining

Language model initialized from Qwen2.5 LM

Vision encoder initialized from:

We trained a Vision Transformer (ViT) from scratch using DataComp ([Gadre et al., 2023](#)) and some in-house datasets as the initialization for the vision encoder

Stages	Visual Pre-Training	Multimodal Pre-Training	Long-Context Pre-Training
Data	Image Caption Knowledge OCR	Pure text Interleaved Data VQA, Video Grounding, Agent	+ Long Video Long Agent Long Document
Tokens	1.5T	2T	0.6T
Sequence length	8192	8192	32768
Training	ViT	ViT & LLM	ViT & LLM

Post-training

Post-training

Supervised finetuning (SFT)

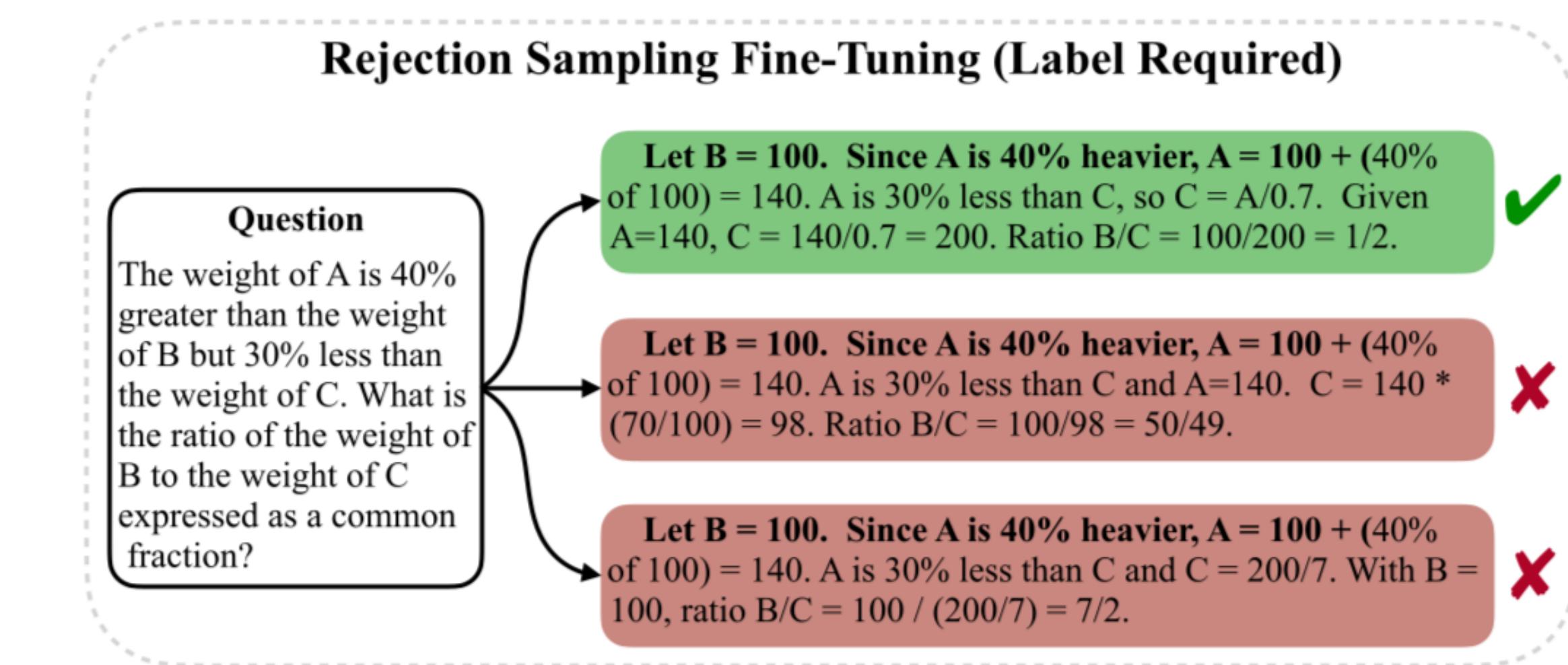
- ▶ Human data

Post-training

Supervised finetuning (SFT)

- ▶ Human data
- ▶ Rejection sampling data:

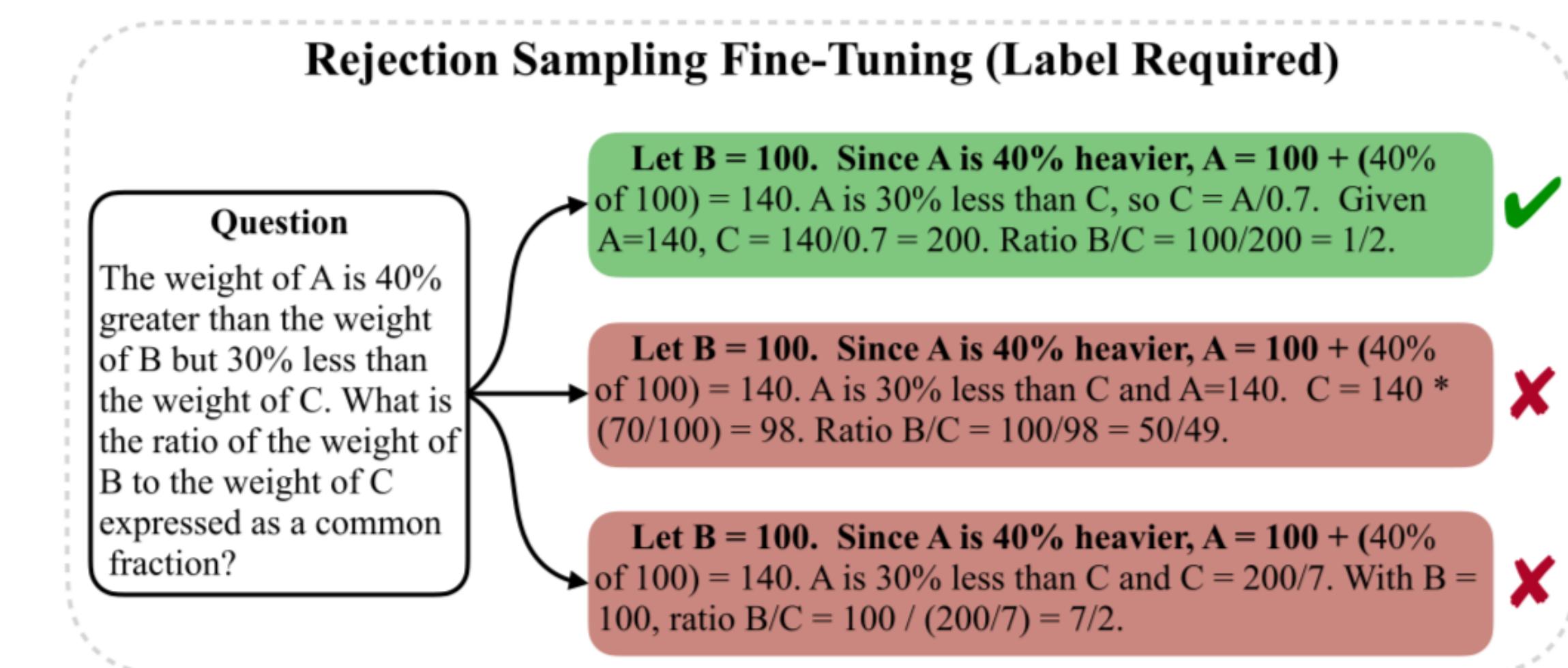
- ▶ Rollout the chain-of-thought + answer from the model
- ▶ Only keep those with the correct answer



Post-training

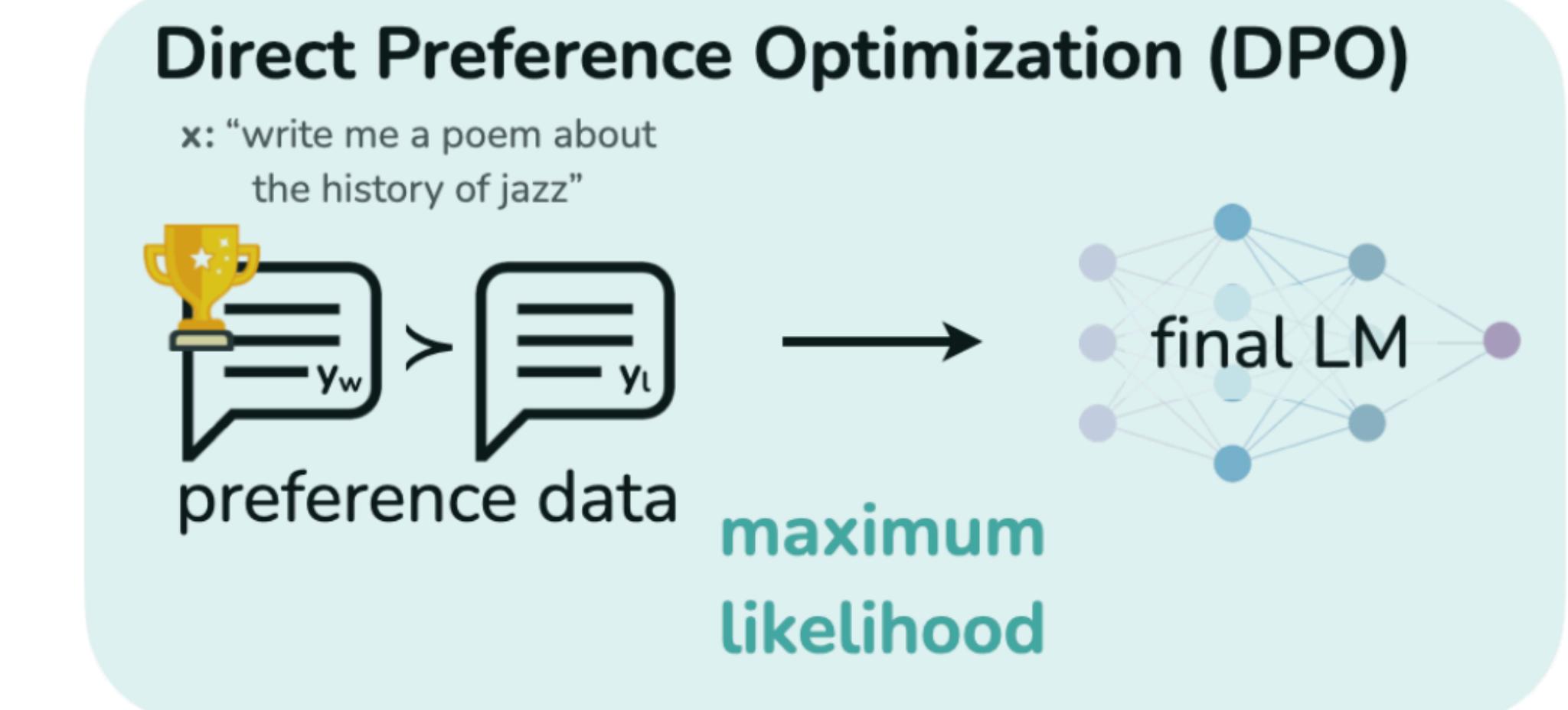
Supervised finetuning (SFT)

- ▶ Human data
- ▶ Rejection sampling data:



- ▶ Rollout the chain-of-thought + answer from the model
- ▶ Only keep those with the correct answer

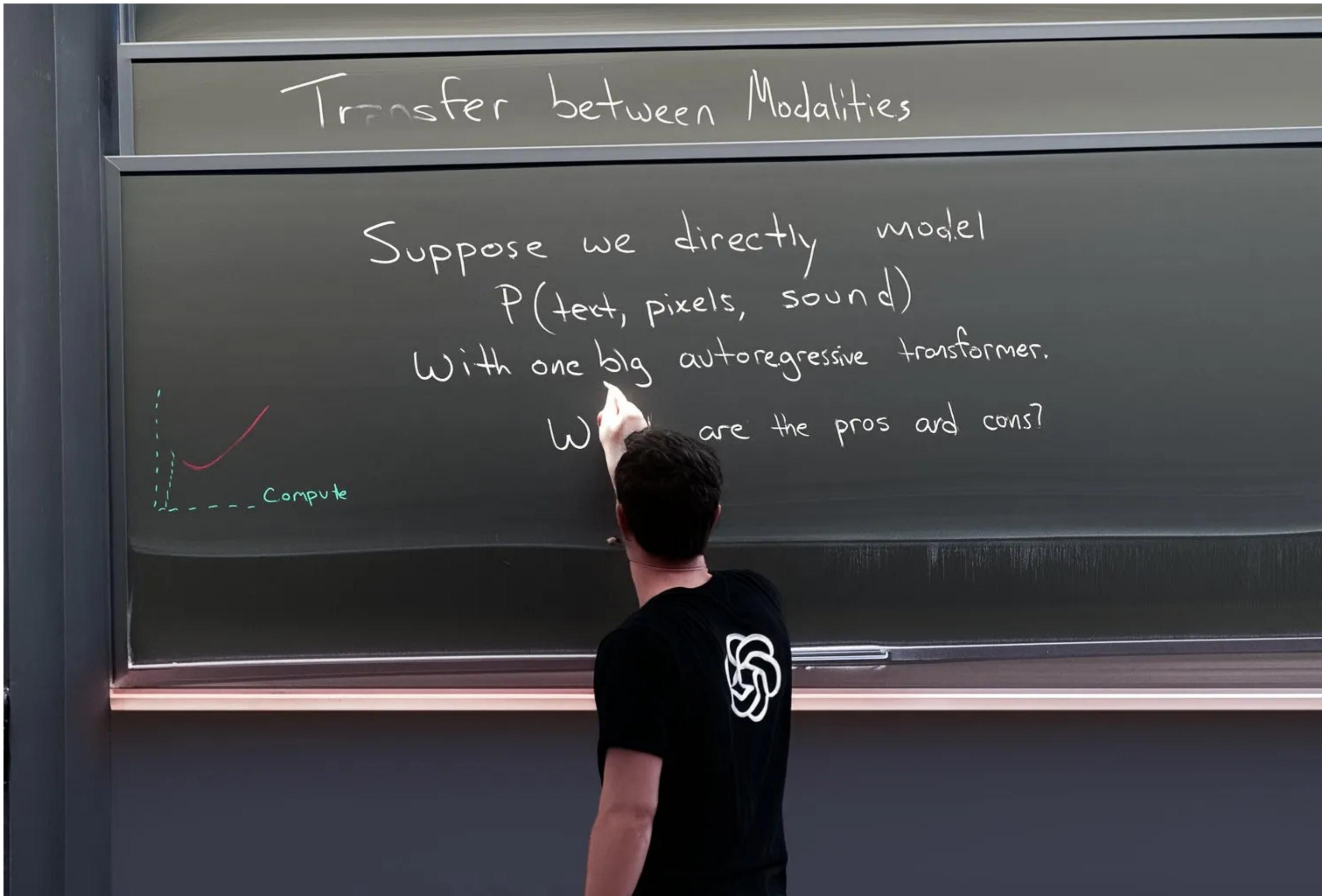
Direct preference optimization (DPO)



What's next?

Natively multimodal LMs

GPT-4o with image generation



A wide image taken with a phone of a glass whiteboard, in a room overlooking the Bay Bridge. The field of view shows a woman writing, sporting a tshirt with a large OpenAI logo. The handwriting looks natural and a bit messy, and we see the photographer's reflection. ...

Read more

Transfer between Modalities:

Suppose we directly model $p(\text{text, pixels, sound})$ with one big autoregressive transformer.

Pros:

- image generation augmented with world knowledge
- next-level text rendering
- native in-context learning
- unified post-training

Cons:

- varying bit-rate of different modalities
- Compute not adapted to different modalities

Fixes:

- model compressed representations
- Compose autoregressive prior with a powerful decoder

tokens → [transformer] → [diffusion] pixels

Best of 8

selfie view of the photographer, as she turns around to high five him

Transfer between Modalities:

Suppose we directly model $p(\text{text, pixels, sound})$ with one big autoregressive prior

Pros:

- image generation augmented with world knowledge
- next-level text rendering
- native in-context learning
- unified post-training

Cons:

- varying bit-rate of different modalities
- Compute not adapted to different modalities

Fixes:

- model compressed representations
- Compose autoregressive prior with a powerful decoder

tokens → [transformer]

A woman in a black t-shirt with the OpenAI logo is writing on a whiteboard. A man in a similar t-shirt is taking a selfie in the foreground, smiling and giving a high-five to the woman. The background shows a window with a view of the Bay Bridge.

Natively multimodal LMs

Transfusion [1]

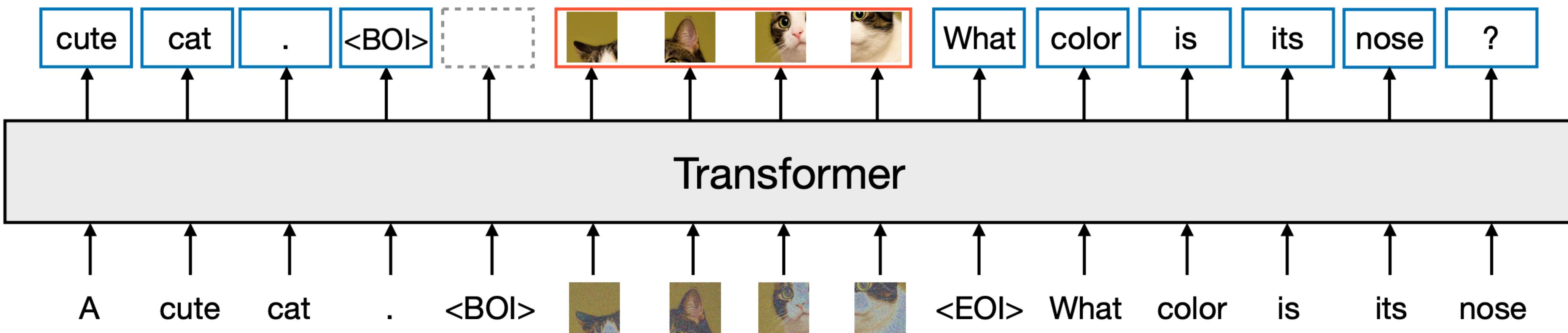
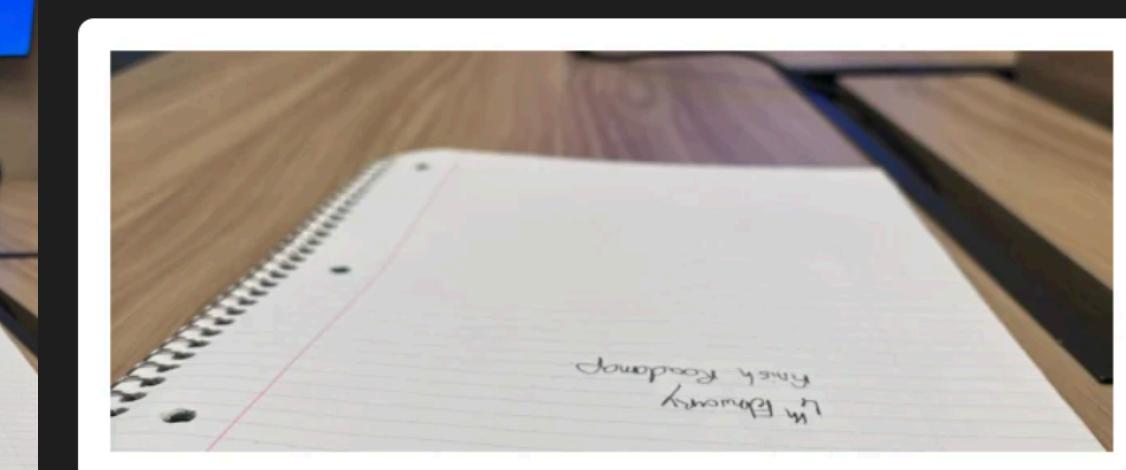
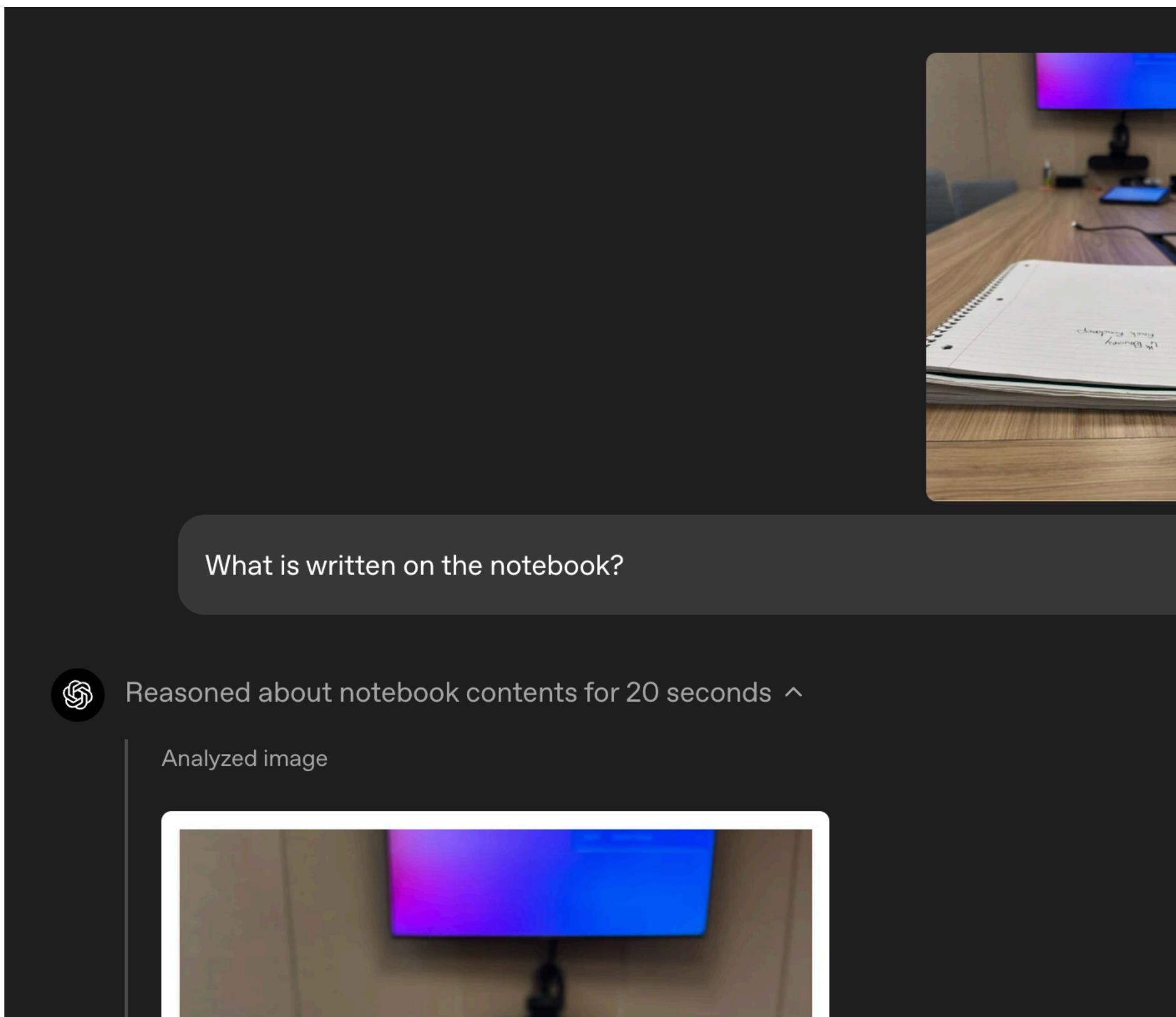


Figure 1: A high-level illustration of Transfusion. A single transformer perceives, processes, and produces data of every modality. Discrete (text) tokens are processed autoregressively and trained on the **next token prediction** objective. Continuous (image) vectors are processed together in parallel and trained on the **diffusion** objective. Marker BOI and EOI tokens separate the modalities.

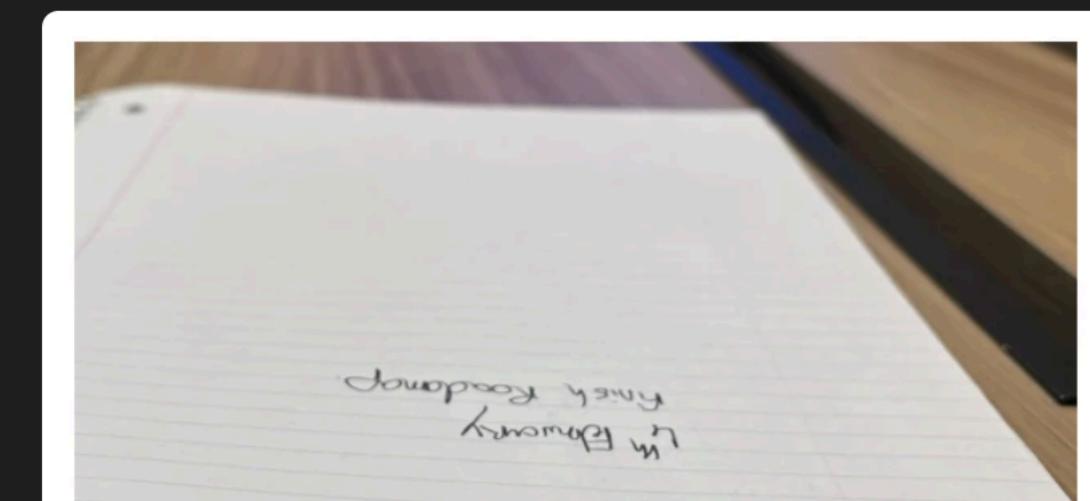
(Complex) visual reasoning

OpenAI-o3

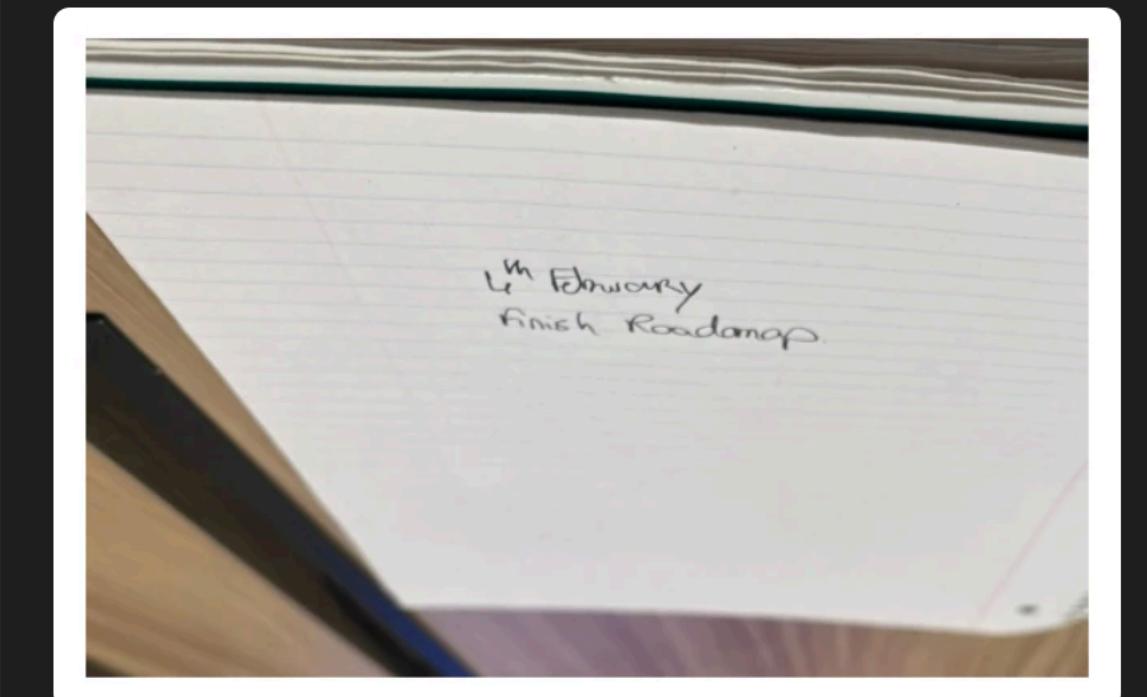


- The user posted an image at /mnt/data/own_notebook.jpg and asked me to inspect what was written on it. I'll need to load the image so I can inspect the text. Once I view it, I notice the text is upside down, so I'll rotate it so it's readable. From there, I can check what's written and share my findings clearly with the user.

Analyzed image



Analyzed image

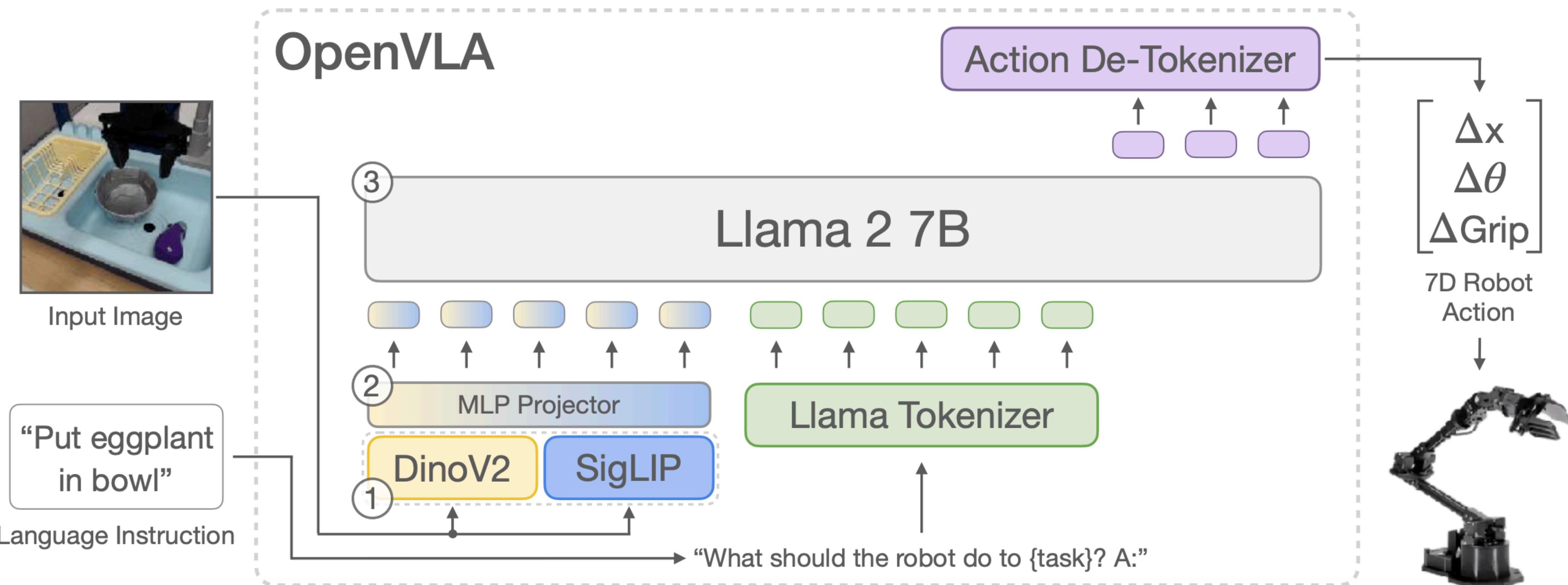


Done

It says: "4th February – finish roadmap."

Vision-language action model

Open-VLA



Vision-language action model

Pi-0

