Vision at Scale

Required

Zhai et al., Sigmoid Loss for Language Image Pre-Training (SigLIP) · arXiv:2303.15343

Optional

Radford et al., CLIP · arXiv:2103.00020 Oquab et al., DINOv2 · arXiv:2304.07193

Wang et al., Qwen2-VL · arXiv:2409.12191 Bai et al., Qwen3-VL Technical Report · arXiv:2511.21631

Christoffer Heckman

Department of Computer Science, University of Colorado Boulder

November 12, 2026

Quick overview

The vision transformer

CLIP vs SigLIP

MAE and DINOv2

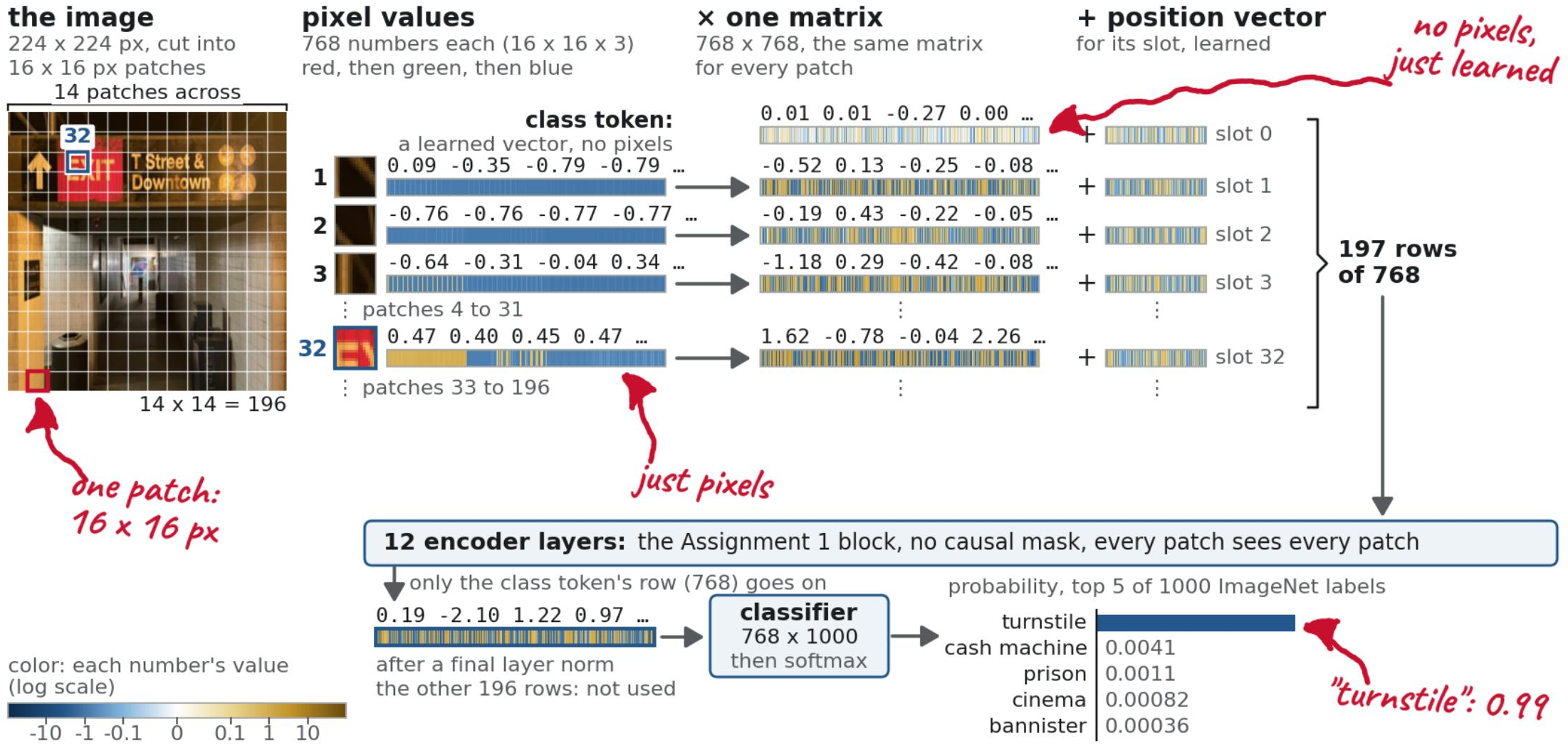
Deep dive into Qwen3-VL

1 THE VISION TRANSFORMER

Sixteen by Sixteen

Patches

my run of ViT-B/16 · Dosovitskiy et al. 2020

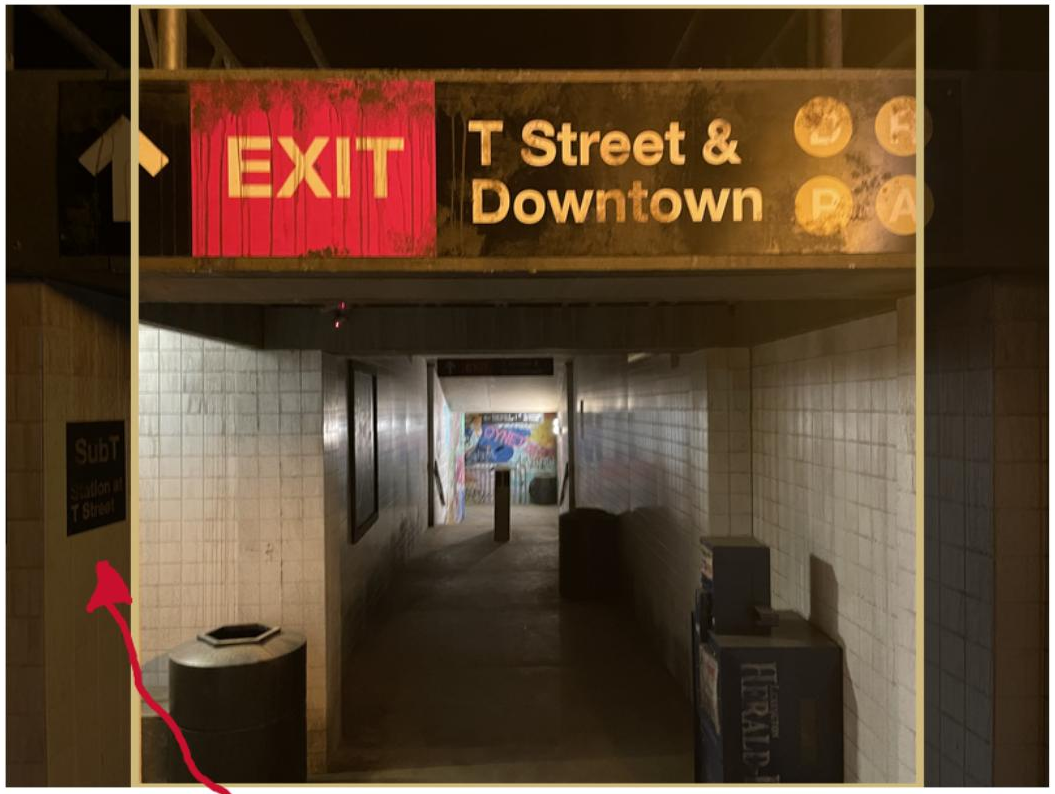


Square Crops

my photo, SubT Final 2021

the photo, 4032 x 3024 px

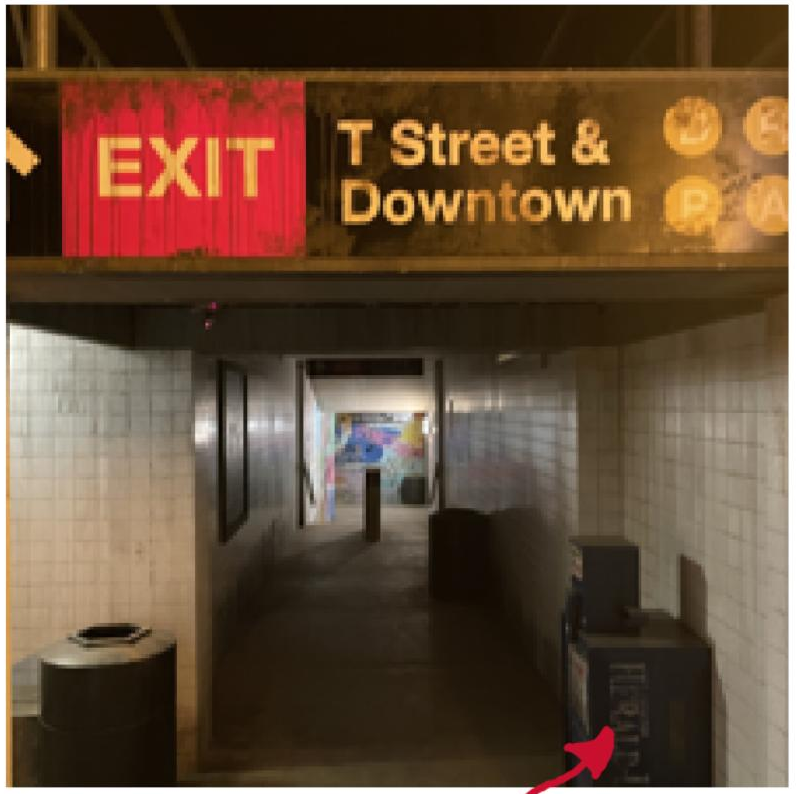
gold box: the square that is kept



sides cut off

resize
and crop

what the encoder sees: 224 x 224



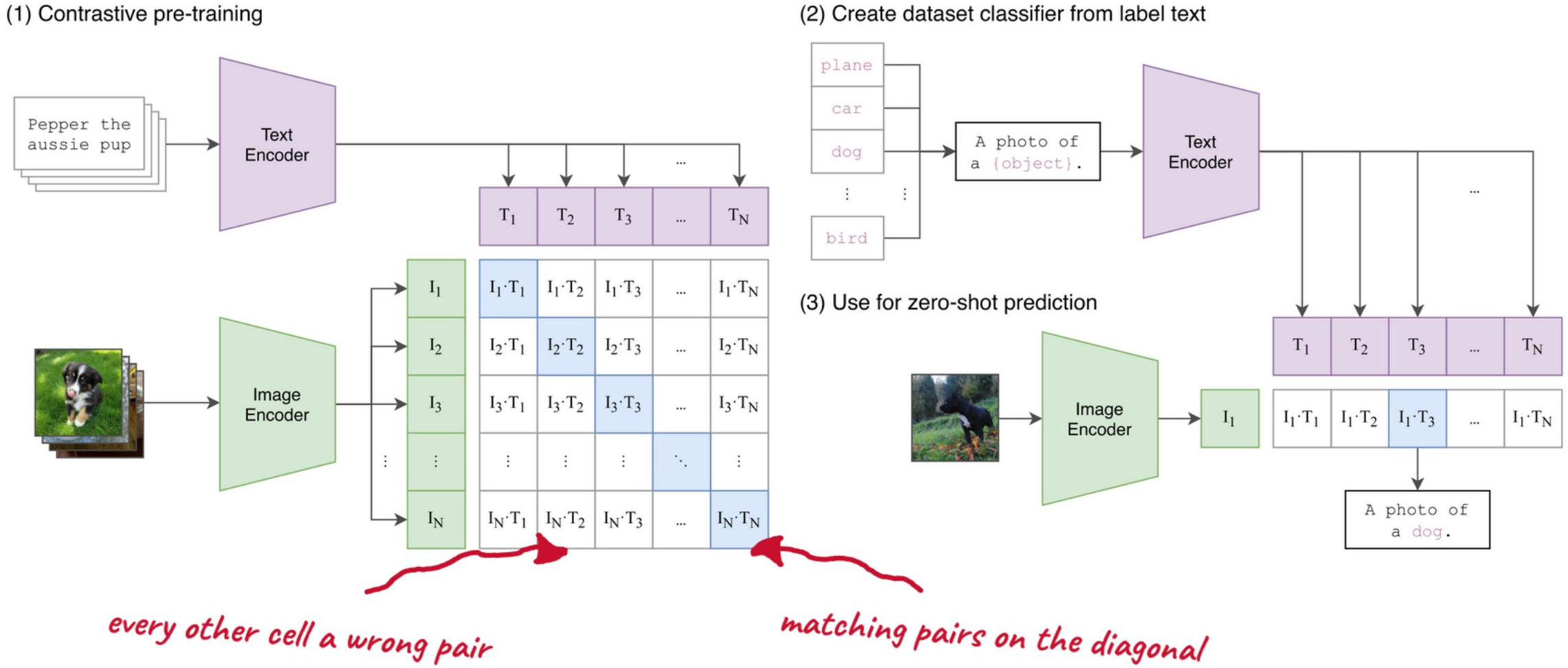
small print gone at 224

2 CONTRASTIVE PRETRAINING

Captions as Labels

CLIP

Radford et al. 2021 · arXiv:2103.00020



The Batch Matrix

my run of SigLIP base

	1	2	3	4	5	6	7	8
	0.15	−0.04	0.02	0.00	−0.12	0.01	0.01	−0.03
	−0.04	0.11	−0.06	−0.02	−0.07	−0.07	−0.08	−0.10
	−0.05	−0.11	0.14	−0.03	−0.15	−0.04	−0.04	−0.04
	−0.03	−0.08	0.06	0.17	−0.07	0.00	−0.01	−0.01
	−0.03	−0.01	−0.04	0.05	0.14	−0.07	0.01	0.00
	0.02	0.02	−0.02	−0.01	−0.04	0.15	0.13	0.14
	0.00	−0.06	−0.01	0.01	−0.07	0.09	0.17	0.11
	−0.04	−0.05	−0.03	0.02	−0.04	0.07	0.20	0.22

true pairs

captions

1 a drone with round propeller guards
2 a group of people standing on grass
3 a red laser beam across a table
4 a lab filled with thick fog
5 cooling towers in a pine forest
6 a small robot in the desert
7 a four-legged robot in a mine tunnel
8 a wheeled robot in a mine tunnel

cosine similarity s

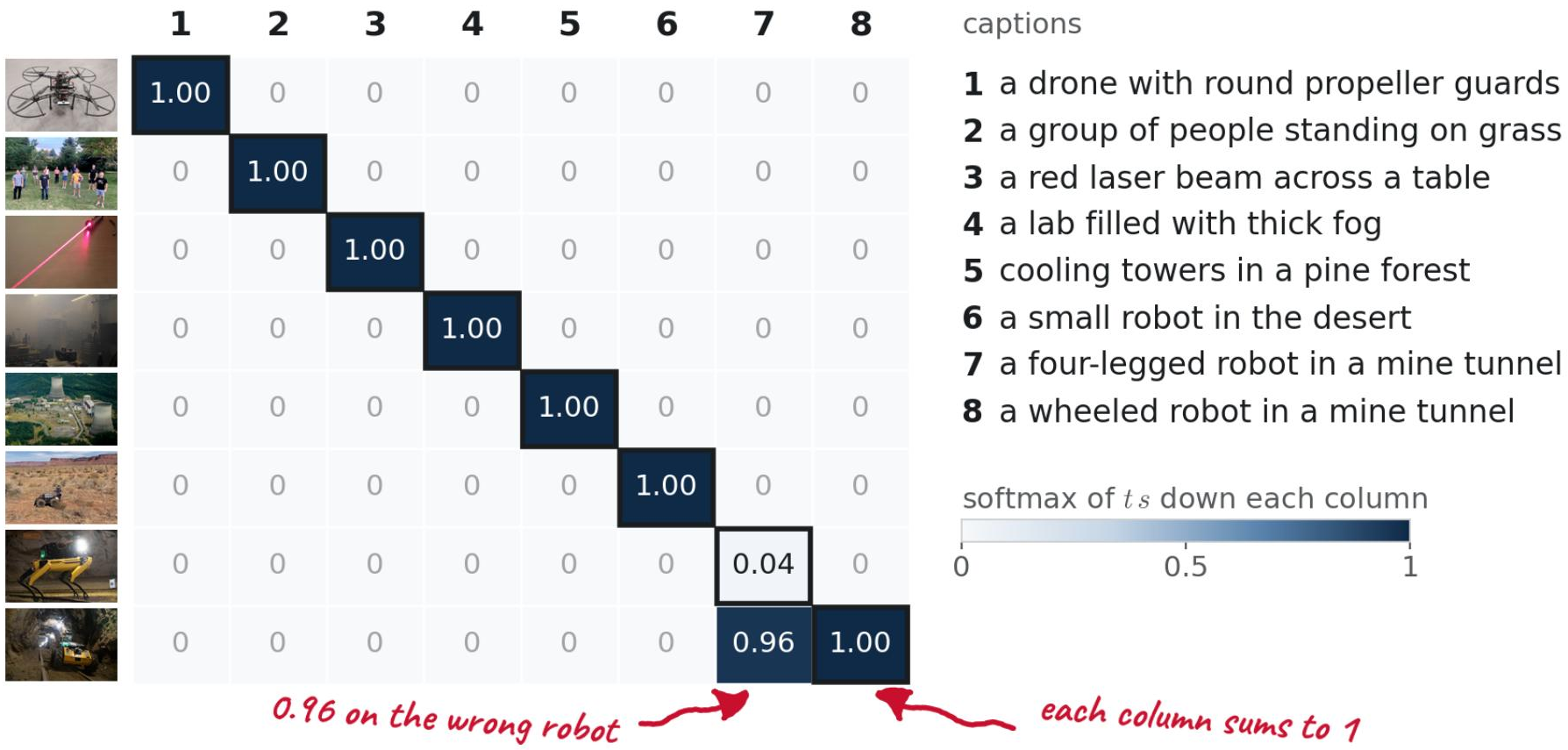
−0.1 0.0 0.1 0.2

0.22 beats 0.20

Softmax Loss

my run of SigLIP base · Radford et al. 2021

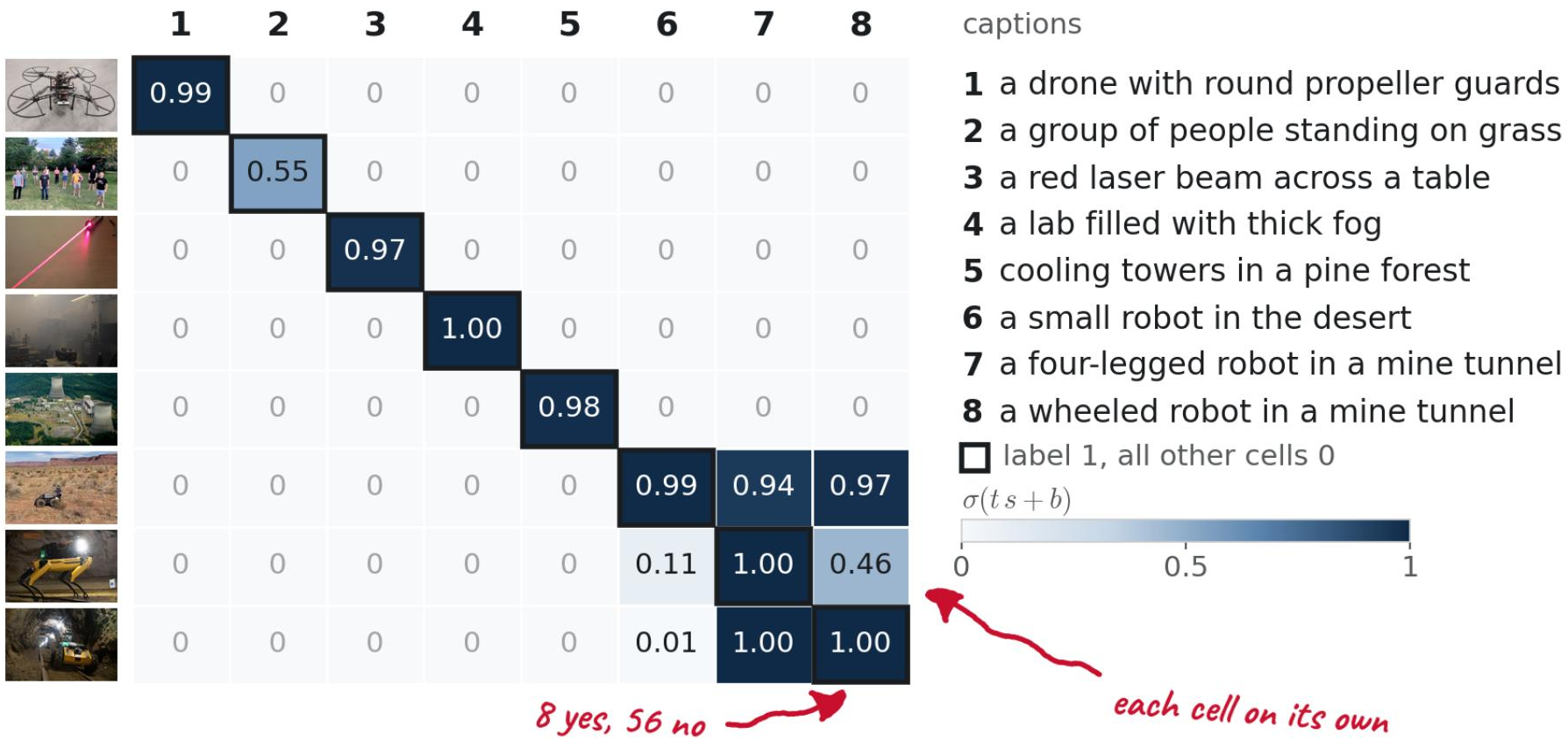
$$\mathcal{L}_{\text{softmax}} = -\frac{1}{2N}\sum_{i=1}^{N}\left[\log\frac{e^{t\,\mathbf{x}_i\cdot\mathbf{y}_i}}{\sum_j e^{t\,\mathbf{x}_i\cdot\mathbf{y}_j}} + \log\frac{e^{t\,\mathbf{x}_i\cdot\mathbf{y}_i}}{\sum_j e^{t\,\mathbf{x}_j\cdot\mathbf{y}_i}}\right]$$



Sigmoid Loss

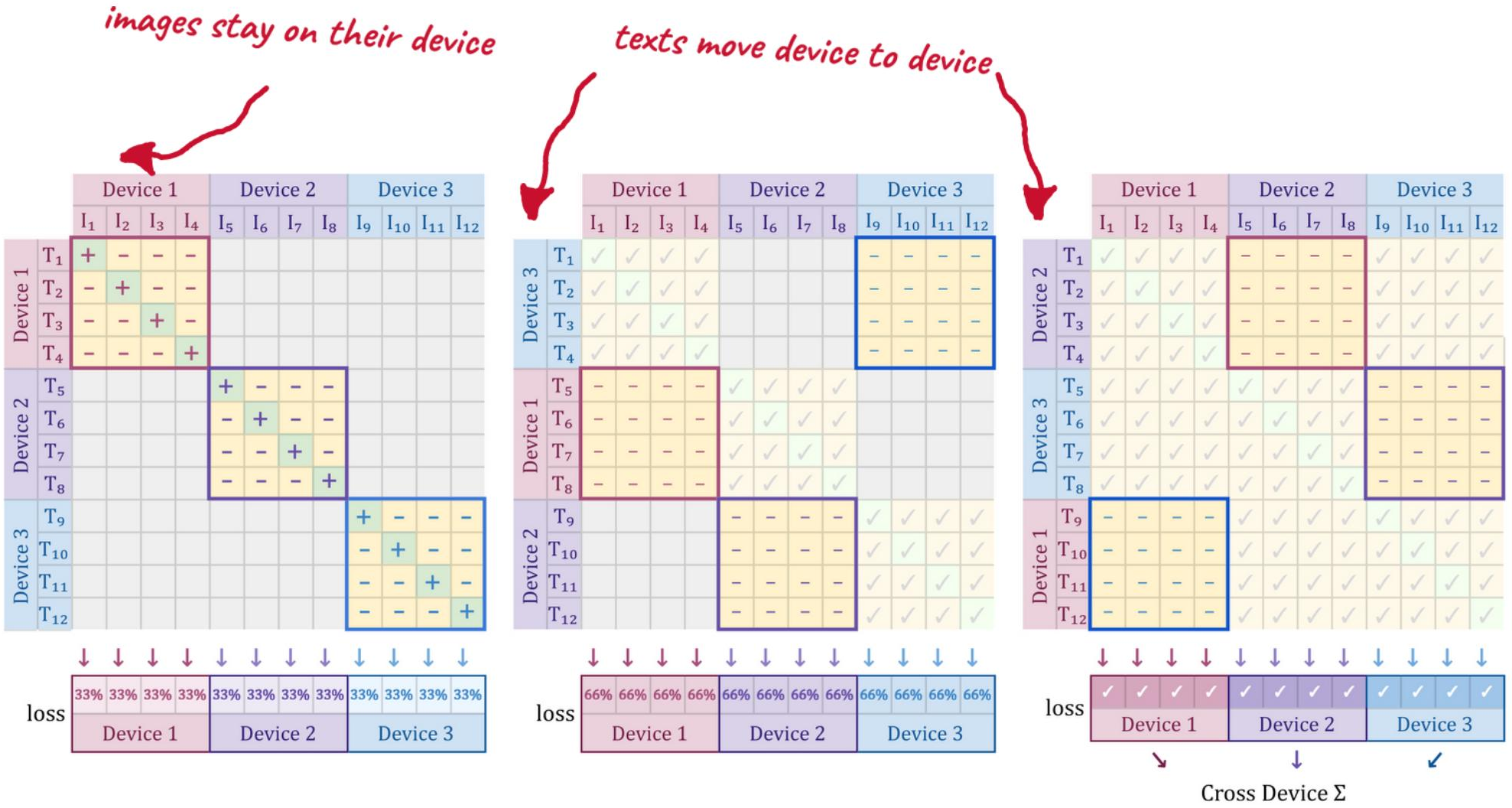
my run of SigLIP base · Zhai et al. 2023

$$\mathcal{L}_{\text{sigmoid}} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N}\log\sigma\big(z_{ij}\,(t\,\mathbf{x}_i\cdot\mathbf{y}_j+b)\big), \qquad z_{ij}=\begin{cases}+1 & i=j\\ -1 & i\neq j\end{cases}$$



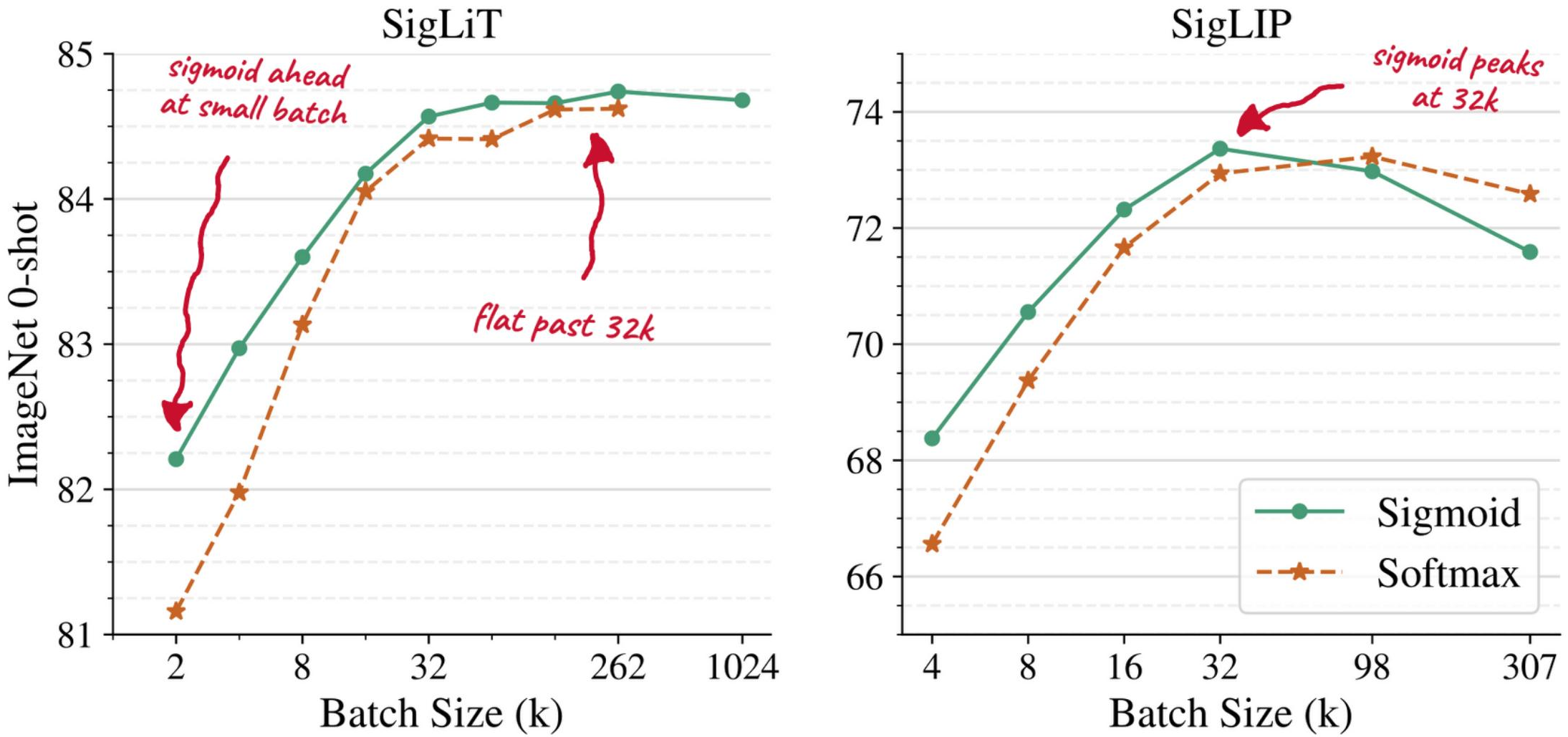
Chunked Batches

Zhai et al. 2023 · arXiv:2303.15343



Batch Size

Zhai et al. 2023 · arXiv:2303.15343

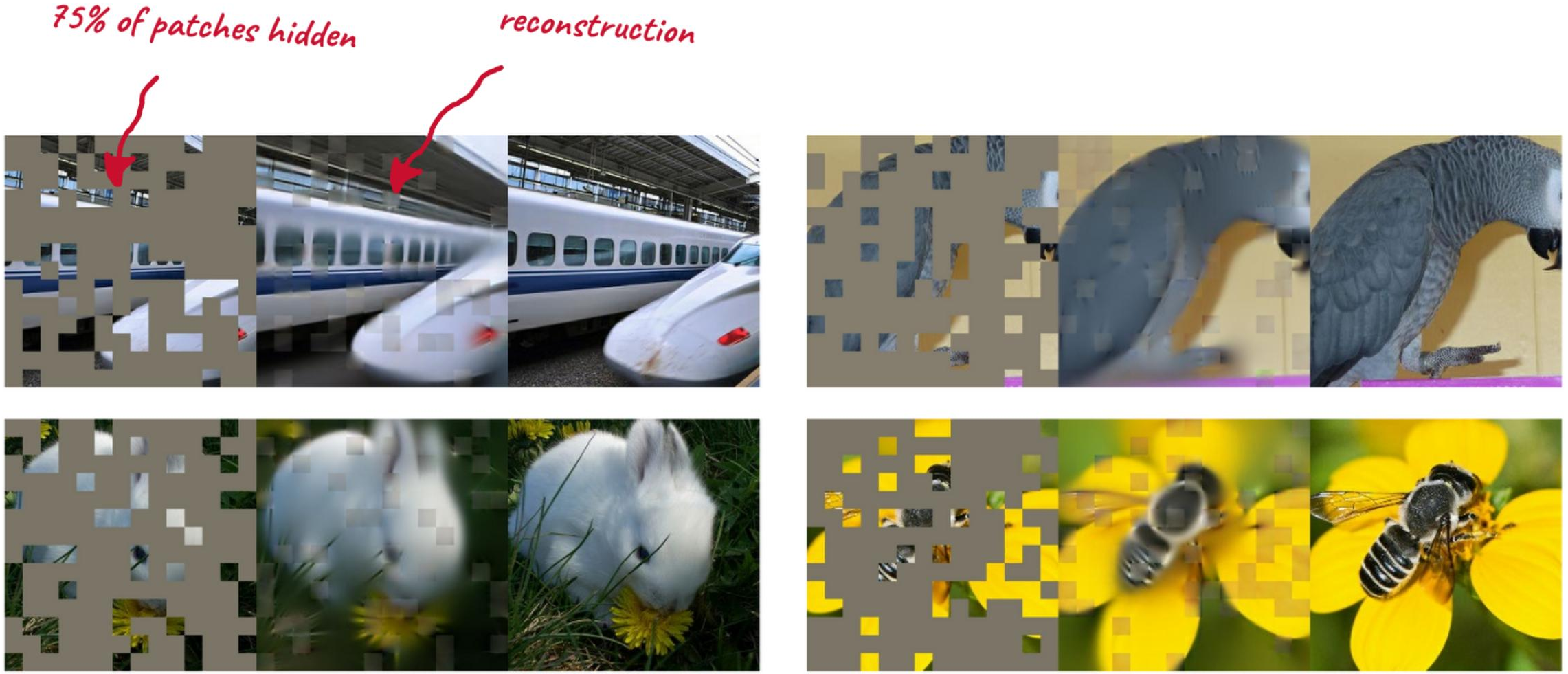


3 SELF-SUPERVISED FEATURES

No Captions at All

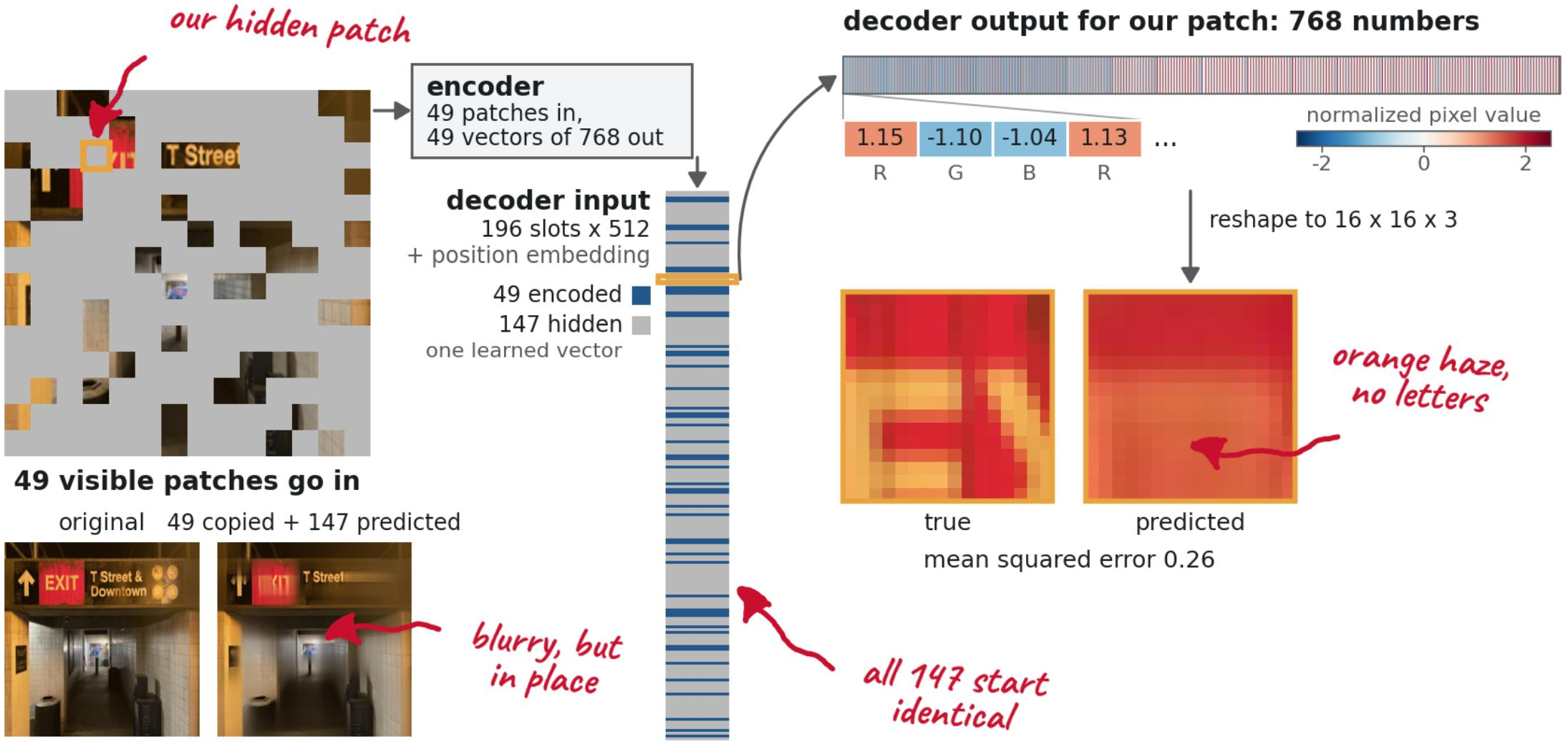
Masked Patches

He et al. 2021 · arXiv:2111.06377



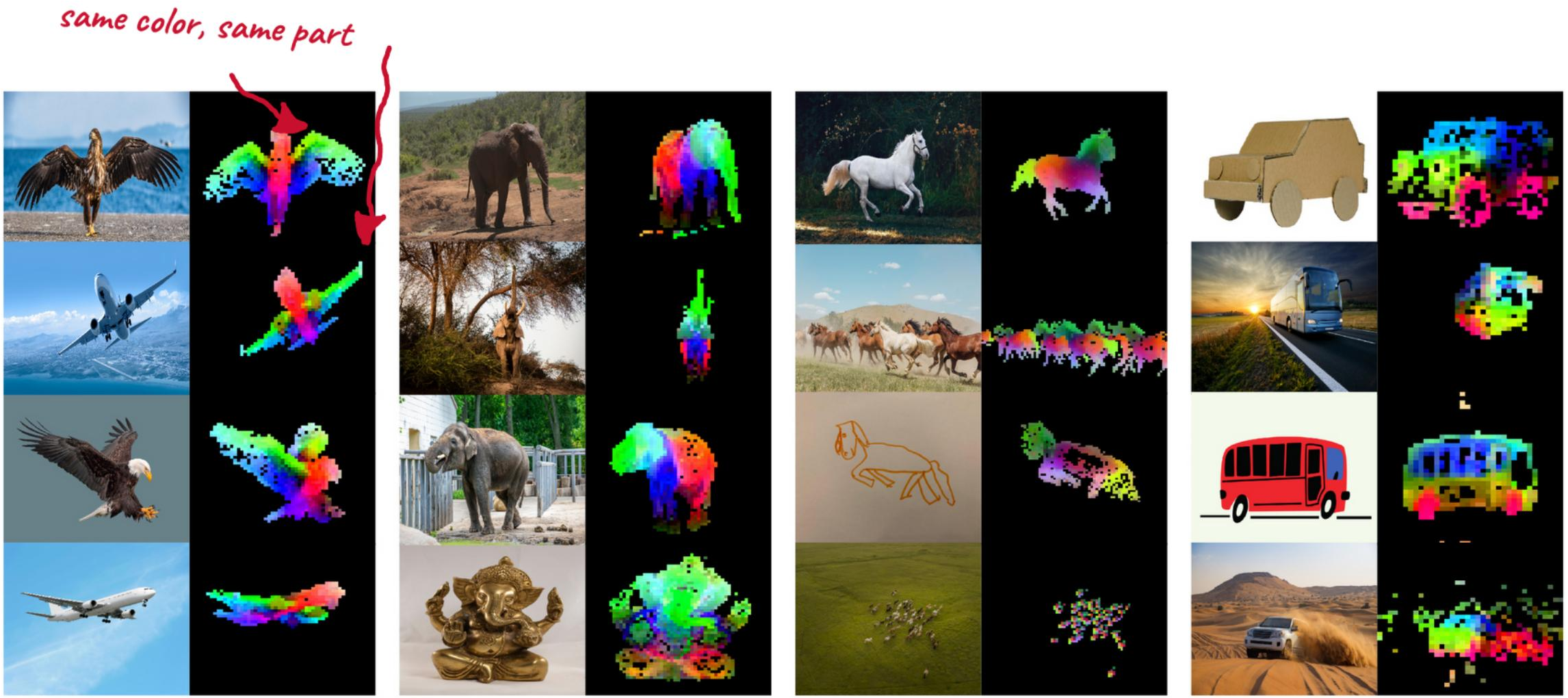
One Patch Back

my run of MAE base



DINOv2

Oquab et al. 2023 · arXiv:2304.07193

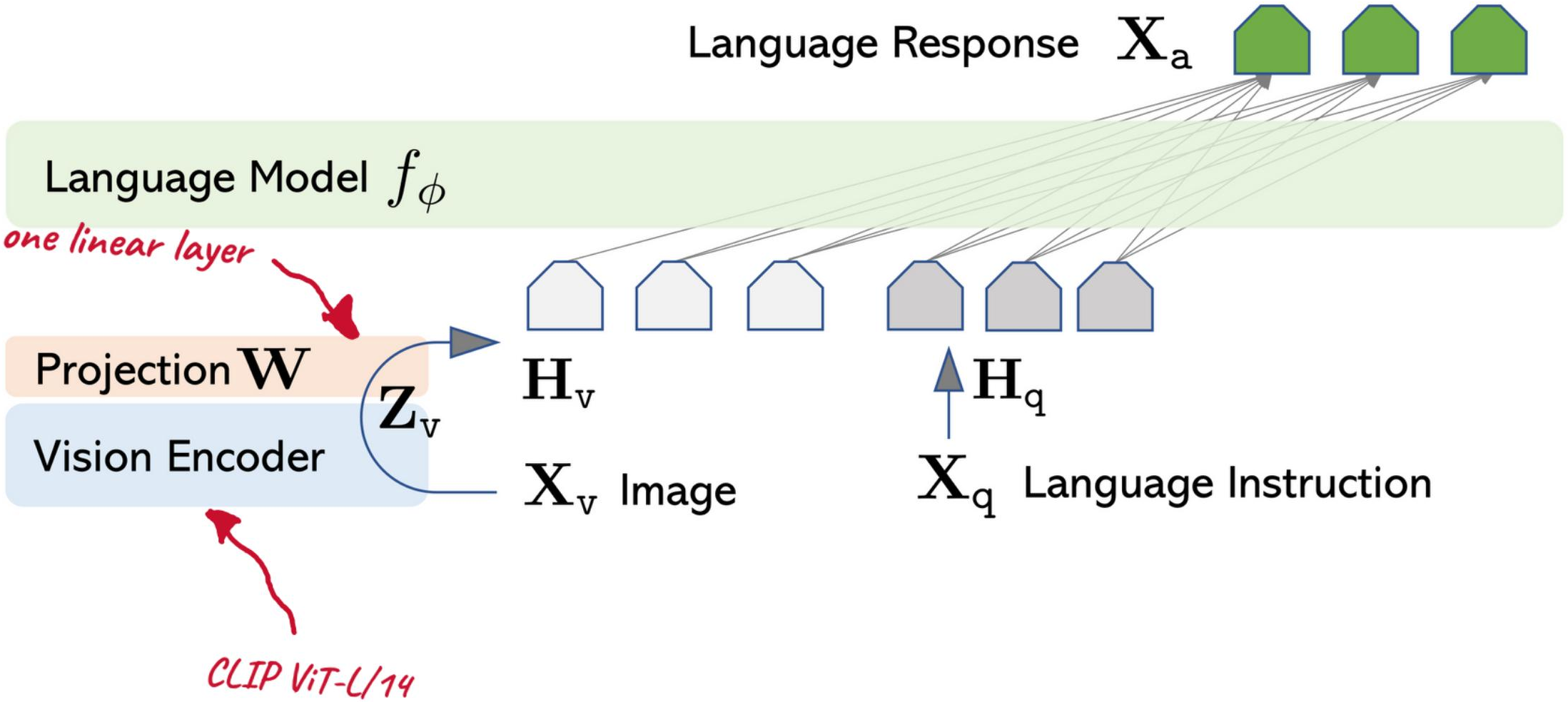


4 A VLM, LAYER BY LAYER

Inside Qwen3-VL

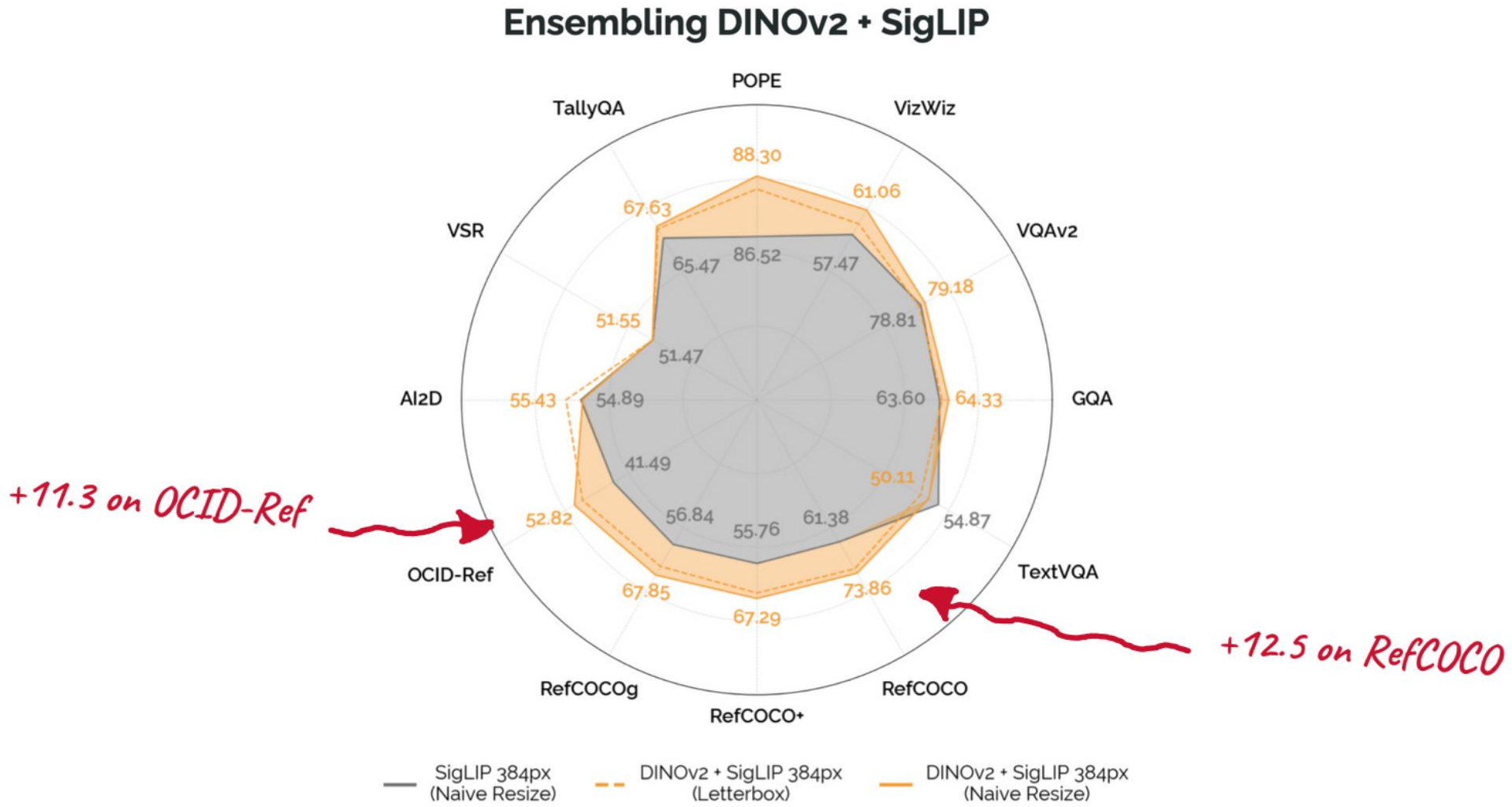
LLaVA

Liu et al. 2023 · arXiv:2304.08485



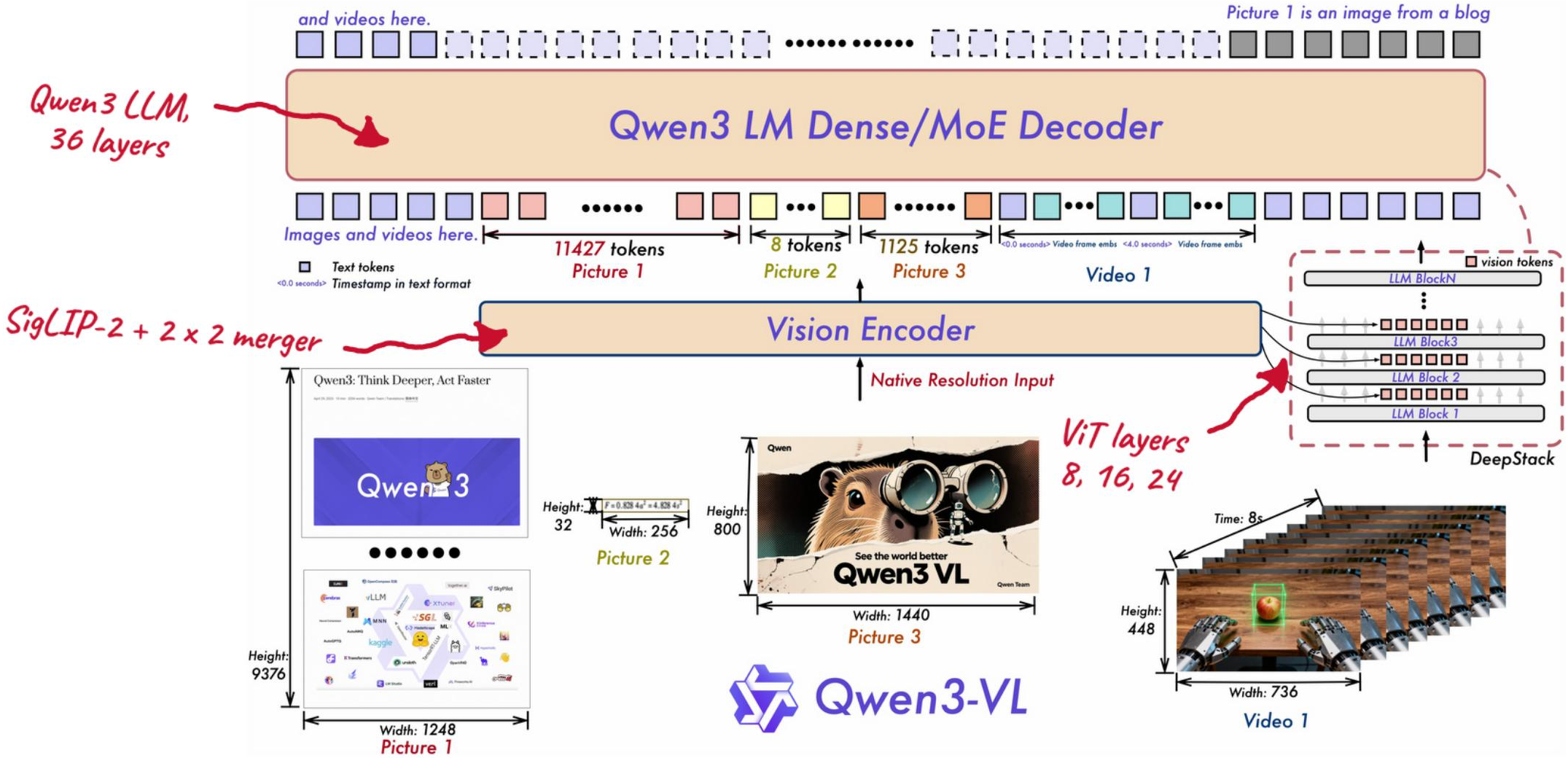
Two Encoders

Karamcheti et al. 2024 · arXiv:2402.07865



Qwen3-VL

Bai et al. 2025 · arXiv:2511.21631



The Config File

Qwen3-VL-8B config.json, as released

config.json, Qwen3-VL-8B-Instruct

```
{
  "vision_config": {
    "depth": 27,
    "hidden_size": 1152,
    "patch_size": 16,
    "spatial_merge_size": 2,
    "temporal_patch_size": 2,
    "out_hidden_size": 4096,
    "deepstack_visual_indexes": [8, 16, 24],
    ...
  },
  "text_config": {
    "hidden_size": 4096,
    "num_hidden_layers": 36,
    "head_dim": 128,
    "num_key_value_heads": 8,
    "rope_theta": 5000000,
    "rope_scaling": {
      "mrope_section": [24, 20, 20],
      "mrope_interleaved": true,
      ...
    },
    ...
  },
  ...
}
```

SigLIP-2 encoder, 27 layers

4 patches into 1 token

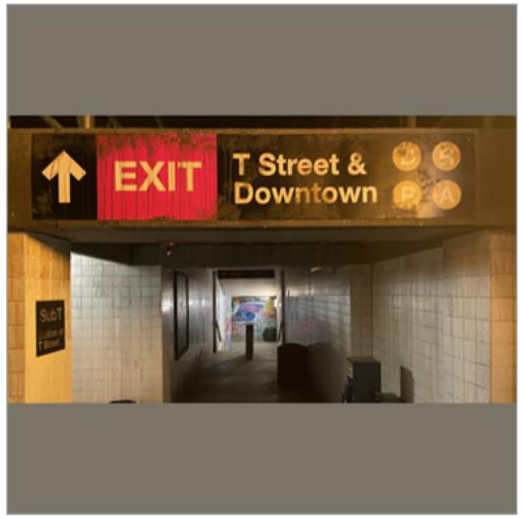
vision into LLM layers

position: 3 axes

Resolution Three Ways

my run of the LLaVA and Qwen3-VL processors

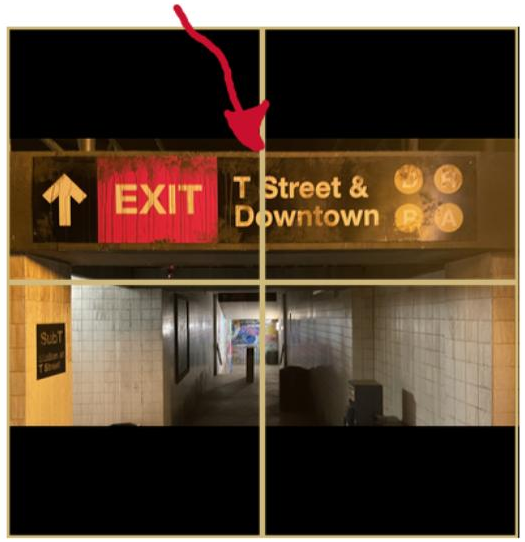
LLaVA-1.5
one 336 x 336 square



padding

576 tokens

LLaVA-1.6 (AnyRes)
2 x 2 tiles of 336 px, plus the whole frame



tile edges

whole frame
at 336 x 336



1,948 tokens

Qwen3-VL
native resolution, 32 x 32 px per token

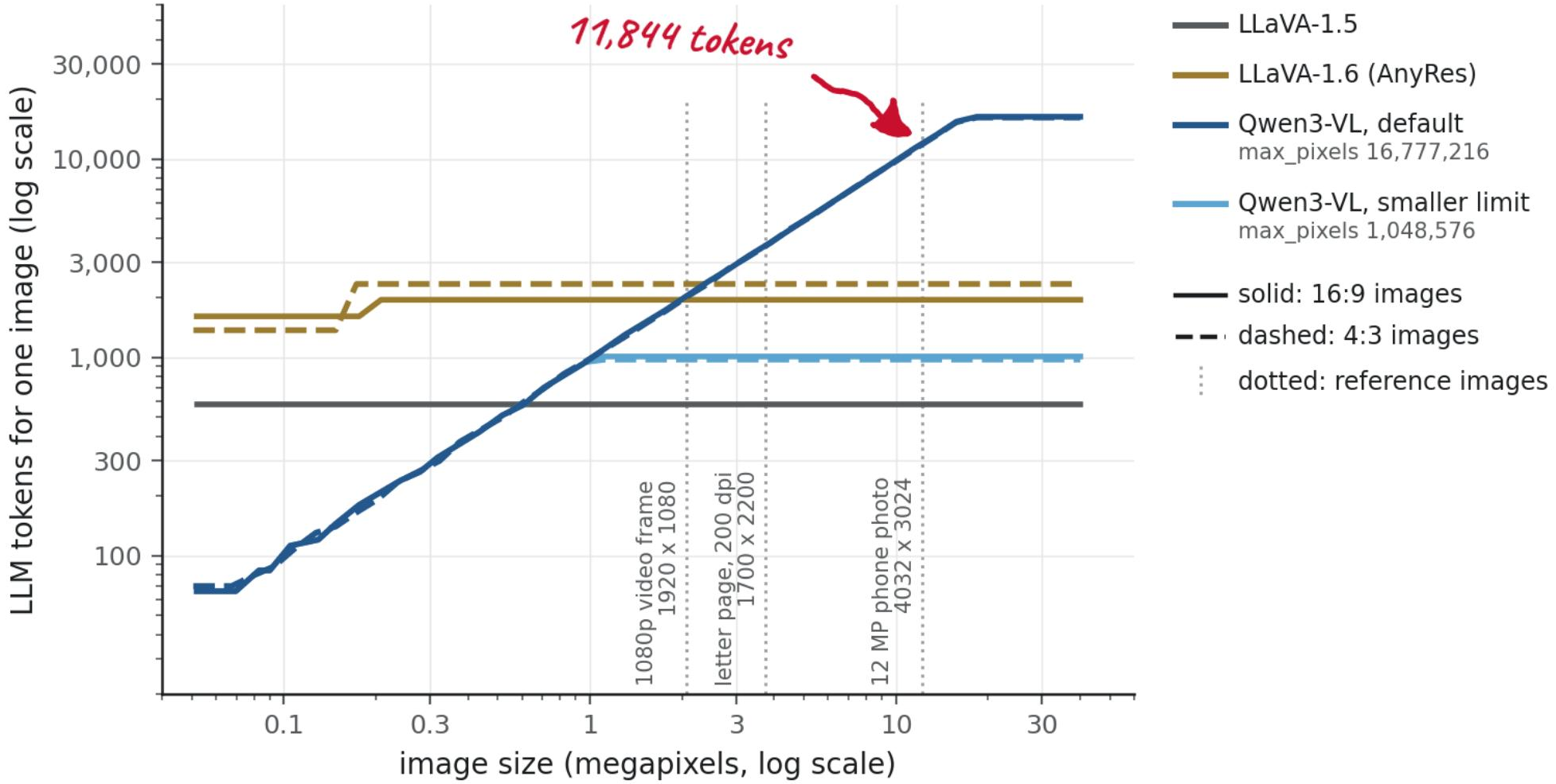


40 x 22 tokens

880 tokens

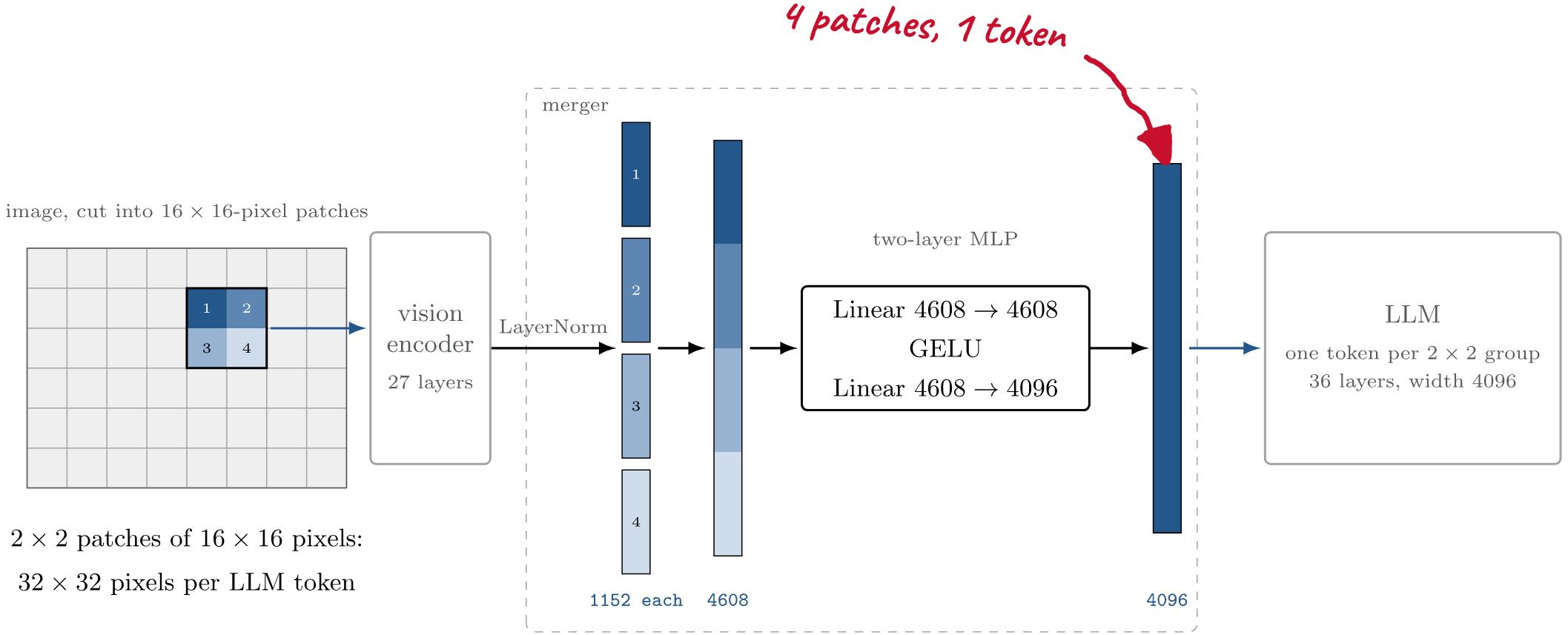
Pixels per Token

my run of the LLaVA and Qwen3-VL processors



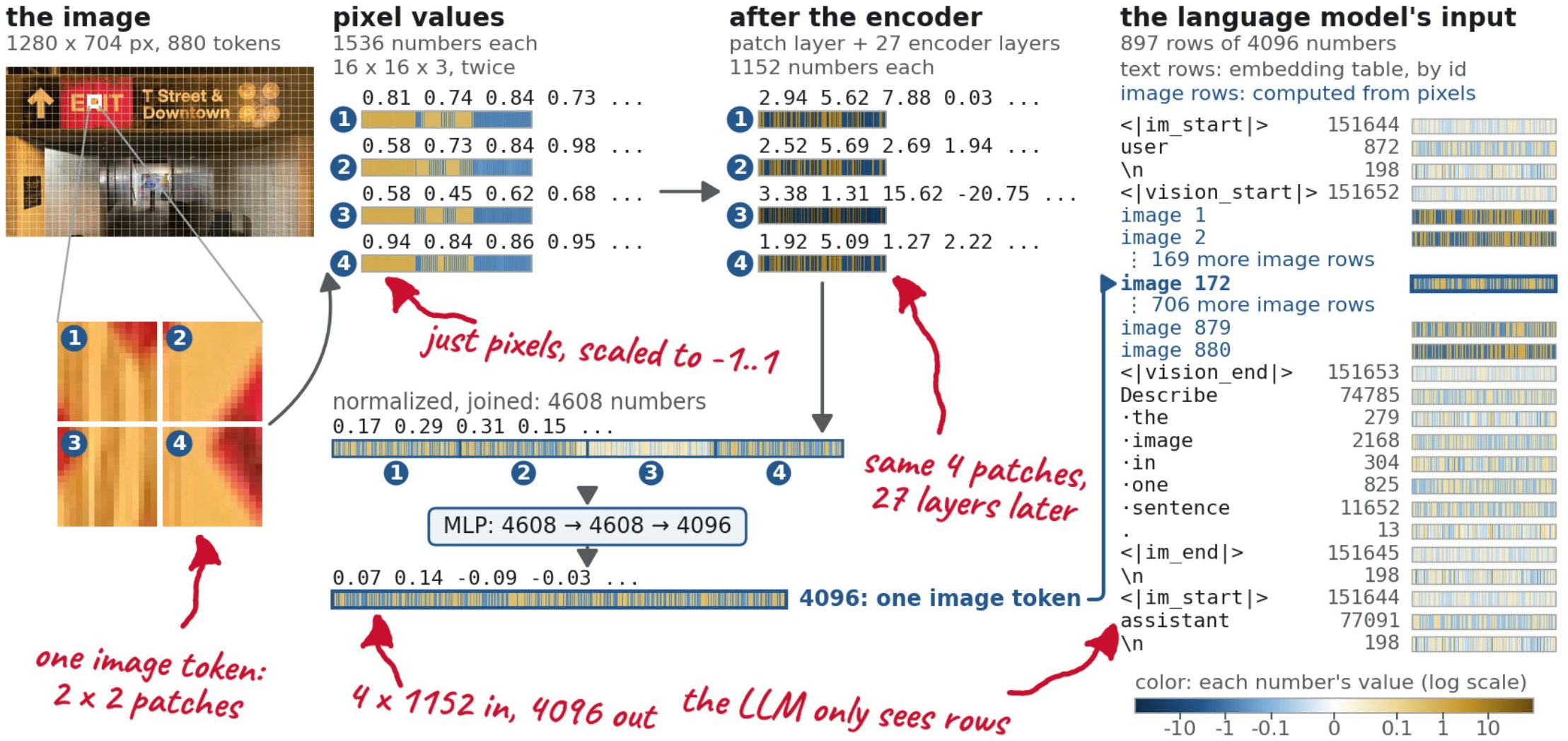
The 2 x 2 Merge

from the Qwen3-VL code



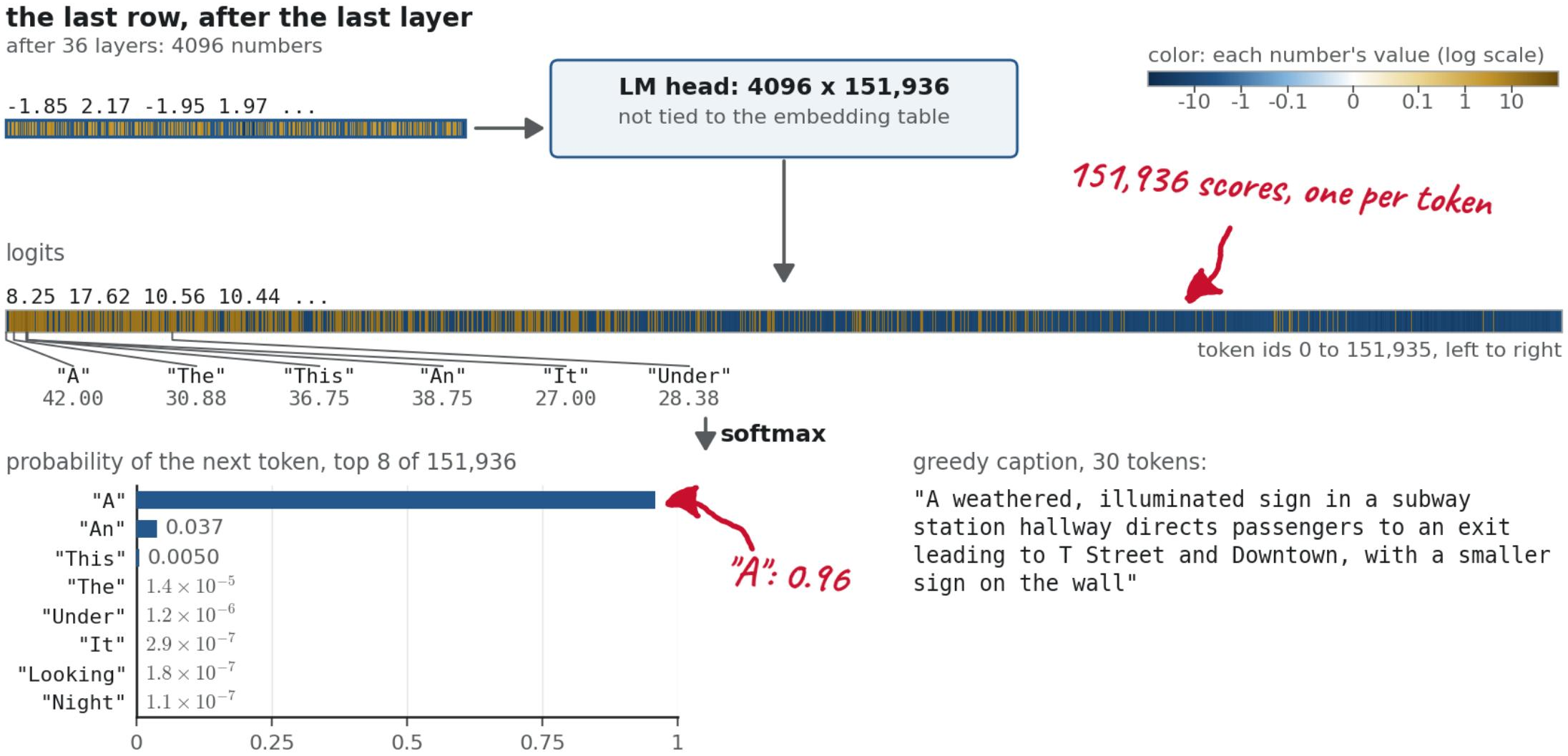
One Token In

my run on Qwen3-VL-8B



One Word Out

my run on Qwen3-VL-8B

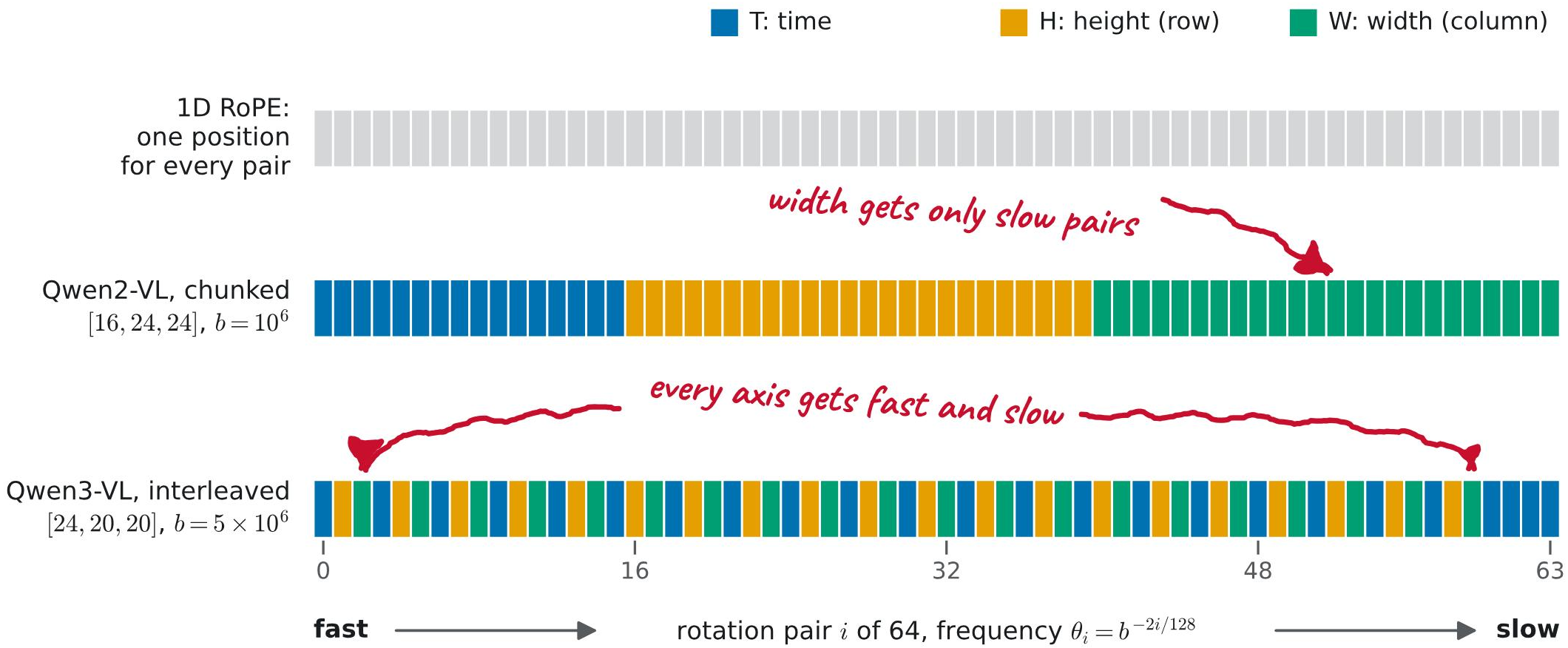


greedy caption, 30 tokens:

```
"A weathered, illuminated sign in a subway
station hallway directs passengers to an exit
leading to T Street and Downtown, with a smaller
sign on the wall"
```

MRoPE

pair layouts from the Qwen2-VL and Qwen3-VL code

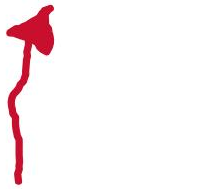


angle of pair i: $p\theta_i$ with $p = t$, h or w by the pair's color

Position IDs

my run on Qwen3-VL

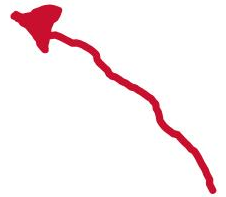
	text		image: a 2×3 grid of tokens						text			
	...	"Describe"	<|vision_start|>	<|image_pad|> × 6					<|vision_end|>	" briefly"	"."	...
t time	...	3	4	5	5	5	5	5	5	8	9	10	...
h height	...	3	4	5	5	5	6	6	6	8	9	10	...
w width	...	3	4	5	6	7	5	6	7	8	9	10	...



text: t = h = w



image: one t, grid of h and w



next text starts at max + 1

DeepStack

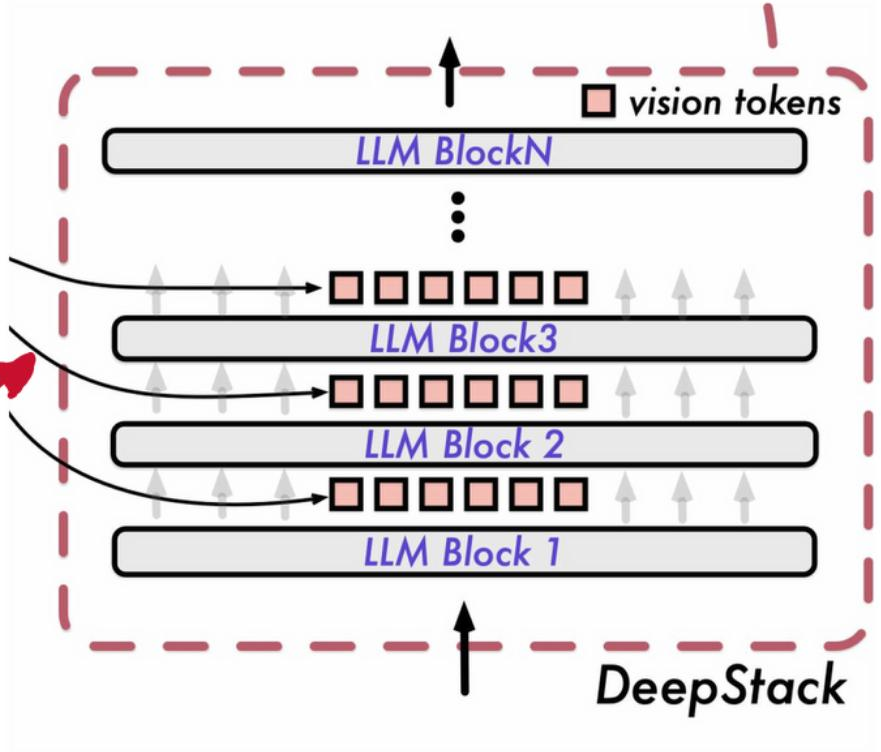
Bai et al. 2025 · arXiv:2511.21631

Method	AVG	AI2D	OCRB	TVQA	InfoVQA	ChartQA	DocVQA	MMMU	MMStar	RLWDQA	MMB_{EN}	MMB_{CN}
Baseline	74.7	81.8	81.0	80.6	71.9	81.5	89.5	52.9	55.5	67.7	81.0	78.1
DeepStack	76.0	83.2	83.6	80.5	74.2	83.3	91.1	54.1	57.7	68.1	81.2	78.5

+2.6 on OCRBench

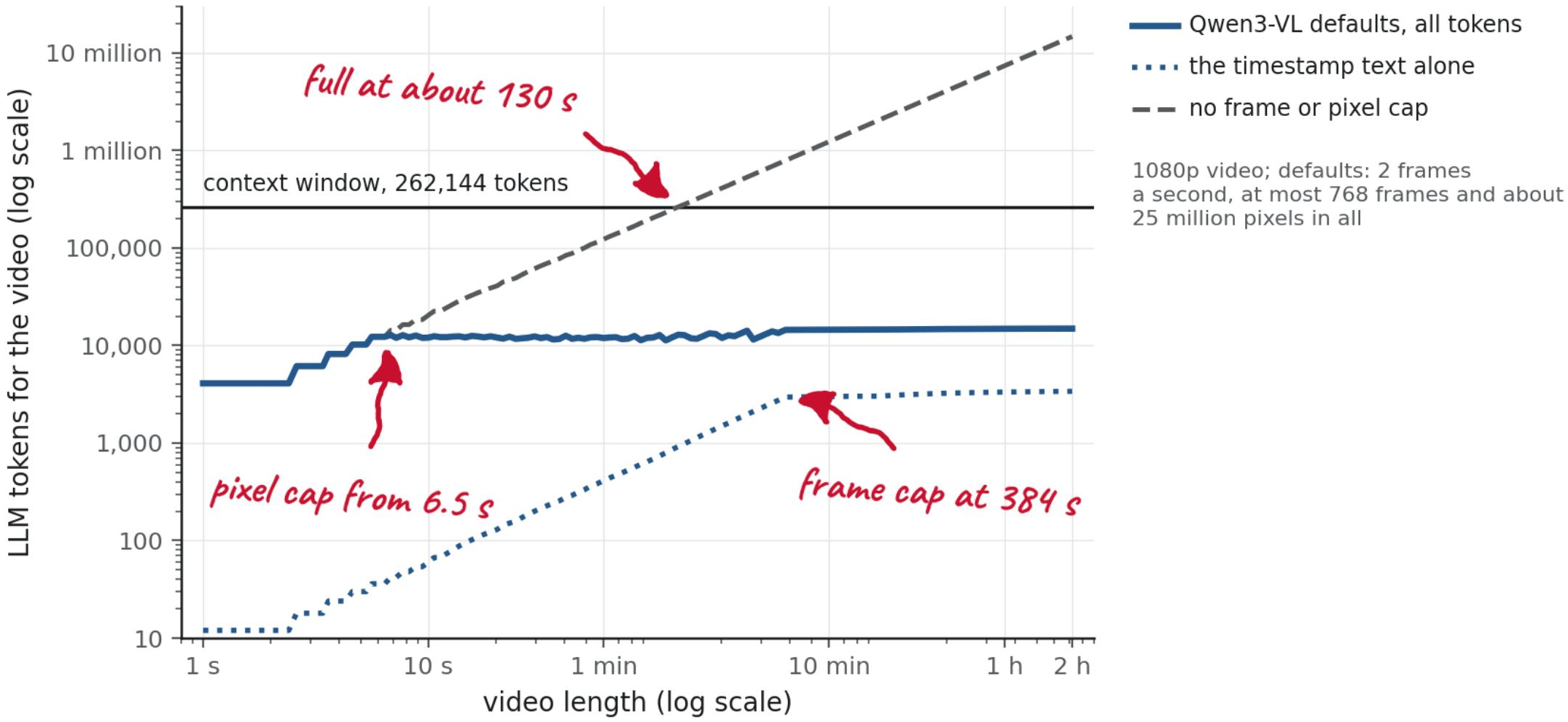
+2.3 on InfoVQA

ViT layers 8, 16, 24



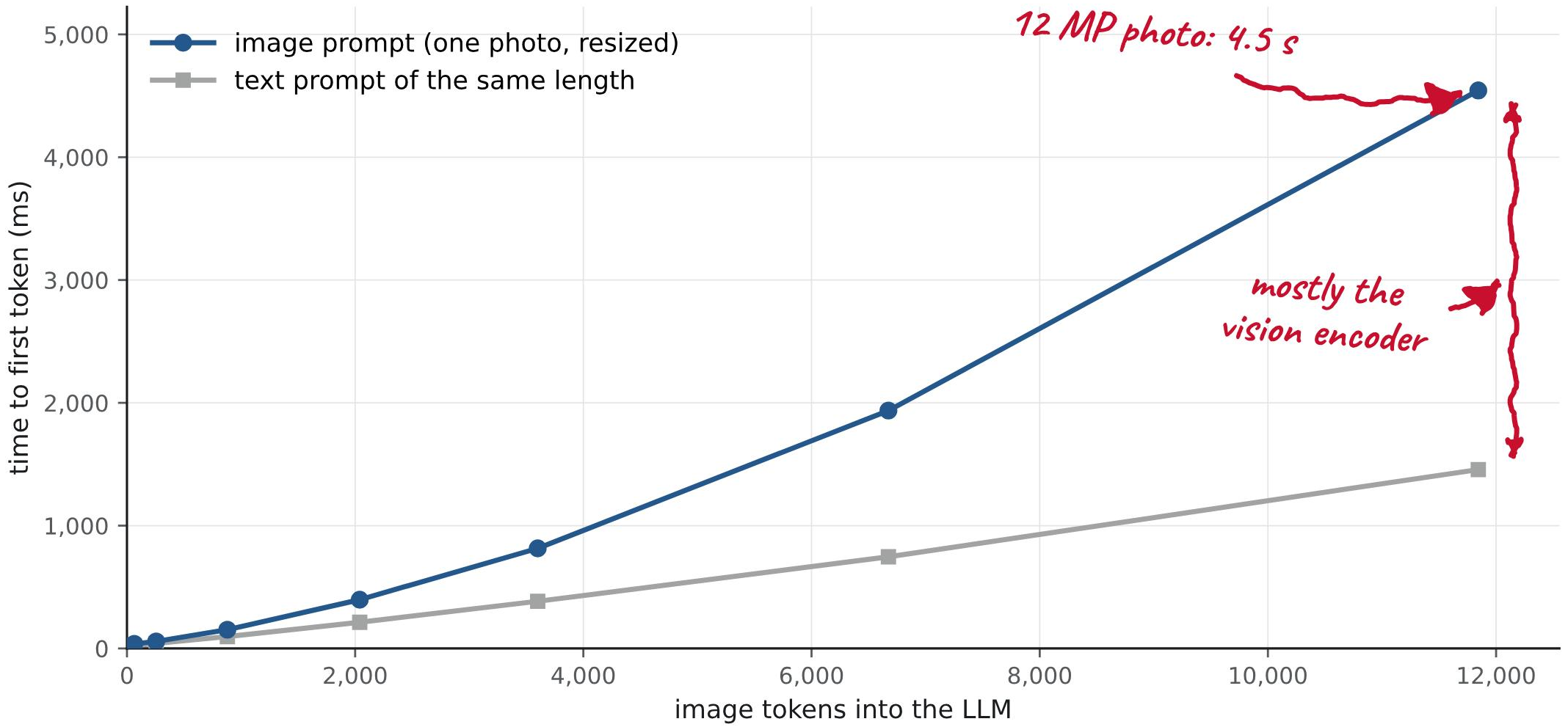
Video

my run of the Qwen3-VL video processor



Prefill

my runs on my 4090, Qwen3-VL-8B



Reading

Required

Zhai et al., Sigmoid Loss for Language Image Pre-Training (SigLIP) · arXiv:2303.15343

Optional

Radford et al., CLIP · arXiv:2103.00020 Oquab et al., DINOv2 · arXiv:2304.07193

Wang et al., Qwen2-VL · arXiv:2409.12191 Bai et al., Qwen3-VL Technical Report · arXiv:2511.21631

Assignment 7 is due Tuesday, and its quiz is in class that day with Nolan.