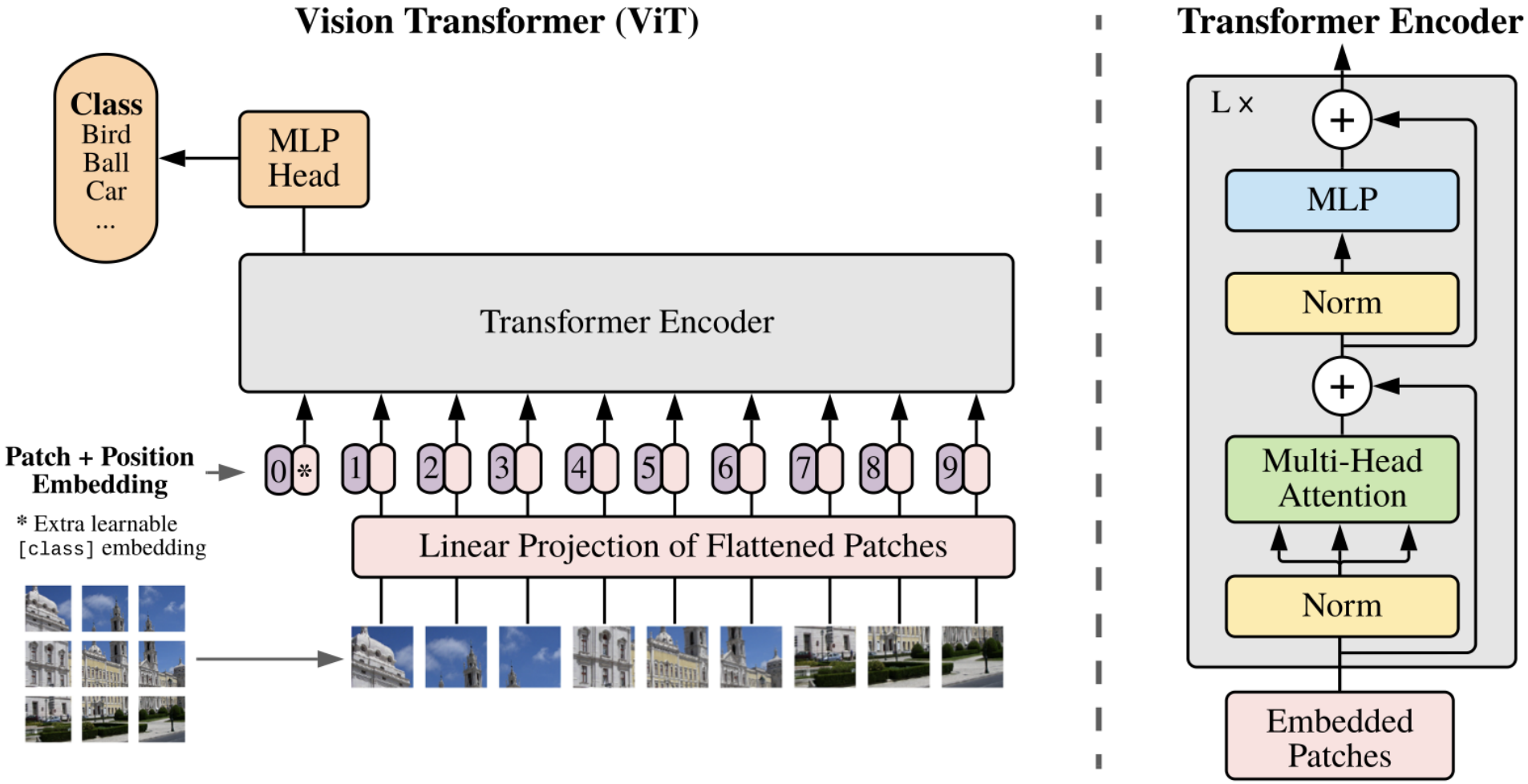
Vision Language Models

Курцев Дмитрий

- **Vision Transformers**
- VLMs
- Benchmarks
- Architectures

ViT



https://arxiv.org/pdf/2010.11929

AN IMAGE IS WORTH 16X16 WORDS: TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE

22.10.2020

Swin

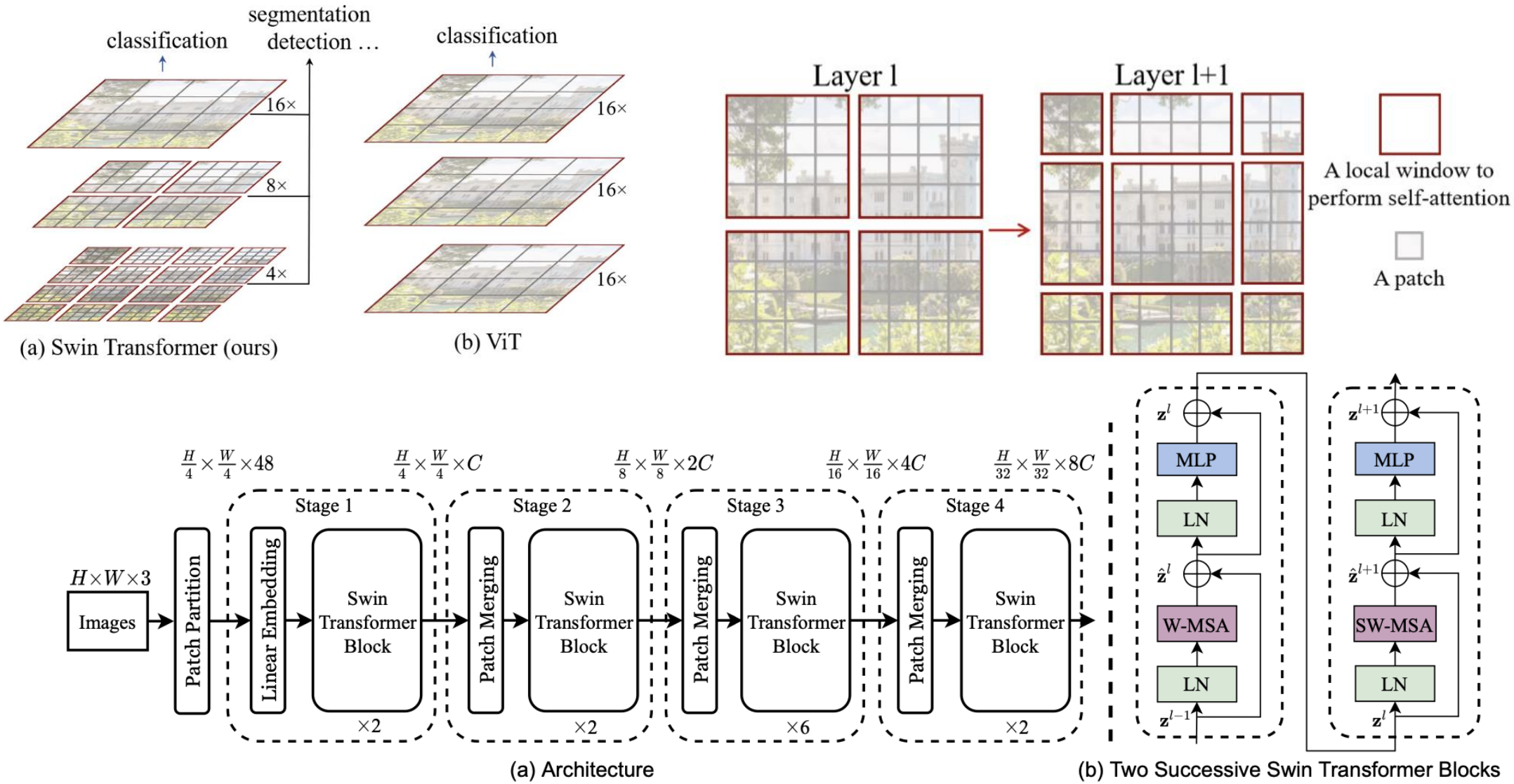


Figure 3. (a) The architecture of a Swin Transformer (Swin-T); (b) two successive Swin Transformer Blocks (notation presented with Eq. (3)). W-MSA and SW-MSA are multi-head self attention modules with regular and shifted windowing configurations, respectively.

https://arxiv.org/pdf/2103.14030

Swin Transformer: Hierarchical Vision Transformer using Shifted Windows

25.03.2021

VOLO

Table 2. Architecture information of different variants of VOLO. The resolution information is based on an input image of size 224×224. The number of parameters includes that of weights for both the network backbone and the classifier head. 'Layer' refers to either a Outlooker block or a Transformer block.

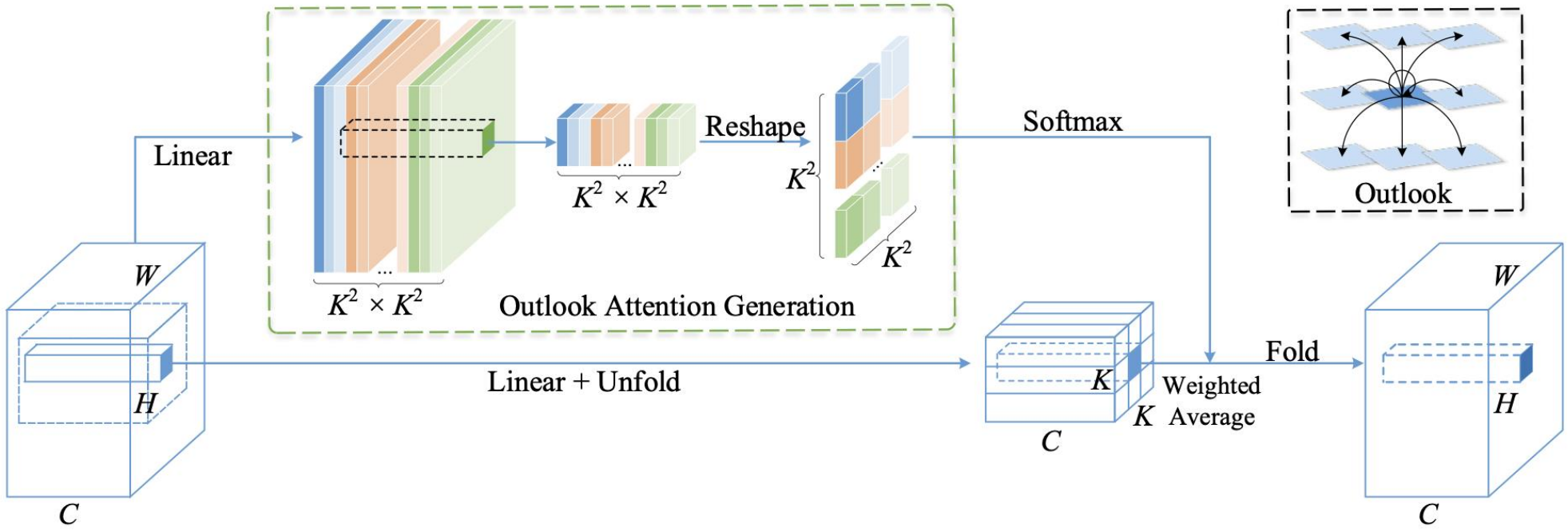
Specification	VOLO-D1	VOLO-D2	VOLO-D3	VOLO-D4	VOLO-D5
Patch Embedding	8×8	8×8	8×8	8×8	8×8
Stage 1 (28×28)	[head: 6, stride: 2 kernel: 3×3 mlp: 3, dim: 192] $\times 4$	[head: 8, stride: 2 kernel: 3×3 mlp: 3, dim: 256] $\times 6$	[head: 8, stride: 2 kernel: 3×3 mlp: 3, dim: 256] $\times 8$	[head: 12, stride: 2 kernel: 3×3 mlp: 3, dim: 384] $\times 8$	[head: 12, stride: 2 kernel: 3×3 mlp: 4, dim: 384] $\times 12$
Patch Embedding	2×2	2×2	2×2	2×2	2×2
Stage 2 (14×14)	[#heads: 12 mlp: 3, dim: 384] $\times 14$	[#heads: 16 mlp: 3, dim: 512] $\times 18$	[#heads: 16 mlp: 3, dim: 512] $\times 28$	[#heads: 16 mlp: 3, dim: 768] $\times 28$	[#heads: 16 mlp: 4, dim: 768] $\times 36$
Total Layers	18	24	36	36	48
Parameters	26.6M	58.7M	86.3M	193M	296M

https://arxiv.org/pdf/2106.13112

VOLO: Vision Outlooker for Visual Recognition

24.06.2021

VOLO



Formally, given the input $\mathbf{X}$, each C-dim token is first projected, using two linear layers of weights $\mathbf{W}_A \in \mathbb{R}^{C\times K^4}$ and $\mathbf{W}_V \in \mathbb{R}^{C\times C}$, into outlook weights $\mathbf{A} \in \mathbb{R}^{H\times W\times K^4}$ and value representation $\mathbf{V} \in \mathbb{R}^{H\times W\times C}$, respectively. Let $\mathbf{V}_{\Delta_{i,j}} \in \mathbb{R}^{C\times K^2}$ denote all the values within the local window centered at (i,j), *i.e.*,

$$\mathbf{V}_{\Delta_{i,j}} = \{\mathbf{V}_{i+p-\lfloor\frac{K}{2}\rfloor,j+q-\lfloor\frac{K}{2}\rfloor}\}, \quad 0 \le p,q < K. \tag{3}$$

Outlook attention The outlook weight at location (i,j) is directly used as the attention weight for value aggregation, by reshaping it to $\hat{\mathbf{A}}_{i,j} \in \mathbb{R}^{K^2\times K^2}$, followed by a Softmax function. Thus, the value projection procedure can be written as

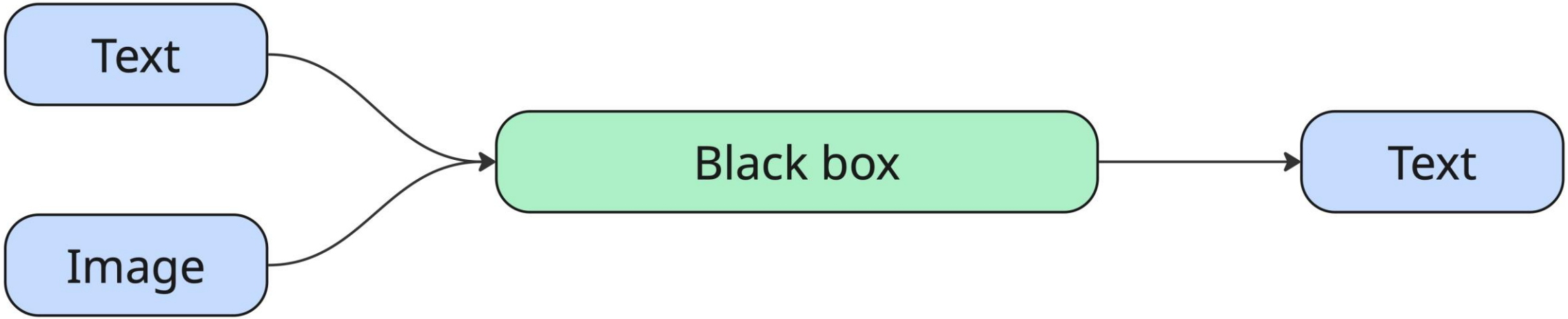
$$\mathbf{Y}_{\Delta_{i,j}} = \mathrm{MatMul}(\mathrm{Softmax}(\hat{\mathbf{A}}_{i,j}), \mathbf{V}_{\Delta_{i,j}}). \tag{4}$$

Dense aggregation Outlook attention aggregates the projected value representations densely. Summing up the different weighted values at the same location from different local windows yields the output

$$\tilde{\mathbf{Y}}_{i,j} = \sum_{0\le m,n<K} \mathbf{Y}^{i,j}_{\Delta_{i+m-\lfloor\frac{K}{2}\rfloor,j+n-\lfloor\frac{K}{2}\rfloor}}. \tag{5}$$

- Vision Transformers
- **VLMs**
- Benchmarks
- Architectures

Что такое VLM?



Зачем нужны мультимодальные модели

Опиши изображение кратко, менее, чем в 50 словах

В чём смысл мема?

Какой штраф грозит водителю за поворот направо после этого знака?

1. Найти сумму или разность:

1) а) $\frac{1}{4}+\frac{3}{8}$; б) $\frac{5}{8}-\frac{1}{16}$; в) $3\frac{3}{4}+6\frac{5}{12}$;

2) а) $\frac{2}{17}-\frac{1}{15}$; б) $\frac{7}{23}+\frac{2}{5}$; в) $3\frac{3}{11}-1\frac{2}{7}$;

3) а) $\frac{3}{44}-\frac{31}{41}$; б) $2\frac{1}{20}-5\frac{7}{80}$; в) $5\frac{1}{8}-11\frac{3}{16}$.

2. Найти значение выражения:

1) а) $8+4{,}61+7+16{,}29$; б) $53{,}7+9{,}43+6{,}89+0{,}08$;

2) а) $7{,}52-(2{,}38+4{,}641)$; б) $(7{,}36+5{,}62)-0{,}002$.

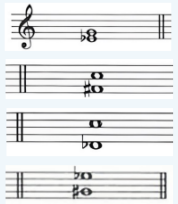
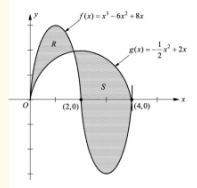
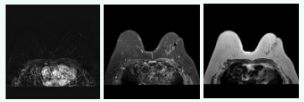
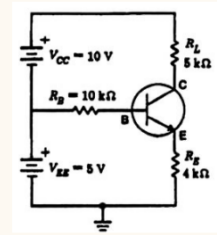
Реши пример 1.2 под пунктом В

- Vision Transformers
- VLMs
- **Benchmarks**
- Architectures

MMMU

11.5K meticulously collected multimodal questions from college exams, quizzes, and textbooks.

Охватывает 30 тем по всему миру 6 дисциплин, включая Искусство, Бизнес, здравоохранение и медицину, естественные науки, гуманитарные и социальные науки, технологии и инжиниринг, а также более 183 подотраслей.

Art & Design	Business	Science
Question: Among the following harmonic intervals, which one is constructed incorrectly? **Options:** (A) Major third <*image 1*> (B) Diminished fifth <*image 2*> **(C) Minor seventh <*image 3*>** (D) Diminished sixth <*image 4*> 	**Question:** ...The graph shown is compiled from data collected by Gallup <*image 1*>. Find the probability that the selected Emotional Health Index Score is between 80.5 and 82? **Options:** (A) 0 (B) 0.2142 **(C) 0.3571** (D) 0.5 	**Question:** <*image 1*> The region bounded by the graph as shown above. Choose an integral expression that can be used to find the area of R. **Options:** **(A)** $\int_0^{1.5}[f(x)-g(x)]dx$ (B) $\int_0^{1.5}[g(x)-f(x)]dx$ (C) $\int_0^{2}[f(x)-g(x)]dx$ (D) $\int_0^{2}[g(x)-x(x)]dx$ 
Subject: Music; **Subfield:** Music; **Image Type:** Sheet Music; **Difficulty:** Medium	**Subject:** Marketing; **Subfield:** Market Research; **Image Type:** Plots and Charts; **Difficulty:** Medium	**Subject:** Math; **Subfield:** Calculus; **Image Type:** Mathematical Notations; **Difficulty:** Easy
Health & Medicine	**Humanities & Social Science**	**Tech & Engineering**
Question: You are shown subtraction <*image 1*>, T2 weighted <*image 2*> and T1 weighted axial <*image 3*> from a screening breast MRI. What is the etiology of the finding in the left breast? **Options:** (A) Susceptibility artifact (B) Hematoma **(C) Fat necrosis** (D) Silicone granuloma 	**Question:** In the political cartoon, the United States is seen as fulfilling which of the following roles? <*image* 1> **Option**: (A) Oppressor (B) Imperialist **(C) Savior** (D) Isolationist 	**Question:** Find the VCE for the circuit shown in <*image 1*>. Neglect VBE **Answer: 3.75** **Explanation**: …IE = [(VEE) / (RE)] = [(5 V) / (4 k-ohm)] = 1.25 mA; VCE = VCC - IERL = 10 V - (1.25 mA) 5 k-ohm; VCE = 10 V - 6.25 V = 3.75 V 
Subject: Clinical Medicine; **Subfield:** Clinical Radiology; **Image Type:** Body Scans: MRI, CT.; **Difficulty:** Hard	**Subject:** History; **Subfield:** Modern History; **Image Type:** Comics and Cartoons; **Difficulty:** Easy	**Subject:** Electronics; **Subfield:** Analog electronics; **Image Type:** Diagrams; **Difficulty:** Hard

https://arxiv.org/pdf/2311.16502

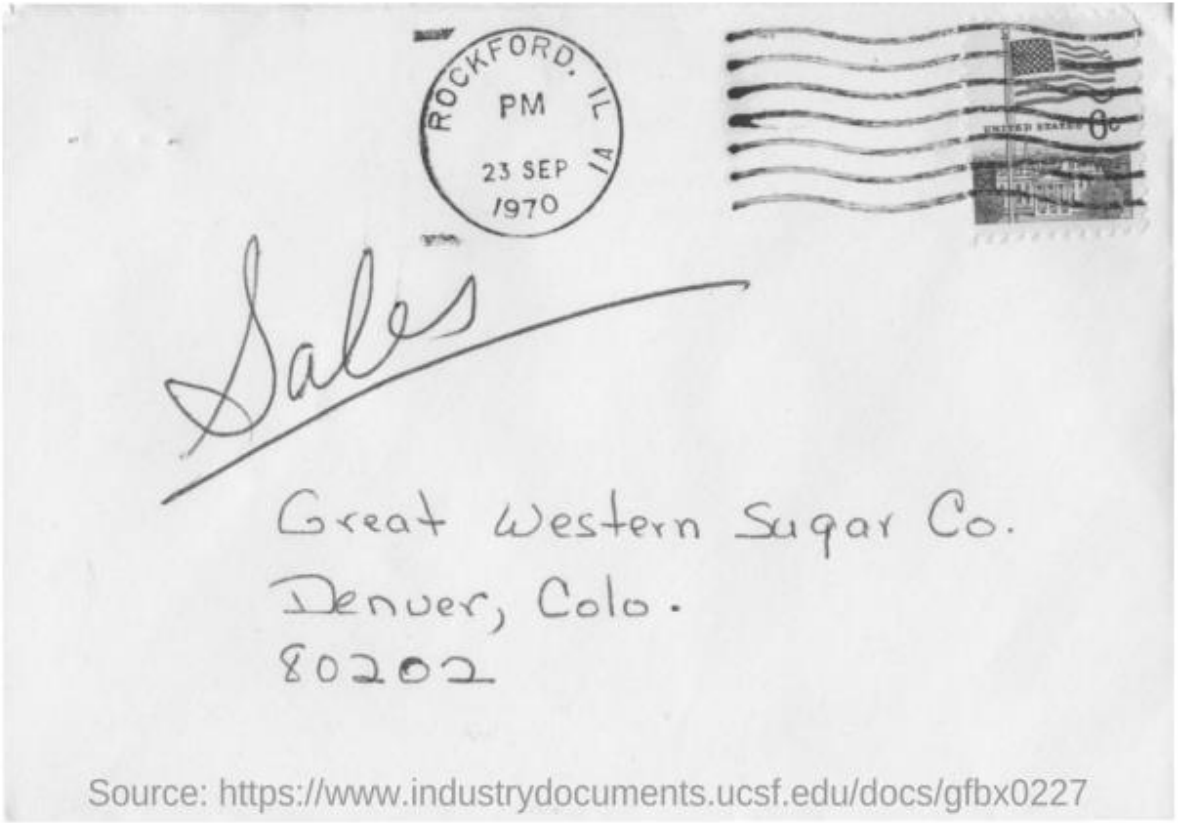
MMMU: A Massive Multi-discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI

27.11.2023

DocVQA

50.000 вопросов на 12.767 изображений.

Тестируют умение читать с документов. Берут документы с UCSF Industry Documents Library. Вопросы-ответы собирали с пользователей онлайн, если по каким-то причинам текст - выброс, его выкидывали.



Q: Mention the ZIP code written?
A: 80202

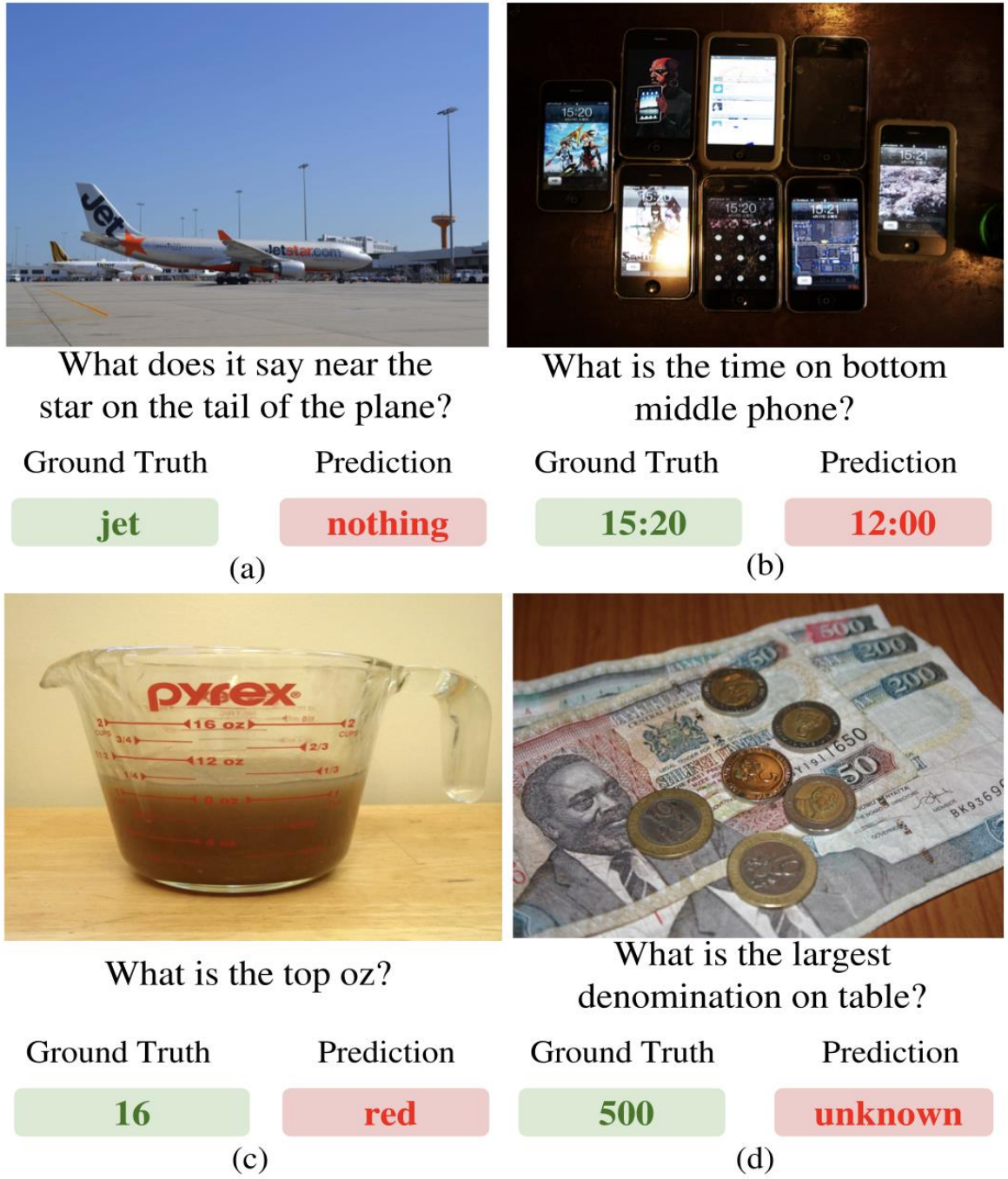
Q: What date is seen on the seal at the top of the letter?
A: 23 sep 1970

Q: Which company address is mentioned on the letter?
A: Great western sugar Co.

https://arxiv.org/pdf/2007.00398
DocVQA: A Dataset for VQA on Document Images
01.07.2020

TextVQA

45.336 вопросов на 28.408 изображений.

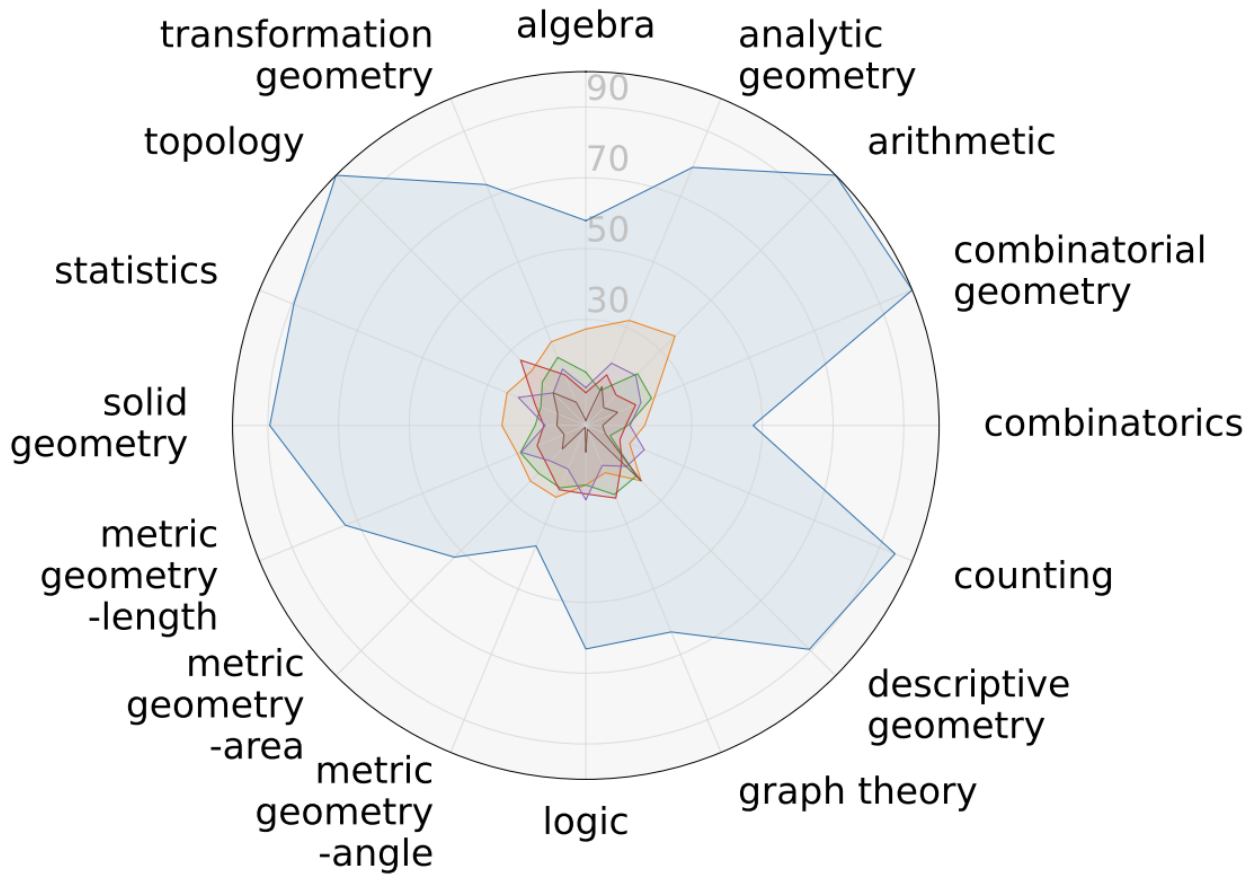


https://arxiv.org/pdf/2007.00398
Towards VQA Models That Can Read
18.04.2019

MathVision

3,040 задач с визуальным контекстом.

The dataset is well-balanced, featuring 1,532 problems in an open-ended format and 1,508 in a multiple-choice format. Данные брались с различных математических олимпиад, но не международного уровня. Убирались дубликаты с помощью подсчета расстояния Левинштейна. В сравнении с MathVista, тут вопросы в среднем длиннее (42.3 против 15.6) и больше категорий (18 против 7). Ну и само собой, что вопросы "новые" и собраны с реальных математических олимпиад. Также важно, что тут именно проверяются скиллы решения и визуального восприятия, а не просто знание каких-то сложных теорий (как в MMMU).

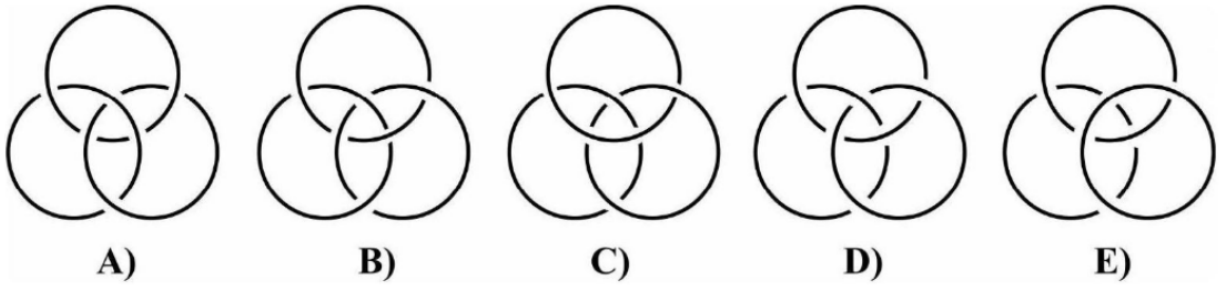


https://arxiv.org/pdf/2402.14804
Measuring Multimodal Mathematical Reasoning with MATH-Vision Dataset
22.02.2024

MathVision

▷ *Borromaic Rings*

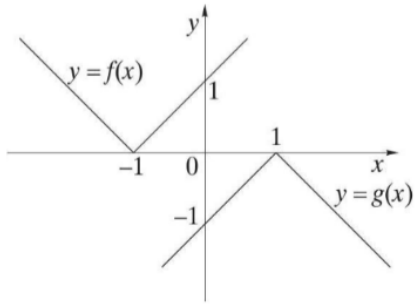
Image:



Question: The "Borromaic Rings" have an extraordinary property. Although no two are interlocked, they are strongly connected within each other. If one ring is cut through, the other two fall apart. Which of the following diagrams shows the picture of "Borromaic Rings"?

▷ *mutual symmetry of functions*

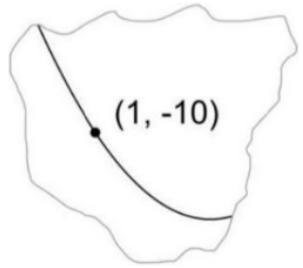
Image:



Question: The figure shows graphs of functions f and g defined on real numbers. Each graph consists of two perpendicular halflines. Which is satisfied for every real number x?
(A) $f(x) = -g(x) + 2$
(B) $f(x) = -g(x) - 2$
(C) $f(x) = -g(x + 2)$
(D) $f(x + 2) = -g(x)$
(E) $f(x + 1) = -g(x - 1)$

▷ *quadratic function discriminant*

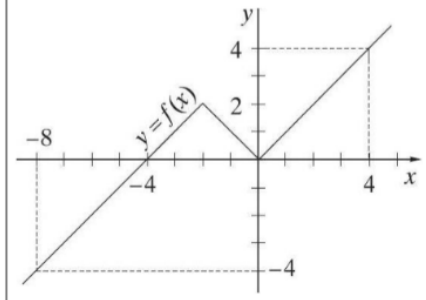
Image:



Question: In the (x,y)-plane the coordinate axes are positioned as usual. Point $A(1, -10)$ which is on the parabola $y = ax^2 + bx + c$ was marked. Afterwards the coordinate axis and the majority of the parabola were deleted. Which of the following statements could be false?
(A) $a > 0$ (B) $b < 0$ ...

▷ *find roots of iterative functions*

Image:



Question: The graph of the function $f(x)$, defined for all real numbers, is formed by two half-lines and one segment, as illustrated in the picture. Clearly, -8 is a solution of the equation $f(f(x)) = 0$, because $f(f(-8)) = f(-4) = 0$. Find all the solutions of the equation $f(f(f(x))) = 0$.

- Vision Transformers
- VLMs
- Benchmarks
- **Architectures**

CLIP

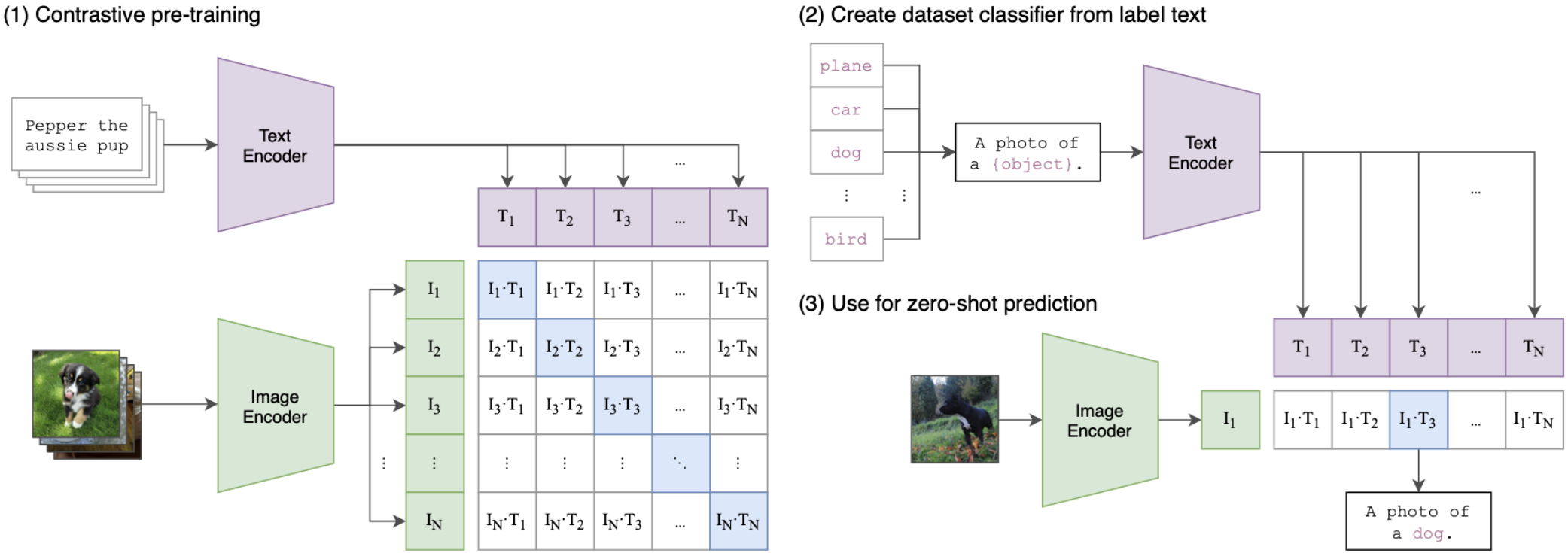


Figure 1. Summary of our approach. While standard image models jointly train an image feature extractor and a linear classifier to predict some label, CLIP jointly trains an image encoder and a text encoder to predict the correct pairings of a batch of (image, text) training examples. At test time the learned text encoder synthesizes a zero-shot linear classifier by embedding the names or descriptions of the target dataset's classes.

https://arxiv.org/pdf/2103.00020

Learning Transferable Visual Models From Natural Language Supervision 26.02.2021

SigLIP

CLIP loss

$$-\frac{1}{2|\mathcal{B}|}\sum_{i=1}^{|\mathcal{B}|}\left(\overbrace{\log\frac{e^{t\mathbf{x}_i\cdot\mathbf{y}_i}}{\sum_{j=1}^{|\mathcal{B}|}e^{t\mathbf{x}_i\cdot\mathbf{y}_j}}}^{\text{image}\to\text{text softmax}}+\overbrace{\log\frac{e^{t\mathbf{x}_i\cdot\mathbf{y}_i}}{\sum_{j=1}^{|\mathcal{B}|}e^{t\mathbf{x}_j\cdot\mathbf{y}_i}}}^{\text{text}\to\text{image softmax}}\right)$$

Sigmoid loss

$$-\frac{1}{|\mathcal{B}|}\sum_{i=1}^{|\mathcal{B}|}\sum_{j=1}^{|\mathcal{B}|}\underbrace{\log\frac{1}{1+e^{z_{ij}(-t\mathbf{x}_i\cdot\mathbf{y}_j+b)}}}_{\mathcal{L}_{ij}}$$

https://arxiv.org/pdf/2303.15343

Sigmoid Loss for Language Image Pre-Training

20.03.2023

Perception Encoder

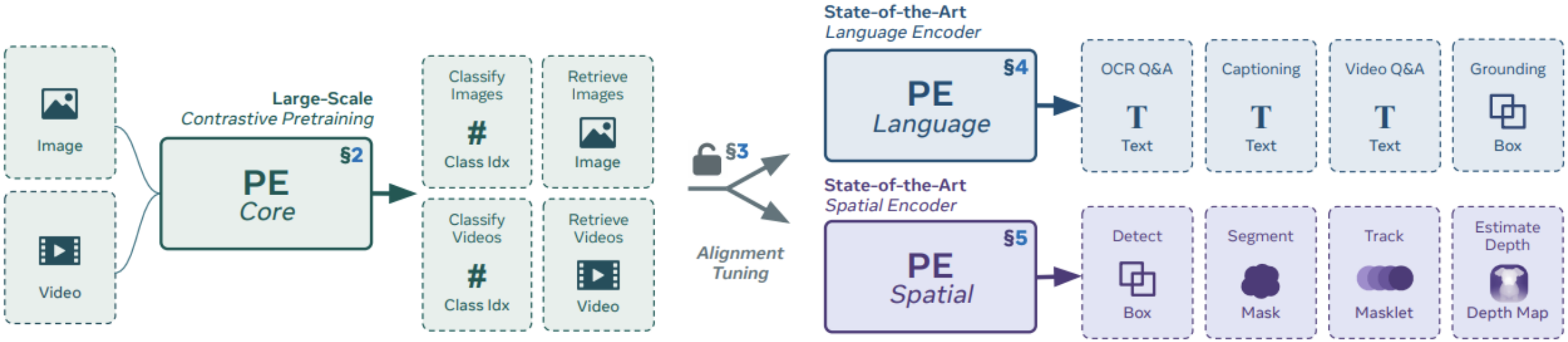


Figure 1 Perception Encoder (PE) is a family of large-scale vision encoder models with state-of-the-art performance on a large variety of vision tasks. By using a robust contrastive pretraining recipe and finetuning on synthetically aligned videos, PE not only outperforms all existing models on classification and retrieval (§2), but it also internally produces strong, general features that *scale* for downstream tasks (§3). PE unlocks the ability for large-scale contrastive pretraining to transfer to downstream tasks with alignment tuning to capitalize on those general features (§4, §5).

https://arxiv.org/pdf/2504.13181

Perception Encoder: The best visual embeddings are not at the output of the network

17.04.2025

LLaVA

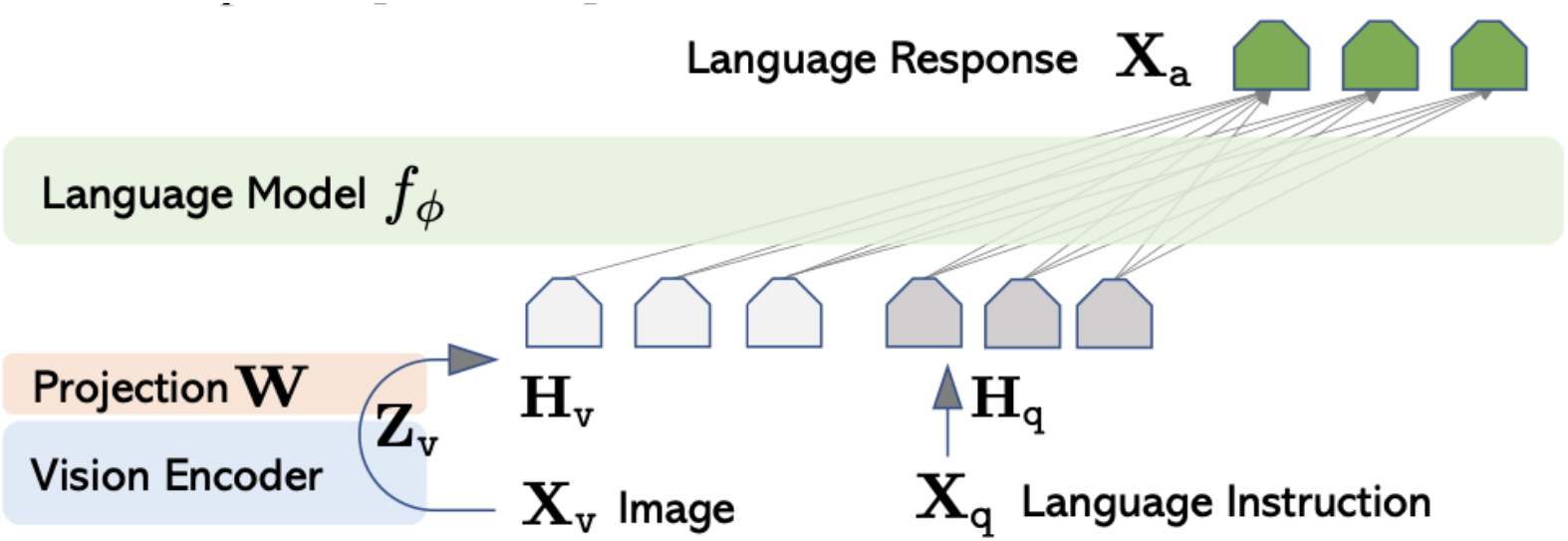


Figure 1: LLaVA network architecture.

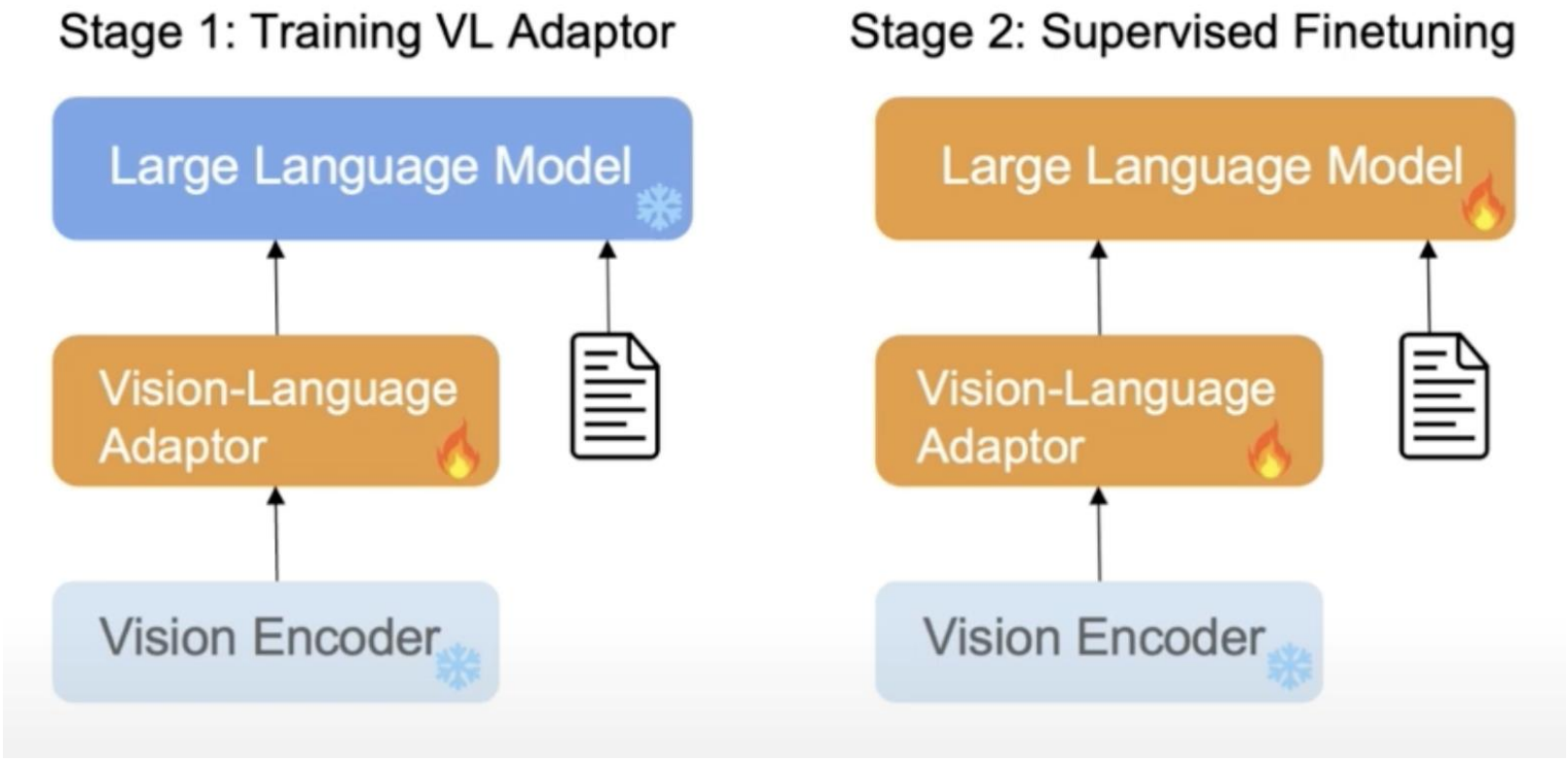
Архитектура модели
1)LLM
2)Vision Encoder
3)Adapter

https://arxiv.org/pdf/2304.08485 Visual Instruction Tuning 17.04.2023

Процесс обучения VLM

Стадии обучения:

1) Pretraining
2) Supervised Fine-tuning (SFT)



Pretrain

caption: На изображении представлена площадь Пьяцца Сан Доменико Маджоре в Неаполе, Италия. В центре площади находится обелиск, установленный на высоком постаменте. Обелиск украшен скульптурами и орнаментами. На заднем плане видно старинное здание, выполненное в классическом архитектурном стиле с множеством окон и балконов. Перед зданием расположены столики с зонтиками, под которыми сидят люди, что создаёт атмосферу кафе или ресторана.

Supervised Fine-tuning (SFT)

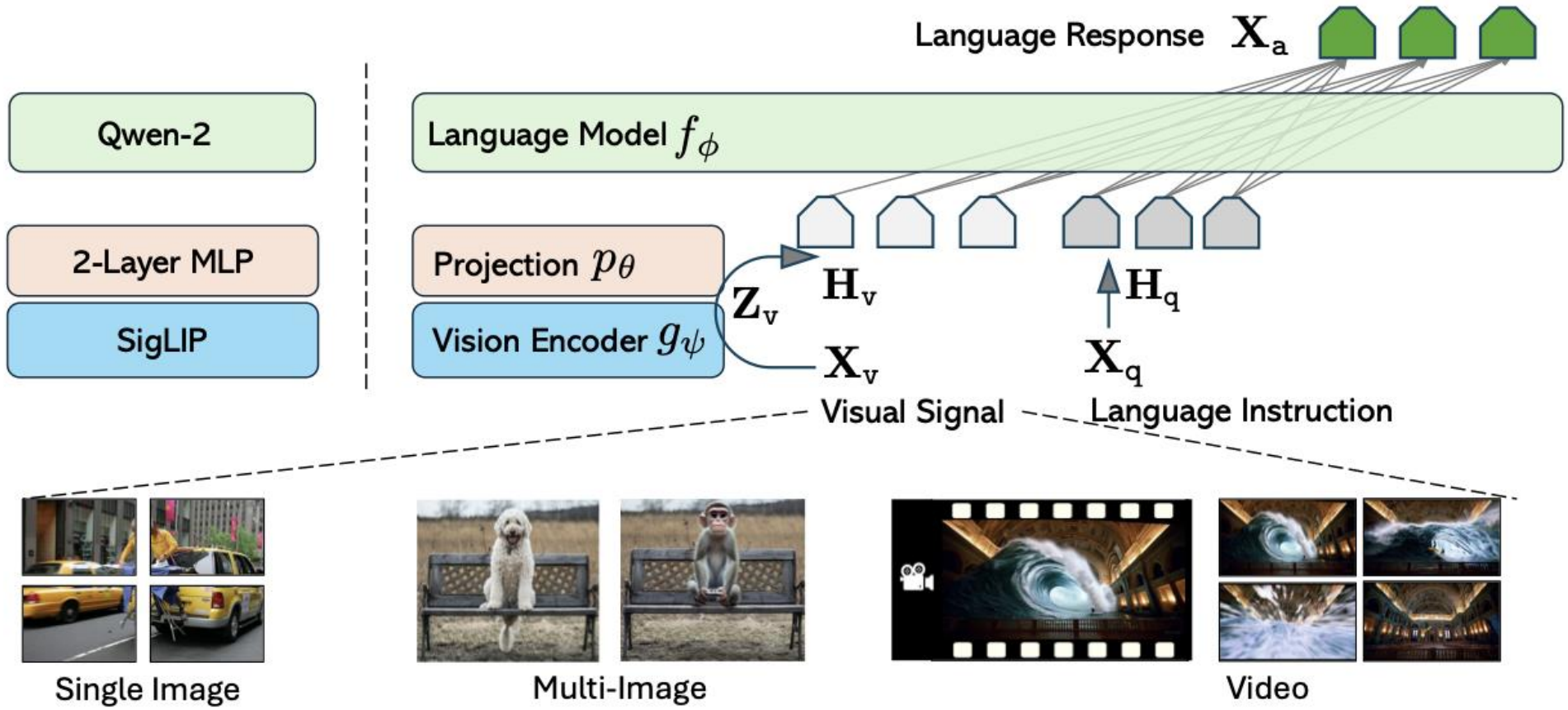
instruct: Какие выводы можно сделать о распределении населения в Доминиканской Республике, посмотрев на концентрацию дорог и скоростных автомагистралей на карте?

answer: Концентрация дорог и скоростных магистралей часто может быть показателем распределения населения, поскольку более густонаселенным районам обычно требуется больше инфраструктуры для удовлетворения транспортных потребностей.

На карте видна заметная концентрация дорог и скоростных трасс вокруг Санто-Доминго, столицы страны, что указывает на то, что это, вероятно, густонаселенный район. Кроме того, регионы Сантьяго и Ла-Вега в центральной части страны также имеют густую сеть дорог, что говорит о том, что это также густонаселенные регионы.

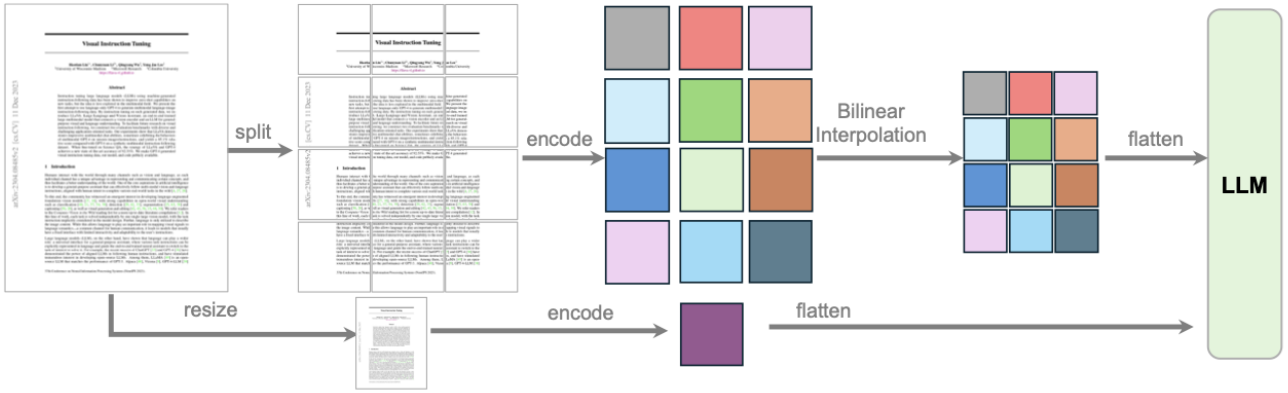
Напротив, в более горных районах, таких как центральный северный регион близ Монте-Кристи и восточная часть страны, меньше дорог, что означает меньшую плотность населения. Таким образом, можно сделать вывод, что население, вероятно, в большей степени сосредоточено в центральных и юго-восточных регионах Доминика, особенно вокруг крупных городов и столиц, соединенных основной транспортной инфраструктурой.

LLaVA One Vision

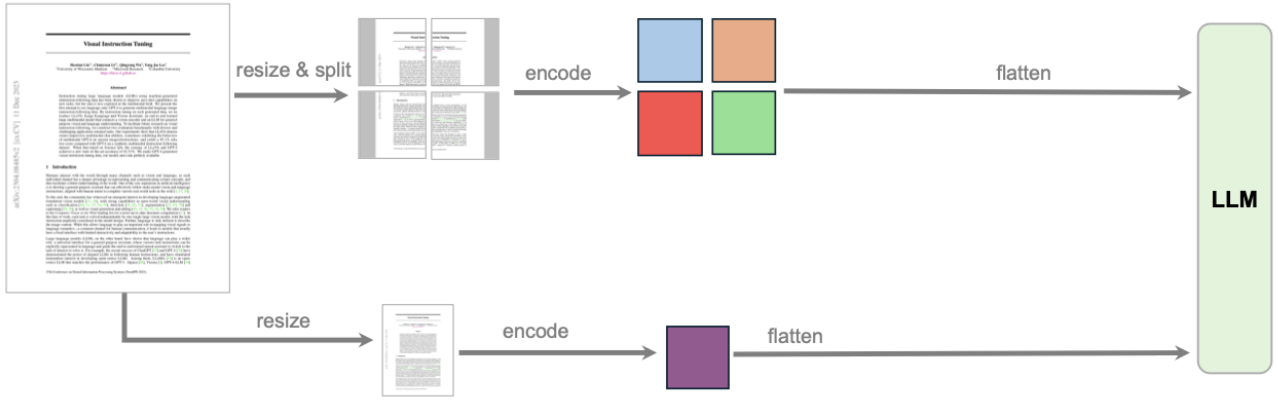


https://arxiv.org/pdf/2408.03326 LLaVA-OneVision: Easy Visual Task Transfer 06.08.2024

LLaVA One Vision



(a) Higher AnyRes with Bilinear Interpolation



(b) The original AnyRes

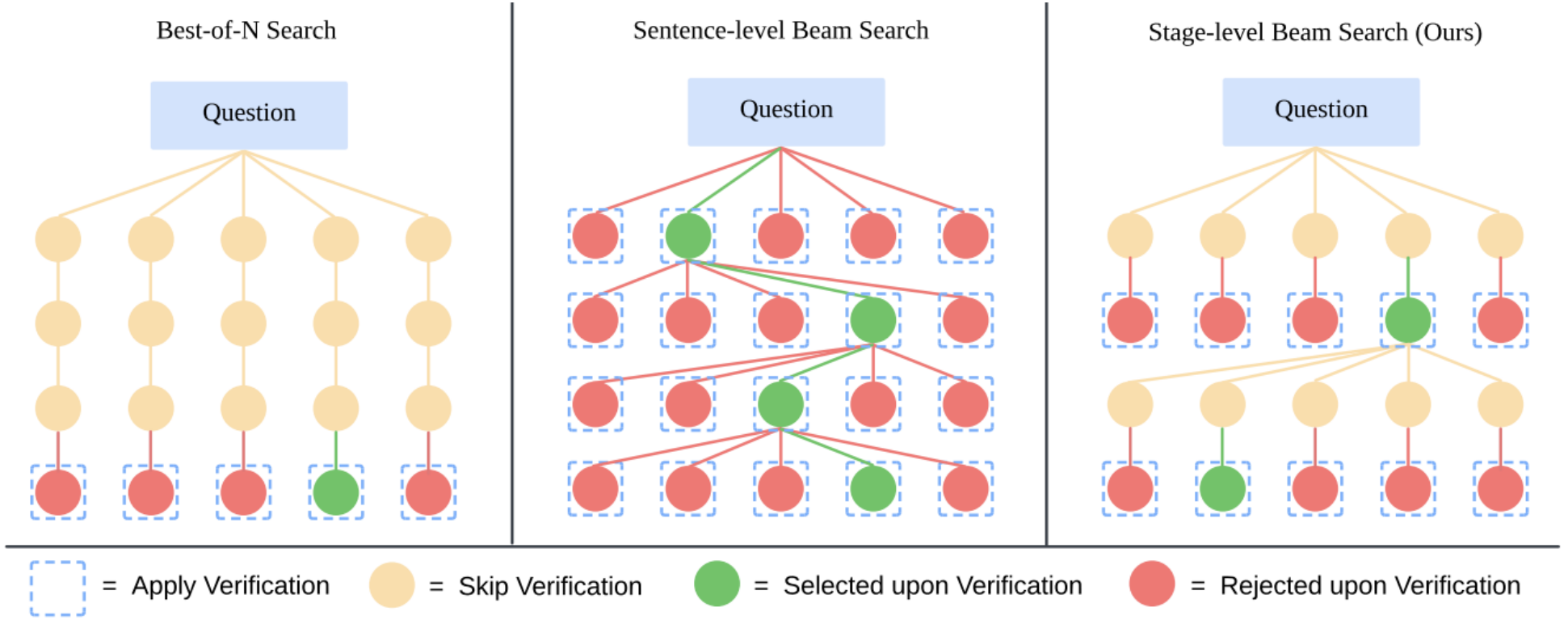
Single-Image		… N Crops	(1 + 9) * 729 = 7290 Tokens
	729 + N * 729 Tokens		
Multi-Image		… N Images	12 * 729 = 8748 Tokens
	N * 729 Tokens		
Video		… N Frames	32 * 196 = 6272 Tokens
	N * 196 Tokens		
	Example on Token Strategy		Max Tokens

LLaVA-CoT

https://arxiv.org/pdf/2408.03326

LLaVA-CoT: Let Vision Language Models Reason Step-by-Step

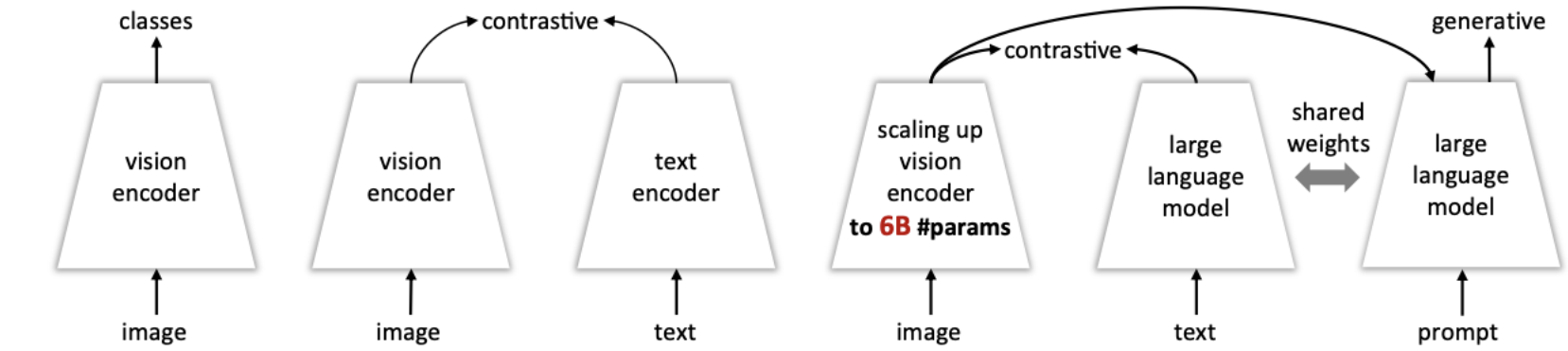
15.11.2024



- **Summary**: A brief outline in which the model summarizes the forthcoming task.
- **Caption**: A description of the relevant parts of an image (if present), focusing on elements related to the question.
- **Reasoning**: A detailed analysis in which the model systematically considers the question.
- **Conclusion**: A concise summary of the answer, providing a final response based on the preceding reasoning.

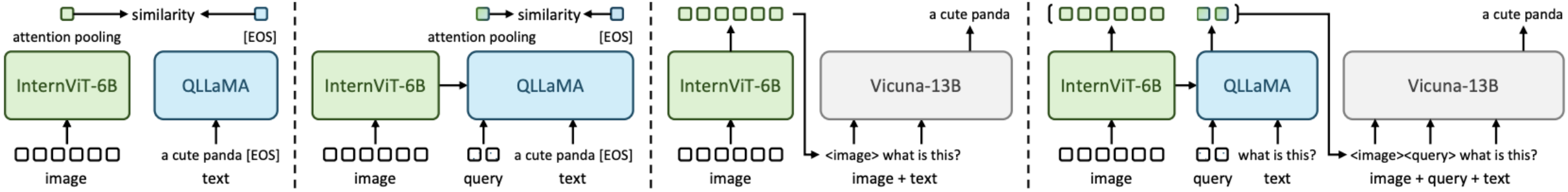
InternVL

https://arxiv.org/pdf/2312.14238
InternVL: Scaling up Vision Foundation Models and Aligning for Generic Visual-Linguistic Tasks
21.12.2023



(a) Supervised pre-training (b) Contrastive pre-training (c) InternVL: Scaling up vision encoder and aligning with LLM (ours)

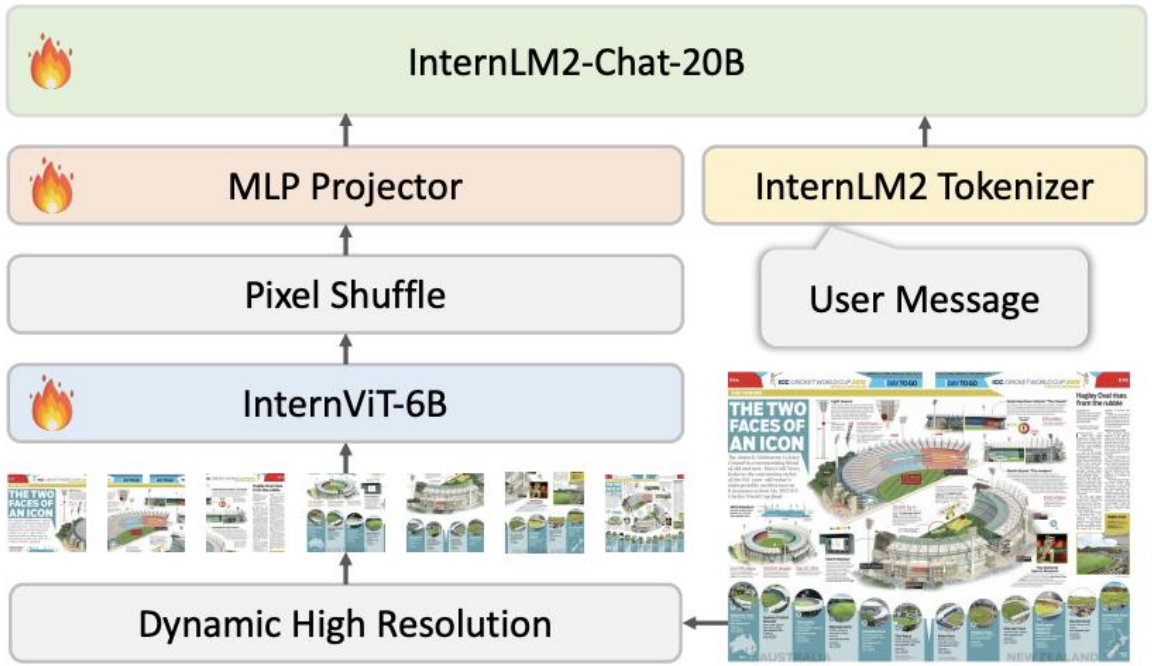
Figure 1. **Comparisons of different vision and vision-language foundation models.** (a) indicates the traditional vision foundation model, *e.g.* ResNet [57] pre-trained on classification tasks. (b) represents the vision-language foundation models, *e.g.* CLIP [117] pre-trained on image-text pairs. (c) is our InternVL, which presents a workable way to align the large-scale vision foundation model (*i.e.*, InternViT-6B) with the large language model and is versatile for both contrastive and generative tasks.

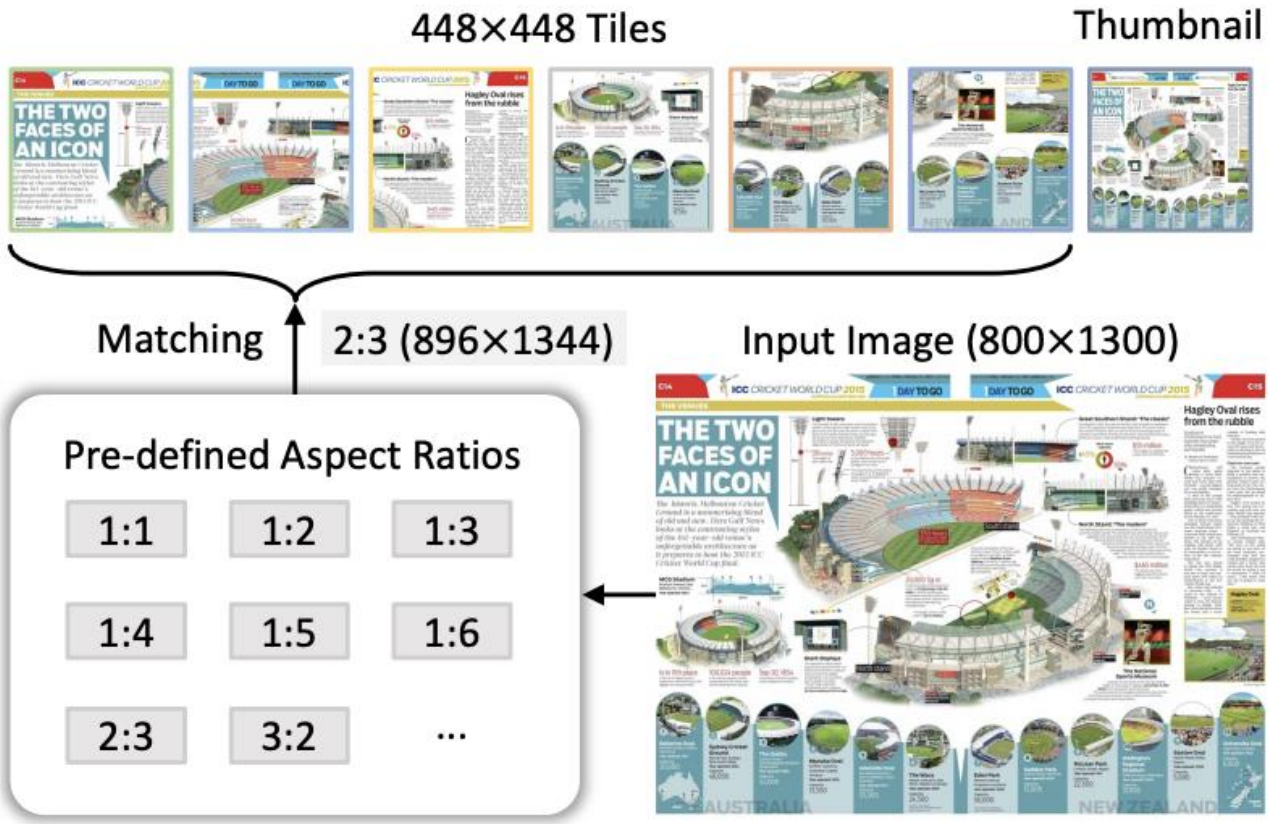


(a) InternVL-C (b) InternVL-G (c) InternVL-Chat (w/o QLLaMA) (d) InternVL-Chat (w/ QLLaMA)

Figure 4. **Different ways to use InternVL.** By flexibly combining the vision encoder and the language middleware, InternVL can support various vision-language tasks, including contrastive tasks, generative tasks, and multi-modal dialogue.

InternVL 1.5



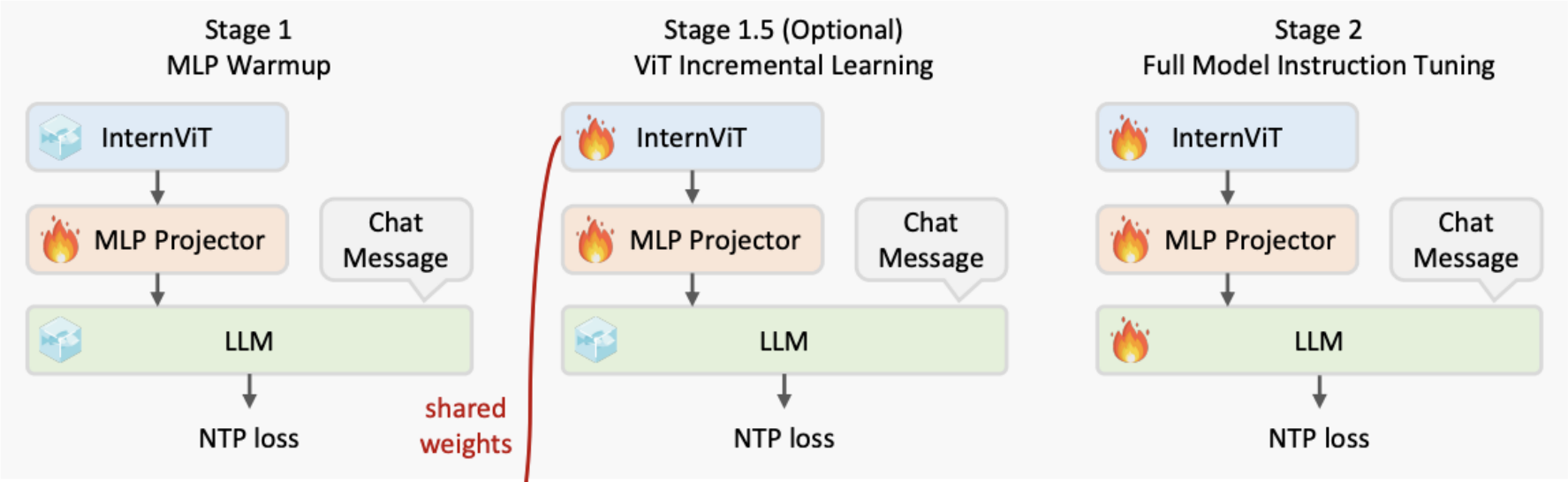


https://arxiv.org/pdf/2404.16821

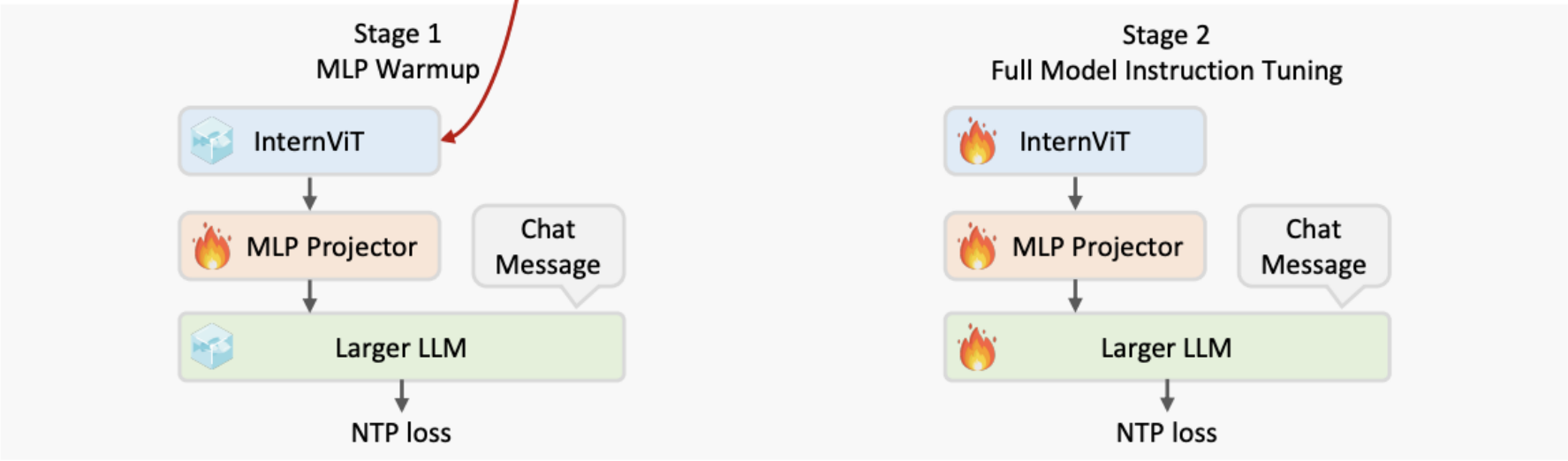
How Far Are We to GPT-4V? Closing the Gap to Commercial Multimodal Models with Open-Source Suites

25.04.2024

InternVL 2.5



(a) Single Model Training Pipeline



(b) Progressive Scaling Strategy

https://arxiv.org/pdf/2412.05271

Expanding Performance Boundaries of Open-Source Multimodal Models with Model, Data, and Test-Time Scaling

13.01.2025

InternVL 2.5

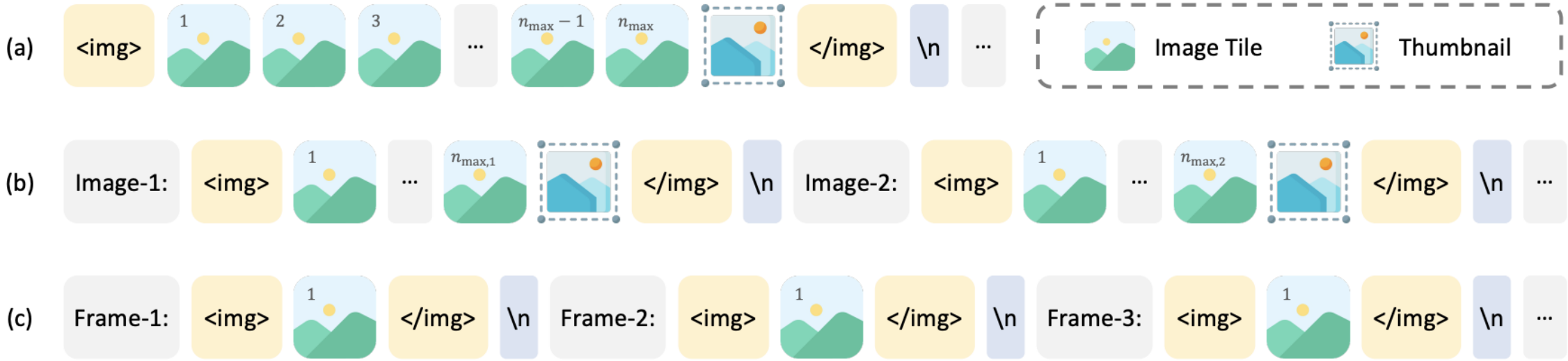


Figure 3: **Illustration of the data formats for various data types.** (a) For single-image datasets, the maximum number of tiles n_{max} is allocated to a single image, ensuring maximum resolution for the input. (b) For multi-image datasets, the total number of tiles n_{max} is distributed proportionally across all images within the sample. (c) For video datasets, the method simplifies the approach by setting $n_{\text{max}} = 1$, resizing individual frames to a fixed resolution of 448×448.

InternVL 3

	InternVL2.5 78B	InternVL3 8B	InternVL3 78B	Qwen2.5-VL 72B	Other Open-Source MLLMs	Claude-3.5 Sonnet	ChatGPT-4o-latest	Gemini-2.5 Pro
Model Weights	✓	✓	✓	✓	✓	✗	✗	✗
Training Data	✗	✓	✓	✗	-	✗	✗	✗
MMMU Multi-discipline	70.1%	65.6%	72.2% (2.1 ↑)	70.2%	64.5%	66.4%	72.9%	**74.7%**
MathVista Math	72.3%	75.2%	**79.6%** (7.3 ↑)	74.8%	70.5%	65.1%	71.6%	80.9%
AI2D Diagrams	89.1%	85.2%	**89.7%** (0.6 ↑)	88.7%	88.1%	81.2%	86.3%	89.5%
ChartQA Charts	88.3%	86.6%	89.7% (1.4 ↑)	89.5%	88.3%	**90.8%**	-	-
DocVQA Documents	95.1%	92.7%	95.4% (0.3 ↑)	96.4%	**96.5%**	95.2%	-	-
InfographicVQA infographics	84.1%	76.8%	85.2% (1.1 ↑)	**87.3%**	84.7%	74.3%	-	-
HallusionBench Hallucination	57.4%	49.9%	59.1% (1.7 ↑)	55.2%	58.1%	55.5%	57.0%	**64.1%**
OCRBench OCR	854	880	**906** (52↑)	885	877	-	894	862
LongVideoBench Video	63.6%	58.8%	**65.7%**(2.1↑)	60.7%	61.3%	-	-	-

Figure 1: **Multimodal performance of the InternVL series and other advanced MLLMs.** The InternVL series has consistently exhibited progressive enhancements in multimodal capabilities. The newly released InternVL3 significantly outperforms existing open-source MLLMs. Moreover, even in comparison with state-of-the-art closed-source commercial models, InternVL3 continues to demonstrate highly competitive performance.

https://arxiv.org/pdf/2504.10479

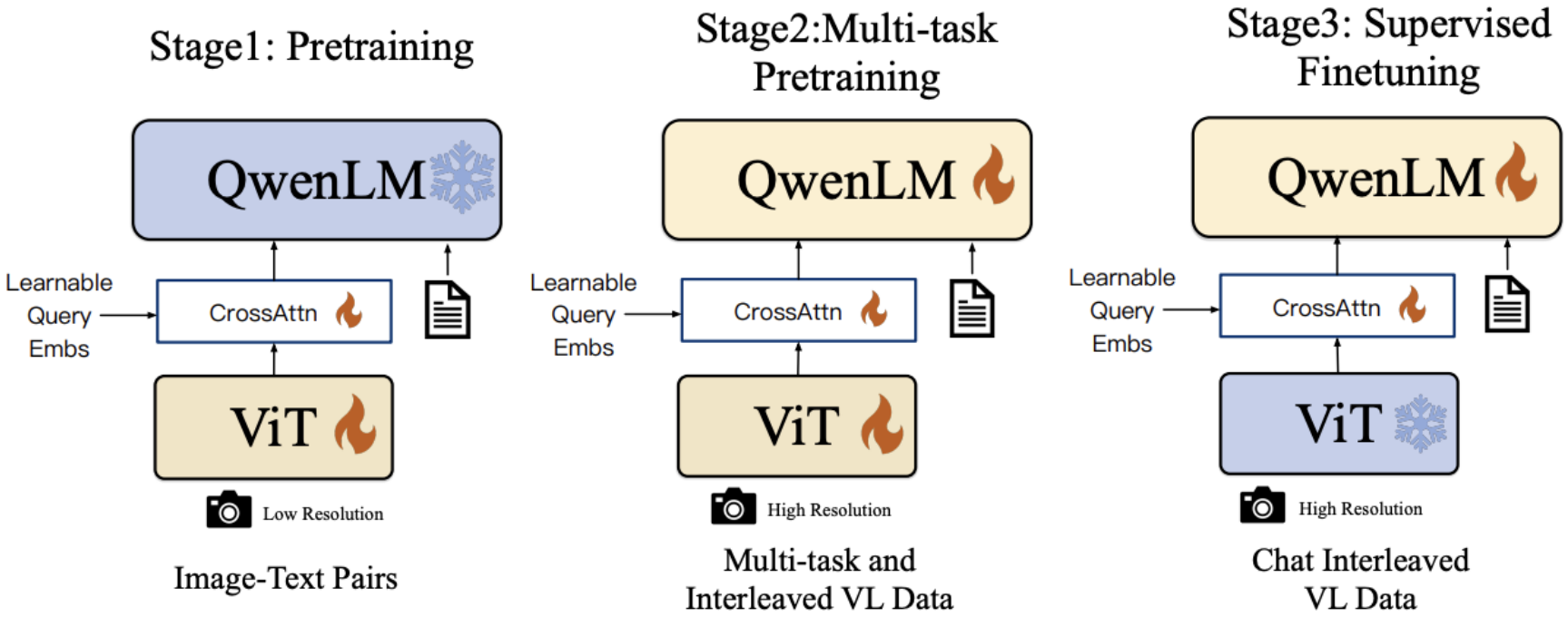
InternVL3: Exploring Advanced Training and Test-Time Recipes for Open-Source Multimodal Models

11.04.2025

Qwen-VL

Table 1: Details of Qwen-VL model parameters.

Vision Encoder	VL Adapter	LLM	Total
1.9B	0.08B	7.7B	9.6B



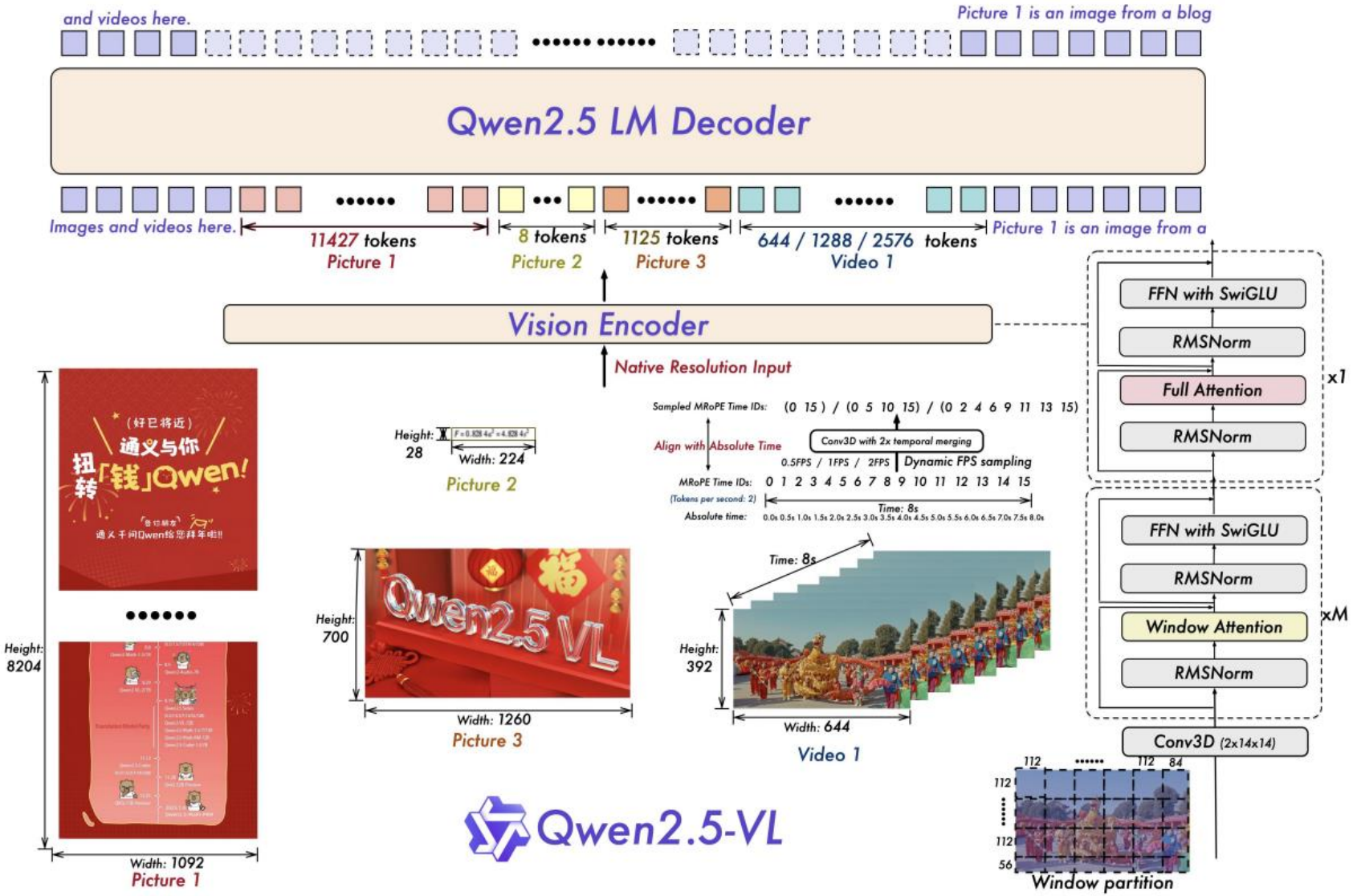
https://arxiv.org/pdf/2308.12966

Qwen-VL: A Versatile Vision-Language Model for Understanding, Localization, Text Reading, and Beyond

24.08.2023

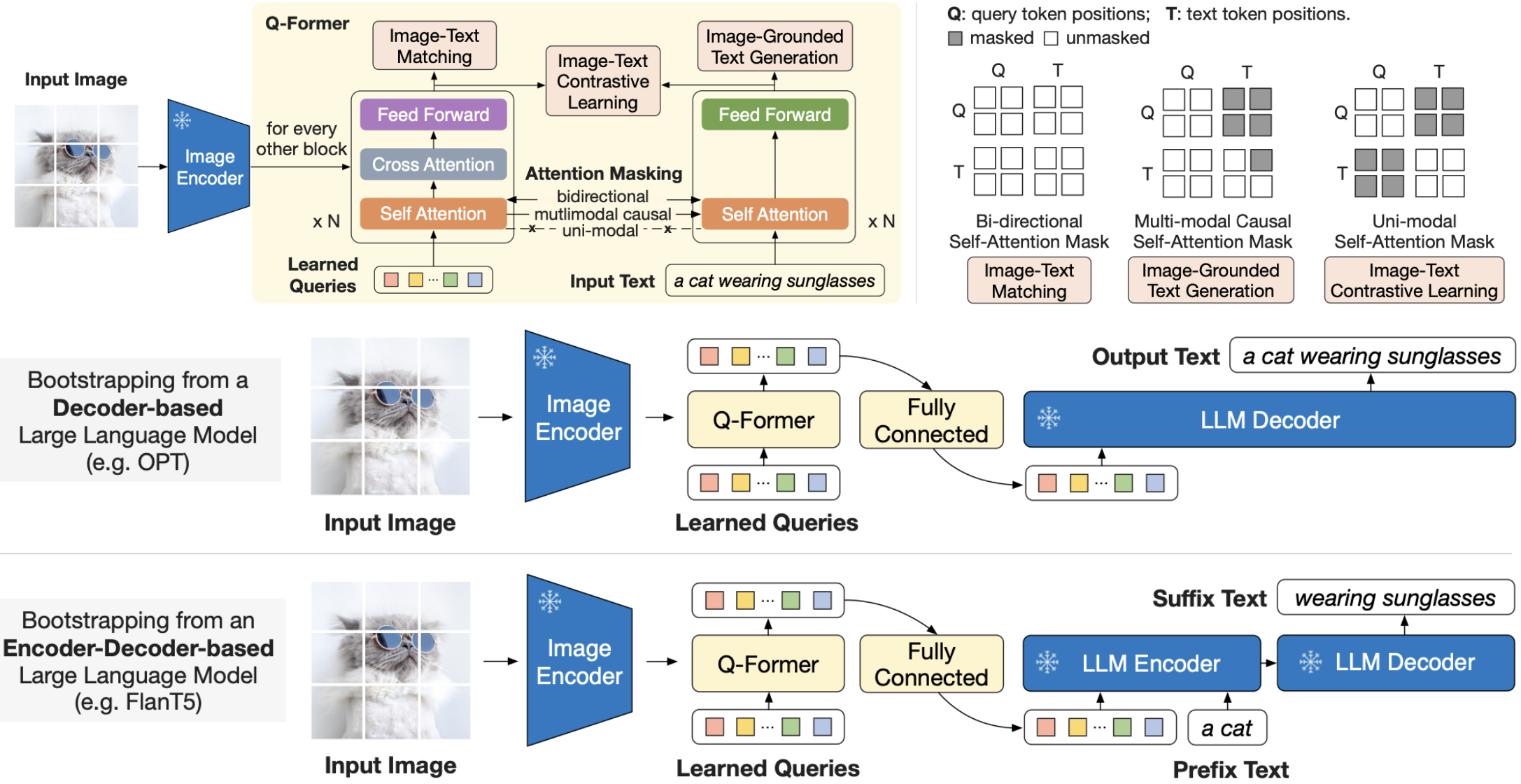
Qwen2.5-VL

- Window attention в визуальном энкодере, что снижает вычислительные затраты
- Динамический сэмплинг кадров при анализе видео, а при работе с видео два соседних кадра группируются вместе.
- Обновленный MRoPE во временном домене, выровненный по абсолютному времени
- Фильтрация данных для стадии претрейна и SFT, включая создание собственной синтетики, и interleaved text-image данных. Датасет претрейна расширен с **1.2T** токенов до **4.1T**.



https://arxiv.org/pdf/2502.13923
Qwen2.5-VL Technical Report
19.02.2025

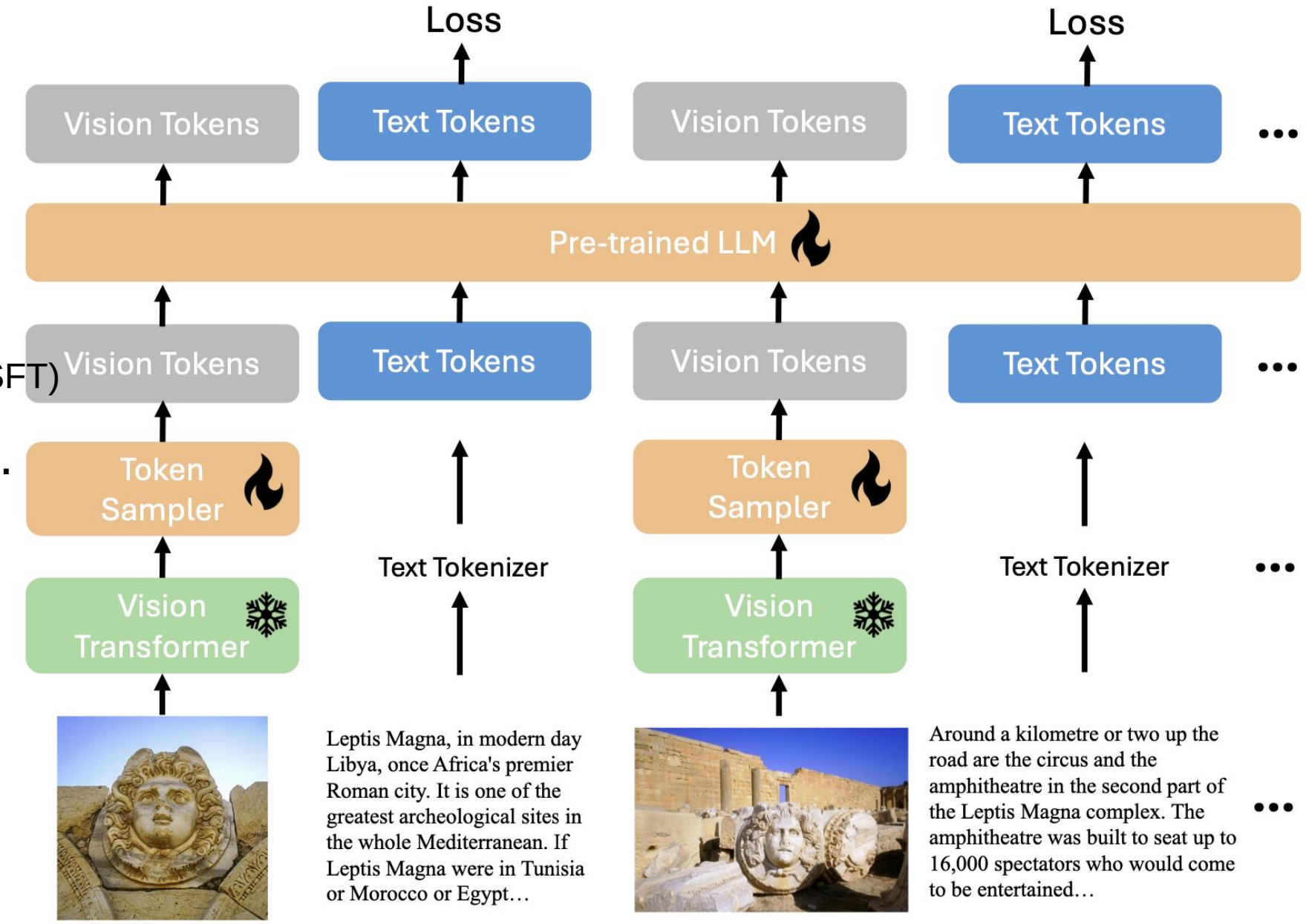
BLIP-2



https://arxiv.org/pdf/2301.12597 BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models 30.01.2023

BLIP-3

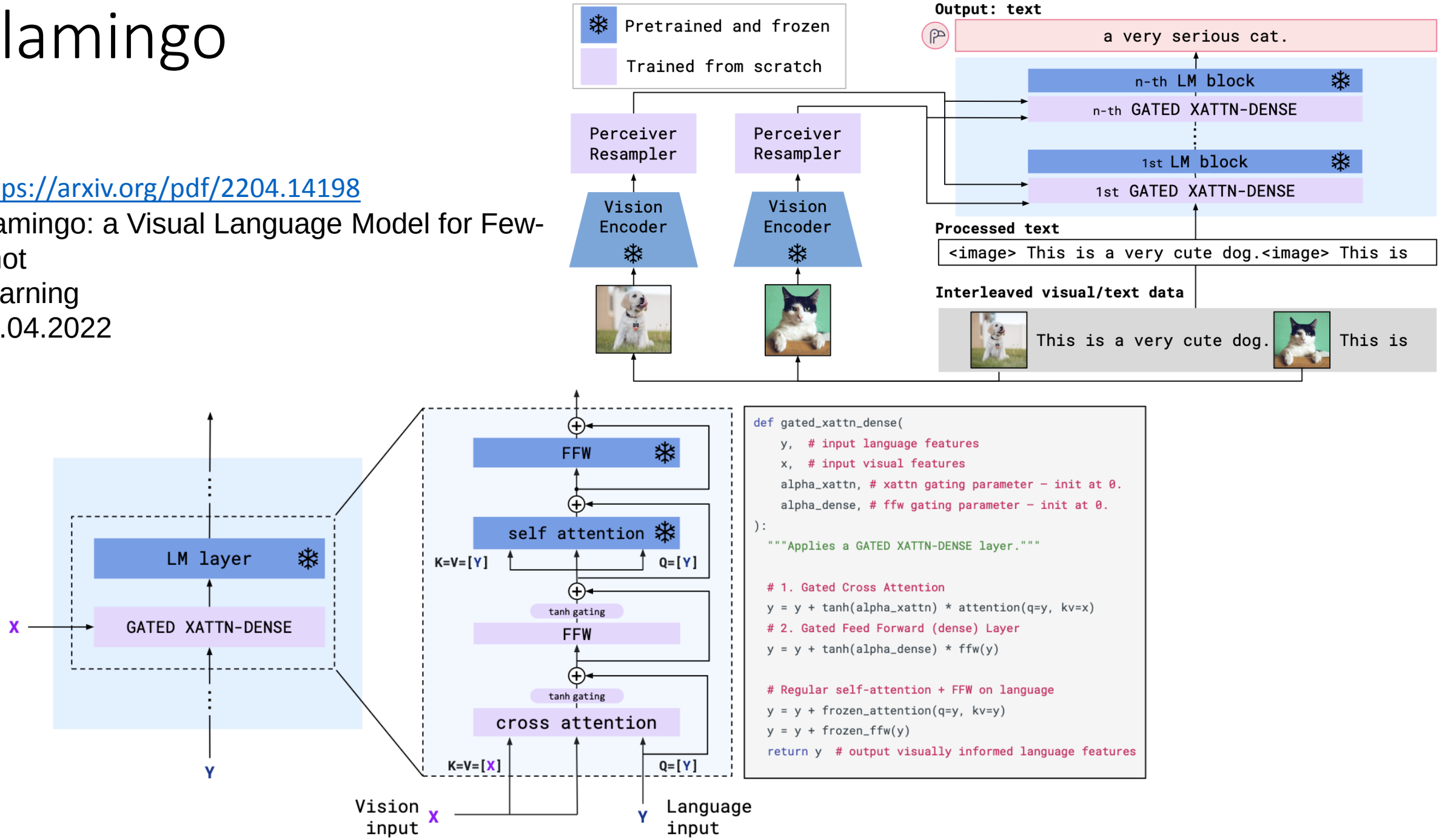
Ещё одна стадия обучения:
1) Pre-training
2) Supervised Fine-tuning (SFT)
3) Interleaved Multi-Image Supervised Fine-tuning.
4) Post-training.



https://arxiv.org/pdf/2408.08872 xGen-MM (BLIP-3): A Family of Open Large Multimodal Models 16.08.2024

Flamingo

https://arxiv.org/pdf/2204.14198
Flamingo: a Visual Language Model for Few-Shot
Learning
29.04.2022

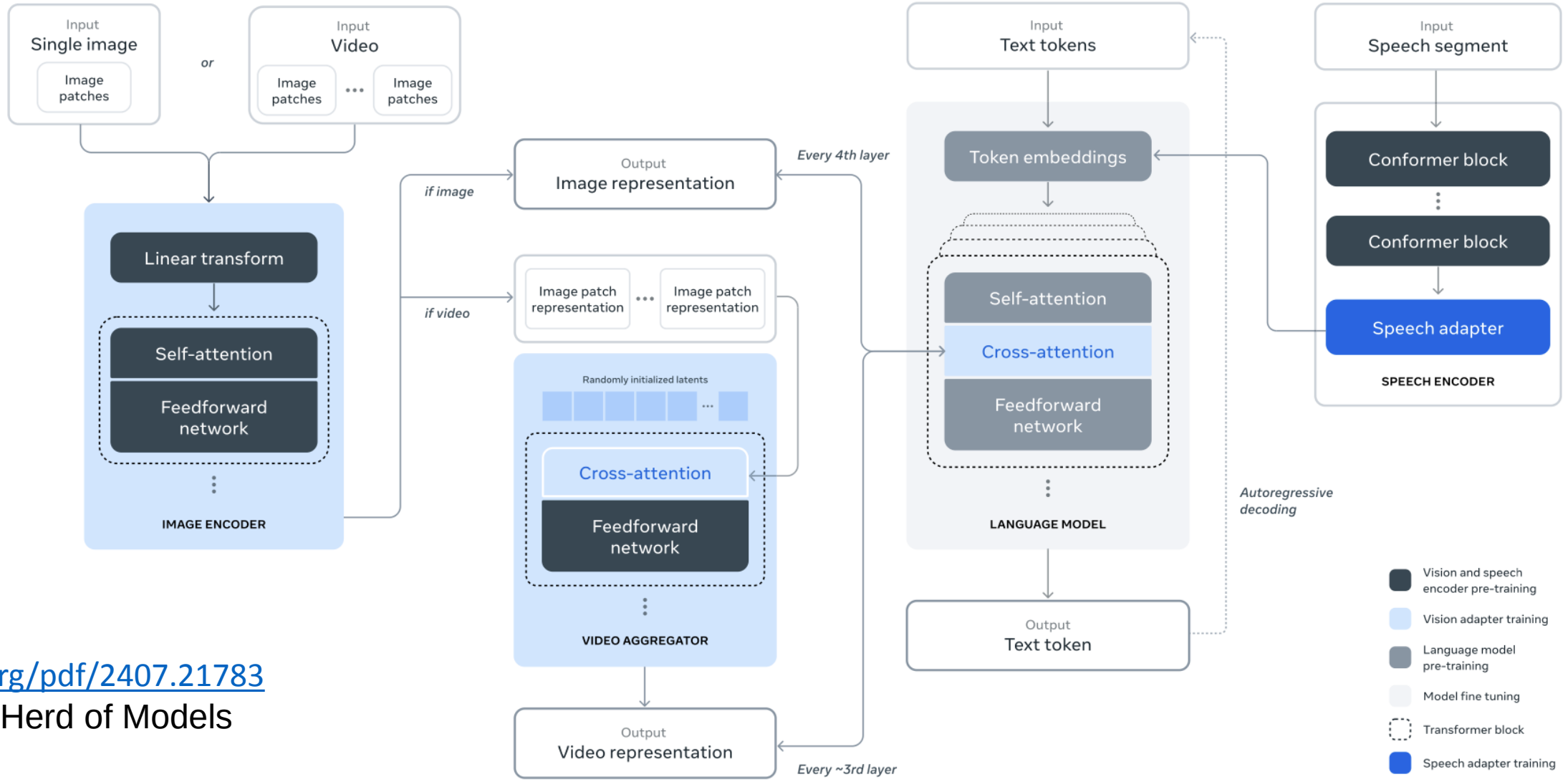



```
def gated_xattn_dense(
    y,  # input language features
    x,  # input visual features
    alpha_xattn, # xattn gating parameter – init at 0.
    alpha_dense, # ffw gating parameter – init at 0.
):
  """Applies a GATED XATTN-DENSE layer."""

  # 1. Gated Cross Attention
  y = y + tanh(alpha_xattn) * attention(q=y, kv=x)
  # 2. Gated Feed Forward (dense) Layer
  y = y + tanh(alpha_dense) * ffw(y)

  # Regular self-attention + FFW on language
  y = y + frozen_attention(q=y, kv=y)
  y = y + frozen_ffw(y)
  return y  # output visually informed language features
```

Llama 3-v

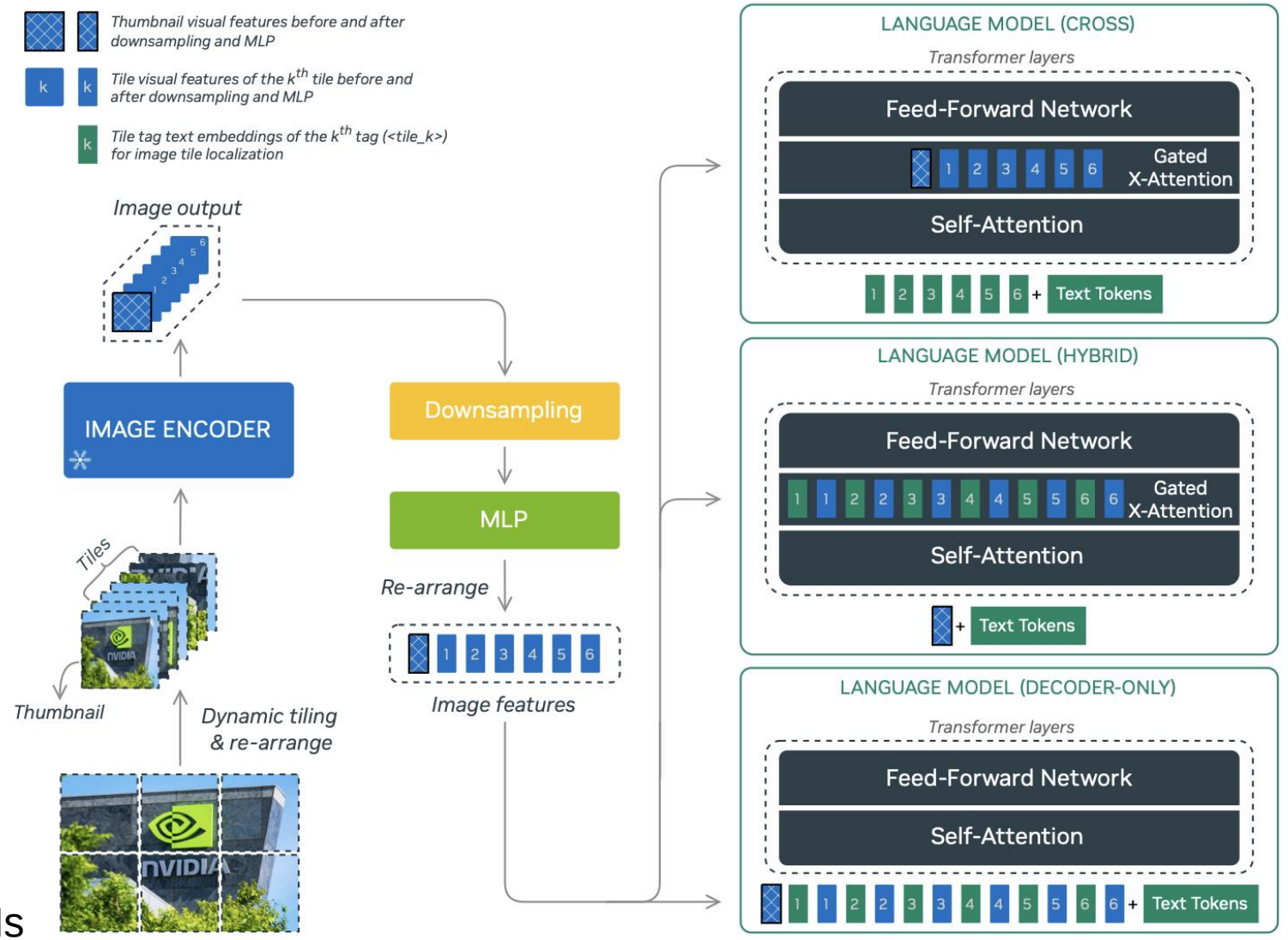


https://arxiv.org/pdf/2407.21783

The Llama 3 Herd of Models

31.07.2024

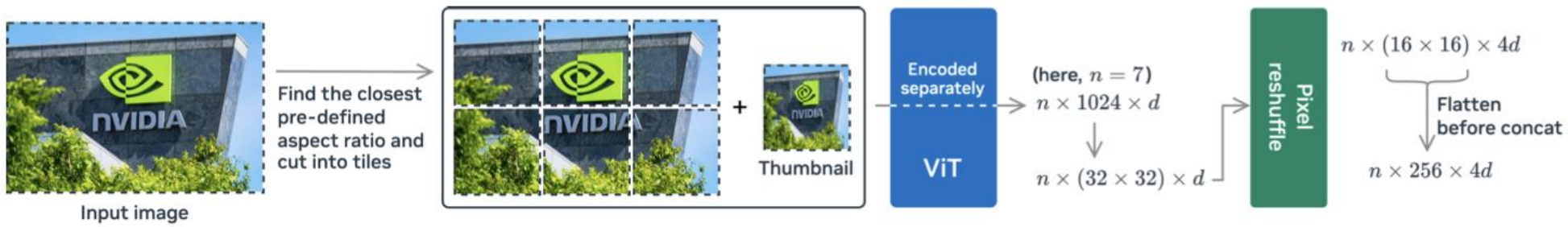
NVLM



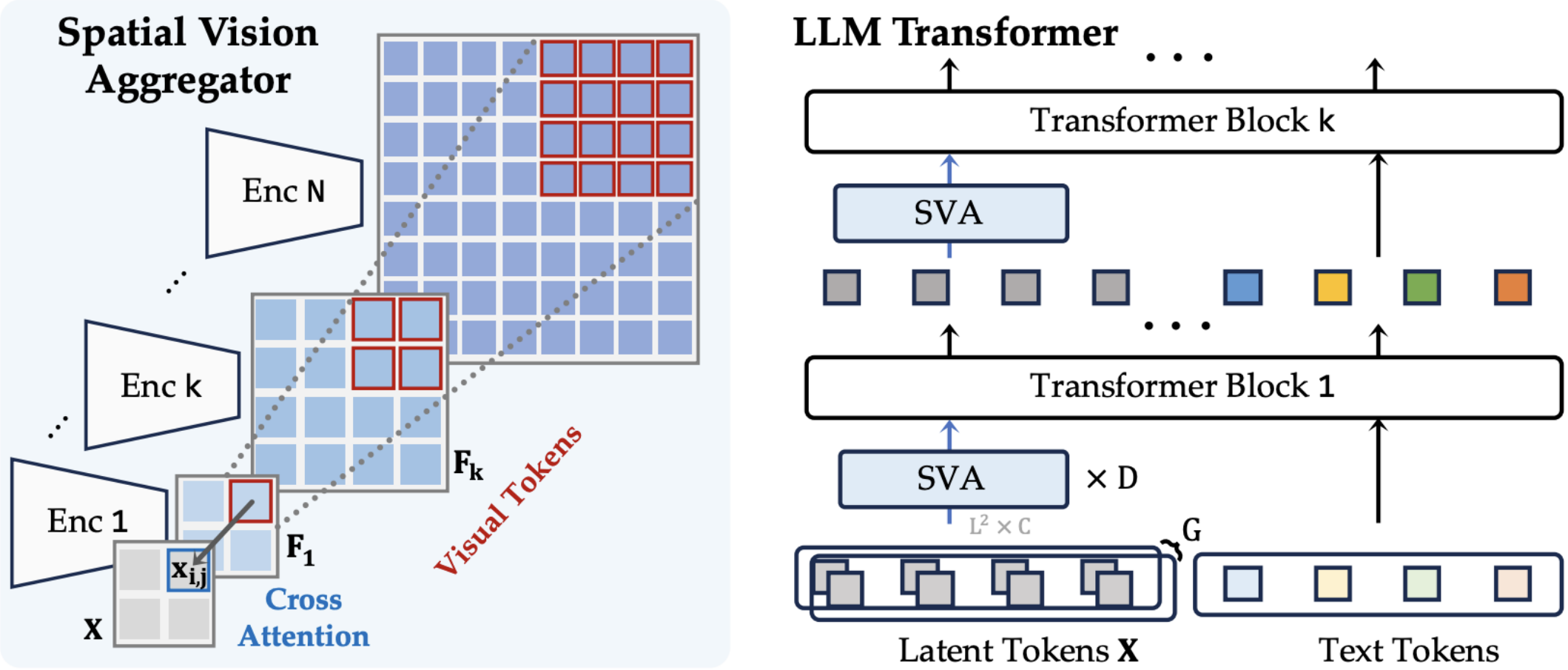
https://arxiv.org/pdf/2409.11402

NVLM: Open Frontier-Class Multimodal LLMs

17.09.2024



Cambrian



https://arxiv.org/pdf/2406.16860 Cambrian-1: A Fully Open, Vision-Centric Exploration of Multimodal LLMs 24.07.2024