DATA 8005 Advanced Natural Language Processing

Contrastive Language-Image Pre-training (CLIP)

Yi Zhang

Fall 2024

Outline

- A Brief Overview of Method
- Discussion 1: Start from an empirical Finding from Experimental Results
  - We observe that different Amount of Supervision (AoS) may lead to performance variance among evaluation datasets.
  - We give an empirical outlook to quantify AoS through an intrinsic score-matching nature in CLIP objective.
- Discussion 2: An close look into the CLIP objective: symmetric InfoNCE
  - A theoretical equivalence to (inversed) optimal transport objective
  - A theoretical understanding of softmax temperature
  - The necessity of "Symmetric" in a theoretical perspective

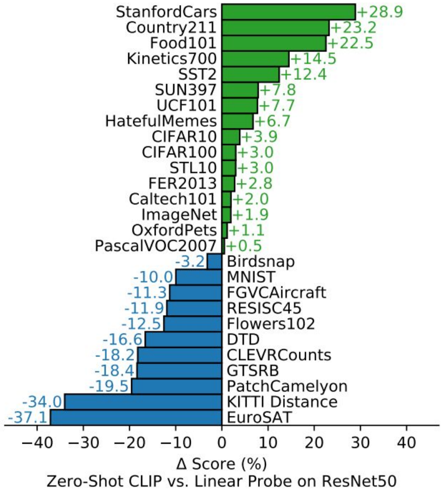
Method Overview

- Dataset Composition: (image, text) pairs
  - Search (image, text) pairs whose text includes one the query words occurring at least 100 times in English version of Wikipedia.
  - Filter the dataset:
    - 20,000 (images, text) pairs for each of the 50,000 query words.
    - This is to balance the supervision.
  - Using random crop augmentation.
- Image Encoder: ResNet, VIT, Text Encoder: Transformer
- Objective: Symmetric InfoNCE with learnable temperature.
  - It is targeted to align the embeddings of images and text.

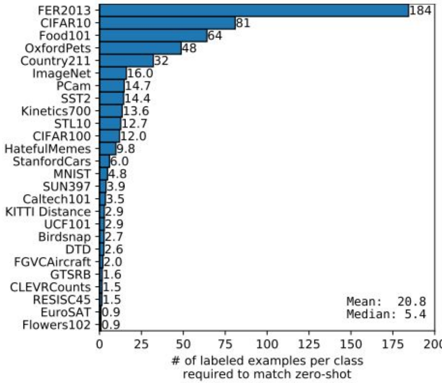
Discussion 1: A Finding in Experimental Results

- The performance of CLIP varies among evaluation datasets.
- The author suspects this is induced by the different Amount of Supervision (AoS).
- We ask a question: Is it possible to validate this congestion?

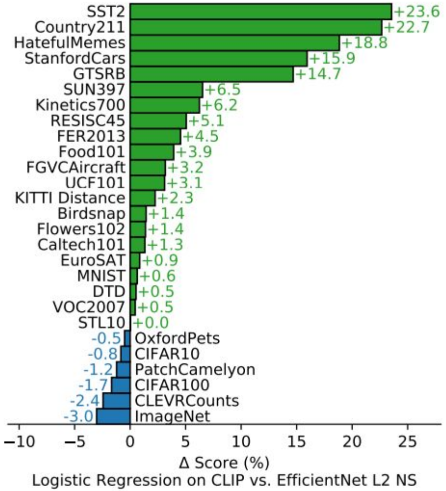
zero-shot CLIP v.s. linear probing a pretrained ResNet.



how many data are required for a few-shot CLIP to match zero-shot



Logistic regression on CLIP feature v.s. SOTA representation.



Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. (2021, July). Learning transferable visual models from natural language supervision. In *International conference on machine learning* (pp. 8748-8763). PMLR.

Score Match Nature induced by Polysemy

- To answer this question, we need some preliminaries on score matching:
  - Recall the training objective of a score-matching model $E_{x_t,\epsilon\sim N(0,I)}|s_\theta(x_t)-\epsilon|^2$
  - The interpretation is to require a model to predict all Gaussian noisy version of a data to its clear version.
  - Recall a property in score matching: a high loss of a data indicates that it is in high probability region, otherwise it is in low probability region.
    - This is because the network can easily denoise an isolated data, but it is hard to denoise a mixed one. (one noise refers to multiple clear ones.)
- In this discussion, we reason that CLIP has an intrinsic score matching nature due to the polysemy:
  - Polysemy here refers that one image can refer to infinite text.
  - Due the polysemy, we can assume each text embedding is a noisy version of an ideal text embedding, and the model is required to learn the expectation of the noisy embedding, which is the ideal text embedding.

Song, Y., & Ermon, S. (2019). Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, *32*.

Proof Sketch: Using a Minor Assumption on Data

- Recall the training objective of a score-matching model $E_{x_t,\epsilon\sim N(0,I)}|s_\theta(x_t)-\epsilon|^2$
- A more general score matching formula is $E_{x\sim D,\epsilon\sim N(0,I)}L(f_\theta(x+\epsilon),x)$, noticing ϵ is sampled independently.
- We rewrite the CLIP objective into a general formula: $E_{(x,y)\sim D}\mathrm{L}(\mathrm{f}_\theta(x),y)$
- Because of polysemy, we assume that the text embedding deviates a gaussian noise $N(0,\sigma I)$ from ideal one, we can rewrite the objective of CLIP into:$E_{(x,y'+\epsilon)\sim D}L(f_\theta(x),y'+\epsilon)$, where $\epsilon\sim N(0,\sigma I)$ dependent on y'.
- It is trivial to prove that when x is dense enough in dataset, we can find a ϵ small ball around $f_\theta(x)$, such that for all$f_\theta(x_i)\sim B_\epsilon\big(f_\theta(x)\big)=\{f_\theta(x'):|f_\theta(x')\ -f_\theta(x)|<\epsilon\}$, $\mathrm{y_i}\sim N(\mathrm{y_i'},\sigma I)$,where $y_i'=f'(x_i)$, f' is the ideal embedding function.
- Then we can rewrite the CLIP objective to $E_{x\sim D,y'=f'(x),\epsilon\sim N(0,\sigma I)}L(f_\theta(x),y'+\epsilon)$,the optimal is achieved when $f_\theta(x)=y'$, where shares the same solution of $E_{x\sim D,y'=f'(x),\epsilon\sim N(0,\sigma I)}L(f_\theta(x),y')$
- Since for all x_i, it is paired with a $y_i'+\epsilon$, by reparameterization, the CLIP objective can be reformulated to $E_{x\sim D,y'=f'(x),\epsilon\sim N(0,\sigma I)}L(g_\theta(\mathrm{y}'+\epsilon),y')$
- The objective falls in the general score-matching formula.

Score Matching Gives Estimation of AoS

- As we have shown the CLIP objective take a intrinsic score-matching, we can utilize a nice property of score matching.
- As reasoned before, in score matching perspective, a small loss of a training (image, text) indicates small AoS, while a large loss indicates the larger one.
- This might be a new interpretation to the Modality Gap phenomenon:
    - "modality gap" is a phenomenon that the trained text embedding and image embedding do not necessarily have a high cosine similarity.
        - Some work explain in Cone effect and local minimum
    - At the angle of score matching, modality gap is an empirical risk, which is a loss by taking expectation on the similarity between all possible noisy data and the clean one.
- Furthermore, nontrivially, one can also run Langevin Dynamics to sample the embedding distribution of training data, which gives quantitative measure of AoS.

Liang, V. W., Zhang, Y., Kwon, Y., Yeung, S., & Zou, J. Y. (2022). Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. *Advances in Neural Information Processing Systems*, *35*, 17612-17625.

Shi, P., Welle, M. C., Björkman, M., & Kragic, D. (2023). Towards understanding the modality gap in CLIP. In *ICLR 2023 Workshop on Multimodal Representation Learning: Perks and Pitfalls*.

Discussion 2: Symmetric InfoNCE

- The loss of CLIP is a combination of symmetric InfoNCEs.

$$\mathcal{L}_I = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{\exp(\mathrm{sim}(z_i^I, z_i^T)/\tau)}{\sum_{j=1}^{N} \exp(\mathrm{sim}(z_j^I, z_j^T)/\tau)}. \qquad \min_{\boldsymbol{\Theta}} \mathcal{L}_{CLIP} = \frac{1}{2}(\mathcal{L}_I + \mathcal{L}_T).$$

- Here, z^I and z^T means the image and text embedding respectively. /tau is the softmax temperature.
- Without loss of generality, we analyze L_I:
    - Q1: Why using log likelihood, and softmax?
    - Q2: Why softmax temperature is initialized by a very small number in CLIP.
    - Q3: Why symmetric? How about just using L_I.

Preliminary: Optimal Transport Objective

To answer these questions, we introduce the objective of Optimal Transport:

- Given two marginal distributions and their transport cost, what is the transport plan with a minimum cost?
- An inverse one: Given a transport plan and marginals, what is the mini cost?

To solve the inverse Optimal Transport, we have A bilevel minimization Objective:

$$\min_{\theta} KL(\tilde{\mathbf{P}}|\mathbf{P}^{\theta})$$

$$\text{where} \quad \mathbf{P}^{\theta} = \arg\min_{\mathbf{P}\in U} < \mathbf{C}^{\theta}, \mathbf{P} > -\epsilon H(\mathbf{P}).$$

Where P tilde here refers to the observed transport plans, P theta refers to the parameterized plan guided by a learned transport cost with a marginal constraint.

Q1: The Equivalence Between InfoNCE and IOT

- The proof is trivial.
- The derivation only constraints transport plan to one marginal distribution, i.e., image marginal distribution.
- We observe that
  - Softmax serves as a regularization in OT.
  - Log likelihood comes from parts of outer KL divergence.
  - One more question: Why KL?
    - An interesting direction to dig.

B. Lagrangian for Regularized OT

Our proof is as follows, which is technically akin to (Cuturi, 2013) in terms of using Lagrangian duals for Regularized OT, which in fact has been well adopted in OT literature.

B.1. Lagrangian under $U(a)$

Now we show the point set matching framework for CL with the simplified constraints:

$$U(\mathbf{a}) = \{\mathbf{P} \in \mathbb{R}_+^{n\times m} | \mathbf{P}\mathbf{1}_m = \mathbf{a}\} \tag{27}$$

Should be n here

where $\mathbf{a} = \mathbf{1}/m$ and $\mathbf{1}_m$ is the m-dimensional column vector whose elements are all ones. With the objective of the regularized OT:

$$\mathbf{P}^\theta = \arg \min_{P\in U(\mathbf{a})} < \mathbf{C}^\theta, \mathbf{P} > -\epsilon H(\mathbf{P}), \tag{28}$$

We introduce the dual variable $\mathbf{f} \in R^n$. The Lagrangian of the above equation is:

$$L(\mathbf{P}, \mathbf{f}) =< \mathbf{C}^\theta, \mathbf{P} > -\epsilon H(\mathbf{P}) - \sum_{i=1}^{n} \mathbf{f}_i \cdot \left(\sum_{j=1}^{m} \mathbf{P}_{ij} - \frac{1}{n} \right) \tag{29}$$

The first order conditions then yield by:

$$\frac{\partial L(\mathbf{P}, \mathbf{f})}{\partial \mathbf{P}_{ij}} = \mathbf{C}_{ij}^\theta + \epsilon \log \mathbf{P}_{ij} - \mathbf{f}_i = 0 \tag{30}$$

Thus we have $\mathbf{P}_{ij} = e^{(\mathbf{f}_i - C_{ij}^\theta)/\epsilon}$ for every i and j, for optimal $\mathbf{P}$ coupling to the regularized problem. Due to $\sum_j \mathbf{P}_{ij} = 1/n$ for every i, we can calculate the Lagrangian parameter $\mathbf{f}_i$ and the solution of the coupling is given by:

$$\mathbf{P}_{ij} = \frac{\exp\left(-\mathbf{C}_{ij}^\theta/\epsilon\right)}{n \sum_{t=1}^{m} \exp\left(-\mathbf{C}_{it}^\theta/\epsilon\right)} \tag{31}$$

Then in outer minimization, if we set $\tilde{P}_{ii} = \frac{1}{n}$ for each i and $\tilde{P}_{ij} = 0$ when $i \neq j$, we get the contrastive loss under $U(\mathbf{a})$

$$\mathcal{L}_{\text{IOT-CL}} = -\frac{1}{n} \sum_{i=1}^{n} \log \left(\frac{\exp(-\mathbf{C}_{ii}^\theta/\epsilon)}{\sum_{j=1}^{m} \exp(-\mathbf{C}_{ij}^\theta/\epsilon))} \right) + \text{Constant} \tag{32}$$

We have therefore got the loss of IOT-CL under $U(\mathbf{a})$.

Shi, L., Zhang, G., Zhen, H., Fan, J., & Yan, J. (2023, July). Understanding and generalizing contrastive learning from the inverse optimal transport perspective. In *International conference on machine learning* (pp. 31408-31421). PMLR.

Q2: Softmax only Benefits Optimization

- We declare that Softmax regularization should be small but not zero.
- From Q1, Softmax Temperature is exactly the entropy regularization coefficient.
- Ideally, there is no need for a regularization in contrastive learning,
    - The matching should be as sharp as possible to ensure discriminess.
- But we cannot cancel this regularization, since it is useful in optimization
    - Negative \epsilon H(x) Ensuring \epsilon-strong convexity.

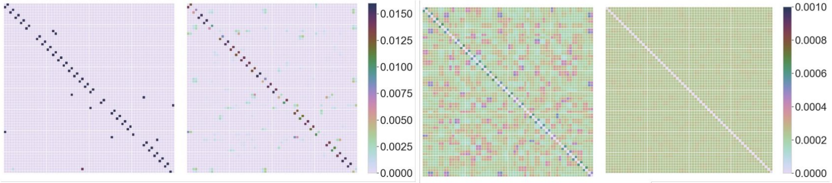


Figure 3. Results of couplings $\mathbf{P}^\theta$ by varying ϵ given 64 trained features on CIFAR-10 based on the SimCLR framework (Chen et al., 2020). When $\epsilon \to 0$, $\mathbf{P}^\theta$ becomes sharper for probability prediction. With the increment of ϵ, $\mathbf{P}^\theta$ becomes more uniform and when $\epsilon \to +\infty$, $\mathbf{P}^\theta$ approximates to a uniform distribution, which has nothing to do with the quality of the learned features.

Shi, L., Zhang, G., Zhen, H., Fan, J., & Yan, J. (2023, July). Understanding and generalizing contrastive learning from the inverse optimal transport perspective. In *International conference on machine learning* (pp. 31408-31421). PMLR.

Q3: The Necessity of Symmetric Loss

- Recall CLIP is a symmetric InfoNCE. $\min_{\boldsymbol{\Theta}} \mathcal{L}_{CLIP} = \frac{1}{2}(\mathcal{L}_I + \mathcal{L}_T).$
- By looking into the equivalent OT objective, L_I ensure the marginal of images, L_t ensure the marginal of text.

$$\min_{\theta} KL(\tilde{\mathbf{P}}|\mathbf{P}^{\theta})$$

$$\text{where} \quad \mathbf{P}^{\theta} = \arg\min_{\mathbf{P}\in U} < \mathbf{C}^{\theta}, \mathbf{P} > -\epsilon H(\mathbf{P}).$$

U here follows either image marginal, or text marginal.

- Then the necessity of a symmetric InfoNCE is clear:
  - Intuitively, if only ensuring image marginal, it does not prohibit one text matches to two images.
  - A symmetric loss ensures both image and text marginal, ensuring a one-one image-text mapping, and intuitively inducing better discriminess.

Take-home Messages

- With minor assumptions, theoretically we show the CLIP objective has an intrinsic score-matching nature. The loss value of a training (image-text) pair has a potential to evaluate the amount of supervision, which induces performance variance in evaluation datasets.
- We show the CLIP objective is theoretically equivalent to a combination of two inverse optimal transport objectives. With this insight, we show a small temperature ensures hard matching and strong convexity. We also show the necessity of a symmetric loss, which ensures both image and text marginal distribution.

Flamingo: a Visual Language Model for Few-Shot Learning

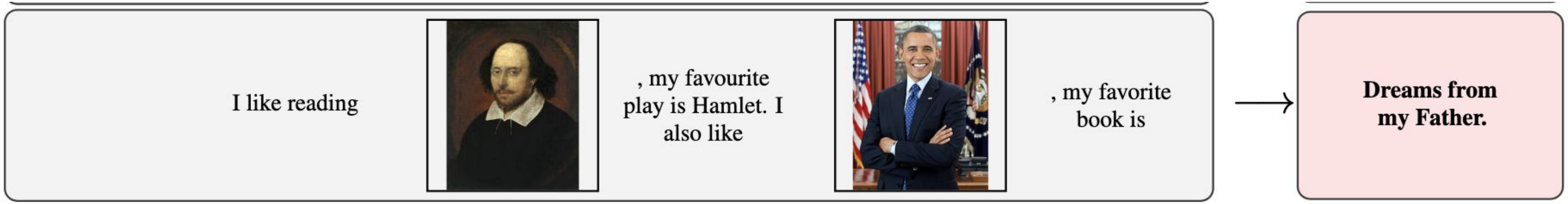
DATA 8005

ZHOU Pei,

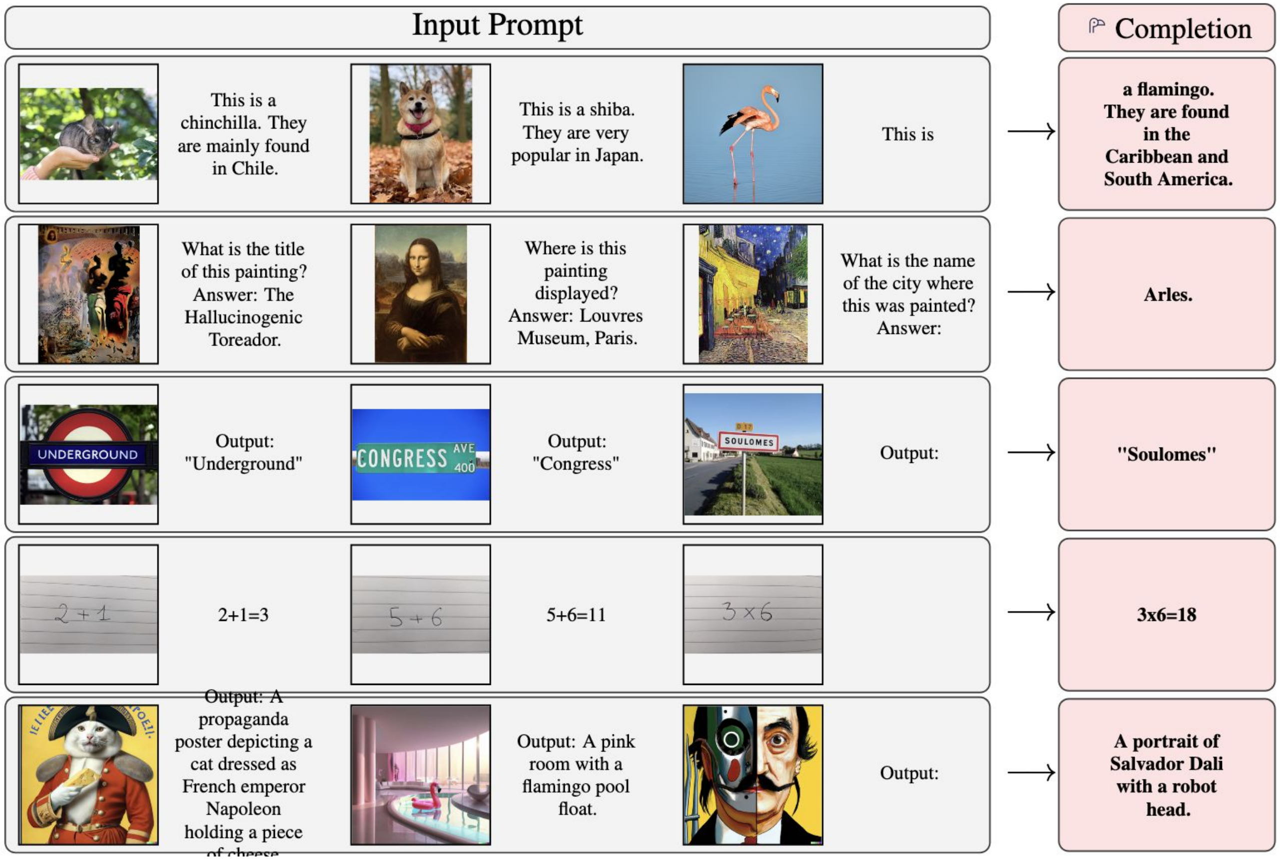
Fall 2024

Key innovations

- Bridge powerful pretrained vision-only and language-only models
- Handle sequences of arbitrarily interleaved visual and textual data
- Seamlessly ingest images or videos as inputs

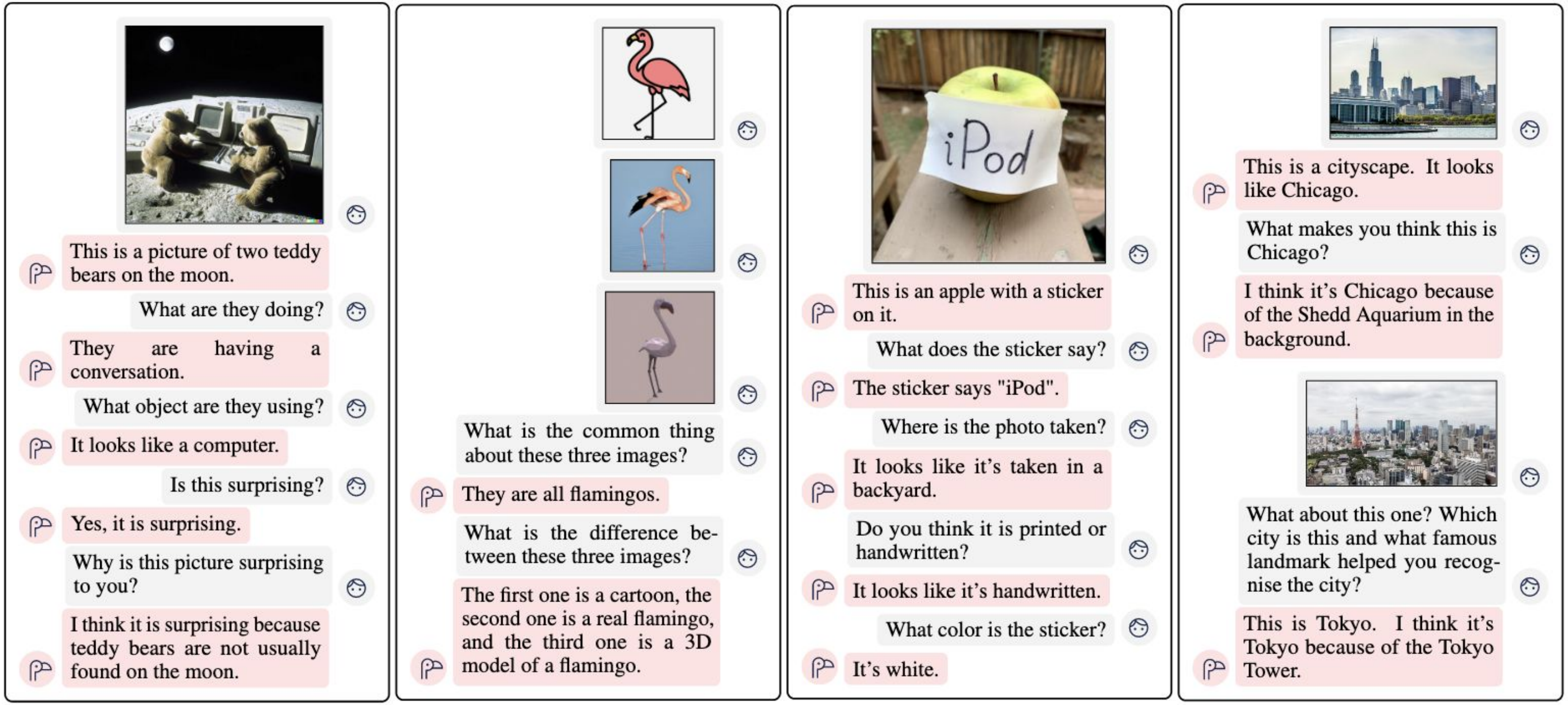


Few-shot prompting



Video Understanding

Dialogue



This is a picture of two teddy bears on the moon.
What are they doing?
They are having a conversation.
What object are they using?
It looks like a computer.
Is this surprising?
Yes, it is surprising.
Why is this picture surprising to you?
I think it is surprising because teddy bears are not usually found on the moon.
What is the common thing about these three images?
They are all flamingos.
What is the difference between these three images?
The first one is a cartoon, the second one is a real flamingo, and the third one is a 3D model of a flamingo.
iPod
This is an apple with a sticker on it.
What does the sticker say?
The sticker says "iPod".
Where is the photo taken?
It looks like it's taken in a backyard.
Do you think it is printed or handwritten?
It looks like it's handwritten.
What color is the sticker?
It's white.
This is a cityscape. It looks like Chicago.
What makes you think this is Chicago?
I think it's Chicago because of the Shedd Aquarium in the background.
What about this one? Which city is this and what famous landmark helped you recognise the city?
This is Tokyo. I think it's Tokyo because of the Tokyo Tower.

Next-token prediction task

- Flamingo models the likelihood of text y conditioned on interleaved images and videos x, and past text y<l.

$$p(y|x) = \prod_{\ell=1}^{L} p(y_\ell | y_{<\ell}, x_{\leq \ell}),$$

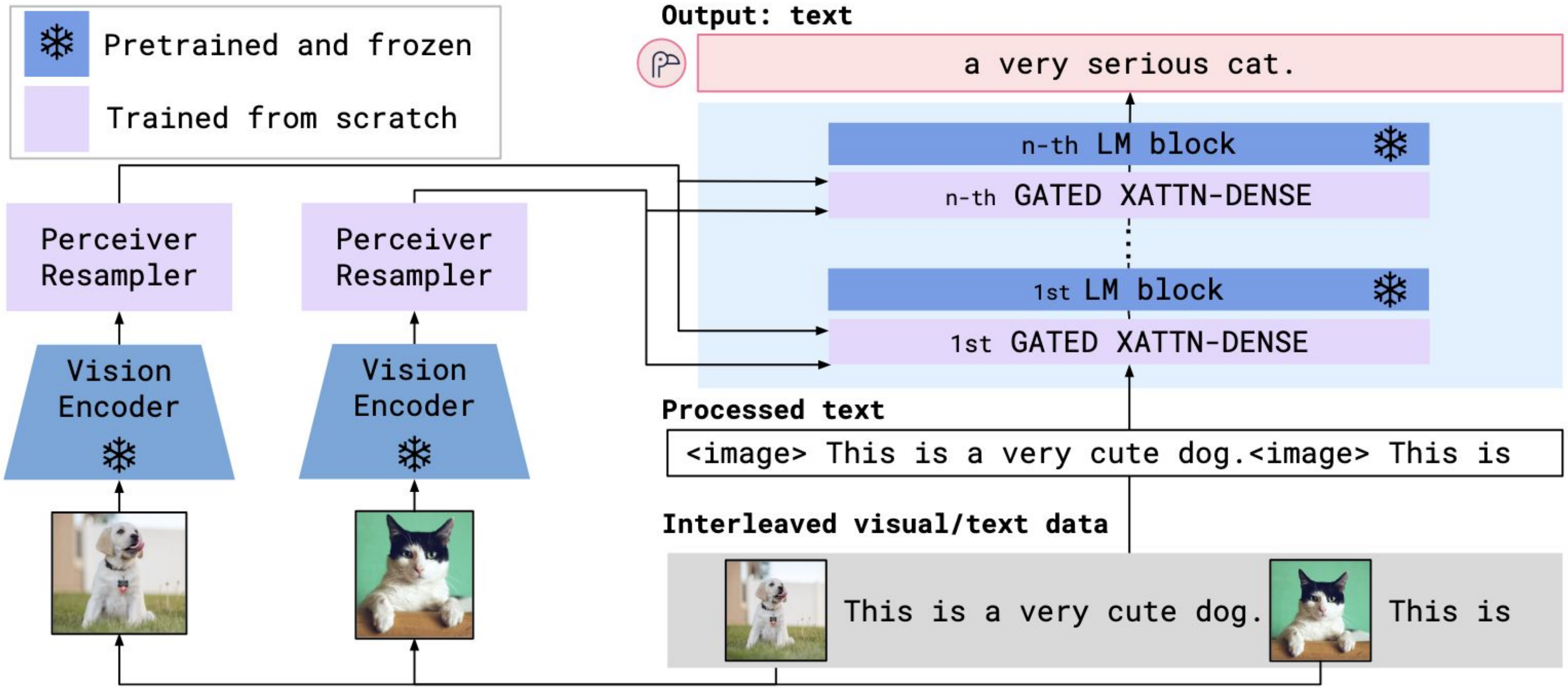
Mixture of vision and language datasets

- Interleaved image and text dataset.
- Image-text pairs.
- Video-text pairs.

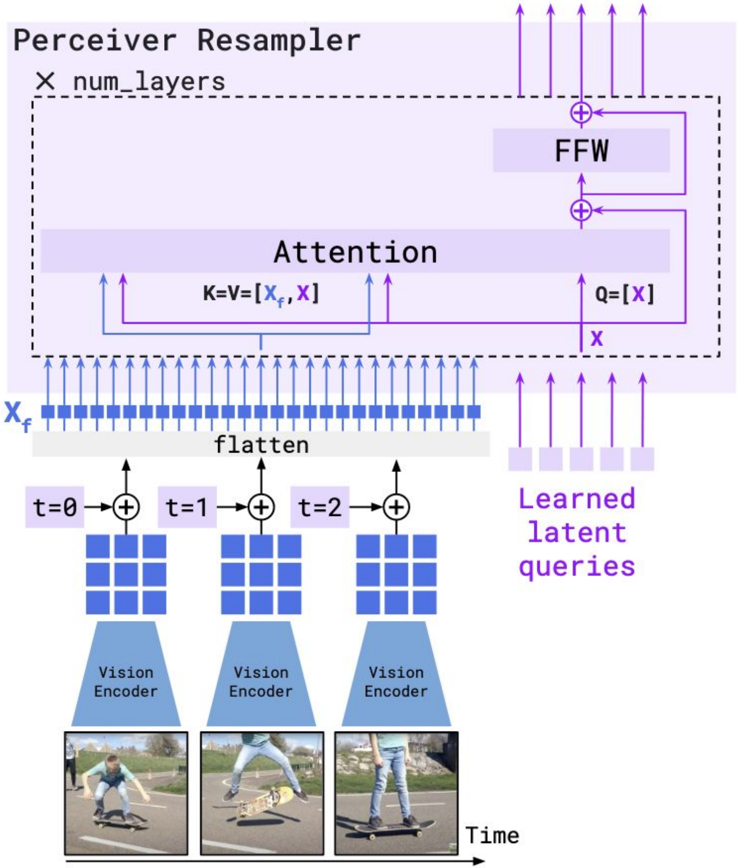
- Multi-objective training and optimisation strategy

$$\sum_{m=1}^{M} \lambda_m \cdot \mathbb{E}_{(x,y)\sim\mathcal{D}_m}\left[-\sum_{\ell=1}^{L} \log p(y_\ell | y_{<\ell}, x_{\leq\ell})\right]$$

Flamingo architecture overview



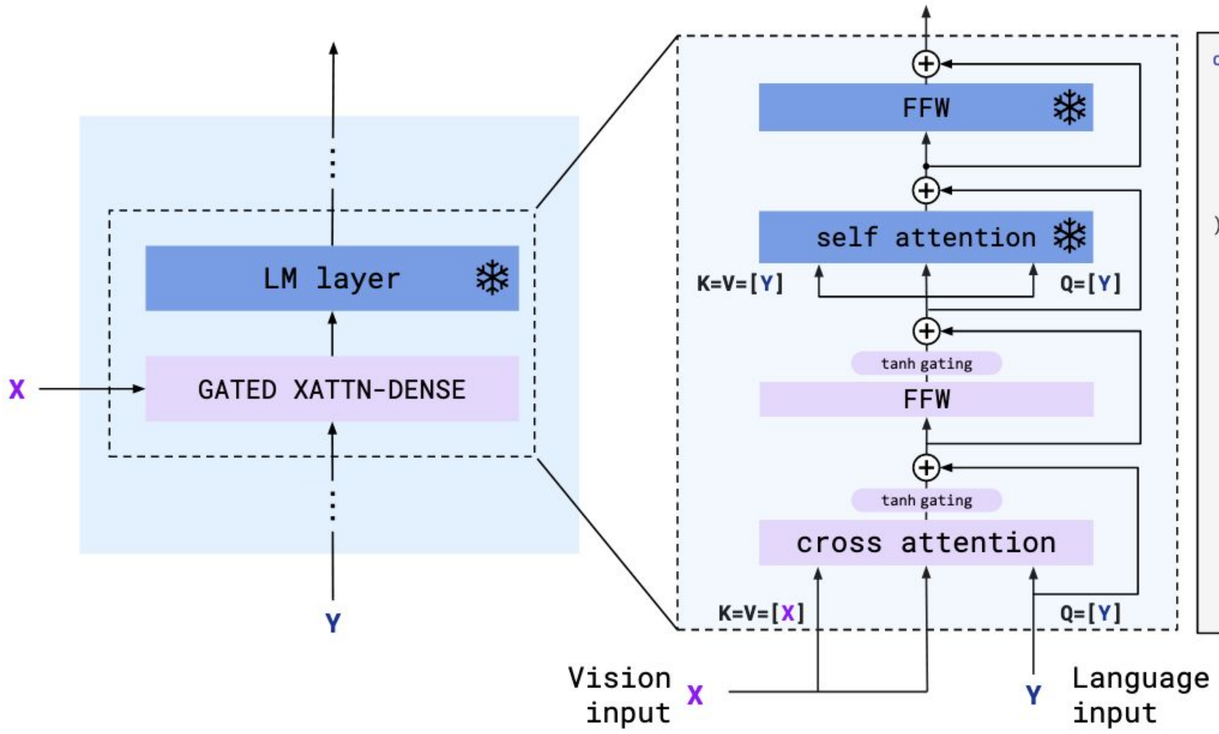
Perceiver Resampler




```python
def perceiver_resampler(
    x_f,  # The [T, S, d] visual features (T=time, S=space)
    time_embeddings,  # The [T, 1, d] time pos embeddings.
    x,  # R learned latents of shape [R, d]
    num_layers,  # Number of layers
):
  """The Perceiver Resampler model."""

  # Add the time position embeddings and flatten.
  x_f = x_f + time_embeddings
  x_f = flatten(x_f)  # [T, S, d] -> [T * S, d]
  # Apply the Perceiver Resampler layers.
  for i in range(num_layers):
    # Attention.
    x = x + attention_i(q=x, kv=concat([x_f, x]))
    # Feed forward.
    x = x + ffw_i(x)
  return x
```

GATED XATTN-DENSE layers

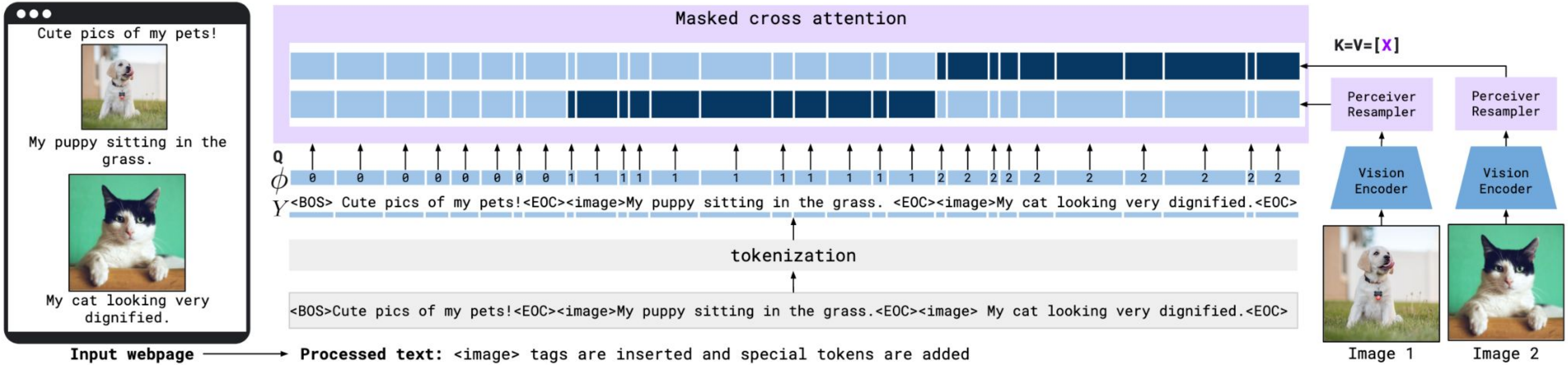



```python
def gated_xattn_dense(
    y,  # input language features
    x,  # input visual features
    alpha_xattn, # xattn gating parameter – init at 0.
    alpha_dense, # ffw gating parameter – init at 0.
):
  """Applies a GATED XATTN-DENSE layer."""

  # 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
```

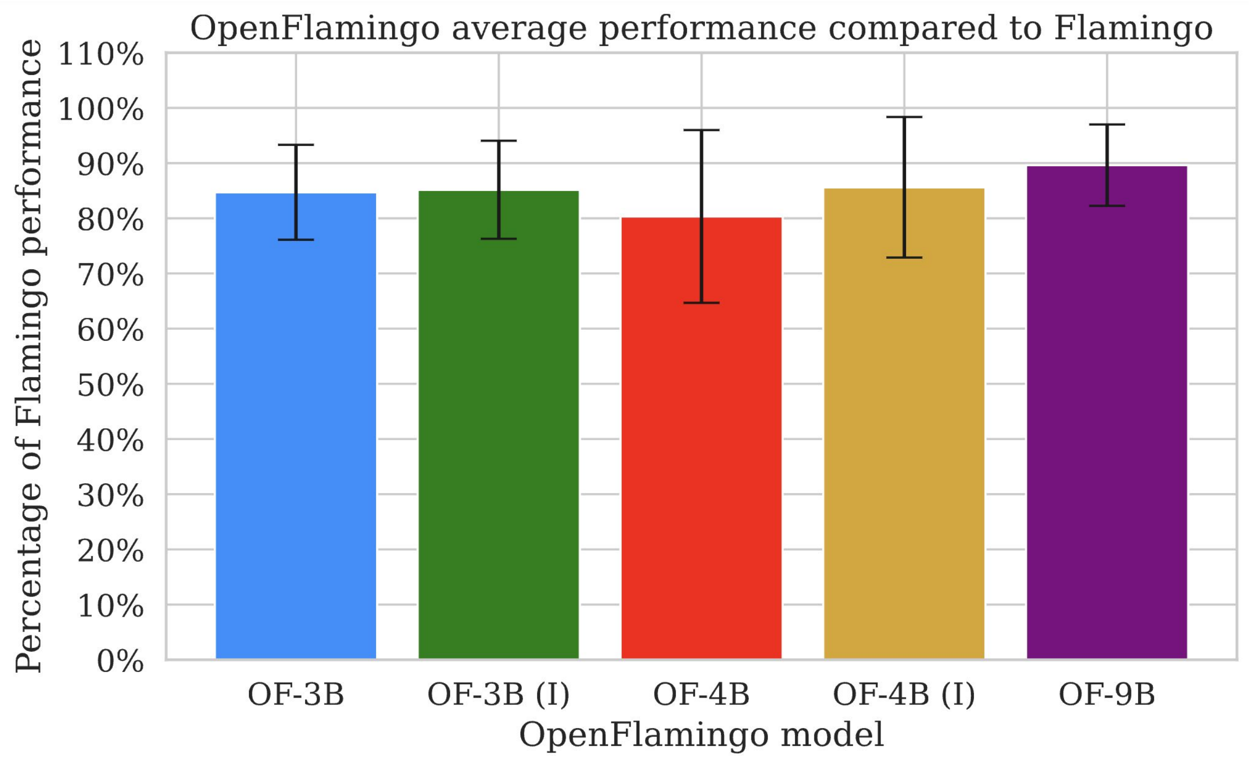
Interleaved visual data and text support



Experiments

Method	FT	Shot	OKVQA (I)	VQAv2 (I)	COCO (I)	MSVDQA (V)	VATEX (V)	VizWiz (I)	Flick30K (I)	MSRVTTQA (V)	iVQA (V)	YouCook2 (V)	STAR (V)	VisDial (I)	TextVQA (I)	NextQA (I)	HatefulMemes (I)	RareAct (V)
Zero/Few shot SOTA	✗	(X)	[34] 43.3 (16)	[114] 38.2 (4)	[124] 32.2 (0)	[58] 35.2 (0)	-	-	-	[58] 19.2 (0)	[135] 12.2 (0)	-	[143] 39.4 (0)	[79] 11.6 (0)	-	-	[85] 66.1 (0)	[85] 40.7 (0)
Flamingo-3B	✗	0	41.2	49.2	73.0	27.5	40.1	28.9	60.6	11.0	32.7	55.8	39.6	46.1	30.1	21.3	53.7	58.4
	✗	4	43.3	53.2	85.0	33.0	50.0	34.0	72.0	14.9	35.7	64.6	41.3	47.3	32.7	22.4	53.6	-
	✗	32	45.9	57.1	99.0	42.6	59.2	45.5	71.2	25.6	37.7	76.7	41.6	47.3	30.6	26.1	56.3	-
Flamingo-9B	✗	0	44.7	51.8	79.4	30.2	39.5	28.8	61.5	13.7	35.2	55.0	41.8	48.0	31.8	23.0	57.0	57.9
	✗	4	49.3	56.3	93.1	36.2	51.7	34.9	72.6	18.2	37.7	70.8	**42.8**	50.4	33.6	24.7	62.7	-
	✗	32	51.0	60.4	106.3	47.2	57.4	44.0	72.8	29.4	40.7	77.3	41.2	50.4	32.6	28.4	63.5	-
Flamingo	✗	0	50.6	56.3	84.3	35.6	46.7	31.6	67.2	17.4	40.7	60.1	39.7	52.0	35.0	26.7	46.4	**60.8**
	✗	4	57.4	63.1	103.2	41.7	56.0	39.6	75.1	23.9	44.1	74.5	42.4	**55.6**	36.5	30.8	68.6	-
	✗	32	**57.8**	**67.6**	**113.8**	**52.3**	**65.1**	**49.8**	**75.4**	**31.0**	**45.3**	**86.8**	42.2	**55.6**	**37.9**	**33.5**	**70.0**	-
Pretrained FT SOTA	✔	(X)	54.4 [34] (10K)	80.2 [140] (444K)	143.3 [124] (500K)	47.9 [28] (27K)	76.3 [153] (500K)	57.2 [65] (20K)	67.4 [150] (30K)	46.8 [51] (130K)	35.4 [135] (6K)	138.7 [132] (10K)	36.7 [128] (46K)	75.2 [79] (123K)	54.7 [137] (20K)	25.2 [129] (38K)	79.1 [62] (9K)	-

OpenFlamingo



Awadalla, Anas, et al. "Openflamingo: An open-source framework for training large autoregressive vision-language models." *arXiv preprint arXiv:2308.01390* (2023).

Discussions

Limitations:

- Inheriting Weaknesses of Language Models.
- Poor Classification Performance
- Limitations of In-Context Learning

Conclusion:

- Flamingo is a general-purpose family of models that can be applied to image and video tasks with minimal task-specific training data.
- Interactive abilities of Flamingo such as "chatting" with the model, demonstrating flexibility beyond traditional vision benchmarks

Development Roadmap from LLaVA to LLaVA-OneVision

Chenming Zhu

Fall 2024

LLaVA Roadmap

- LLaVA (Visual Instruction Tuning)
- LLaVA-NeXT Blog:

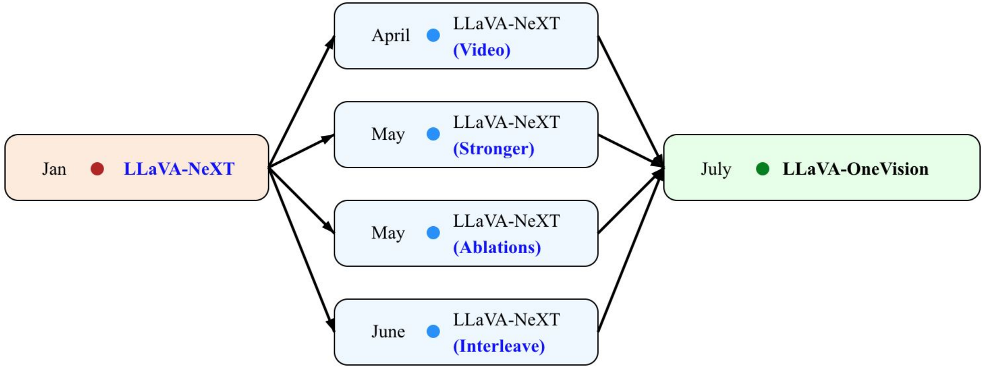
  Improved reasoning, OCR, and world knowledge

  A Strong Zero-shot Video Understanding Model

  Stronger LLMs Supercharge Multimodal Capabilities in the Wild

  What Else Influences Visual Instruction Tuning Beyond Data?

- LLaVA-OneVision



LLaVA: Visual Instruction Tuning

Model Architecture

- Vision Encoder + Projection Layer + LLM

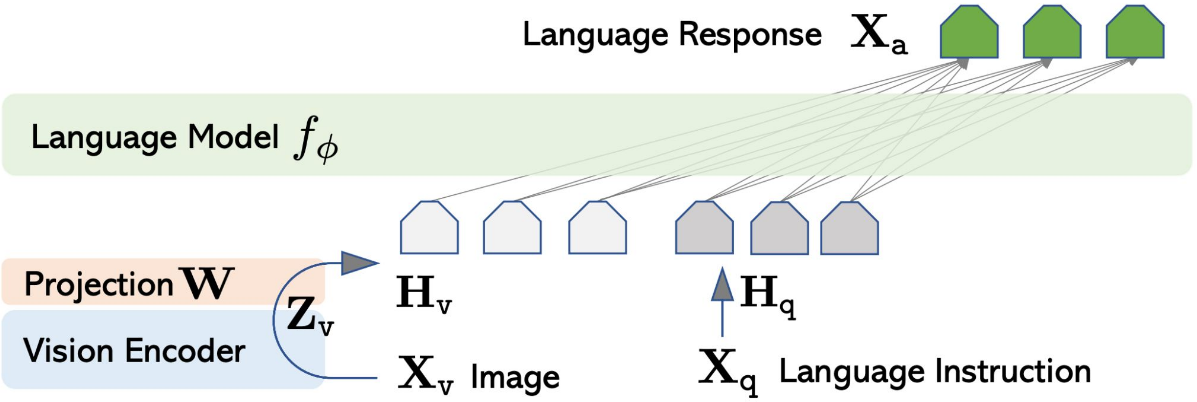
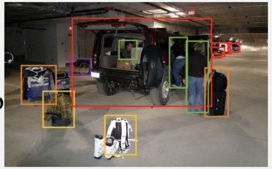


Figure 1: LLaVA network architecture.

Visual Instruction Tuning Data Construction

Context type 1: Captions
A group of people standing outside of a black vehicle with various luggage.
Luggage surrounds a vehicle in an underground parking area
People try to fit all of their luggage in an SUV.
The sport utility vehicle is parked in the public garage, being packed for a trip
Some people with luggage near a van that is transporting it.



Context type 2: Boxes
person: [0.681, 0.242, 0.774, 0.694], backpack: [0.384, 0.696, 0.485, 0.914], suitcase: ...<omitted>

Response type 1: conversation
Question: What type of vehicle is featured in the image?
Answer: The image features a black sport utility vehicle (SUV) ...<omitted>
Response type 2: detailed description
The image is an underground parking area with a black sport utility vehicle (SUV) parked. There are three people in the scene, with one person standing closer to the left side of the vehicle, another person in the middle, and the third person on the right side. They are all working together to pack their luggage into the SUV for a trip. ...<omitted>
Response type 3: complex reasoning
Question: What challenges do these people face?
Answer: In the image, a group of people is standing outside a black SUV in a parking area, surrounded by various pieces of luggage, including suitcases and backpacks. They are facing the challenge of fitting all their luggage into the black SUV. There are multiple suitcases and backpacks to be packed, which suggests that the group has a significant amount of belongings ...<omitted>

Use GPT-4 to convert the COCO dataset with Caption and Bounding Boxes information to data:

Conversation: Dialogue data, totaling 58K samples.

Detailed description: Rich and comprehensive descriptions of images, totaling 23K samples.

Complex reasoning: Complex reasoning data, totaling 77K samples.

Training Recipe

Stage 1: Pre-training for Feature Alignment [projection layer]

Dataset: filter CC3M to 595K image-text pairs

Stage 2: Fine-tuning End-to-End for Instruction Following [projection layer and LLM]

Dataset: Instruction Tuning Dataset

Limitations

- The simple connector may limit the model's ability to deeply understand complex visual information.
- Limited training data scale and diversity
- Potential hallucination and misinformation

LLaVA-1.5: Improved Baselines with Visual Instruction Tuning

Model Architecture Modification:

- Replacing the original CLIP-ViT-L/14 visual encoder with the CLIP-ViT-L-336px visual encoder.
- Replacing the original single linear layer with an MLP layer (two linear layers).

Incorporating VQA data oriented towards academic tasks and specifying response format in prompts: This enhances LLaVA's performance on academic task benchmarks.

	Method	LLM	Res.	GQA	MME	MM-Vet
	InstructBLIP	14B	224	49.5	1212.8	25.6
Only using a subset of InstructBLIP training data						
0	**LLaVA**	7B	224	–	809.6	25.5
1	+VQA-v2	7B	224	47.0	1197.0	27.7
2	+Format prompt	7B	224	46.8	1323.8	26.3
3	+MLP VL connector	7B	224	47.3	1355.2	27.8
4	+OKVQA/OCR	7B	224	50.0	1377.6	29.6
Additional scaling						
5	+Region-level VQA	7B	224	50.3	1426.5	30.8
6	+Scale up resolution	7B	336	51.4	1450	30.3
7	+GQA	7B	336	62.0*	1469.2	30.7
8	+ShareGPT	7B	336	62.0*	1510.7	31.1
9	+Scale up LLM	13B	336	**63.3***	**1531.3**	**36.1**

Table 2. **Scaling results** on data, model, and resolution. We choose to conduct experiments on GQA [20], MME [16], and MM-Vet [52] to examine the representative capabilities of VQA with short answers, VQA with output formatting, and natural visual conversations, respectively. *Training images of GQA were observed during training.

LLaVA-NeXT: Improved reasoning, OCR, and world knowledge

Dynamic High Resolution

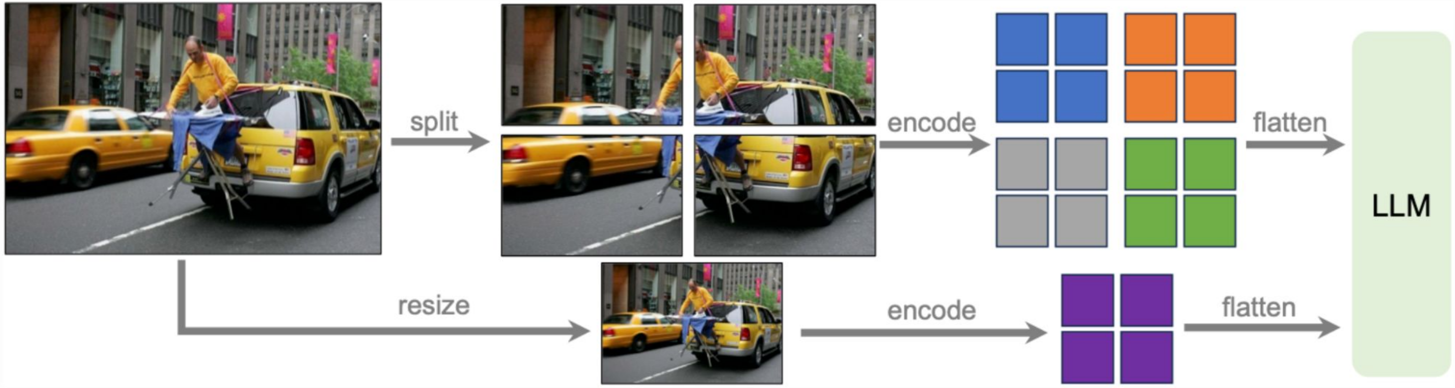


Illustration of dynamic high resolution scheme: a grid configuration of 2×2

AnyRes technique is designed to accommodate images of various high resolutions. It employs a grid configuration of {2×2,1×{2,3,4},{2,3,4}×1}, balancing performance efficiency with operational costs for the high-resolution image.

Data Mixture

- **High-quality User Instruct Data**.

First, the diversity of task instructions, ensuring adequately represent a broad spectrum of user intents that are likely to be encountered in real-world scenarios, particularly during the model's deployment phase. Second, the superiority of responses is critical, with the objective of soliciting favorable user feedback. To achieve this, it considers two data sources:

(1) Existing GPT-V data. LAION-GPT-V and ShareGPT-4V.

(2) To further facilitate better visual conversation for more scenarios, it collects a small 15K visual instruction tuning dataset covering different applications. The instructions and images come from LLaVA demo, which are real-world users requests. They carefully filter samples that may have privacy concerns or are potentially harmful, and generate the response with GPT-4V.

- **Multimodal Document/Chart Data.**

LLaVA-NeXT: Improved reasoning, OCR, and world knowledge

Compared with LLaVA-1.5, LLaVA-NeXT has several improvements:

1. Increasing the input image resolution to 4x more pixels. This allows it to grasp more visual details. It supports three aspect ratios, up to 672x672, 336x1344, 1344x336 resolution.
2. Better visual reasoning and OCR capability with an improved visual instruction tuning data mixture.
3. Better visual conversation for more scenarios, covering different applications. Better world knowledge and logical reasoning.

LLaVA-NeXT: A Strong Zero-shot Video Understanding Model

Zero-shot video representation capabilities with AnyRes

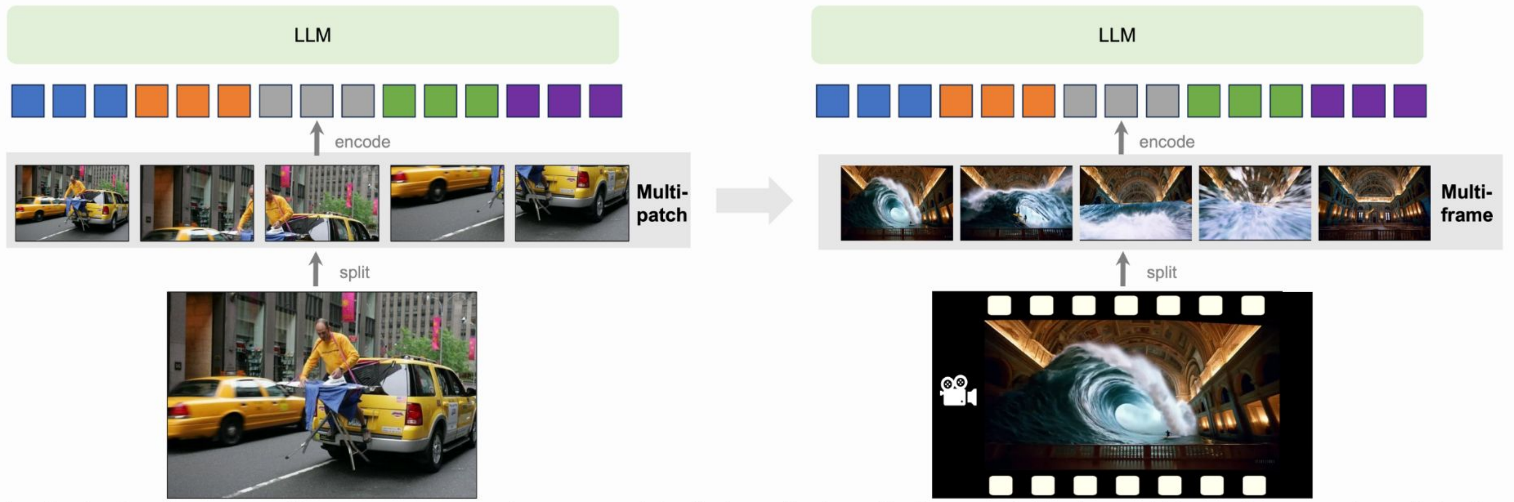
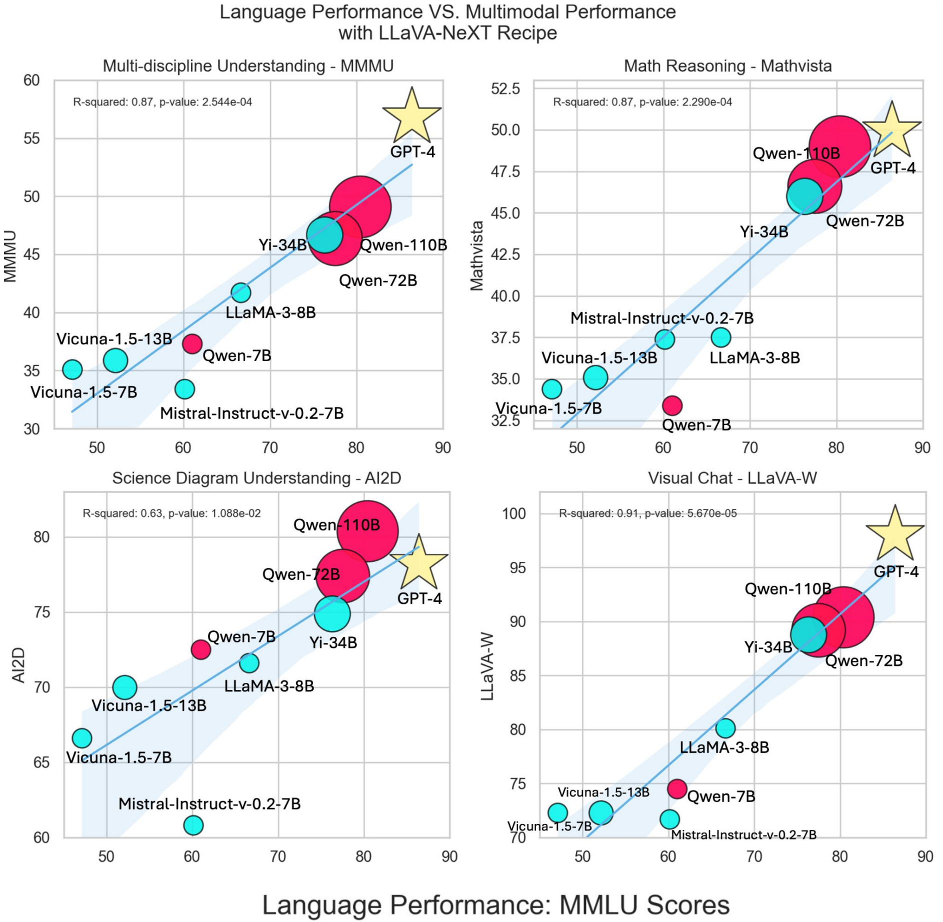


Illustration that AnyRes digests a set of image as a sequence of concatenated visual tokens, allowing unified image and video input, which natually suppots the evolution from multi-image to multi-frame

With minor code adjustments, LLaVA-NeXT can process N video frames arranged in a {1xN} grid. Assuming each frame comprises 24x24 tokens, the total token count for a video would be 24x24xN. However, considering the "max_token_length" limit of 4096 for the LLM, it is crucial to ensure that 24x24xN + the number of text tokens < 4096 to avoid nonsensical outputs.

How effectively can the language capabilities of LLMs be transferred to multimodal settings?

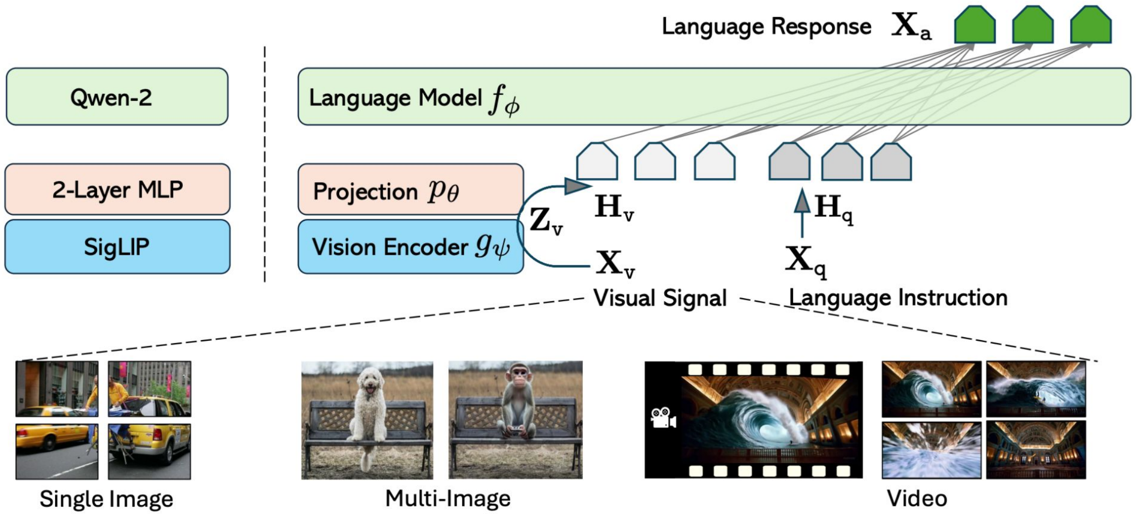


Improved Language Capability:
Across LLMs of comparable sizes (e.g., 7B Mistral/Vicuna, 7B Qwen, 8B LLaMa3), there exists a consistent pattern where higher language proficiency, as measured by MMMU scores, corresponds to improved multimodal capabilities.

Influence of Model Size:
Within the same LLM family (e.g., Qwen LLM: 7B, 72B, 110B), larger models consistently demonstrate superior performance on multimodal benchmarks.

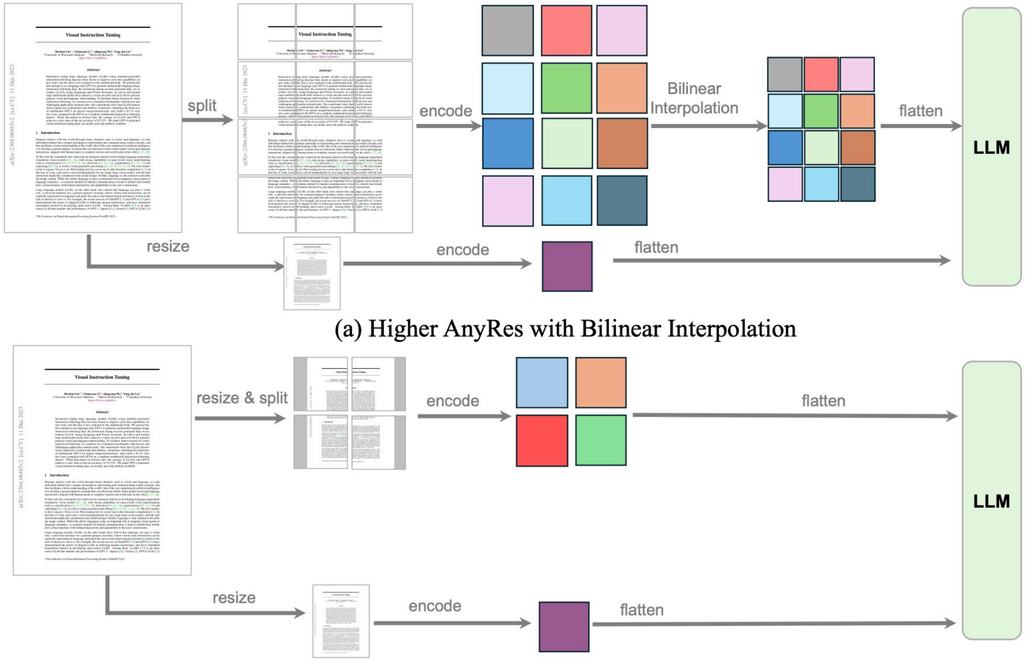
LLaVA-OneVision: Easy Visual Task Transfer

LLaVA-OneVision is the first single model that can simultaneously push the performance boundaries of open LMMs in three important computer vision scenarios: single-image, multi-image, video scenarios.



Visual Representation

To strike a balance of performance and cost, the author observe that the scaling of resolution is more effective than the scaling of token numbers, and recommend an AnyRes strategy with pooling.



(a) Higher AnyRes with Bilinear Interpolation

(b) The original AnyRes

Visual Representation

Single-Image: consider a large maximum spatial configuration (a, b) for single-image representation to maintain the original image resolution without resizing.

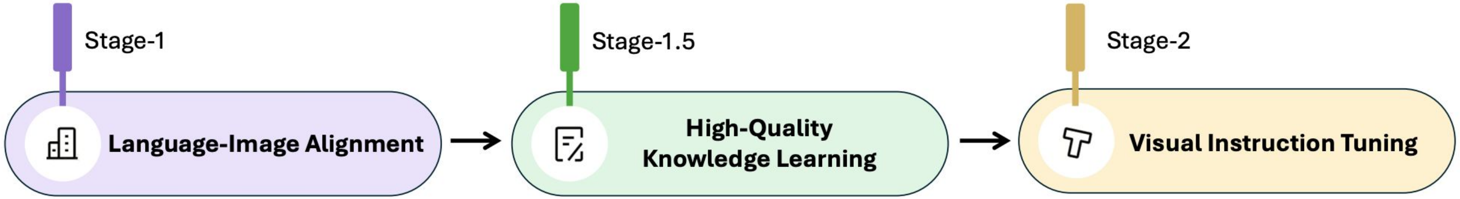
Multi-image: Only the base image resolution is considered.

Video: Each frame of the video is resized to the base image resolution and processed by the vision encoder to generate feature maps. Bilinear interpolation is employed to reduce the number of tokens

Insights on Training Strategies

Prior LLaVA models mainly explore Stage-2 for new scenarios and improved performance. However, the first two functionalities are less frequently investigated and therefore constitute the primary focus of this section.

- Stage-1: Language-Image Alignment.
- Stage-1.5: High-Quality Knowledge Learning.
- Stage-2: Visual Instruction Tuning.



High-Quality Knowledge

To illustrate high-quality knowledge, we consider data from three major categories:

- **Re-Captioned Detailed Description Data:** LLaVA-NeXT-34B is known for its strong detailed caption ability among open-source LMMs. We used the model to generate new captions for the images from the following datasets: COCO118K, BLIP558K, and CC3M.

- **Document / OCR Data:** We utilized the Text Reading subset from the UReader dataset, totaling 100K, which is easily accessible through PDF rendering. We used this text reading data along with the SynDOG EN/CN 1M datasets.

- **ShareGPT4V Chinese Detailed Caption:** We used the original ShareGPT4V[3] images and utilized GPT-4V provided by the Azure API to generate detailed Chinese caption data, aiming to improve the model's capability in Chinese.

High-Quality Knowledge

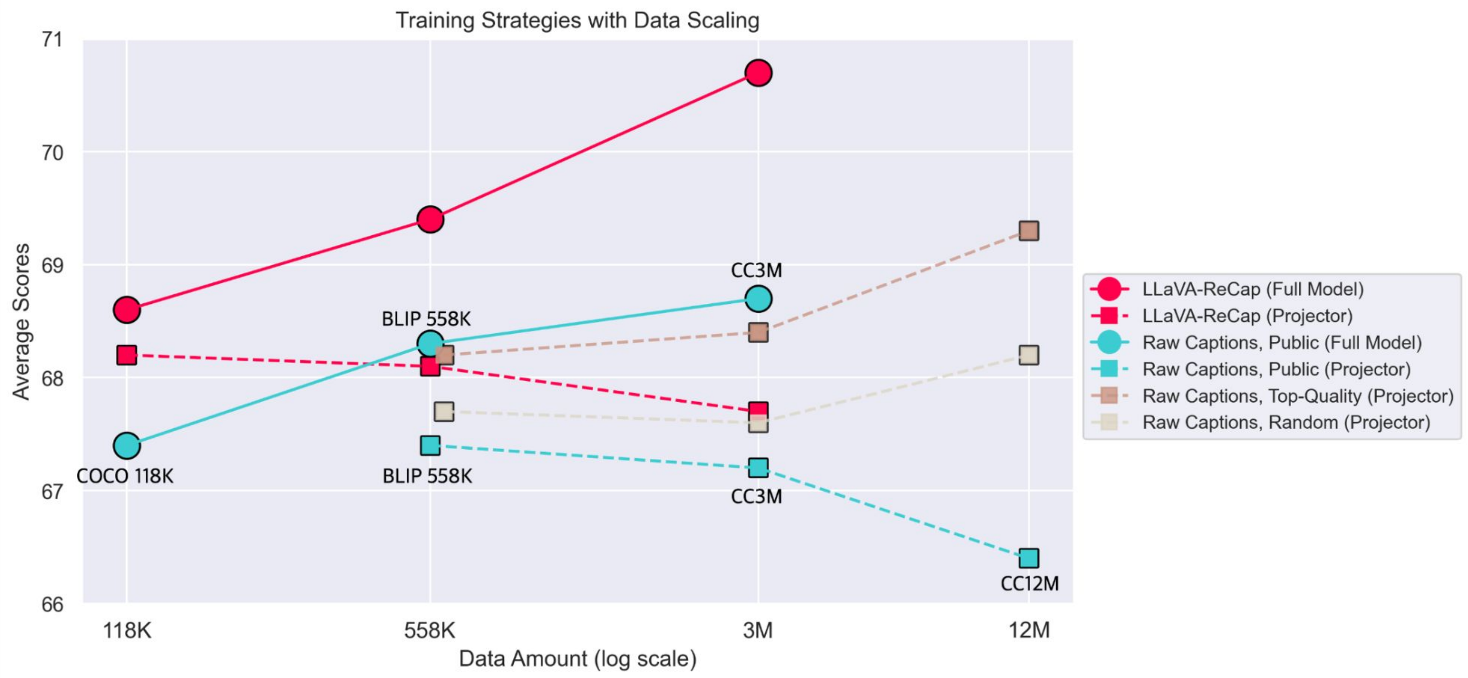


Figure 2. This figure shows that using LLaVA-ReCap Data in Stage 1.5 training yields the most significant improvements (red circles). Performance with raw captions data like COCO18K, BLIP558K, and CC3M is also strong (blue circles). We also include the results from Section 3.1 (squared shape), where only the projector was trained using raw captions data at various scales (e.g. from BLIP558K to Web 12M).

Discussions

- Why LLaVA could be so popular in the community compared with others?
- How can the hallucination problem in multimodal large models be mitigated?
- Why are fewer works adopting the Q-Former architecture, instead utilizing the simple MLP layer?