DeepMind

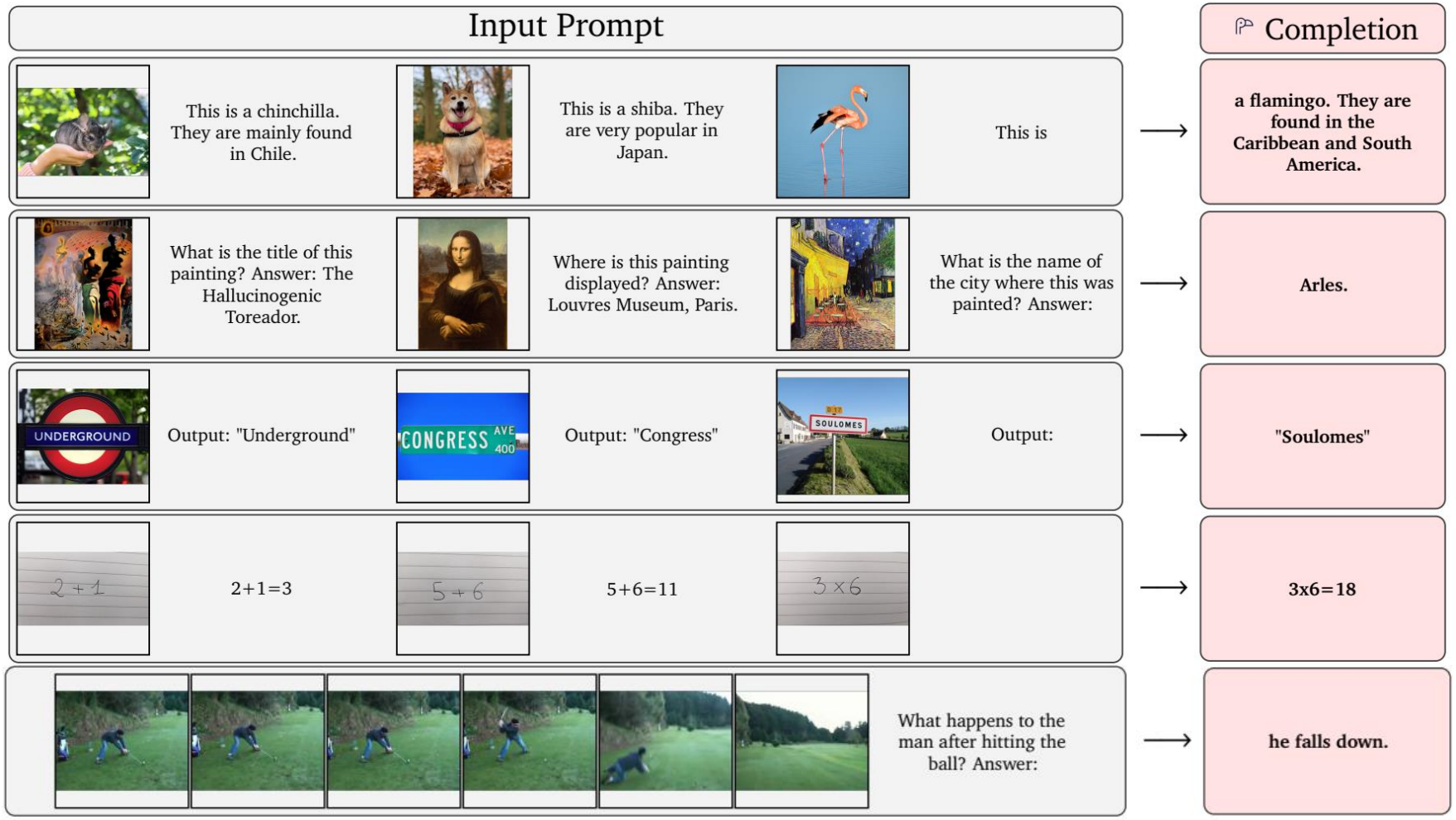
28-04-2022

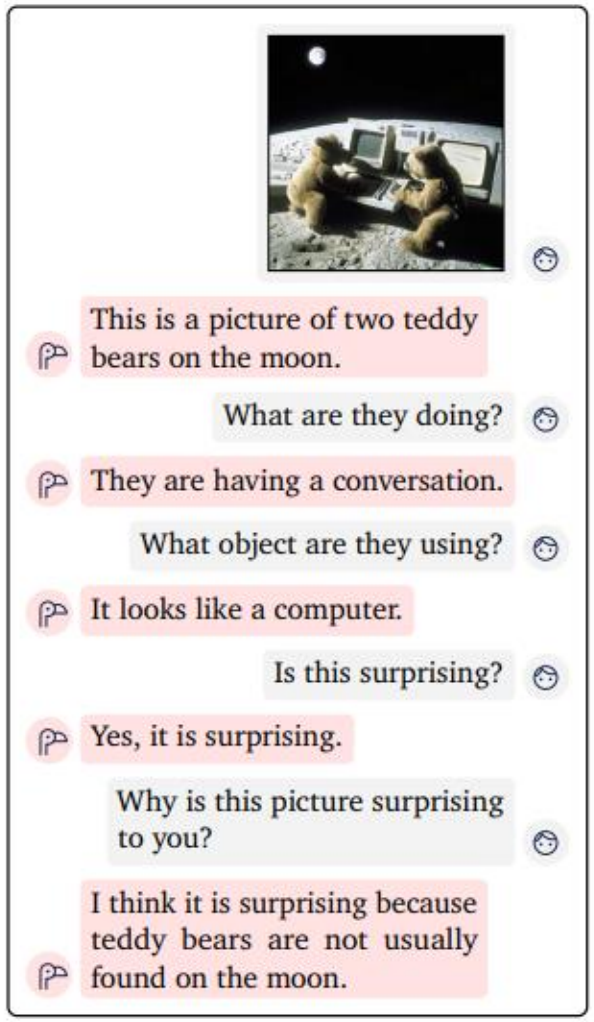
# Flamingo: a Visual Language Model for Few-Shot Learning

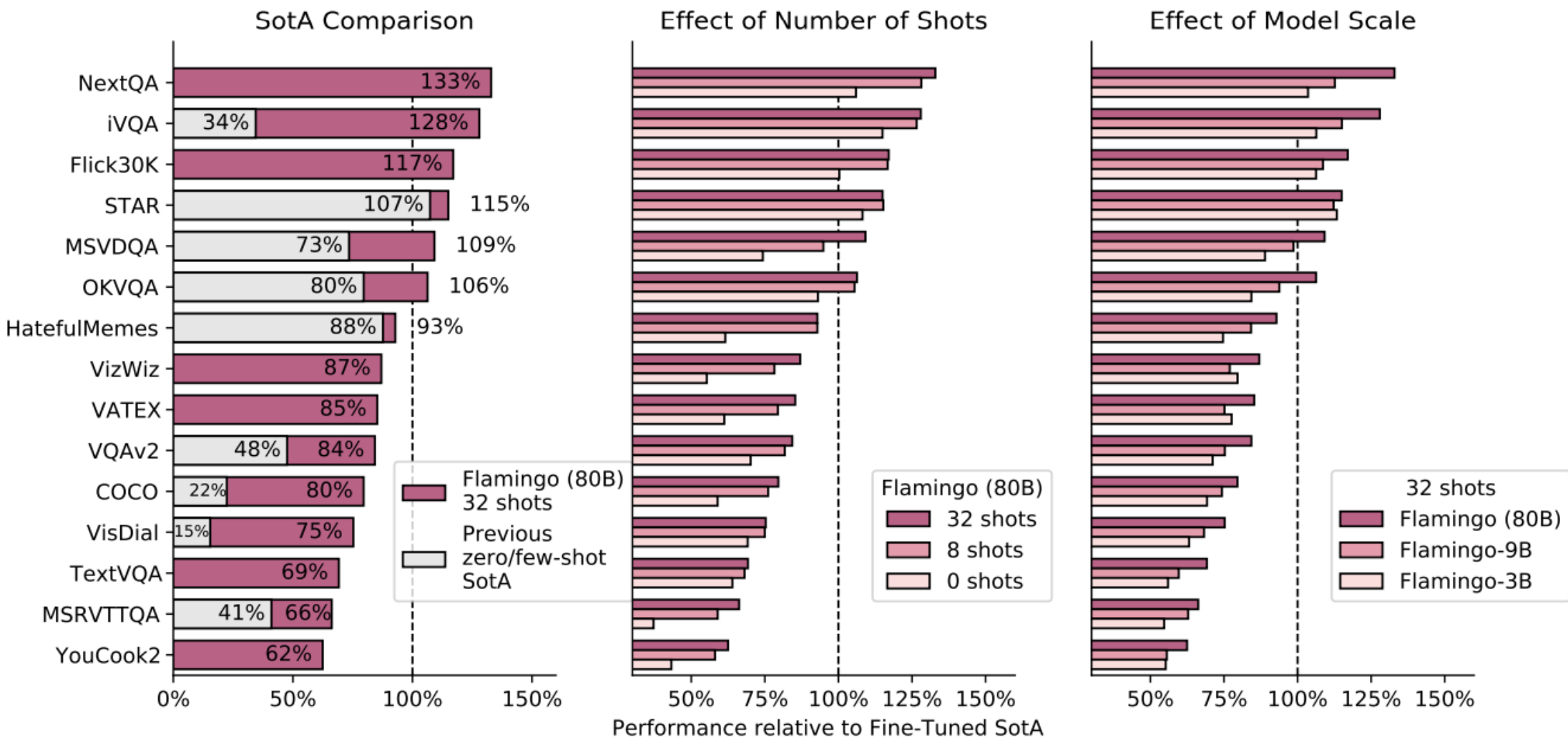
**Jean-Baptiste Alayrac**[*,‡], **Jeff Donahue**[*], **Pauline Luc**[*], **Antoine Miech**[*], **Iain Barr**[†], **Yana Hasson**[†], **Karel Lenc**[†], **Arthur Mensch**[†], **Katie Millican**[†], **Malcolm Reynolds**[†], **Roman Ring**[†], **Eliza Rutherford**[†], **Serkan Cabi**, **Tengda Han**, **Zhitao Gong**, **Sina Samangooei**, **Marianne Monteiro**, **Jacob Menick**, **Sebastian Borgeaud**, **Andrew Brock**, **Aida Nematzadeh**, **Sahand Sharifzadeh**, **Mikolaj Binkowski**, **Ricardo Barreira**, **Oriol Vinyals**, **Andrew Zisserman**, **Karen Simonyan**[*,‡]

[*]Equal contributions, ordered alphabetically, [†]Equal contributions, ordered alphabetically, [‡]Equal senior contributions

# Flamingo: a Visual Language Model for Few-Shot Learning




SotA Comparison
Effect of Number of Shots
Effect of Model Scale
NextQA
iVQA
Flick30K
STAR
MSVDQA
OKVQA
HatefulMemes
VizWiz
VATEX
VQAv2
COCO
VisDial
TextVQA
MSRVTTQA
YouCook2
133%
34%
128%
117%
107%
115%
73%
109%
80%
106%
88%
93%
87%
85%
48%
84%
22%
80%
15%
75%
69%
41%
66%
62%
Flamingo (80B)
32 shots
Previous
zero/few-shot
SotA
Flamingo (80B)
32 shots
8 shots
0 shots
32 shots
Flamingo (80B)
Flamingo-9B
Flamingo-3B
0%
50%
100%
150%
50%
75%
100%
125%
150%
Performance relative to Fine-Tuned SotA

# Flamingo: a Visual Language Model for Few-Shot Learning

Dominant computer vision paradigm:

Large-scale petraining + Task-specific fine-tuning

Limitation:

- Thousands of training samples
- Careful per-task hyperparameter tuning
- Significant computational resource

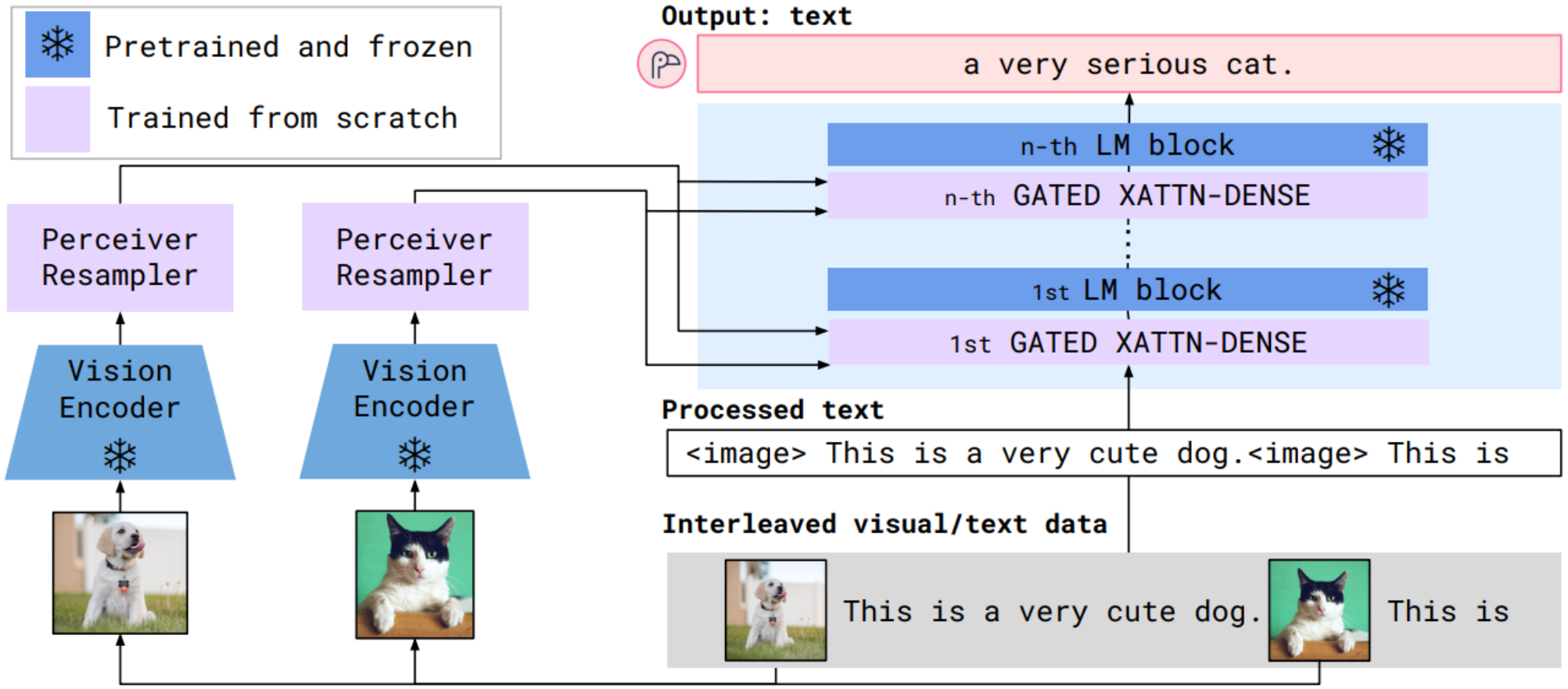
Current multimodal models (CLIP, VL-T5)

- Show promising zero-shot performance
- The trade-off between flexibility and few-shot performance

# Challenges and conttributions

- Flamingo
  - Using strong unimodal models
    - Fixed pretrained language model and vision model
    - Train fusion layers.
  - Supporting images and videos
    - Perceiver-based architecture with a fixed number of visual tokens
  - Heterogeneous training data
    - Combined web scraping with existing paired (image, text) and (video, text) datasets

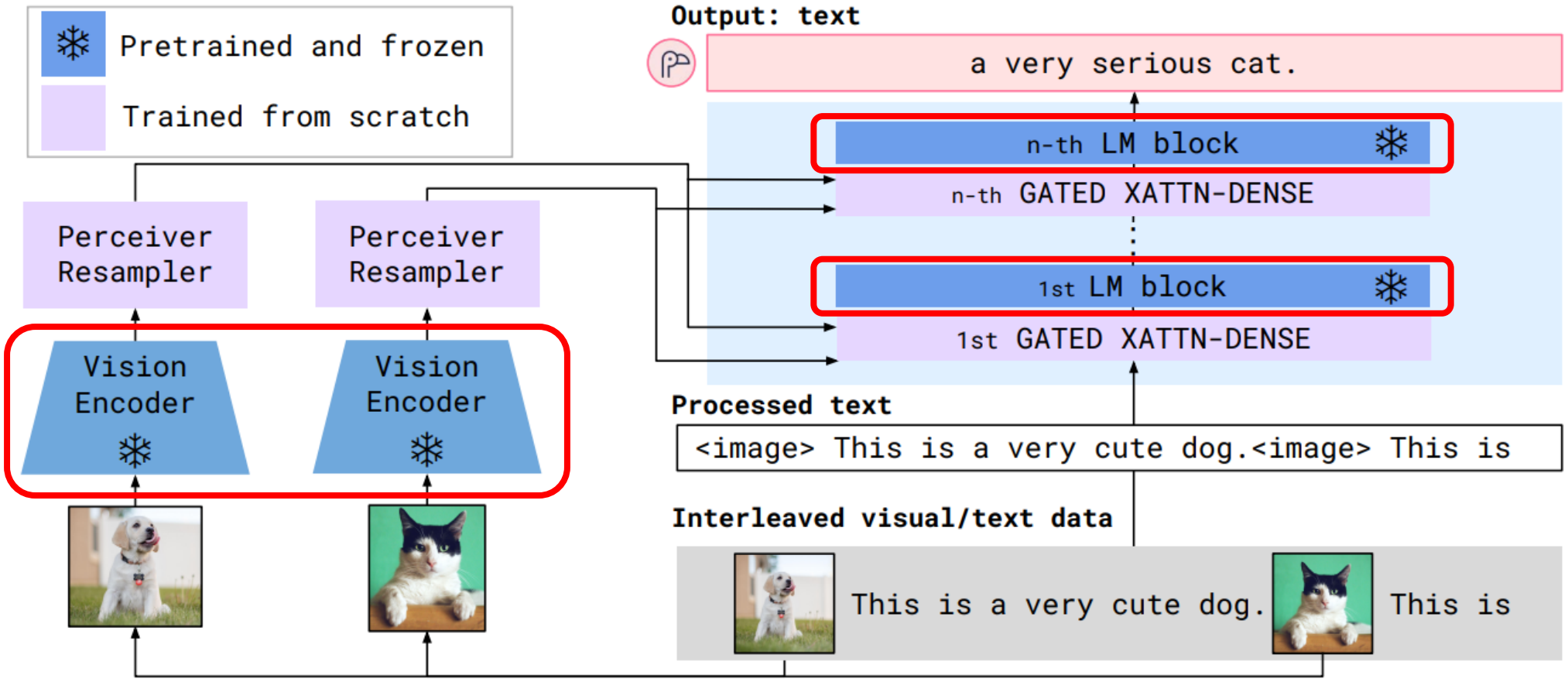
# Overview




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

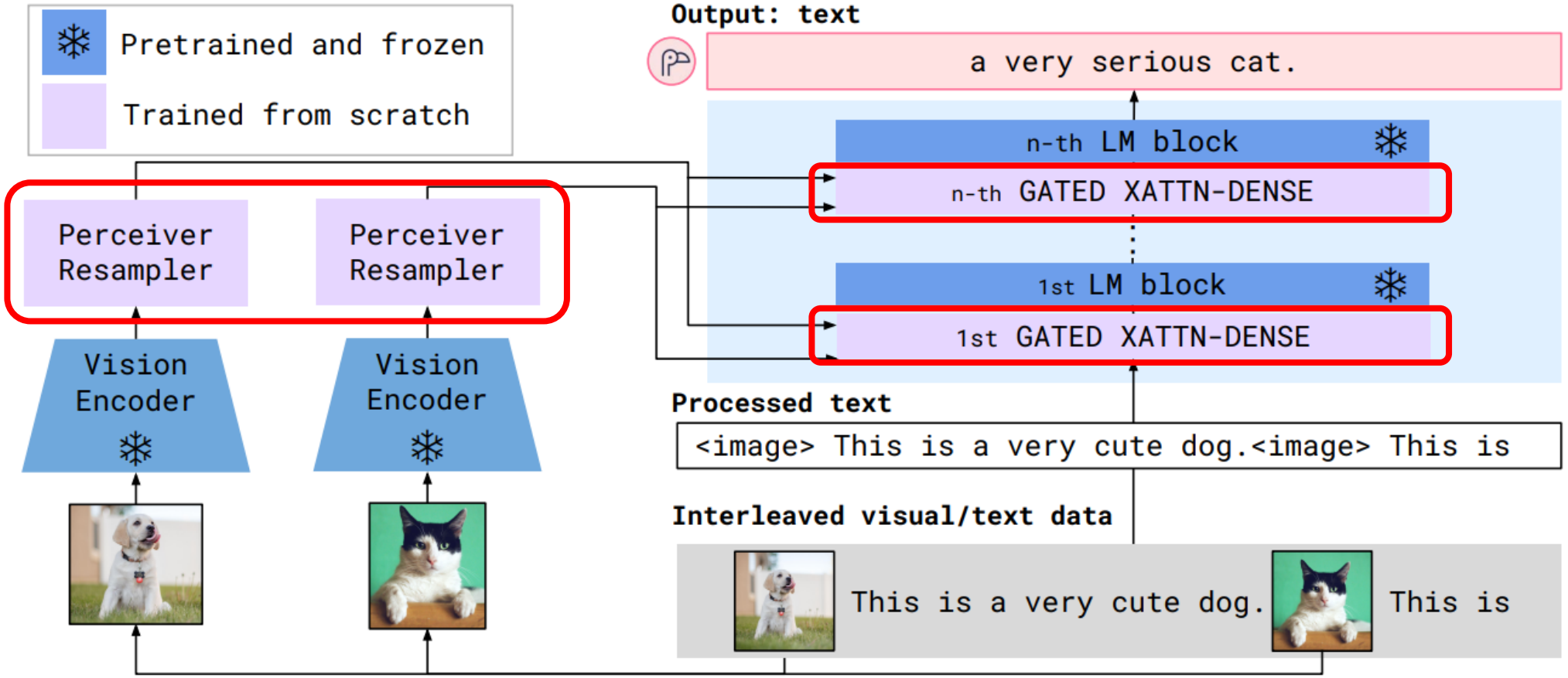
# Overview

Leverage pretrained models to save compute



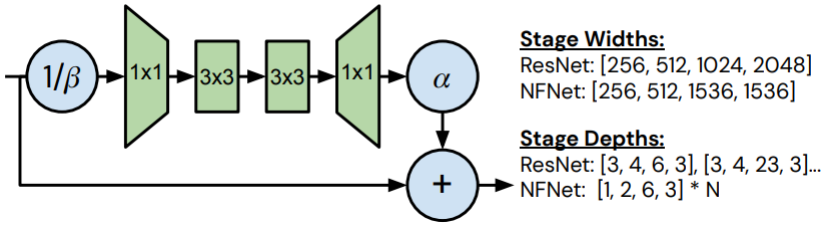
# Overview

Bridge pretrained models harmoniously



# Vision Encoder

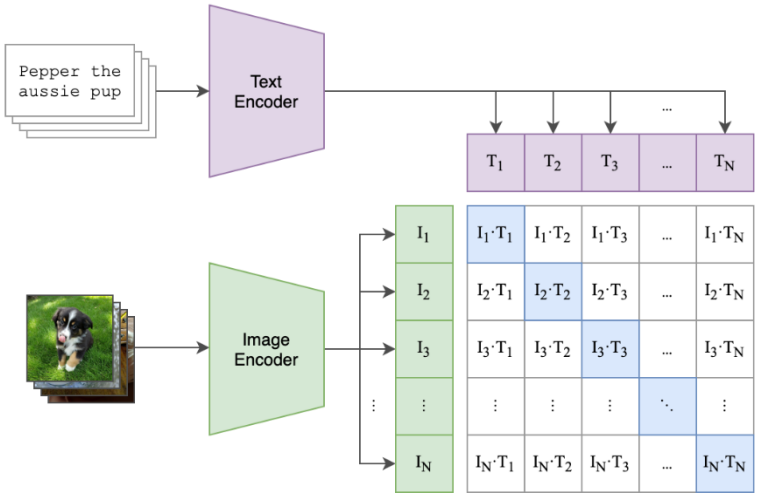
**NFNet**




**Stage Widths:**
ResNet: [256, 512, 1024, 2048]
NFNet: [256, 512, 1536, 1536]

**Stage Depths:**
ResNet: [3, 4, 6, 3], [3, 4, 23, 3]...
NFNet: [1, 2, 6, 3] * N

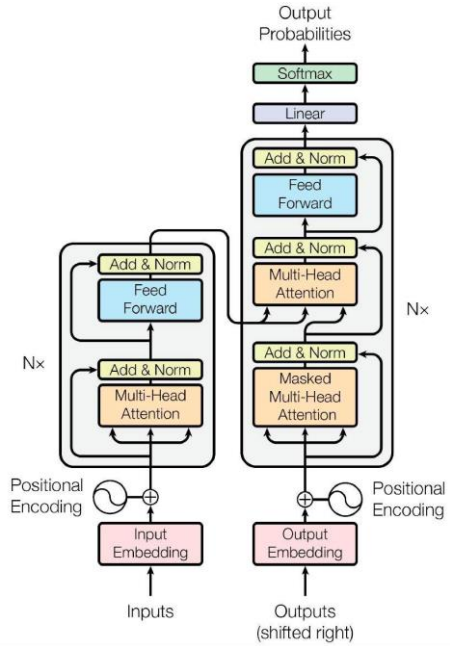
Contrastive training





- Brock, A., De, S., Smith, S. L., & Simonyan, K. (2021, July). High-performance large-scale image recognition without normalization. In *International Conference on Machine Learning* (pp. 1059-1071). PMLR.
- 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.

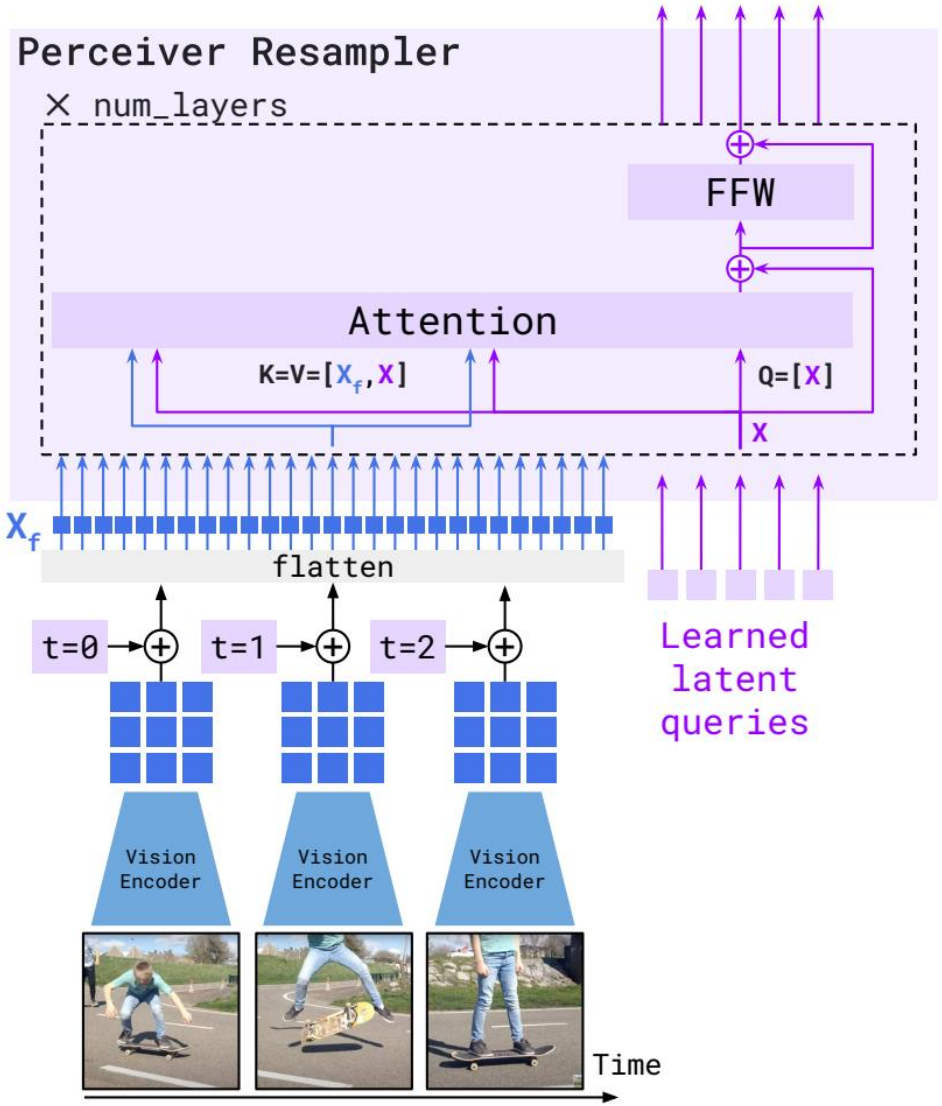

# Pretrained language model




| Model | Size (# Parameters) | Training Tokens |
|---|---|---|
| LaMDA (Thoppilan et al., 2022) | 137 Billion | 168 Billion |
| GPT-3 (Brown et al., 2020) | 175 Billion | 300 Billion |
| Jurassic (Lieber et al., 2021) | 178 Billion | 300 Billion |
| *Gopher* (Rae et al., 2021) | 280 Billion | 300 Billion |
| MT-NLG 530B (Smith et al., 2022) | 530 Billion | 270 Billion |
| *Chinchilla* | 70 Billion | 1.4 Trillion |


- Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., ... & Sifre, L. (2022). Training Compute-Optimal Large Language Models. arXiv preprint arXiv:2203.15556.

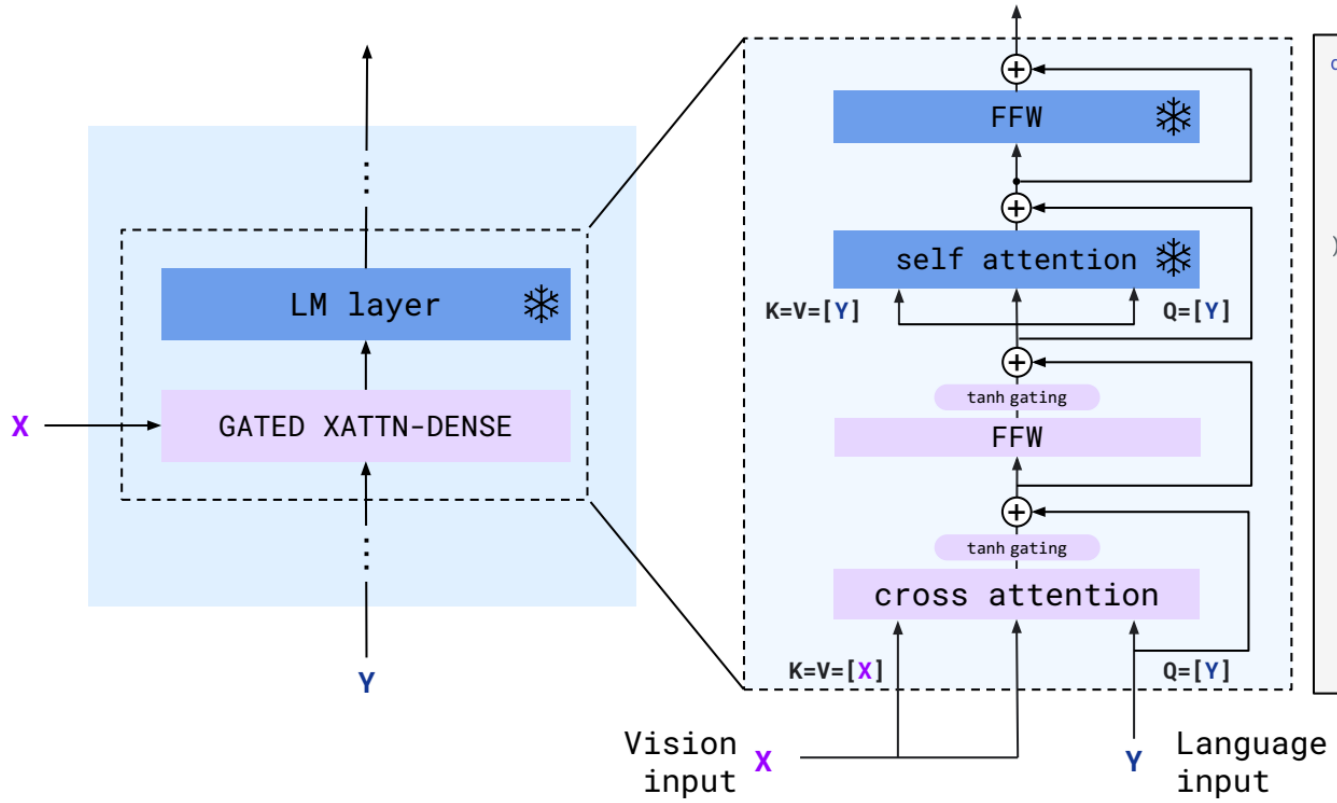
# Perceiver Resampler




```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
```

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

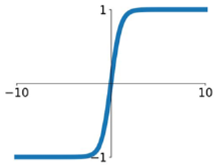
# Gated xattn-dense block



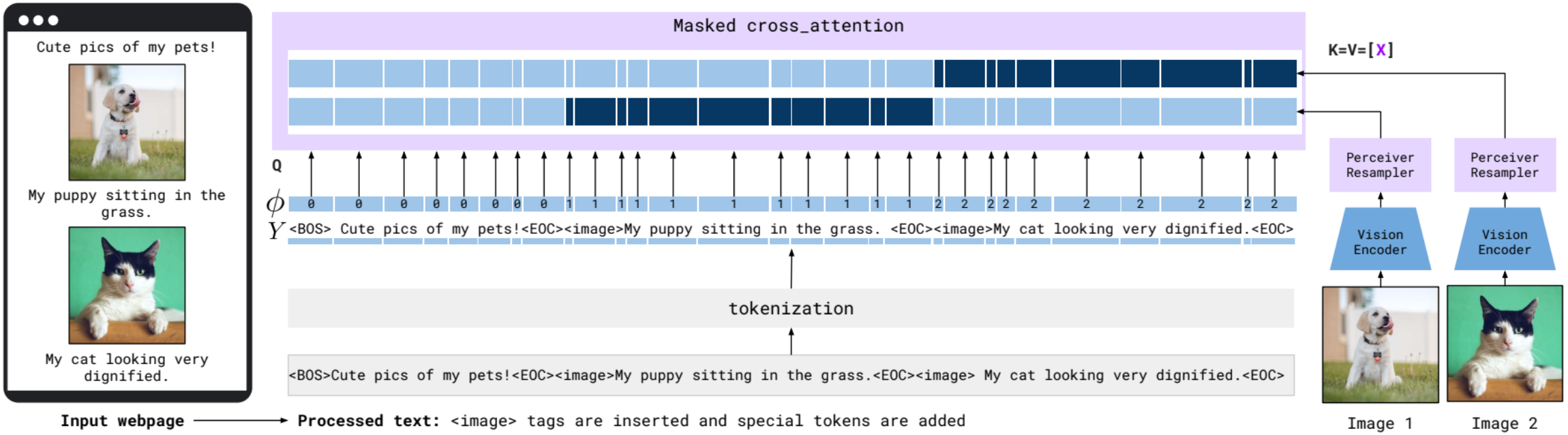

```
def gated_xattn_dense(
    y,  # input language features
    x,  # input visual features
    alpha_xattn, # xattn gating parameter – init at 0.
    alpha_dense, # ffw gating parameter – init at 0.
):
  """Applies a GATED XATTN-DENSE layer."""

  # 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
```

**tanh**

$\tanh(x)$



# Per-image/video attention masking



# Training data


Image-Text Pairs dataset
[N=1, T=1, H, W, C]

Video-Text Pairs dataset
[N=1, T>1, H, W, C]

Multi-Modal Massive Web (M3W) dataset
[N>1, T=1, H, W, C]

ALIGN：1.8B
LTIP：312M

VTP：27M

M3W：185M images
and 182GB of text

# Training objective

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

# Experiments

- Model

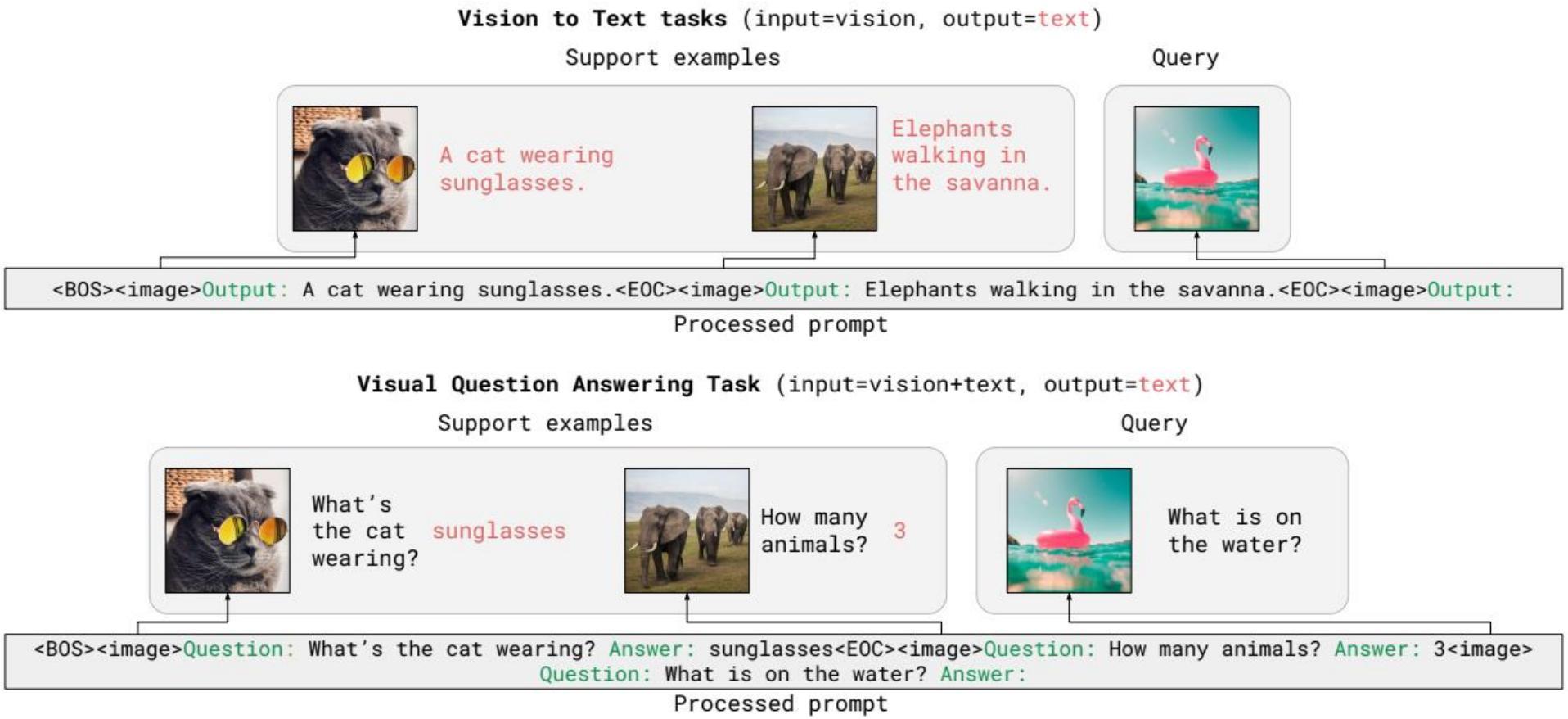
| | Requires model sharding | Frozen | | Trainable | | Total count |
|---|---|---|---|---|---|---|
| | | Language | Vision | GATED XATTN-DENSE | Resampler | |
| *Flamingo*-3B | ✗ | 1.4B | 435M | 1.2B (every) | 194M | **3.2B** |
| *Flamingo*-9B | ✗ | 7.1B | 435M | 1.6B (every 4th) | 194M | **9.3B** |
| *Flamingo* | ✓ | 70B | 435M | 10B (every 7th) | 194M | **80B** |

| | Resampler | | | | xattn dense | | | | Frozen LM | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | L | D | H | Act. | L | D | H | Act. | L | D | H | Act. |
| *Flamingo*-3B | 6 | 1536 | 16 | Sq. ReLU | 24 | 2048 | 16 | Sq. ReLU | 24 | 2048 | 16 | GeLU |
| *Flamingo*-9B | 6 | 1536 | 16 | Sq. ReLU | 10 | 4096 | 32 | Sq. ReLU | 40 | 4096 | 32 | GeLU |
| *Flamingo* | 6 | 1536 | 16 | Sq. ReLU | 12 | 8192 | 64 | Sq. ReLU | 80 | 8192 | 64 | GeLU |

1536 TPUv4 for 15 days

# Experiments

- Few-shot interleaved prompt generation.



# Experiments

- Few-shot: comparison to SotA

| Method | FT | Shot | OKVQA | VQAv2 | COCO | MSVDQA | VATEX | VizWiz | Flick30K | MSRVTTQA | iVQA | YouCook2 | STAR | VisDial | TextVQA | NextQA | HatefulMemes | RareAct |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Zero/Few shot SOTA | ✗ | (X) | [39] 43.3 (16) | [124] 38.2 (4) | [134] 32.2 (0) | [64] 35.2 (0) | - | - | - | [64] 19.2 (0) | [145] 12.2 (0) | - | [153] 39.4 (0) | [87] 11.6 (0) | - | - | [94] 66.1 (0) | [94] 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 | - |
| | ✗ | 8 | 44.6 | 55.4 | 90.6 | 37.0 | 54.5 | 38.4 | 71.7 | 19.6 | 36.8 | 68.0 | 40.6 | 47.6 | 32.4 | 23.9 | 54.7 | - |
| | ✗ | 16 | 45.6 | 56.7 | 95.4 | 40.2 | 57.1 | 43.3 | 73.4 | 23.4 | 37.4 | 73.2 | 40.1 | 47.5 | 31.8 | 25.2 | 55.3 | - |
| | ✗ | 32 | 45.9 | 57.1 | 99.0 | 42.6 | 59.2 | 45.5 | 71.2 | 25.6 | 37.7 | 76.7 | 41.6 | OOC | 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 | - |
| | ✗ | 8 | 50.0 | 58.0 | 99.0 | 40.8 | 55.2 | 39.4 | 73.4 | 23.9 | 40.0 | 75.0 | **43.4** | 51.2 | 33.6 | 25.8 | 63.9 | - |
| | ✗ | 16 | 50.8 | 59.4 | 102.2 | 44.5 | 58.5 | 43.0 | 72.7 | 27.6 | 41.5 | 77.2 | 42.4 | 51.3 | 33.5 | 27.6 | 64.5 | - |
| | ✗ | 32 | 51.0 | 60.4 | 106.3 | 47.2 | 57.4 | 44.0 | 72.8 | 29.4 | 40.7 | 77.3 | 41.2 | OOC | 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 | - |
| | ✗ | 8 | 57.5 | 65.6 | 108.8 | 45.5 | 60.6 | 44.8 | 78.2 | 27.6 | 44.8 | 80.7 | 42.3 | 56.4 | 37.3 | 32.3 | **70.0** | - |
| | ✗ | 16 | 57.8 | 66.8 | 110.5 | 48.4 | 62.8 | 48.4 | **78.9** | 30.0 | 45.2 | 84.2 | 41.1 | **56.8** | 37.6 | 32.9 | **70.0** | - |
| | ✗ | 32 | **57.8** | **67.6** | **113.8** | **52.3** | **65.1** | **49.8** | 75.4 | **31.0** | **45.3** | **86.8** | 42.2 | OOC | **37.9** | **33.5** | **70.0** | - |
| Pretrained FT SOTA | ✔ | (X) | 54.4 [39] (10K) | 80.2 [150] (444K) | 143.3 [134] (500K) | 47.9 [32] (27K) | 76.3 [165] (500K) | 57.2 [70] (20K) | 67.4 [162] (30K) | 46.8 [57] (130K) | 35.4 [145] (6K) | 138.7 [142] (10K) | 36.7 [138] (46K) | 75.2 [87] (123K) | 54.7 [147] (20K) | 25.2 [139] (38K) | 75.4 [60] (9K) | - |

# Experiments

- Fine-tuning Flamingo: Flamingo + base visual encoder

| Method | VQAV2 | | COCO | VATEX | VizWiz | | MSRVTTQA | VisDial | | YouCook2 | TextVQA | | HatefulMemes |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | test-dev | test-std | test | test | test-dev | test-std | test | valid | test-std | valid | valid | test-std | test seen |
| *Flamingo* - 32 shots | 67.6 | - | 113.8 | 65.1 | 49.8 | - | 31.0 | 56.8 | - | 86.8 | 36.0 | - | 70.0 |
| SimVLM [134] | 80.0 | 80.3 | **143.3** | - | - | - | - | - | - | - | - | - | - |
| OFA [129] | 79.9 | 80.0 | 149.6 | - | - | - | - | - | - | - | - | - | - |
| Florence [150] | 80.2 | 80.4 | - | - | - | - | - | - | - | - | - | - | - |
| *Flamingo* Fine-tuned | **82.0** | **82.1** | 138.1 | **84.2** | **65.7** | **65.4** | **47.4** | 61.8 | 59.7 | 118.6 | **57.1** | 54.1 | **86.6** |
| Restricted SotA† | 80.2 [150] | 80.4 [150] | **143.3** [134] | 76.3 [165] | - - | - - | 46.8 [57] | **75.2** [87] | **74.5** [87] | **138.7** [142] | 54.7 [147] | **73.7** [92] | 75.4 [60] |
| Unrestricted SotA | 81.3 [143] | 81.3 [143] | 149.6 [129] | 81.4 [165] | 57.2 [70] | 60.6 [70] | - - | - - | 75.4 [133] | - - | - - | - - | 84.6 [164] |

# Experiments

- Ablation studies

| | Ablated setting | Flamingo 3B value | Changed value | Param. count ↓ | Step time ↓ | COCO CIDEr↑ | OKVQA top1↑ | VQAv2 top1↑ | ImageNet top1↑ | MSVDQA top1↑ | VATEX CIDEr↑ | Kinetics top1-top5↑ | Overall score↑ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **Flamingo 3B model (short training)** | | | 3.2B | 1.74s | 86.5 | 42.1 | 55.8 | 59.9 | 36.3 | 53.4 | 49.4 | **68.4** |
| **(i)** | Training data | All data | M3W | 3.2B | 0.68s | 58.0 | 37.2 | 48.6 | 35.7 | 29.5 | 33.6 | 34.0 | 50.7 |
| | | | w/o VTP | 3.2B | 1.42s | 84.2 | 43.0 | 53.9 | 59.6 | 34.5 | 46.0 | 45.8 | 65.4 |
| | | | w/o LTIP/ALIGN | 3.2B | 0.95s | 66.3 | 39.2 | 51.6 | 41.4 | 32.0 | 41.6 | 38.2 | 56.5 |
| | | | w/o M3W | 3.2B | 1.02s | 54.1 | 36.5 | 52.7 | 24.9 | 31.4 | 23.5 | 28.3 | 46.9 |
| **(ii)** | Optimisation | Grad. accumulation | Round Robin | 3.2B | 1.68s | 76.1 | 39.8 | 52.1 | 50.7 | 33.2 | 40.8 | 39.7 | 59.7 |
| **(iii)** | Tanh gating | ✓ | ✗ | 3.2B | 1.74s | 78.4 | 40.5 | 52.9 | 54.0 | 35.9 | 47.5 | 46.4 | 64.0 |
| **(iv)** | Cross-attention architecture | GATED XATTN-DENSE | VANILLA XATTN | 2.4B | 1.16s | 80.6 | 41.5 | 53.4 | 59.0 | 32.9 | 50.7 | 46.8 | 65.2 |
| | | | GRAFTING | 3.3B | 1.74s | 79.2 | 36.1 | 50.8 | 47.5 | 32.2 | 47.8 | 27.9 | 57.4 |
| **(v)** | Cross-attention frequency | Every | Single in middle | 2.0B | 0.87s | 71.5 | 38.1 | 50.2 | 44.0 | 29.1 | 42.3 | 28.3 | 54.6 |
| | | | Every 4th | 2.3B | 1.02s | 82.3 | 42.7 | 55.1 | 57.1 | 34.6 | 50.8 | 45.5 | 65.9 |
| | | | Every 2nd | 2.6B | 1.24s | 83.7 | 41.0 | 55.8 | 59.6 | 34.5 | 49.7 | 47.4 | 66.2 |
| **(vi)** | Resampler | Perceiver | MLP | 3.2B | 1.85s | 78.6 | 42.2 | 54.7 | 53.6 | 35.2 | 44.7 | 42.1 | 63.3 |
| | | | Transformer | 3.2B | 1.81s | 83.2 | 41.7 | 55.6 | 59.0 | 31.5 | 48.3 | 47.4 | 65.1 |
| **(vii)** | Resampler size | Medium | Small | 3.1B | 1.58s | 81.1 | 40.4 | 54.1 | 60.2 | 36.0 | 50.2 | 48.9 | 66.4 |
| | | | Large | 3.4B | 1.87s | 84.4 | 42.2 | 54.4 | 60.4 | 35.1 | 51.4 | 49.4 | 67.3 |

# Experiments

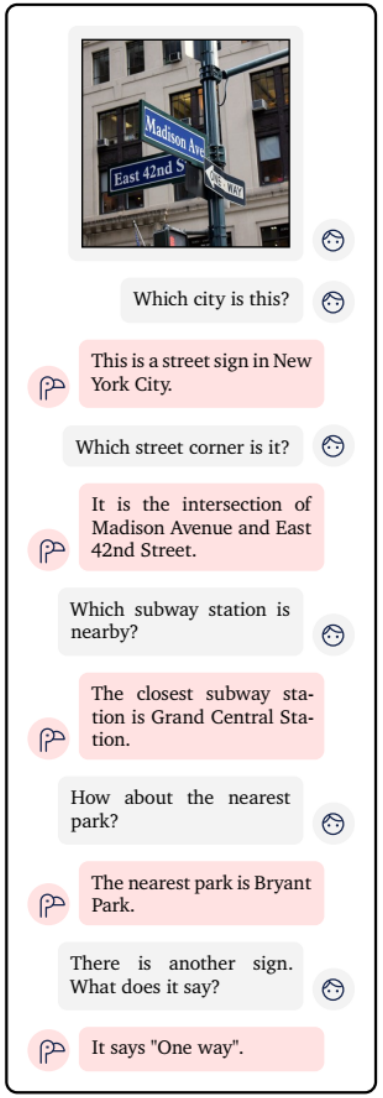
- Ablation studies

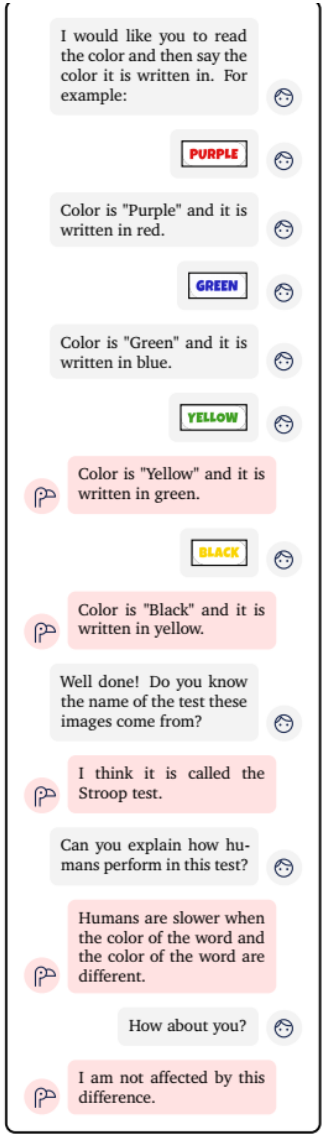
| | Ablated setting | Flamingo 3B value | Changed value | Param. count ↓ | Step time ↓ | COCO CIDEr↑ | OKVQA top1↑ | VQAv2 top1↑ | ImageNet top1↑ | MSVDQA top1↑ | VATEX CIDEr↑ | Kinetics top1-top5↑ | Overall score↑ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **Flamingo 3B model (short training)** | | | 3.2B | 1.74s | 86.5 | 42.1 | 55.8 | 59.9 | 36.3 | 53.4 | 49.4 | **68.4** |
| **(viii)** | Multi-Img att. | Only last | All previous | 3.2B | 1.74s | 70.0 | 40.9 | 52.0 | 52.3 | 32.1 | 46.8 | 42.0 | 60.8 |
| **(ix)** | $p_{next}$ | 0.5 | 0.0 | 3.2B | 1.74s | 85.0 | 41.6 | 55.2 | 60.3 | 36.7 | 50.6 | 49.9 | 67.8 |
| | | | 1.0 | 3.2B | 1.74s | 81.3 | 43.3 | 55.6 | 57.8 | 36.8 | 52.7 | 47.8 | 67.6 |
| **(x)** | Vision encoder | NFNet-F6 | CLIP ViT-L/14 | 3.1B | 1.58s | 76.5 | 41.6 | 53.4 | 49.5 | 33.2 | 44.5 | 42.3 | 61.4 |
| | | | NFNet-F0 | 2.9B | 1.45s | 73.8 | 40.5 | 52.8 | 49.8 | 31.1 | 42.9 | 36.6 | 58.9 |
| **(xi)** | LM pretraining | MassiveText | C4 | 3.2B | 1.74s | 81.3 | 34.4 | 47.1 | 60.6 | 30.9 | 53.9 | 46.9 | 62.5 |
| **(xii)** | Freezing Vision | ✓ | ✗ (random init) | 3.2B | 4.70s* | 74.5 | 41.6 | 52.7 | 45.2 | 31.4 | 35.8 | 32.6 | 56.6 |
| | | | ✗ (pretrained) | 3.2B | 4.70s* | 83.5 | 40.6 | 55.1 | 55.6 | 34.6 | 50.7 | 41.2 | 64.5 |
| **(xiii)** | Freezing LM | ✓ | ✗ (random init) | 3.2B | 2.42s | 74.8 | 31.5 | 45.6 | 59.5 | 26.9 | 50.1 | 43.4 | 58.2 |
| | | | ✗ (pretrained) | 3.2B | 2.42s | 81.2 | 33.7 | 47.4 | 60.7 | 31.0 | 53.9 | 49.9 | 62.9 |
| **(xiv)** | Co-train LM on MassiveText | ✗ | ✓ (random init) | 3.2B | 5.34s* | 69.3 | 29.9 | 46.1 | 59.9 | 28.1 | 45.5 | 46.9 | 57.4 |
| | | | ✓ (pretrained) | 3.2B | 5.34s* | 83.0 | 42.5 | 53.3 | 60.9 | 35.1 | 51.1 | 50.1 | 67.2 |

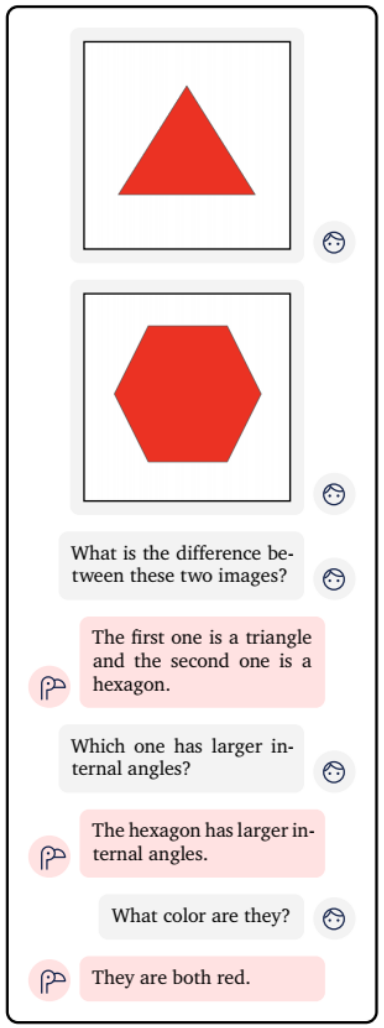
# Experiments

| | | | |
|---|---|---|---|
| Input Prompt |  Question: What do you think the capacities of these are? Answer: |  Question: What is odd about this image? Explain why it is unusual. Answer: |  Question: What country is this? Why do you think so? Answer: |
| Completion | **The floppy disk is 1.44MB and the CD is 700MB.** | **The image is odd because the elephant is in the back of the truck. It is unusual because elephants are not usually transported in the back of a truck.** | **It is Canada. I think so because the flag is the Canadian flag.** |

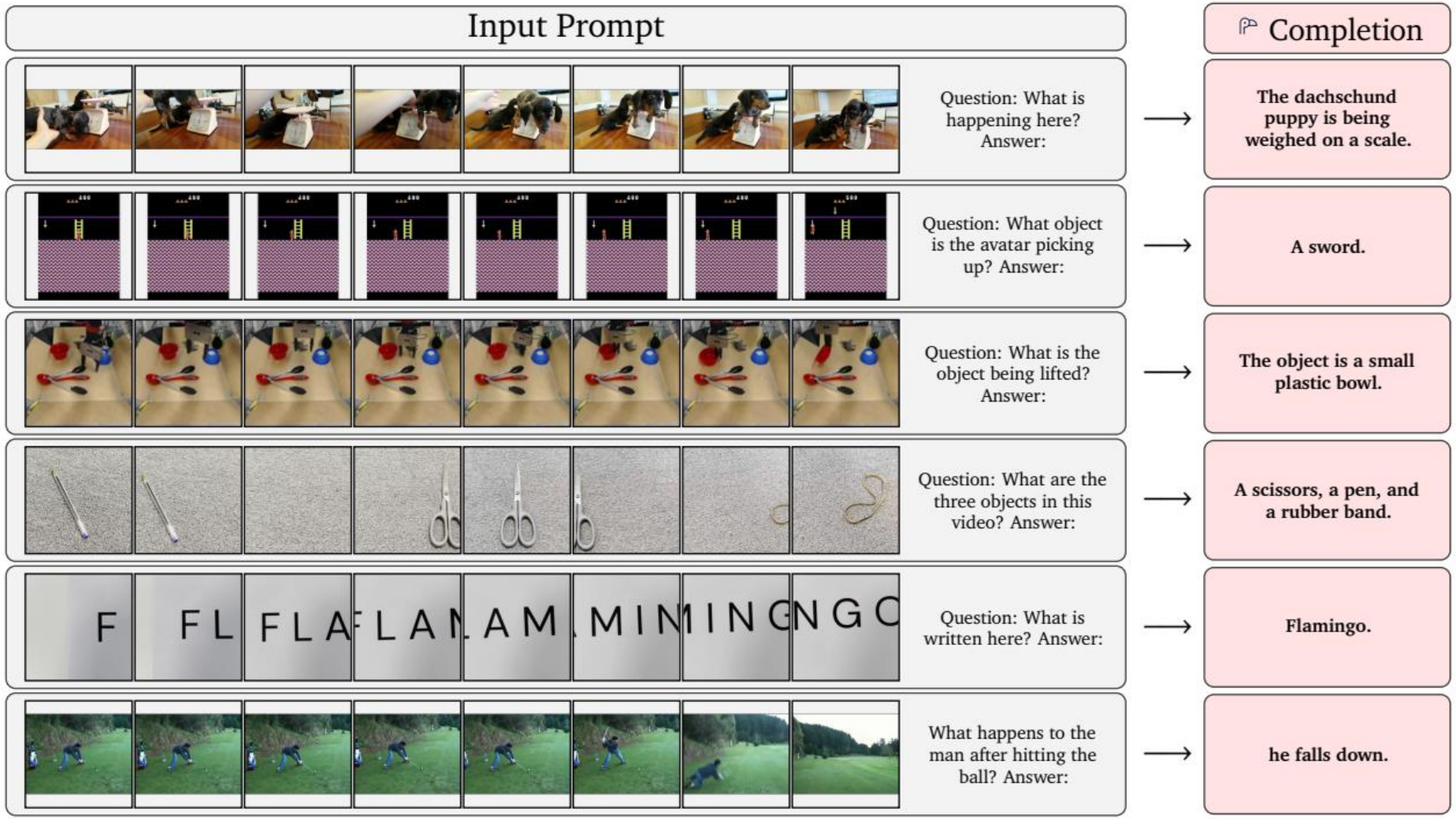
# Experiments

# Experiments



# Experiments

- Limitation on Classification performance.

| Model | Method | Prompt size | shots/class | ImageNet top 1 | Kinetics700 avg top1/5 |
|---|---|---|---|---|---|
| SotA | Fine-tuned | - | full | 91.0 [137] | 89.0 [144] |
| SotA | Contrastive | - | 0 | **85.7** [90] | **69.6** [94] |
| NFNetF6 | Our contrastive | - | 0 | 77.9 | 62.9 |
| *Flamingo*-3B | RICES | 8 | 1 | 70.9 | 55.9 |
| | | 16 | 1 | 71.0 | 56.9 |
| | | 16 | 5 | 72.7 | 58.3 |
| *Flamingo*-9B | RICES | 8 | 1 | 71.2 | 58.0 |
| | | 16 | 1 | 71.7 | 59.4 |
| | | 16 | 5 | 75.2 | 60.9 |
| *Flamingo*-80B | Random | 16 | $\leq 0.02$ | 66.4 | 51.2 |
| | RICES | 8 | 1 | 71.9 | 60.4 |
| | | 16 | 1 | 71.7 | 62.7 |
| | | 16 | 5 | 76.0 | 63.5 |
| | RICES+ensembling | 16 | 5 | 77.3 | 64.2 |

# Experiments

- Limitation

| | | | |
|---|---|---|---|
| Input Prompt |  Question: What is on the phone screen? Answer: |  Question: What can you see out the window? Answer: |  Question: Whom is the person texting? Answer: |
| Output | **A text message from a friend.** | **A parking lot.** | **The driver.** |