



Tencent AI Lab

Recent Advances of Multimodality on Text Generation

Qian Cao 07/28/2022

1. Background: VLP & VLMs
2. Prefix Language Models are Unified Modal Learners
3. Flamingo: a Visual Language Model for Few-Shot Learning
4. A Glance of Some Other Multi-modal Text Generation Tasks
5. Discuss

Background: VLP & VLMs

Why Multi-modal?

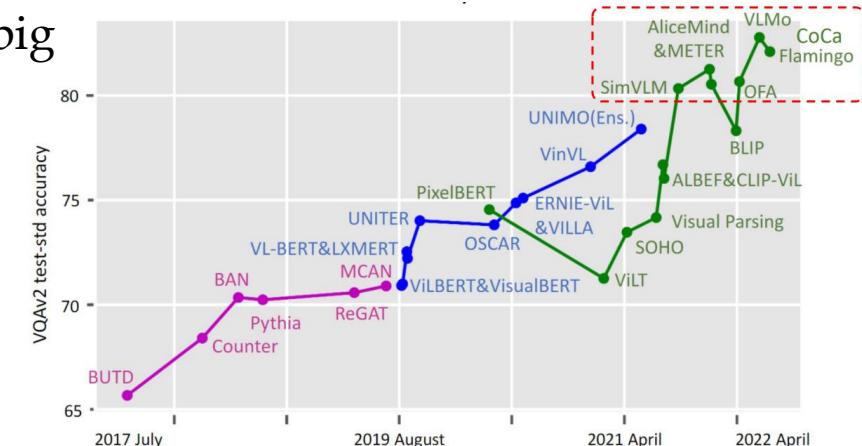
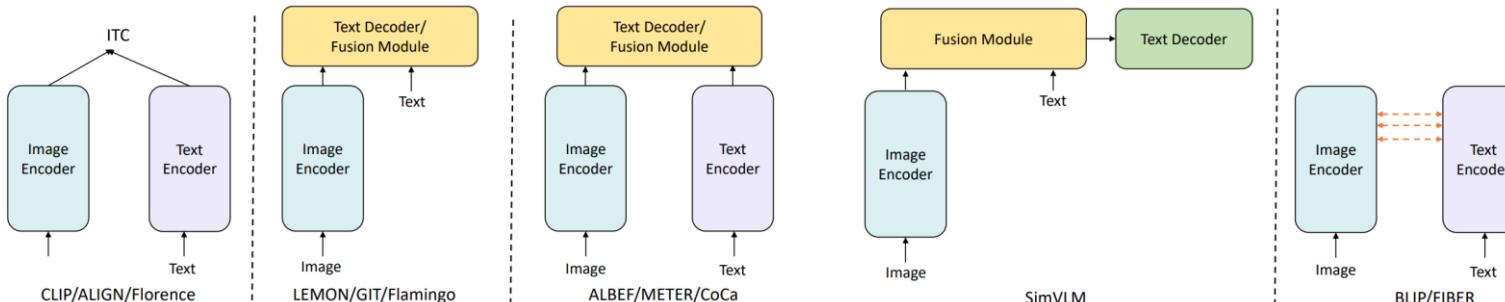
- Humans can align and fuse information collected from multiple channels, to better understand the world.

Most existing VLMs are BERT-like Transformer encoders pre-trained with a combination of different vision-language pre-training (VLP) objectives: *masked multi-modal modeling [ViLBert, UNITER, Oscar, etc.], multi-modal alignment prediction [ViLBert, UNITER, Oscar, etc.], region of interest feature regression [LXMERT, etc.], image-text matching [ALBEF, X-VLM, etc.]*

⇒ a transition pattern from encoder-only models to sequence-to-sequence models.

Trends:

- Network Architecture: from OD(object detector)-based to E2E(end-to-end).
- Tasks or architecture: Towards unified image-text modeling.
- Pre-training Data and model scale: from millions to billions, towards big foundation models.



Prefix Language Models are Unified Modal Learners



Prefix Language Models are Unified Modal Learners

Shizhe Diao*

The Hong Kong University of Science and Technology

sdiaoaa@connect.ust.hk

Wangchunshu Zhou

ByteDance AI Lab

wcszhou@outlook.com

Xinsong Zhang[†]

ByteDance AI Lab

zhangxinsong.0320@bytedance.com

Jiawei Wang

Shanghai Jiao Tong University

wjw_sjt@sjs.edu.cn

Prefix Language Models are Unified Modal Learners



Motivation

Current pre-training paradigm is either

- incapable of targeting all modalities at once (e.g., text generation and image generation), or
- requires multi-fold well-designed tasks which significantly limits the scalability.

The encoder-only architecture and complicated pre-training objectives of most current VLMs inevitably limit the potential towards pre-training more scalable and general VLMs.

Seq2Seq VLP: SOTA results, VL understanding ✓ & generation ✓

⇒ **hard to scale** [VL-T5, OFA]: non-trivial to collect a large number of VL datasets for pre-training.

⇒ capable of a **subset** of image-text modalities tasks [ERNIE-ViLG, SimVLM]: objectives not versatile enough.

Motivated by large-scale generative pre-training of prefix language models => **prefix multi-modal modeling**.

Prefix Language Models are Unified Modal Learners



DAVINCI

- **Model Architecture**

Textual Feature Embedding

$$T = \{t_1, t_2, \dots, t_i, \dots, t_n\}$$

$$t_i = \text{LayerNorm}(e_i + p_i),$$

Visual Feature Embedding

$$V = \{v_1, v_2, \dots, v_i, \dots, v_m\}$$

$$v_i = f_i + p_i,$$

Cross-Modal Transformer

$$X = \{x_1, x_2, \dots, x_l\} = [V, T] = \{v_1, v_2, \dots, v_m, t_1, t_2, \dots, t_n\}$$

Image Tokenizer and Decoder

an image $I \rightarrow$ a sequence of discrete visual tokens

$$Z = \{z_1, z_2, \dots, z_m\}$$

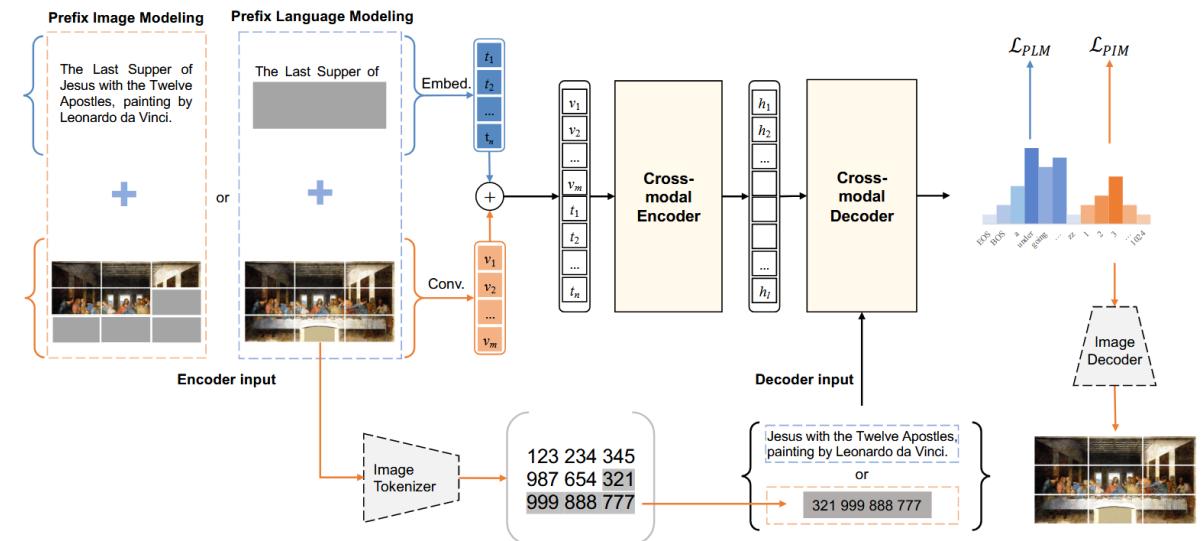


Figure 1: Illustration of the overall architecture and pre-training procedures of DAVINCI, a Transformer-based sequence-to-sequence model. Given an image-text pair, DAVINCI first splits either the word sequence or image token sequence into prefix and suffix. It then concatenates the prefix with the complete sequence in the other modality as input. DAVINCI is trained to recover the suffix with maximum likelihood estimation.

Prefix Language Models are Unified Modal Learners



DAVINCI

- **Pre-training Objectives** $\mathcal{L} = \mathcal{L}_{PLM} + \mathcal{L}_{PIM}$
 \Rightarrow Conduct LM with image supervision and IM with natural language supervision at the same time.

Prefix Language Modeling (PLM)

- a full image and a prefix caption \Rightarrow recover the masked textual tokens
- prefix length is randomly decided during training

$$\mathcal{L}_{PLM} = - \sum_{(I,S) \in D} \log p(\mathbf{Y}_{text} | \mathbf{X}_{image}, \tilde{\mathbf{X}}_{text})$$

prefix length 0: degenerate to “image captioning”

$$\mathcal{L}'_{PLM} = - \sum_{(I,S) \in D} \log p(\mathbf{Y}_{text} | \mathbf{X}_{image})$$

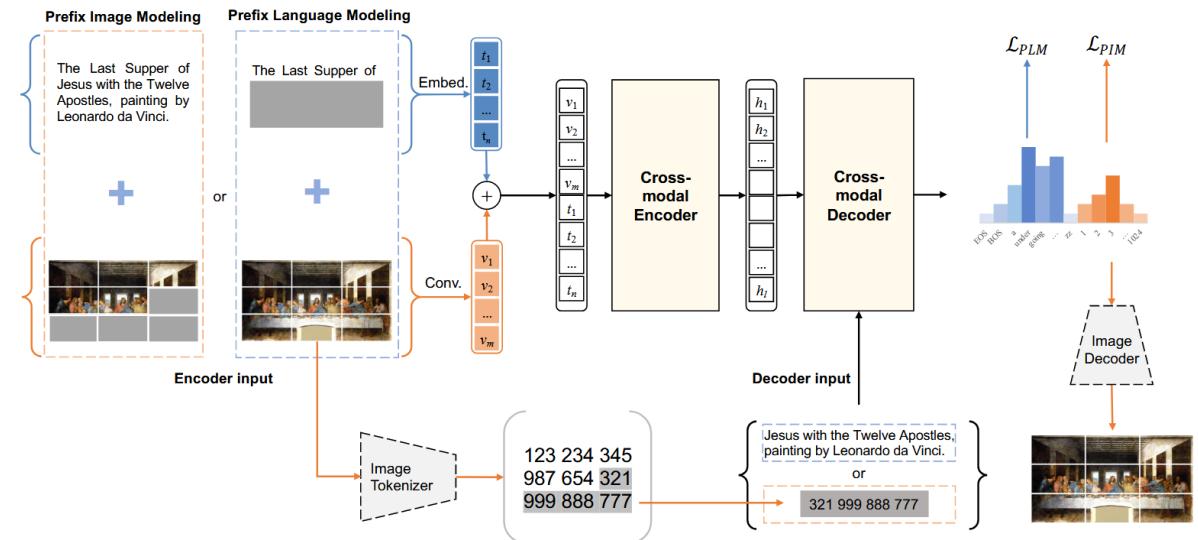


Figure 1: Illustration of the overall architecture and pre-training procedures of DAVINCI, a Transformer-based sequence-to-sequence model. Given an image-text pair, DAVINCI first splits either the word sequence or image token sequence into prefix and suffix. It then concatenates the prefix with the complete sequence in the other modality as input. DAVINCI is trained to recover the suffix with maximum likelihood estimation.

Prefix Language Models are Unified Modal Learners



DAVINCI

- **Pre-training Objectives** $\mathcal{L} = \mathcal{L}_{PLM} + \mathcal{L}_{PIM}$
 \Rightarrow Conduct LM with image supervision and IM with natural language supervision at the same time.

Prefix Image Modeling (PIM)

- a full caption and a corrupted (prefix) image => recover masked visual tokens (continuous image patches at the end, a.k.a., suffix image)

$$\mathcal{L}_{PIM} = - \sum_{(I,S) \in D} \log p(\mathbf{Y}_{\text{image}} | \mathbf{X}_{\text{text}}, \tilde{\mathbf{X}}_{\text{image}})$$

prefix length 0: degenerate to “text-to-image generation”

$$\mathcal{L}'_{PIM} = - \sum_{(I,S) \in D} \log p(\mathbf{Y}_{\text{image}} | \mathbf{X}_{\text{text}})$$

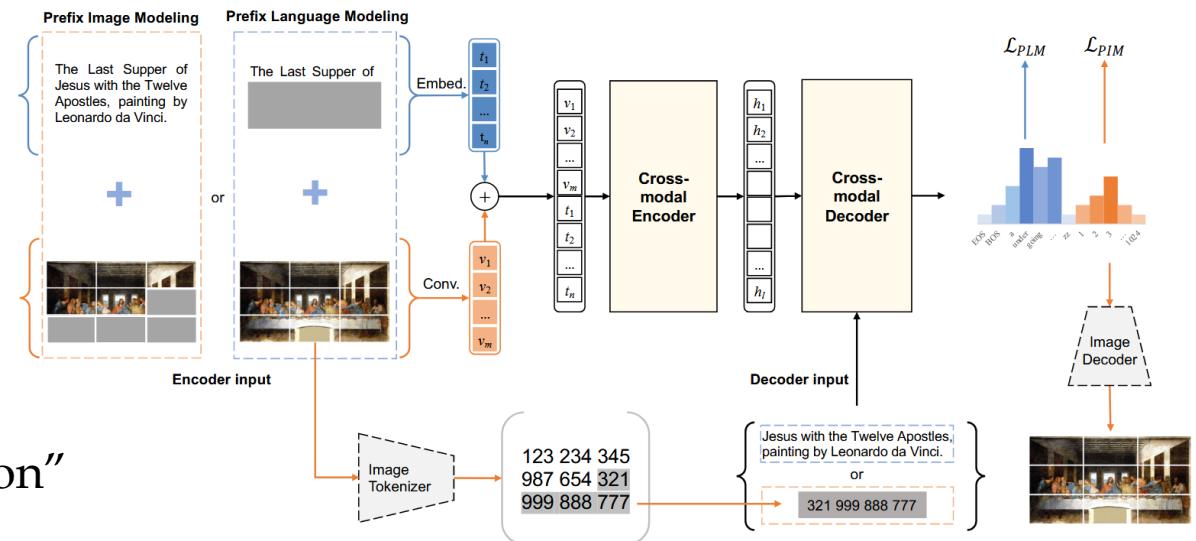


Figure 1: Illustration of the overall architecture and pre-training procedures of DAVINCI, a Transformer-based sequence-to-sequence model. Given an image-text pair, DAVINCI first splits either the word sequence or image token sequence into prefix and suffix. It then concatenates the prefix with the complete sequence in the other modality as input. DAVINCI is trained to recover the suffix with maximum likelihood estimation.

Prefix Language Models are Unified Modal Learners



Experiments

- **Pre-training Datasets**
- **Downstream Tasks**
 - Language Understanding: GLUE benchmark including MNLI, CoLA, MRPC, QQP, SST-2, QNLI, RTE, and STS-B.
 - Vision Understanding: ImageNet, Food101, CIFAR10, CIFAR100, Cars, Aircraft, DTD, Pets, Flowers102, MNIST, STL10, and Country211.
 - Multi-modal Understanding: VQAv2, SNLIVE and NLVR2.
 - Text-to-Image Generation: 30, 000 images sampled from COCO.
 - Image-to-Text Generation: COCO dataset.

Data Type	Dataset	Image Domain	#Images	#Captions	#Total
In-Domain Data (ID)	COCO	COCO	110.3K	551.7K	1.3M
	Visual Genome	COCO	108.2K	759.0K	
Small-scale Web Data (SWD)	SBU	Web	859.7K	859.7K	14.9M
	CC-3M	Web	2.9M	2.9M	
	CC-12M	Web	11.1M	11.1M	
Object-Region Data (ORD)	VG regions	COCO	108.2K	3.6M	17.0M
	VG objects	COCO	108.2K	925.6K	
	COCO objects	COCO	110.3K	736.6K	
	Refcoco	COCO	27.9K	589.9K	
	Open Image	Flickr	1.7M	7.5M	
	Obj365	Flickr	577.6K	3.6M	
Vision Data (VD)	ImageNet-21K	ImageNet	13.2M	13.2M	13.2M
Large-scale Web Data (LWD)	DAVINCI-200M	Web	205.6M	205.6M	601.3M
	LAION-400M	Web	395.7M	395.7M	
Text Data (TD)	C4	Web	–	–	800GB

Table 1: Statistics of the pre-training datasets. #Images, #Captions and #Total denote number of images, number of image-text pairs and the total number of image-text pairs, respectively.

Prefix Language Models are Unified Modal Learners



Experiments

- Experiment Results

		MIM 1	MLM 2	FLAVA 3	CLIP 4	SimVLM 5	DAVINCI 6	SimVLM 7	DAVINCI 8
Task	Eval method	70M	70M	70M	70M	46.4M	46.4M	647.7M	647.7M
MNLI	fine-tuning	—	73.23	80.33	32.85	82.13	82.25	83.27	83.13
CoLA	fine-tuning	—	39.55	50.65	11.02	52.47	52.10	54.22	54.75
MRPC	fine-tuning	—	73.24	84.16	68.74	82.70	83.14	84.26	84.54
QQP	fine-tuning	—	86.68	88.74	59.17	88.39	88.15	89.05	88.92
SST-2	fine-tuning	—	87.96	90.94	83.49	90.65	90.48	91.12	91.37
QNLI	fine-tuning	—	82.32	87.31	49.46	87.55	87.21	88.28	87.90
RTE	fine-tuning	—	50.54	57.76	53.07	59.80	60.72	63.34	64.22
STS-B	fine-tuning	—	78.89	85.67	13.70	86.62	86.27	87.24	87.05
NLP Avg.		—	71.55	78.19	46.44	78.79	78.79	80.10	80.23
ImageNet	linear eval	41.79	—	75.54	72.95	74.31	75.87	76.04	77.65
Food101	linear eval	53.30	—	88.51	85.49	83.41	89.33	85.52	90.12
CIFAR10	linear eval	76.20	—	92.87	91.25	91.56	93.01	92.41	93.96
CIFAR100	linear eval	55.57	—	77.68	74.40	72.51	78.98	75.23	80.11
Cars	linear eval	14.71	—	70.87	62.84	61.44	72.69	68.83	74.57
Aircraft	linear eval	13.83	—	47.31	40.02	41.28	47.42	47.75	49.55
DTD	linear eval	55.53	—	77.29	73.40	72.55	77.12	76.59	78.33
Pets	linear eval	34.48	—	84.82	79.61	78.77	85.52	86.13	88.21
Flowers102	linear eval	67.23	—	96.37	94.94	93.24	96.12	95.41	96.88
MNIST	linear eval	96.40	—	98.42	97.38	96.66	98.67	98.45	99.01
STL10	linear eval	80.12	—	98.89	97.29	97.51	99.03	98.02	99.21
Country211	linear eval	8.87	—	28.92	25.12	26.45	28.99	27.81	29.94
Vision Avg.		49.84	—	78.12	74.56	74.14	78.56	77.34	79.80
VQAv2	fine-tuning	—	—	72.49	59.81	72.12	73.89	75.03	76.44
SNLI-VE	fine-tuning	—	—	78.89	73.53	78.74	79.11	79.63	80.01
NLVR2	fine-tuning	—	—	—	—	77.45	77.91	79.72	80.25
I2T@B4	fine-tuning	—	—	—	—	38.00	38.50	38.10	39.20
I2T@C	fine-tuning	—	—	—	—	126.96	128.66	128.91	130.44
T2I@IS ↑	fine-tuning	—	—	—	—	—	17.55	—	22.41
T2I@FID ↓	fine-tuning	—	—	—	—	—	23.58	—	19.82
T2I@IS ↑	zero-shot	—	—	—	—	—	14.91	—	17.44
T2I@FID ↓	zero-shot	—	—	—	—	—	29.83	—	24.21
Multi-modal Avg.		—	—	—	—	78.65	79.61	80.28	81.27

Table 2: Experimental results on vision, language and multi-modal downstream tasks. MNLI results are average of MNLI-m and MNLI-mm. MRPC and QQP results are average of accuracy and F1. Matthews correlation coefficient (MCC) is reported for CoLA and Pearson correlation coefficient (PCC) is reported for STS-B. @B4, @C denote BLEU@4, CIDEr, respectively. For all other tasks we report accuracy. I2T and T2I denote image-to-text and text-to-image tasks. Multi-modal Avg. is the average score of VQAv2, SNLI-VE, NLVR2, I2T@B4 and I2T@C.

Model	#Params.	Text MNLI Acc	Vision ImageNet LE / FT	Image2Text COCO B@4 / C	Text2Image COCO IS↑ / FID↓	Encoder-only Multi-modal Models		
						Encoder-Decoder Multi-modal Models	Text2Image Models	Unified Models
VisualBERT [80]	170M	81.60	—	—	—	—	70.80 / 71.00	67.40 / 67.00
ViLBERT [18]	274M	79.90	—	—	—	—	70.55 / 70.92	—
VL-BERT [81]	170M	81.20	—	—	—	—	71.16 / —	—
LXMERT [19]	240M	80.40	—	—	—	—	72.42 / 72.54	74.90 / 74.50
UNITER [23]	155M	80.90	—	—	—	—	72.70 / 72.91	77.18 / 77.85
OSCAR [24]	155M	—	—	36.5 / 123.7	—	—	73.16 / 73.44	78.07 / 78.36
VinVL [36]	157M	—	—	38.2 / 129.3	—	—	75.95 / 76.12	82.05 / 83.08
ViLT [82]	88M	—	—	—	—	—	70.85 / —	74.91 / 75.57
ALBEF [25]	210M	—	—	—	—	—	75.84 / 76.04	82.55 / 83.14
UNIMO [83]	155M	—	—	38.8 / 124.4	—	—	73.79 / 74.02	—
X-VLM [26]	240M	—	—	39.6 / 132.6	—	—	78.22 / 78.37	84.41 / 84.76
VLMO [39]	—	—	—	—	—	—	76.64 / 76.89	82.77 / 83.34
UNICORN [84]	—	—	—	35.8 / 119.1	—	—	69.20 / 69.40	— / —
Uni-ENDN [85]	110M	—	—	—	—	—	72.20 / 72.50	— / —
Pixel-BERT [86]	144M	—	—	—	—	—	74.45 / 74.55	76.50 / 77.20
E2E-VLP [87]	94M	—	—	36.2 / 117.3	—	—	73.25 / 73.67	77.25 / 77.96
VL-T5 [28]	220M	—	—	34.5 / 116.5	—	—	/ 70.30	74.60 / 73.60
VL-BART [28]	220M	—	—	35.1 / 116.6	—	—	/ 71.30	71.70 / 70.30
AttnGAN [88]	—	—	—	—	23.30 / 35.20	— / —	— / —	— / —
DM-GAN [89]	—	—	—	—	32.20 / 26.50	— / —	— / —	— / —
DALLE [32] (250M)	12B	—	—	—	17.90 / 27.50	— / —	— / —	— / —
DALLE [32] (640M)†	82M	—	—	—	15.79 / 29.22	— / —	— / —	— / —
CogView [90]	4B	—	—	—	18.20 / 27.10	— / —	— / —	— / —
Unifying [91]	228M	—	—	37.3 / 122.6	— / 29.90	— / —	— / —	— / —
FLAVA [34]	240M	80.33	75.54 / —	—	—	72.80 / 72.49	— / —	— / —
SimVLM [30] (640M)†	153M	83.27	76.04 / —	38.5 / 128.7	—	75.04 / 75.03	78.82 / 79.72	— / —
SimVLM [30] (1.8B)	—	83.40	80.60 / —	39.0 / 134.8	—	77.87 / 78.14	81.72 / 81.77	— / —
OFA [31]	180M	84.30	— / 82.20	— / 135.6	21.50* / 20.80*	76.00 / —	— / —	— / —
DAVINCI	154M	83.13	78.81 / 83.92	39.2 / 130.4	17.44 (22.41*) / 24.21 (19.82*)	76.32 / 76.44	80.03 / 80.25	— / —

Table 3: Comparison with state-of-the-art vision-language models on vision, language and multi-modal downstream tasks. All results are from base-size models. LE and FT denote linear evaluation and fine-tuning performance, respectively. Image2Text results are reported without CIDEr optimization. † are our reproduced models. * are the results after fine-tuning. SimVLM (1.8B) and OFA are pre-trained with much larger corpus or human-labeled data of many downstream tasks, thus they are not comparable and labeled in gray. **bold** denotes the best across unified models.

Prefix Language Models are Unified Modal Learners



Analyses

- Impact of Pre-training Datasets

Settings	Pre-training Data					#Image	#Caption	Models	COCO Captions		VQA	SNLI-VE	NLVR2
	ID	SWD	ORD	VD	LWD				B@4 / C	Acc			
1	✓					0.2M	1.3M	SimVLM DAVINCI	35.2 / 115.06 35.8 / 117.30	68.89 69.25	76.10 76.22	71.21 72.55	
2	✓	✓				15.1M	16.2M	SimVLM DAVINCI	37.0 / 122.63 37.4 / 123.11	71.54 71.88	78.36 78.62	75.50 77.46	
3	✓		✓			2.7M	18.3M	SimVLM DAVINCI	38.2 / 123.85 38.0 / 124.20	69.57 70.02	76.65 76.92	70.50 72.01	
4	✓			✓		13.4M	14.5M	SimVLM DAVINCI	36.2 / 119.73 36.6 / 121.27	70.53 71.23	76.90 77.40	73.25 74.62	
5	✓	✓	✓	✓	✓	30.5M	46.4M	SimVLM DAVINCI	38.5 / 128.12 38.6 / 128.73	71.84 73.53	78.81 79.24	76.75 77.55	
6				✓		601.3M	601.3M	SimVLM DAVINCI	37.3 / 123.81 37.6 / 124.42	73.73 73.95	78.79 79.29	77.69 78.54	
7	✓				✓	601.5M	602.6M	SimVLM DAVINCI	37.9 / 125.50 38.1 / 125.91	74.64 74.91	79.05 79.22	77.68 78.12	
8	✓	✓	✓	✓	✓	631.8M	647.7M	SimVLM DAVINCI	38.5 / 128.25 39.1 / 130.21	75.04 76.32	79.32 80.04	78.82 80.03	

Table 4: Evaluation on downstream tasks using COCO Captions, VQA, SNLI-VE, and NLVR2. #Image and #Caption denote the numbers of images and image-text pairs that are used in the pre-training. Results are reported on the development set.

- Ablation Study

Method	COCO B@4 / C		SNLI-VE Acc	NLVR2 Acc
No Pre-training	32.1 / 96.71		54.23	51.08
Ours	35.8 / 117.30		76.22	72.55
– PLM	33.6 / 111.17		73.91	53.28
– PIM	34.3 / 116.58		75.79	69.78
– Text2Text	34.1 / 115.21		75.38	70.34

Table 5: Ablation study on COCO Captions, SNLI-VE and NLVR2. “–” denotes removing the corresponding objective. Results are reported on development set.

Conclusion

- (1) Introduce prefix multi-modal modeling, a simple unified generative vision-language pre-training framework that is **scalable** for large-scale pre-training and versatile for multiple modalities (vision, language, multi-modal) and tasks (understanding or generation).
- (2) Propose DAVINCI, a vision-language foundation model, and show that it performs competitively across tasks and modalities.
- (3) Conduct an analysis about the impact of different pre-training data sources on the performance of seq2seq VLMs.

Prefix Language Models are Unified Modal Learners



Visualization of Image Generation



Figure 2: Comparison with DALLE and OFA on text-to-image generation.



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

Motivation

- VLMs simply provides a similarity score between a text and an image, but they can only tackle limited use cases such as classification, where a finite set of outcomes is provided beforehand.
- Lack the ability to generate language, less suitable to more open-ended tasks.
- Not yet shown good performance in low data regimes.

Challenges of multimodal generative modelling

- Unifying strong single-modal models.
 - It is crucial to keep the pretrained model's language understanding and generation capabilities.
- Supporting both images and videos.
 - The 2D spatial structure and high dimensionality of images and videos is not immediately amenable to the homogeneous treatment as a 1D sequence commonly used in unimodal text generation.
- Obtaining heterogeneous training data to induce good generalist capabilities.
 - Paired image / caption datasets alone may not be general enough to induce few-shot learning and task induction capabilities like GPT-3.
 - The images and text are often only weakly related.

Flamingo: a Visual Language Model for Few-Shot Learning



Approach

- Visual processing and the Perceiver Resampler
- Conditioning a frozen language model on visual representations
- Multi-visual input support: per-image/video attention masking

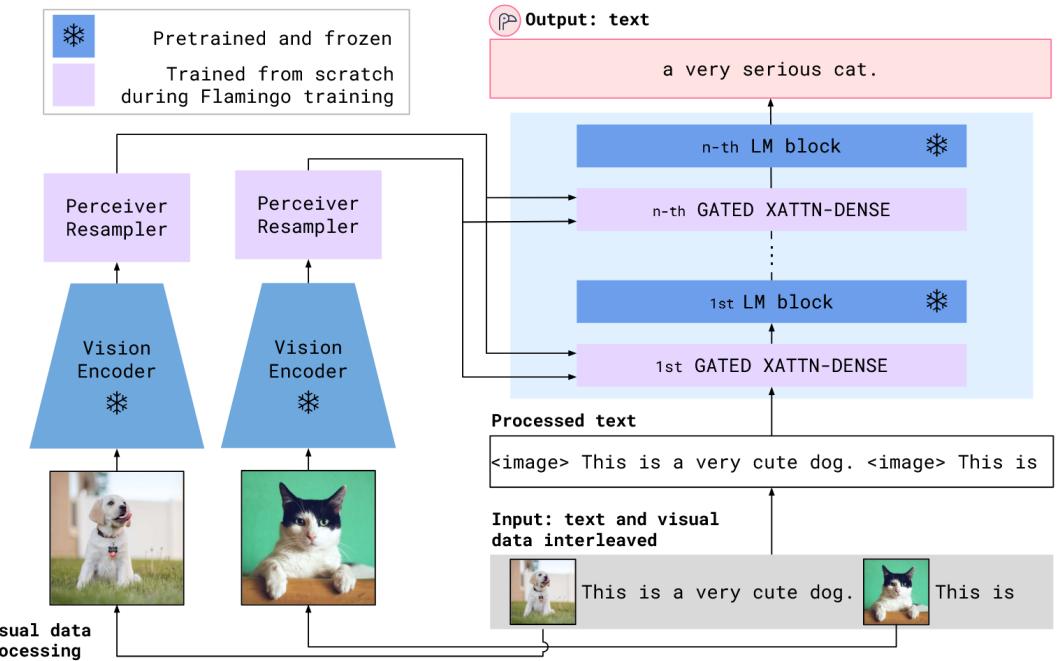


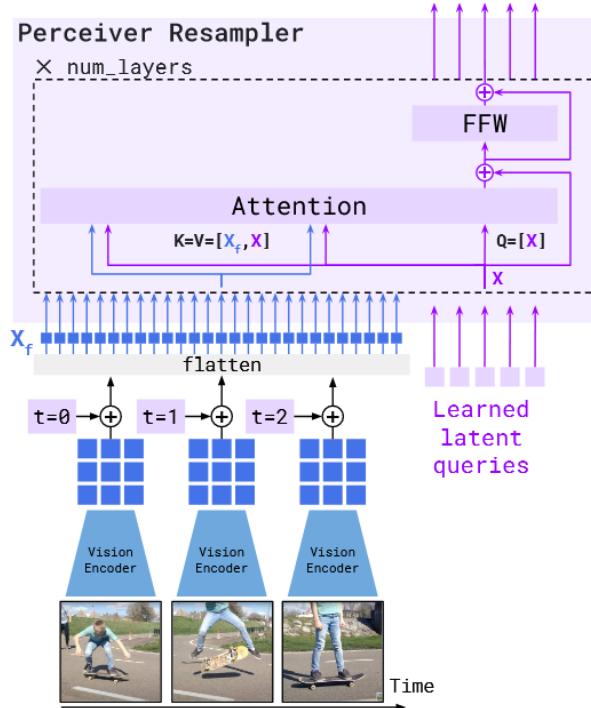
Figure 3 | Overview of the Flamingo model. The Flamingo models are a family of visual language model (VLM) that can take as input visual data interleaved with text and can produce free-form text as output. Key to its performance are novel architectural components and pretraining strategies described in Section 3.

Flamingo: a Visual Language Model for Few-Shot Learning



Approach

- **Visual processing and the Perceiver Resampler**
 - **Vision encoder:** from pixels to features.
 - Normalizer Free ResNet (NFNet).
 - Pre-trained & frozen.
 - **Perceiver Resampler:** from varying-size large feature maps to few visual tokens.
 - a variable number of image or video features => a fixed number of visual outputs.
 - learn a predefined number of latent input queries to cross attend to the flattened visual features.



```
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
```

Figure 4 | The Perceiver Resampler module maps a *variable size* grid of spatio-temporal visual features coming out of the Vision Encoder to a *fixed number* of output tokens (five in the figure), independently of the input image resolution or the number of input video frames. This transformer has a set of learned latent vectors as queries, and the keys and values are a concatenation of the spatio-temporal visual features with the learned latent vectors. More details can be found in Section 3.1.1.

Flamingo: a Visual Language Model for Few-Shot Learning



Approach

- Conditioning a frozen language model on visual representations
 - Interleaving new gated xattn-dense layers within a frozen pretrained LM
 - pretrained blocks from a text-only language model, + blocks trained from scratch that use the output of the Perceiver Resampler as one input.
 - A tanh-gating mechanism: It consists in multiplying the output of a newly added layers by $\tanh(\alpha)$ right before adding it to the input representation from the residual connection, where α is a layer-specific learnable scalar initialized at 0.

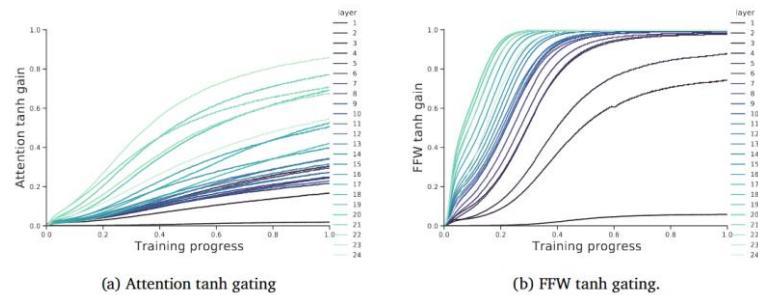


Figure 14 | Evolution of the absolute value of the tanh gating at different layers of *Flamingo-3B*.

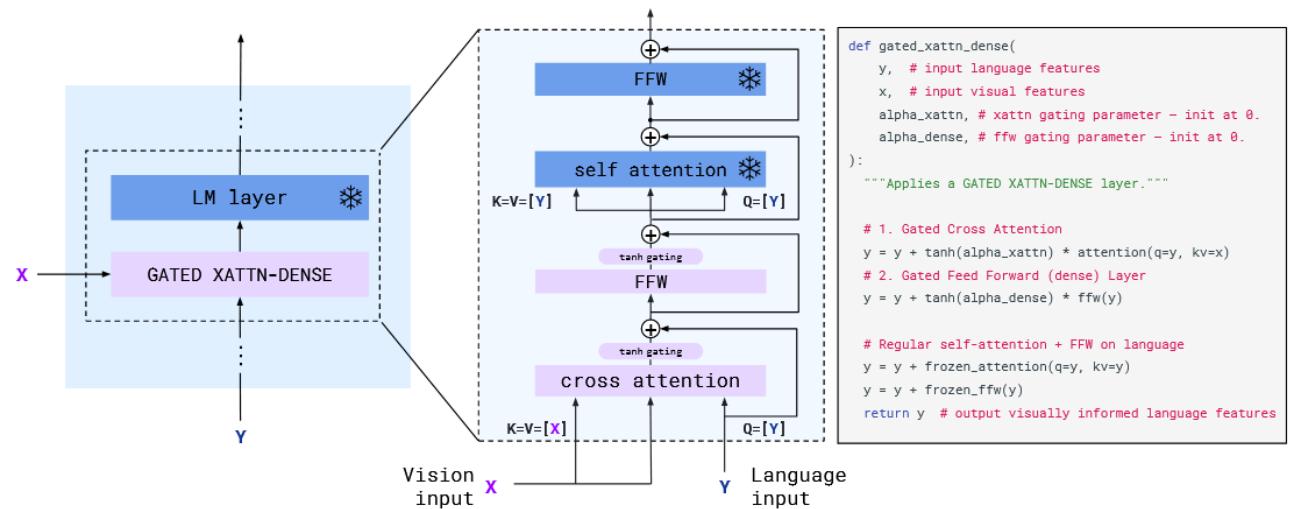


Figure 5 | GATED XATTN-DENSE layers. We insert new cross-attention layers, whose keys and values are obtained from the vision features while using language queries, followed by dense feed forward layers in between existing pretrained and frozen LM layers in order to condition the LM on visual inputs. These layers are gated so that the LM is kept intact at initialization for improved stability and performance.

Flamingo: a Visual Language Model for Few-Shot Learning

Approach

- **Multi-visual input support: per-image/video attention masking**
 - Interleaved sequence of visual data and text.
 - Multi-image attention.

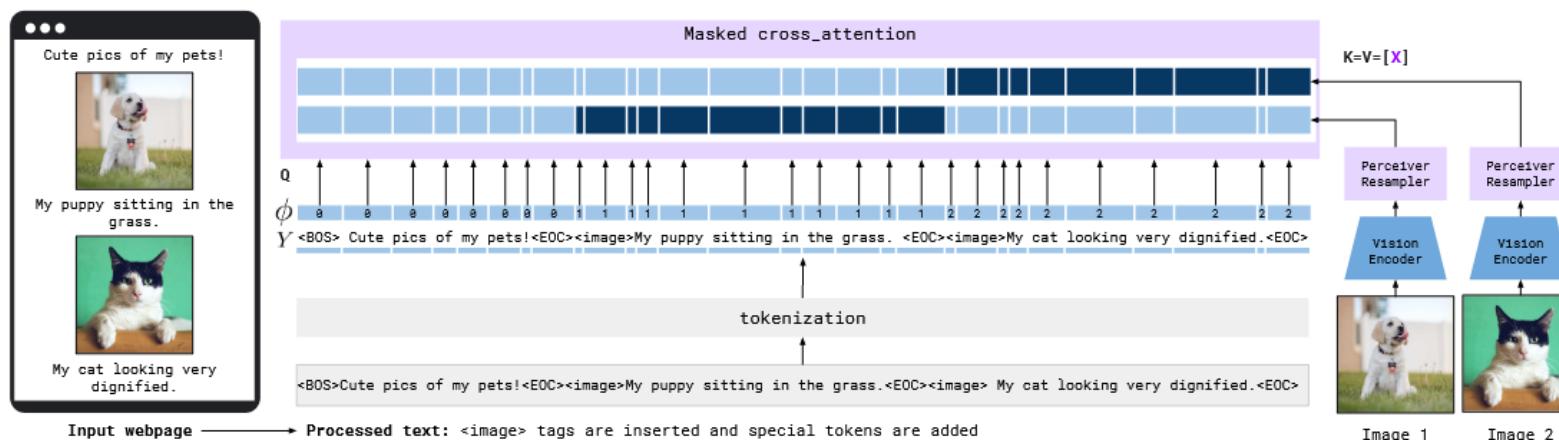


Figure 6 | Interleaved visual data and text support. Given text interleaved with images/videos, e.g. coming from a webpage, we first process the text by inserting `<image>` tags at the location of the visual data in the text as well as special tokens (`<BOS>` for “beginning of sentence” or `<EOC>` for “end of chunk”). The images are processed independently by the Vision Encoder and Perceiver Resampler to extract visual tokens. Following our modeling choice motivated in Section 3.1.3, each text token only cross-attends to the visual tokens corresponding to the last preceding image. The function ϕ illustrated above indicates for each token what is the index of the last preceding image (and 0 if there are no preceding images). In practice, this selective cross-attention is achieved via a masked cross attention mechanism – illustrated here with the dark blue entries (non masked) and light blue entries (masked).

Flamingo: a Visual Language Model for Few-Shot Learning



Approach

- Training on a mixture of vision and language datasets (Interleaved image and text)
- Training objective and optimization strategy
 - minimizing a weighted sum of dataset specific expected negative log likelihood of text given some visual inputs:

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

- Task adaptation with few-shot in-context learning

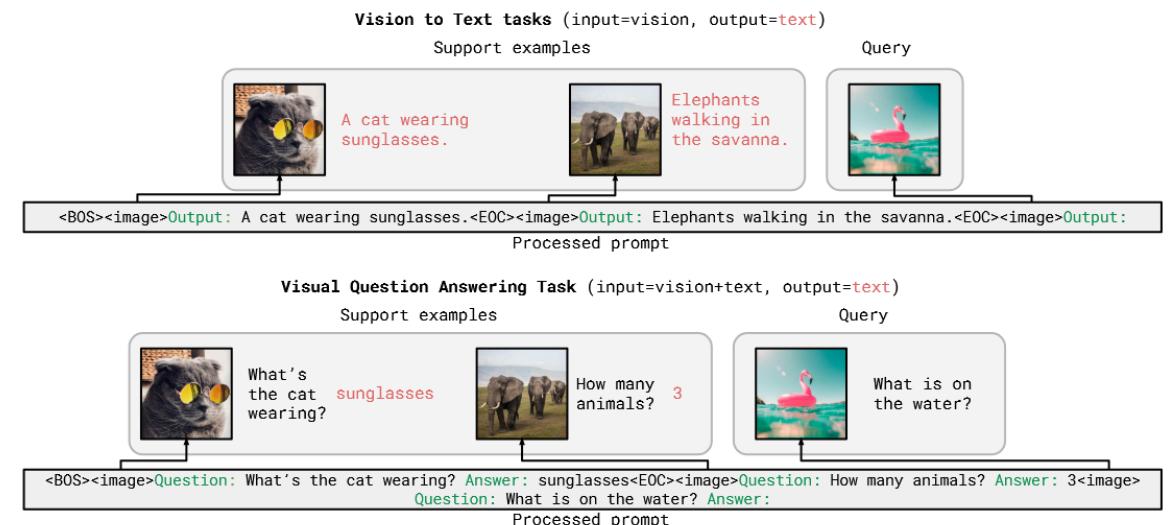


Figure 8 | Few-shot interleaved prompt generation. Given some task-specific few-shot examples (a.k.a. support examples) and a query for which Flamingo models have to make a prediction, we build the prompt by interleaving the image before each corresponding text. We introduce some formatting to do this, e.g. we prepend "**Output:**" to the expected response for all vision to text tasks or use a formatting prompt "**Question:** {question} **Answer:** {answer}" for visual question answering tasks.

Flamingo: a Visual Language Model for Few-Shot Learning



Experiments

Dataset	DEV	Gen.	Custom prompt	Task description	Eval set	Metric
ImageNet-1k [103]	✓			Object classification	Val	Top-1 acc.
MS-COCO [18]	✓	✓		Scene description	Test	CIDEr
VQAv2 [3]	✓	✓		Scene understanding QA	Test-dev	VQA acc. [3]
OKVQA [75]	✓	✓		External knowledge QA	Val	VQA acc. [3]
Flickr30k [149]		✓		Scene description	Test (Karpathy)	CIDEr
VizWiz [40]		✓		Scene understanding QA	Test-dev	VQA acc. [3]
TextVQA [108]		✓		Text reading QA	Val	VQA acc. [3]
VisDial [22]				Visual Dialogue	Val	NDCG
HatefulMemes [60]			✓	Meme classification	Seen Test	ROC AUC
Video						
Kinetics700 2020 [110]	✓			Action classification	Val	Top-1/5 avg
VATEX [132]	✓	✓		Event description	Test	CIDEr
MSVDQA [140]	✓	✓		Event understanding QA	Test	Top-1 acc.
YouCook2 [161]		✓		Event description	Val	CIDEr
MSRVTQA [140]		✓		Event understanding QA	Test	Top-1 acc.
iVQA [145]		✓		Event understanding QA	Test	iVQA acc. [145]
RareAct [81]			✓	Composite action retrieval	Test	mWAP
NextQA [139]	✓			Temporal/Causal QA	Test	WUPS
STAR [138]				Multiple-choice QA	Test	Top-1 acc.

Table 2 | Summary of the evaluation benchmarks. DEV benchmarks were used to validate general design decision of the Flamingo models. Gen. stands for generative task where we sample text from the VLM. If a task is non-generative it means that we use VLM to score answers among a given finite set. For most of our tasks we use a common default prompt, hence minimizing task-specific tuning (see Section 4.1.3).

	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

Table 1 | Parameter counts for Flamingo models. We focus on increasing the parameter count of the frozen LM and the trainable vision-text GATED XATTN-DENSE modules while maintaining the frozen vision encoder and trainable Resampler to a fixed and small size across the different models. The frequency of the GATED XATTN-DENSE with respect to the original language model blocks is given in parenthesis.

Flamingo: a Visual Language Model for Few-Shot Learning



Experiments

Method	FT	Shot	OKVQA	VQAv2	COCO	MSVDQA	VATEX	ViZWiz	Flick30K	MSRVTTQA	iVQA	YouCook2	STAR	VisDial	TextVQA	NextQA	HatefulMemes	RareAct
Zero/Few shot SOTA	✗	[39]	[124]	[134]	[64]					[64]	[145]		[153]	[87]			[94]	[94]
		(X)	43.3 (16)	38.2 (4)	32.2 (0)	35.2 (0)	-	-	-	19.2 (0)	12.2 (0)	-	39.4 (0)	11.6 (0)	-	-	66.1 (0)	40.7 (0)
<i>Flamingo-3B</i>	✗	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	-
<i>Flamingo-9B</i>	✗	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	-
<i>Flamingo</i>	✗	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	<u>60.8</u>
	✗	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	✓		54.4 (X)	80.2 (10K)	143.3 (444K)	47.9 (500K)	76.3 (27K)	57.2 (500K)	67.4 (20K)	46.8 (30K)	35.4 (130K)	138.7 (6K)	36.7 (10K)	75.2 (46K)	54.7 (123K)	25.2 (20K)	75.4 (38K)	- (9K)
			[39]	[150]	[134]	[32]	[165]	[70]	[162]	[57]	[145]	[142]	[138]	[87]	[147]	[139]	[60]	-

Table 3 | Comparison to the state of the art on multimodal benchmarks. A single Flamingo model reaches state-of-the-art on a wide array of image and video tasks with in-context learning from as few as 4 examples per task, beating previous zero-shot or few-shot method by a large margin. More importantly, using only 32 examples and without adapting any model weight, Flamingo *outperforms* the current best methods on 7 tasks, that are fine-tuned on thousands of annotated examples. Best few-shot numbers are in **bold**. Best numbers overall are underlined. See also Figure 2 that illustrate the table. OOC: out-of-context, which happens when the few-shot prompt is longer than the maximum sequence length the model has been trained on.

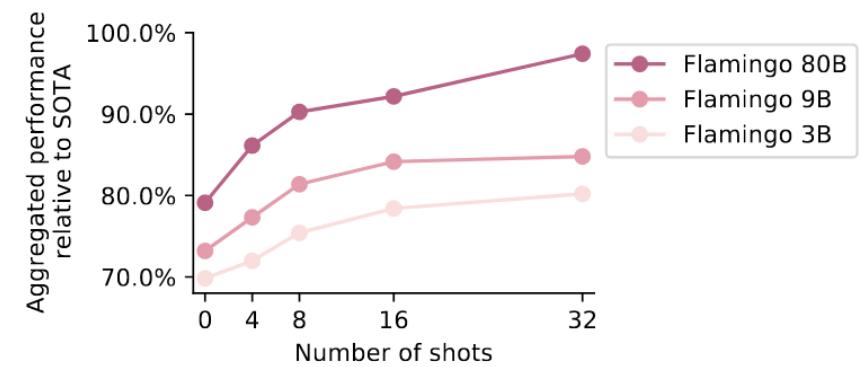
Flamingo: a Visual Language Model for Few-Shot Learning



Experiments

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	valid	valid	test-std	test	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	<u>143.3</u>	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-					
OFA [129]	79.9	80.0	<u>149.6</u>	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-					
Florence [150]	80.2	80.4	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-					
Flamingo Fine-tuned	<u>82.0</u>	<u>82.1</u>	138.1	<u>84.2</u>	<u>65.7</u>	<u>65.4</u>	<u>47.4</u>	61.8	59.7	118.6	<u>57.1</u>	54.1	<u>86.6</u>	-	-	-	-	-	-	-	-	-	-				
Restricted SotA [†]	80.2	80.4	<u>143.3</u>	76.3	-	-	46.8	<u>75.2</u>	74.5	<u>138.7</u>	54.7	<u>73.7</u>	75.4	[150]	[134]	[165]	[57]	[87]	[87]	[142]	[147]	[92]	[60]				
Unrestricted SotA	81.3	81.3	<u>149.6</u>	81.4	57.2	60.6	-	-	<u>75.4</u>	-	-	-	84.6	[143]	[143]	[129]	[165]	[70]	[70]	[133]	-	-	[164]				

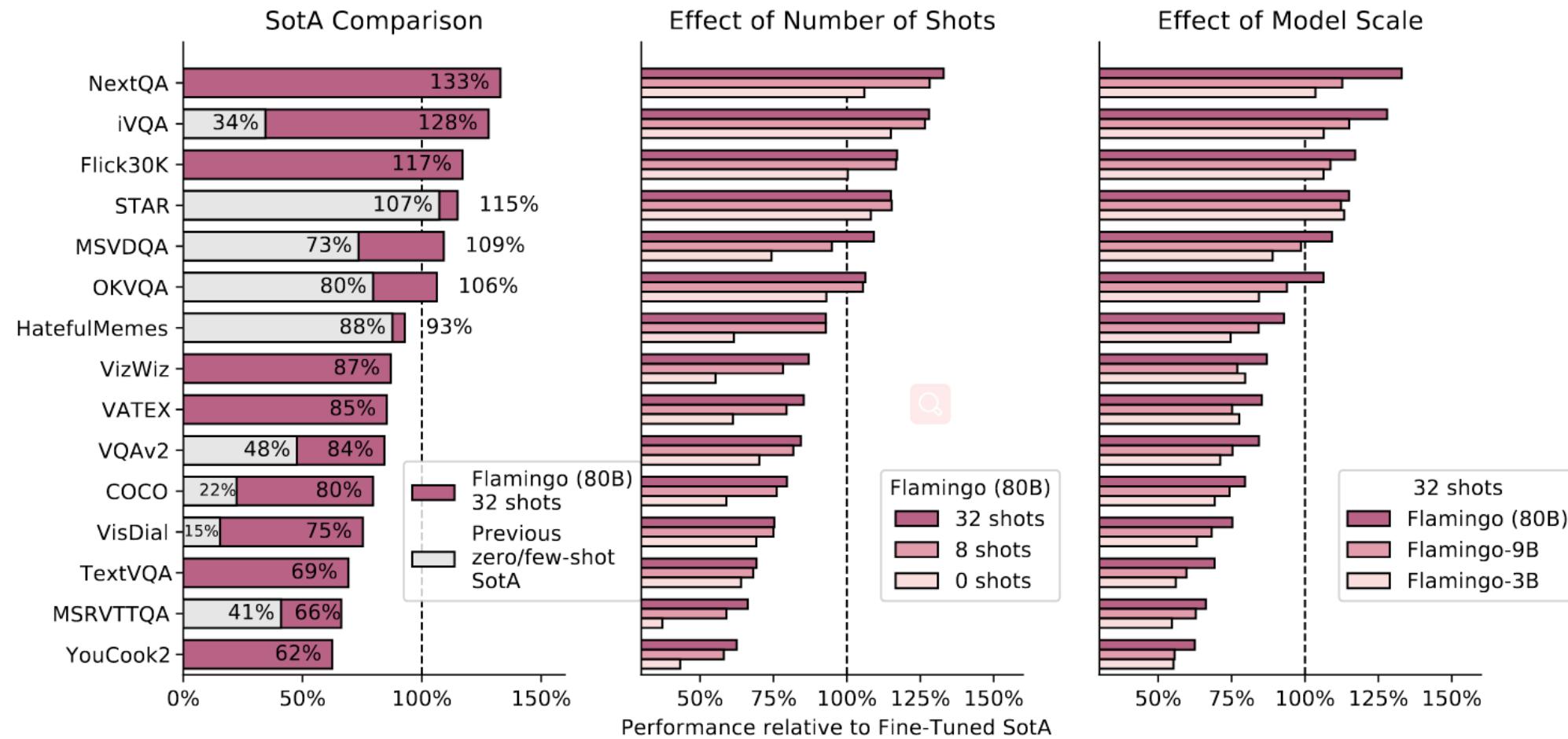
Table 6 | Comparison to SotA when fine-tuning Flamingo. We fine-tune Flamingo on all nine tasks where Flamingo was SotA overall with few-shot learning. Flamingo sets a new SotA on five of these tasks sometimes even beating methods that resorts to known performance optimization tricks such as model ensembling (on VQAv2, VATEX, VizWiz and HatefulMemes). Best numbers among the restricted SotA are in **bold**. Best numbers overall are underlined. Restricted SotA[†]: only includes methods that use a single model (not ensembles) and do not directly optimise the test metric (no CIDEr optimisation).



Flamingo: a Visual Language Model for Few-Shot Learning



Experiments



Flamingo: a Visual Language Model for Few-Shot Learning



Examples

 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?

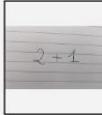
It looks like a computer.

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.

Input Prompt				→	Completion
	This is a chinchilla. They are mainly found in Chile.		This is a shiba. They are very popular in Japan.		This is a flamingo. They are found in the Caribbean and South America.
	What is the title of this painting? Answer: The Hallucinogenic Toreador.		Where is this painting displayed? Answer: Louvre Museum, Paris.		What is the name of the city where this was painted? Answer: Arles.
	Output: "Underground"		Output: "Congress"		Output: "Soulomes"
	$2+1=3$		$5+6=11$		$3 \times 6 = 18$
	Output: A propaganda poster depicting a cat dressed as French emperor Napoleon holding a piece of cheese.		Output: A pink room with a flamingo pool float.		A portrait of Salvador Dali with a robot head.
	Les sanglots longs des violons de l'automne blessent mon cœur d'une langueur monotone.		Pour qui sont ces serpents qui sifflent sur vos têtes?		Je suis un cœur qui bat pour vous.
	pandas: 3		dogs: 2		giraffes: 4
I like reading		, my favourite play is Hamlet. I also like		, my favorite book is	Dreams from my Father.
	What happens to the man after hitting the ball? Answer:				he falls down.

 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.

 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.

Conclusion

- A novel architecture for accepting arbitrarily interleaved visual data and text as input and generating output text in an open-ended manner.
- Architectural innovations and training strategies that effectively leverage large pretrained vision-only and language-only models, preserving the benefits of these initial models while efficiently fusing the modalities.
- Efficient ways to adapt to visual inputs of varying size, making Flamingo applicable to images and videos.
- Flamingo sets a new state of the art in few-shot learning on a wide array of 16 multimodal language and image/video understanding tasks.

A Glance of Some Other Multi-modal Text Generation Tasks



Rough Overview

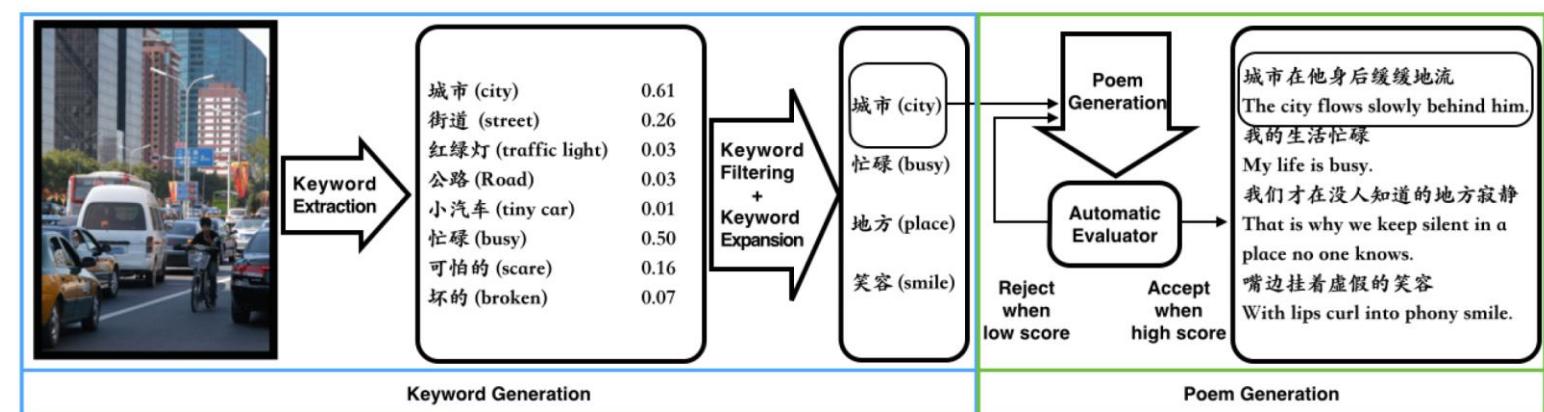
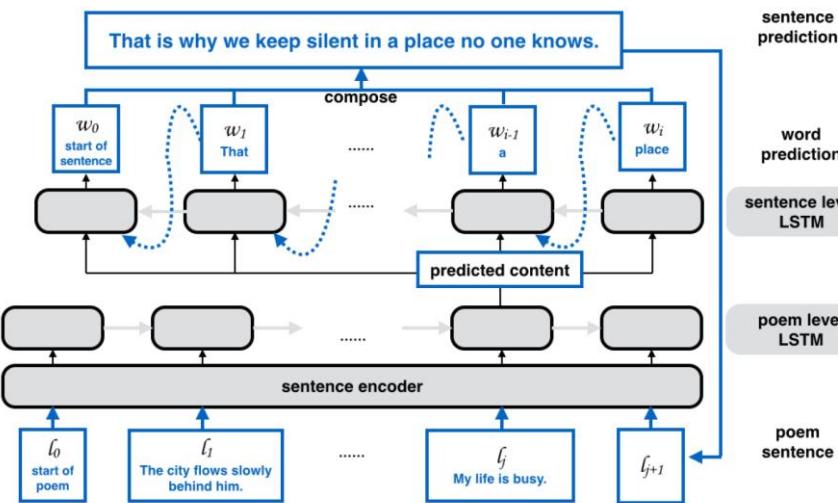
- Description: Image Caption, Event/Scene Description, Visual Storytelling, etc.
- Creation: Image Inspired Poem Generation, etc.
- Dialogue: VQA, Image/Video -grounded Dialogue, AVSD, etc.
- Others: Multi-modal Summarization, Multi-modal Translation, etc.

A Glance of Some Other Multi-modal Text Generation Tasks



Image Inspired Poem Generation

- Image Inspired Poetry Generation in XiaoIce

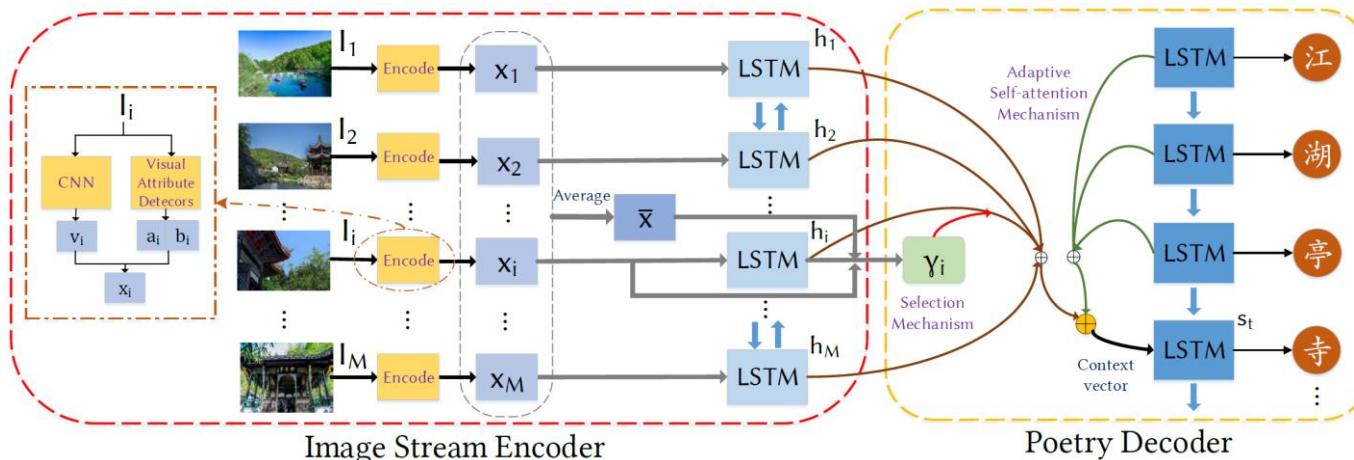
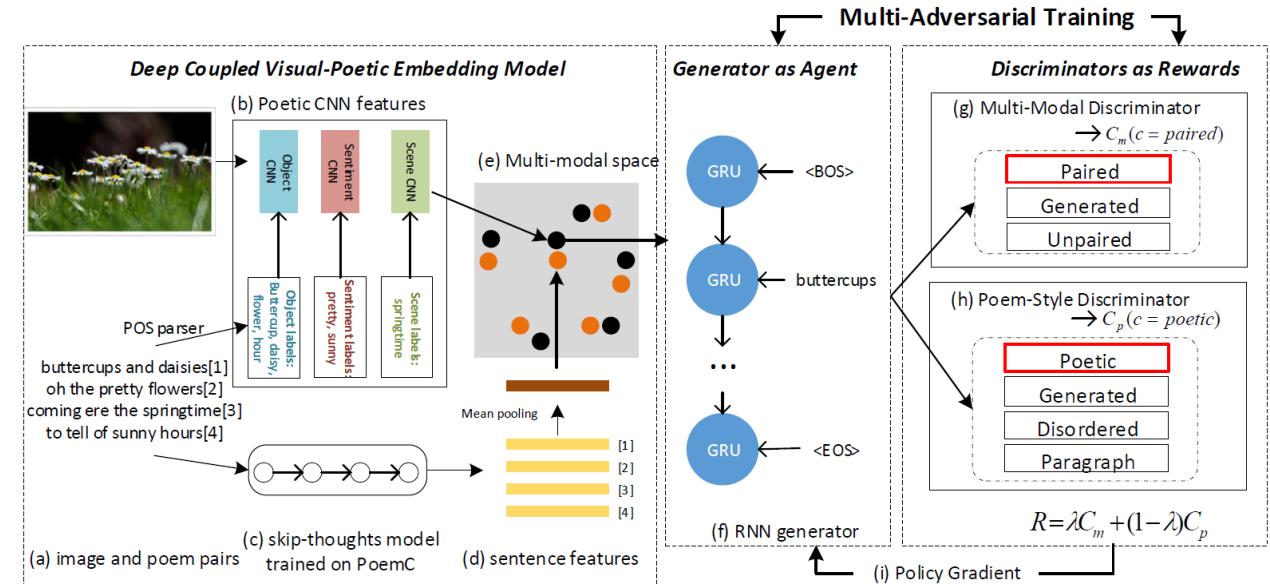


A Glance of Some Other Multi-modal Text Generation Tasks



Image Inspired Poem Generation

- Beyond Narrative Description: Generating Poetry from Images by Multi-Adversarial Training.
- Images2poem: Generating Chinese poetry from image streams.



Bei Liu, Jianlong Fu, Makoto P. Kato, and Masatoshi Yoshikawa. 2018. Beyond Narrative Description: Generating Poetry from Images by Multi-Adversarial Training. CoRRabs/1804.08473 (2018). arXiv:1804.08473 <http://arxiv.org/abs/1804.08473>

Lixin Liu, Xiaojun Wan, and Zongming Guo. 2018. Images2poem: Generating Chinese poetry from image streams. In Proceedings of the 26th ACM international conference on Multimedia. 1967–1975.

A Glance of Some Other Multi-modal Text Generation Tasks



Image Inspired Poem Generation

- Multi-Modal Experience Inspired AI Creation

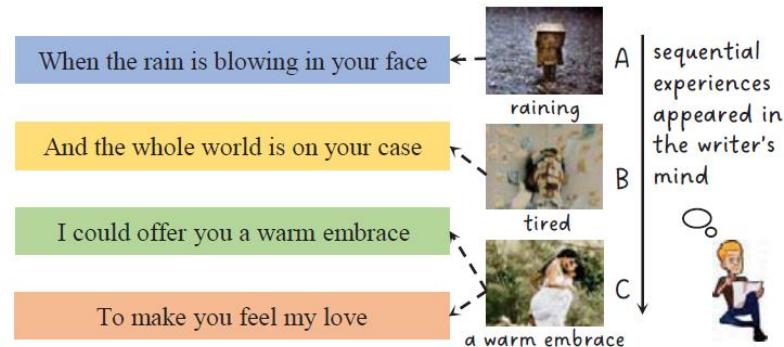


Figure 1: A toy example of the human creation process. The inputs and outputs are sequentially corresponded in a loose manner, that is, each input may influence multiple outputs.

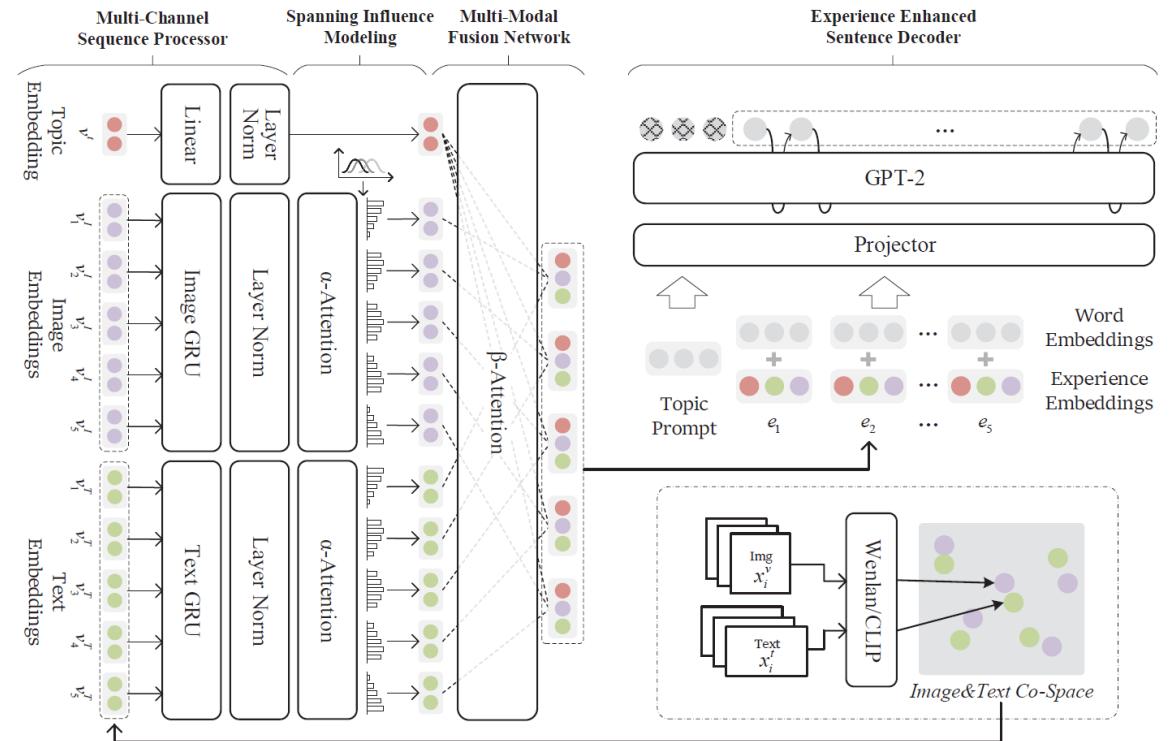


Figure 2: The framework of our proposed MMTG model. Experiences are shown in image and text sequences. An image corresponds to its text at the same time step. The modules of Multi-Channel Sequence Processor, Spanning Influence Modeling, Multi-Modal Fusion Network, and Experience Enhanced Sentence Decoder are presented from left to right.

A Glance of Some Other Multi-modal Text Generation Tasks



Multimodal Dialogue

- **Image Caption => VQA (Visual: image & video)**
- **Visual Dialog (VisDial)**
 - [Visual Dialog], [DMRM], [Are You Talking to Me?]=>GAN+RL, [FlipDial], [Improving Cross-Modal Understanding in VisDial], [VD-BERT], [VU-BERT]
- **Image-Grounded Conversations (IGC)**
 - [Image-Grounded Conversations:...]=>QG+RG, [Image-Chat], [MMChat], [MM Open-Domain Dialogue]
- **Visual Context**
 - [OpenViDial 1.0/2.0], [Modeling Text-visual Mutual Dependency 4 MDG]
 - Multimodal Emotion Recognition (MER): [Emotion-Aware Multimodal Pre-training...], [M3ED], [MELD], [MSCTD], [Modality-Transferable Emotion Embeddings...]
 - MDS (Multimodal Dialogue System) (multi-images, in retail): [MDS: Generating Responses via Adaptive Decoders], [MDS via Capturing Context-aware Dependencies...], [A non-hierarchical attention network...], [Towards Building Large Scale...]
- **Audio Visual Scene-Aware Dialog (AVSD) [aka. Video Dialog / Video-Grounded Dialogue Systems (VGDS)]**
 - [Audio Visual Scene-Aware Dialog], [AVSD Generation with Transformer-based Video Representations], [Bridging Text and Video], [Dynamic Graph Representation Learning for Video Dialog], [End-to-End AVSD using...], [Multimodal Transformer Networks for End-to-End...], [Video-Grounded Dialogues with PGLMs], [VX2TEXT]
- **Multimodal Response**
 - [An animated picture says at least 1k words], [Multimodal Dialogue Response Generation], [PhotoChat], [Towards Expressive Communication with Internet Memes]
- **Misc (Special setting)**
 - VQA on Game: [GuessWhat?!], [Learning Cooperative Visual Dialog Agents with Deep RL]
 - Construct Visual Latency: [Open Domain Dialogue Generation with Latent Images], [Text is NOT Enough], [Maria]
 - Live Comments: [LiveBot], [Response to LiveBot]

A Glance of Some Other Multi-modal Text Generation Tasks



Multimodal Dialogue



VQA

Q: How many people on wheelchairs ?
A: Two

Q: How many wheelchairs ?
A: One

Captioning

Two people are in a wheelchair and one is holding a racket.



Visual Dialog

Q: How many people are on wheelchairs ?
A: Two
Q: What are their genders ?
A: One male and one female
Q: Which one is holding a racket ?
A: Playing a Wii game
Q: Is that a man to her right ?
A: No, it's a woman

Q: What is the gender of the one in the white shirt ?
A: She is a woman
Q: What is she doing ?
A: Playing a Wii game
Q: Is that a man to her right ?
A: No, it's a woman



Place near my house is getting ready for Halloween a little early.

Don't you think Halloween should be year-round, though?

That'd be fun since it's my favorite holiday!

It's my favorite holiday as well!

I never got around to carving a pumpkin last year even though I bought one.

Well, it's a good thing that they are starting to sell them early this year!



Is the photo in color?

Yes

Is the photo close up?

No

Is this at a farm?

Possibly

Do you think it's for Halloween?

That is possible

Do you see anyone?

No

Do you see trees?

No

Any huge pumpkins?

No

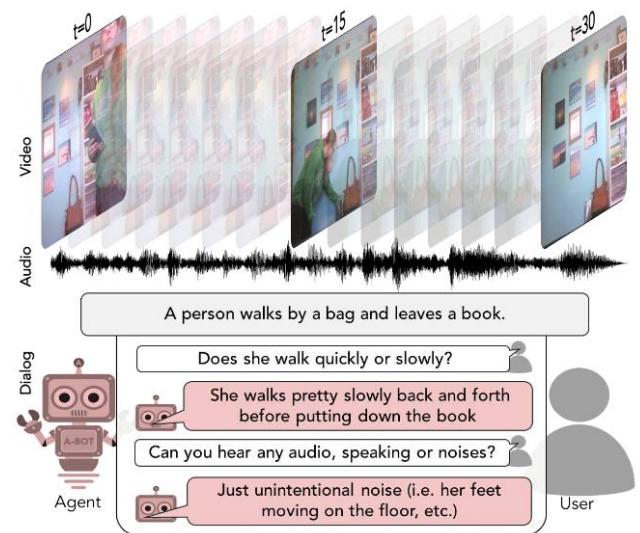


Figure 2: Differences between image captioning, Visual Question Answering (VQA) and Visual Dialog. Two (partial) dialogs are shown from our VisDial dataset, which is curated from a live chat between two Amazon Mechanical Turk workers (Sec. 3).



Context: Go! Lock!



Context: Officer down. Officer down.



NV: No, no, no, no.

CV: Get out of the way!

FV: Hey, hey, hey!

FV+MI: I'm on the phone!

Truth: I need an ambulance at the Girard Street subway.

Figure 2: Typical crowdsourced conversations in IGC (left) and VisDial (right).



A Glance of Some Other Multi-modal Text Generation Tasks



Multimodal Dialogue

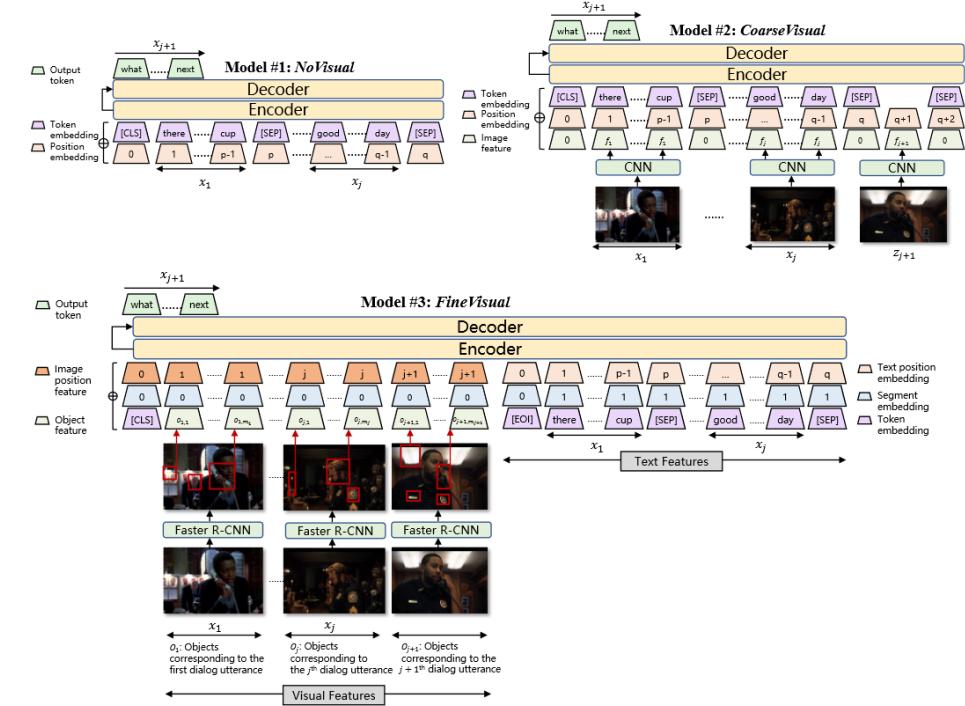
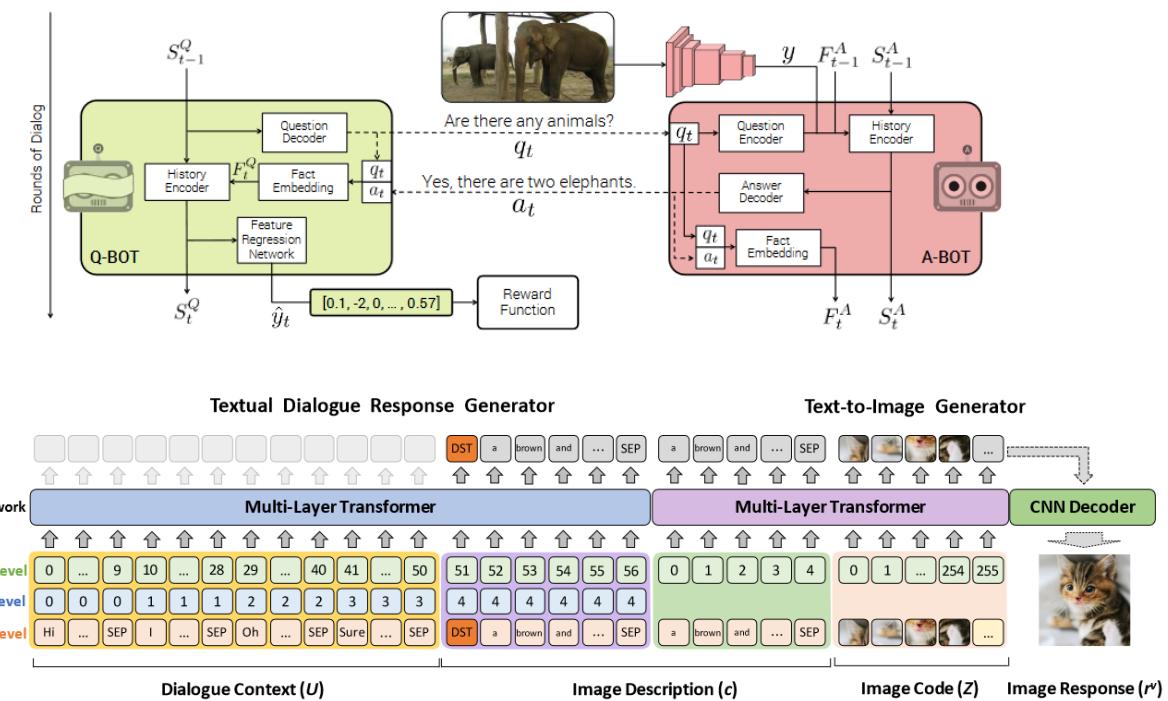


Figure 1: An overview of the proposed models NoVisual, CoarseVisual and FineVisual.

Das, A., Kottur, S., Moura, J.M., Lee, S. and Batra, D., 2017. Learning cooperative visual dialog agents with deep reinforcement learning. In Proceedings of the IEEE international conference on computer vision (pp. 2951-2960).

Sun, Q., Wang, Y., Xu, C., Zheng, K., Yang, Y., Hu, H., Xu, F., Zhang, J., Geng, X. and Jiang, D., 2021. Multimodal dialogue response generation. arXiv preprint arXiv:2110.08515.

Wang, S., Meng, Y., Sun, X., Wu, F., Ouyang, R., Yan, R., Zhang, T. and Li, J., 2021. Modeling Text-visual Mutual Dependency for Multi-modal Dialog Generation. arXiv preprint arXiv:2105.14445.

- Multi-modality really helps. => But how and how much?
- X-Attention, Fusion at different levels, Visual/Textual Prompts => Is there a better way to integrate multi-modal information for text generation?
- VLP models are strong, benefited from both huge data and model size. => How to better use VLP models to help text generation, especially domain specific tasks?

Thanks!