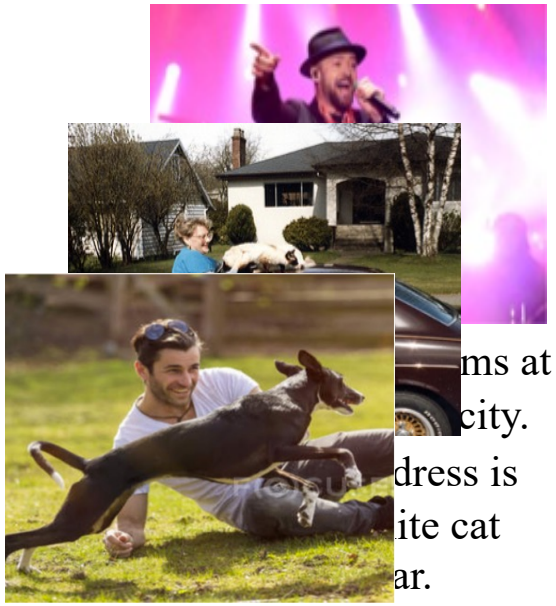
Visual-Language Models

CS6420

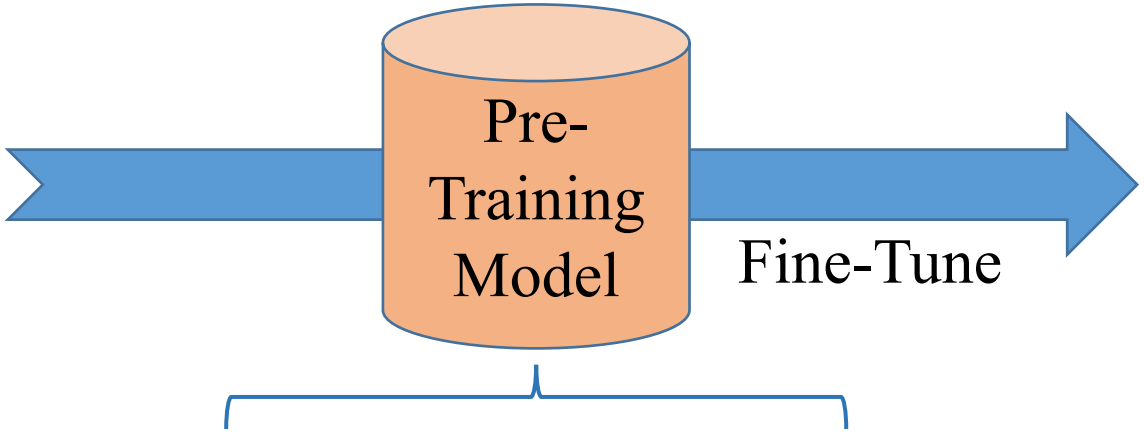
Wei Ji

What is Visual-Language Model?



ms at
city.
dress is
ite cat
ar.

Portrait of man playing with his dog on a meadow.

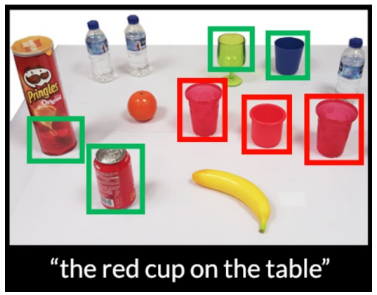


Pre-training Tasks

- Image-Text Matching
- Image-Text Contrastive Learning
- Masked Language Modeling
- ...

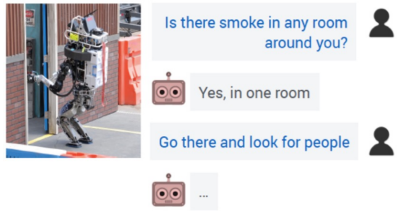
What color are her eyes?
What is the mustache made of?

Visual Question Answering



Visual Grounding

.....

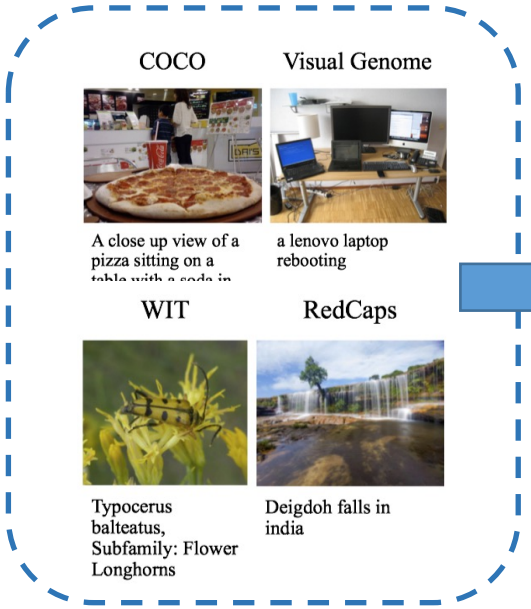


Visual Dialog

- Pretraining with image-text pairs in proxy tasks to learn common cross-modal representation
- Transfer to downstreaming cross-modal reasoning task after finetuning

Why We Need Visual-language Model?

- Human brain can deal with multi-modal information at the same time , studies on brain science indicate human brain exists consistent recognition mechanism about visual concept and linguistic concept
- Large amounts of image-text pairs on the Internet
- Large pretrained model based on multi-modal data can provide common cross-modal representation for downstreaming tasks



Visual Question Answering

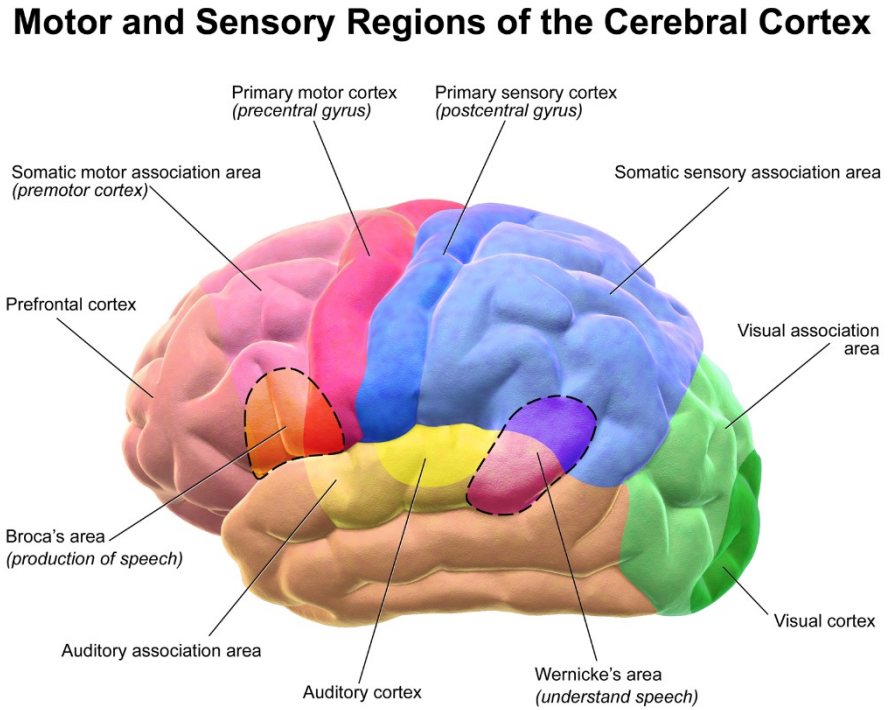
Image-Text Retrieval

Image Captioning

Visual Grounding

.....

Visual Entailment

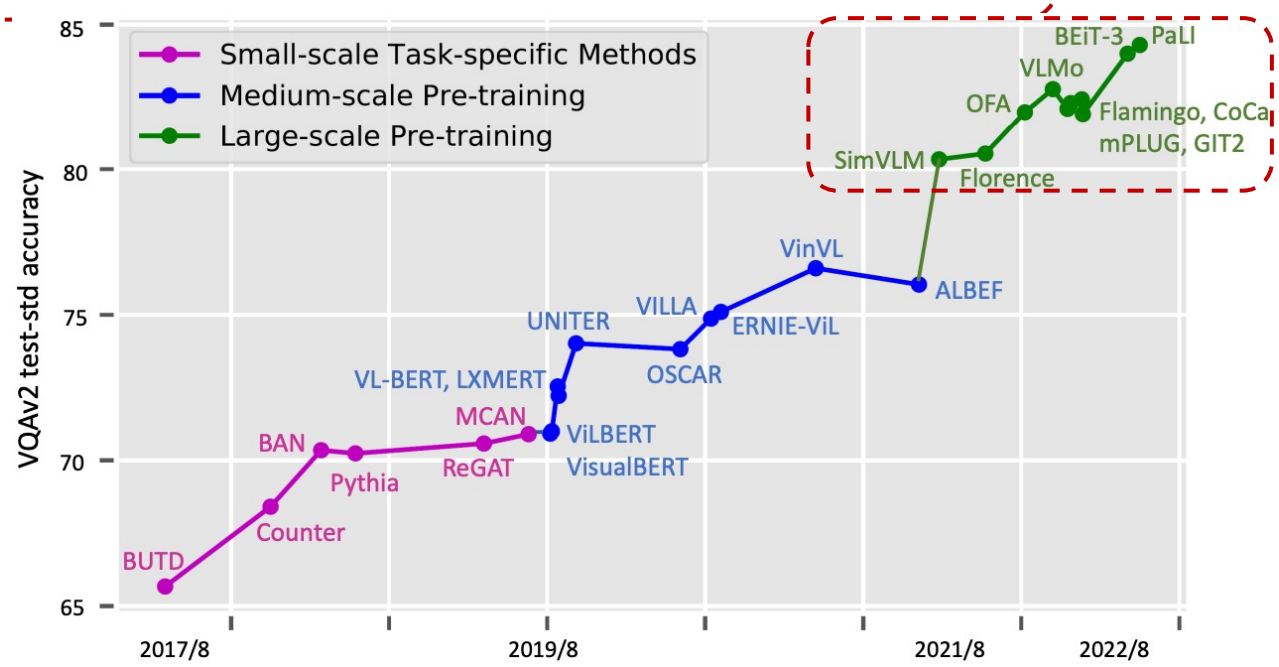


Why We Need Visual-language Model?

- Large-scale Multi-modal Pretraining Models have achieved SOTA performance in multiple downstreaming cross-modal reasoning tasks.
- The size of datasets and models are increasing :

  BEiT-3: 1.9B; GIT2 5.1B; PaLI: 16.9B; Flamingo: 80.2B (188x CLIP)

	VQA dev/std	GQA dev/std	VCR Q/A	VCR QA/R	VCR Q/AR	RefCOCO val/ testA/testB	NLVR2 test-P	IR R1/R5/R10	ZS IR R1/R5/R10
ViLBERT$_{nopt}$(Lu et al. (2019))	68.93/–	–	69.26	71.01	49.48	68.61/75.97/58.44	–	45.50/76.78/85.02	0.00/0.00/0.00
ViLBERT	70.55/70.92	–	73.3	74.6	54.8	72.34/78.52/62.61	–	58.20/84.90/91.52	31.86/61.12/72.80
B2T2(Alberti et al. (2019))	–	–	72.6	75.7	55.0	–	–	–	–
B2T2$_{ensemble}$	–	–	74.0	77.1	57.1	–	–	–	–
VL-BERT$_{nopt}$(Su et al. (2020))	69.58/–	–	–	–	–	66.03/71.87/56.13	–	–	–
VL-BERT	71.16/–	–	–	–	–	71.60/77.72/60.99	–	–	–
VL-BERT$_{large}$	71.79/72.22	–	75.8	78.4	59.7	72.59/78.57/62.30	–	–	–
LXMERT(Tan and Bansal (2019))	–/72.5	60.0/60.3	–	–	–	–	–	–	–
Unicoder-VL$_{zs}$(Li et al. (2020a))	–	–	–	–	–	–	–	48.4/76.0/85.2	–
Unicoder-VL$_{nopt}$	–	–	–	–	–	–	–	57.8/82.2/88.9	–
Unicoder-VL	–	–	73.4	74.4	54.9	–	–	71.5/90.9/94.9	–
VisualBERT(Li et al. (2019))	70.80/71.00	–	71.6	73.2	52.4	–	67.0	–	–
UNITER$_{base}$(Chen et al. (2020c))	72.70/72.91	–	75.0	77.2	58.2	75.31/81.30/65.58	77.85	72.52/92.36/96.08	66.16/88.40/92.94
UNITER$_{large}$	73.82/74.02	–	77.3	80.8	62.8	**75.9/81.45/66.7**	79.98	**75.56/94.08/96.76**	**68.74/89.2/93.86**
MiniVLM(Wang et al. (2020b))	69.39/69.06	–	–	–	–	–	73.93	–	–
PixelBERT$_{r50}$(Huang et al. (2020b))	71.35/71.42	–	–	–	–	–	72.4	59.8/85./91.6	–
PixelBERT$_{x152}$	74.45/74.55	–	–	–	–	–	77.2	71.5/92.1/95.8	–
OSCAR$_{base}$(Li et al. (2020c))	–/73.44	–	–	–	–	–	78.36	–	–
OSCAR$_{large}$	–/73.82	61.58/61.62	–	–	–	–	**80.37**	–	–
UnifiedVLP(Zhou et al. (2019))	70.5/70.7	–	–	–	–	–	–	–	–
InterBERT$_{nopt}$(Lin et al. (2021))	–	–	63.6	63.1	40.3	–	–	53.1/80.6/87.9	–
InterBERT	–	–	73.1	74.8	54.9	–	–	61.9/87.1/92.7	49.2/77.6/86.0
ERNIE-ViL$_{base}$(Yu et al. (2020))	72.62/72.85	–	77.0	80.3	62.1	74.02/80.33/64.74	–	74.44/92.72/95.94	–
ERNIE-ViL$_{large}$	73.78/73.96	–	**79.2**	**83.5**	**66.3**	74.24/80.97/64.70	–	75.10/93.42/96.26	–
DeVLBERT$_{V}$(Zhang et al. (2020))	–	–	–	–	–	–	–	59.3/85.4/91.8	32.8/63.0/74.1
DeVLBERT$_{VL}$	–	–	–	–	–	–	–	60.3/86.7/92.2	34.9/65.5/77.0
DeVLBERT$_{VLC}$	71.1/71.5	–	–	–	–	–	–	61.6/87.1/92.6	36.0/67.1/78.3
SemVLP(Li et al. (2021a))	**74.52/74.68**	**62.87/63.62**	–	–	–	– /–/–	79.55	74.10/92.43/96.12	–
CAPT(Luo et al. (2020a))	72.78/73.03	60.48/60.93	–	–	–	–	75.13	–	–
LAMP$_{coco}$(Guo et al. (2020))	70.85/71.0	–	–	–	–	–	74.34	–	42.5/70.9/80.8
LAMP$_{coco+vg}$	72.48/72.62	–/61.05	–	–	–	–	75.43	–	51.8/77.4/85.3
(Kervadec et al. (2019))	–	–/60.5	–	–	–	–	75.5	–	–
ImageBERT(Qi et al. (2020))	–	–	–	–	–	–	–	73.1/92.6/96.0	54.3/79.6/87.5
Human	–	89.3	91.0	93.0	85.0	–	96.3	–	–



Perspectives and Prospects on Transformer Architecture for Cross-Modal Tasks with Language and Vision.
Vision-Language Pre-training: Basics, Recent Advances, and Future Trends

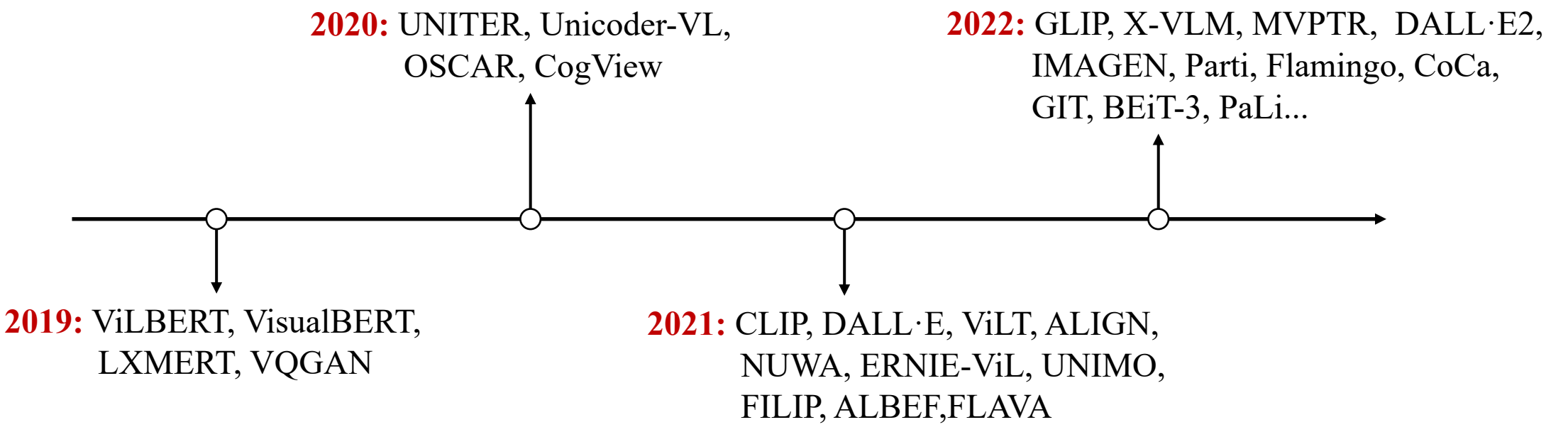
Comparison of Visual-language Pretrained Models

Model	Model Size				PT dataset size	PT Tasks
	Image Enc.	Text Enc.†	Fusion†	Total		
CLIP ViT-L/14 (Radford et al., 2021)	302M	123M	0	425M	400M	ITC
ALIGN (Jia et al., 2021)	480M	340M	0	820M	1.8B	ITC
Florence (Yuan et al., 2021)	637M	256M	0	893M	900M	ITC
SimVLM-huge (Wang et al., 2022k)	300M	39M	600M	939M	1.8B	PrefixLM
METER-huge (Dou et al., 2022b)	637M	125M	220M	982M	900M+20M[1]	MLM+ITM
LEMON (Hu et al., 2022)	147M[2]	39M	636M	822M	200M	MLM
Flamingo (Alayrac et al., 2022)	200M	70B	10B	80.2B	2.1B+27M[3]	LM
GIT (Wang et al., 2022d)	637M	40M	70M	747M	800M	LM
GIT2 (Wang et al., 2022d)	4.8B	40M	260M	5.1B	12.9B	LM
CoCa (Yu et al., 2022a)	1B	477M	623M	2.1B	1.8B+3B[4]	ITC+LM
BEiT-3 (Wang et al., 2022g)	692M[5]	692M[5]	52M[5]	1.9B	21M+14M[6]	MIM+MLM +MVLM
PaLI (Chen et al., 2022e)	3.9B	40M	13B	16.9B	1.6B	LM+VQA[7] +OCR+OD

Cross-modal Reasoning based on Visual-language Pretrained Models

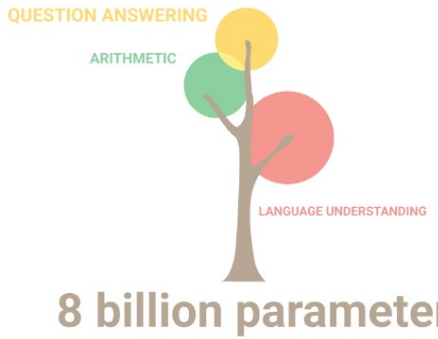
- Large language models based on pretraining (BERT, GPT, etc) have achieved great success in natural language processing area;
- The paradigm of pretraining has been extended to cross-modal reasoning tasks:

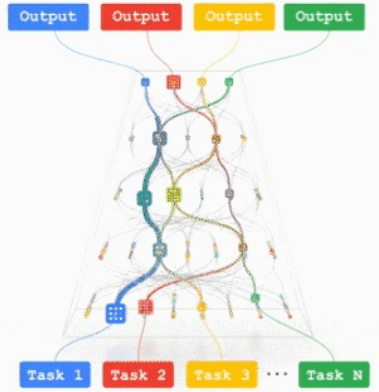
  CLIP (OpenAI) , DALL·E (OpenAI), NUWA (Microsoft) , Parti（Google） , 盘古 (Huawei) , M6 (Alibaba), OFA（Alibaba）, X-VLM (Bytedance) ...



Quantitative Change Leads to Qualitative Change

- In April 2022, Google released a pretrained language model PaLM（Pathways Language Model）with 540 billion non-sparse parameters.
- During the training progress of PaLM，some new capabilities suddenly emerge as the number of model parameters increasing.





s: A single model that can generalize across millions of tasks.

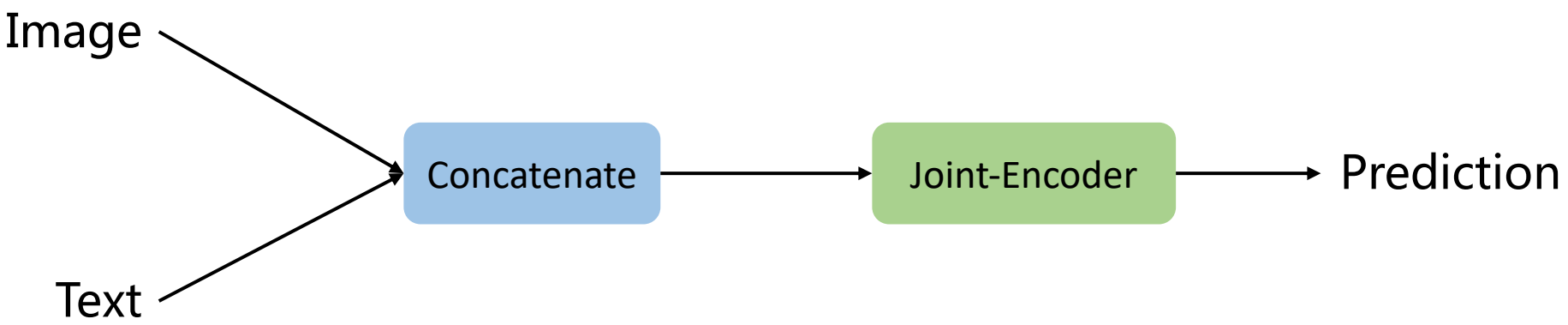
https://ai.googleblog.com/2022/04/pathways-language-model-palm-scaling-to.html

Large-scale Visual-language Pretrained Models

- Multi-modal Pretrained Models
  - Joint-Encoder
  - Dual-Encoder
  - Hybrid Method – Extremely Large Unified Models

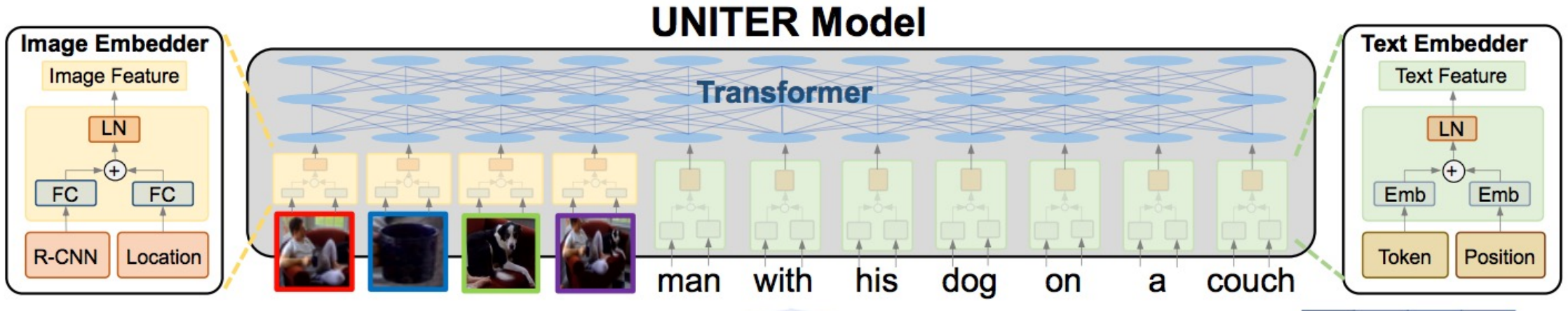
Joint-Encoder

- The image and text features are pieced together into a sequence, which is encoded by a joint cross-modal encoder. The cross-modal self-attention mechanism is used to model the correlation between the modes



UNITER

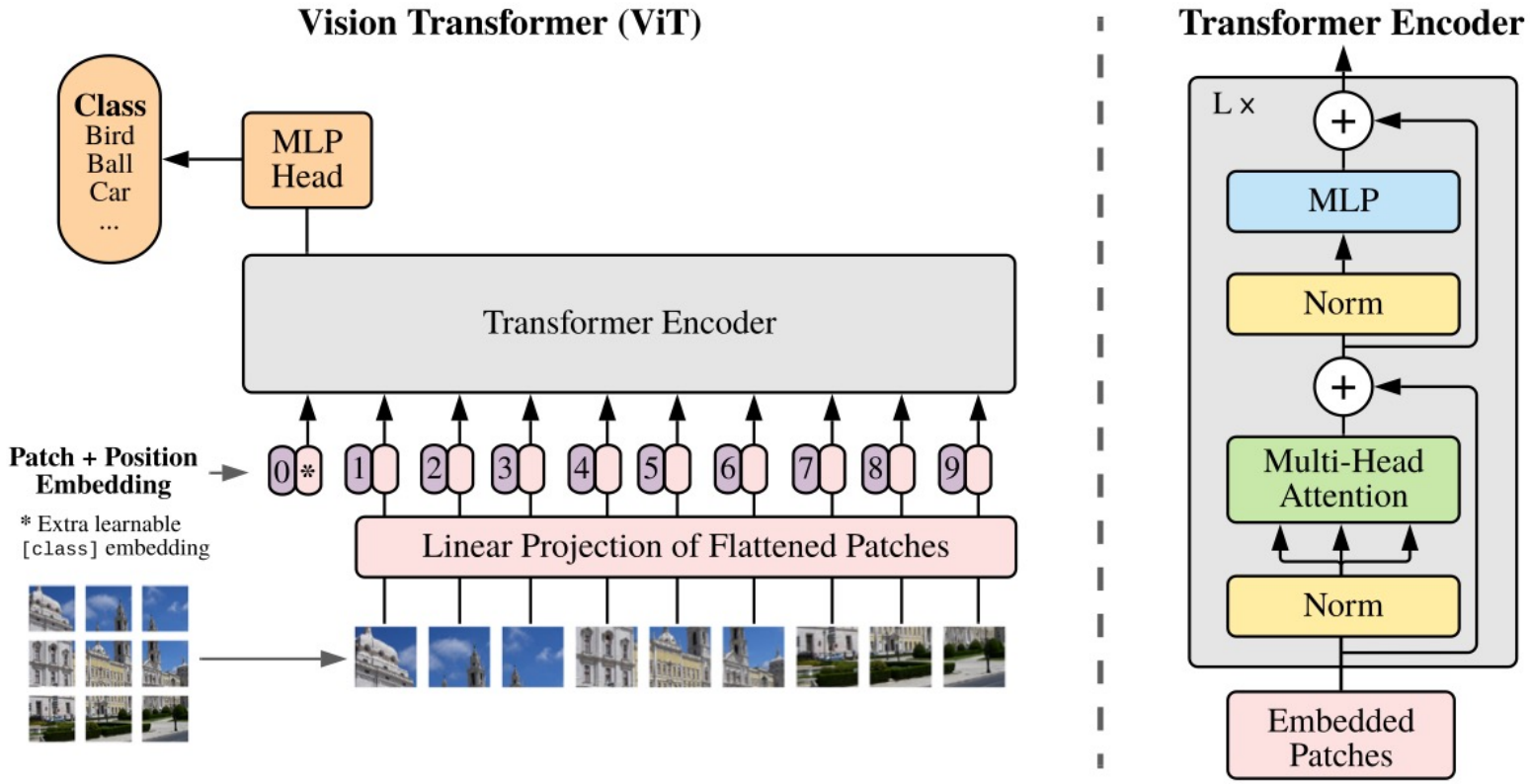
- UNITER- (joint-encoder) : The text and image features are connected as a sequence as input, and the cross-modal transformer is used for joint reasoning
- Image input : Using Faster R-CNN to detect targets, and the detected object region information was used as visual features
- Test input : word sequence



UNITER: UNiversal Image-Text Representation Learning. Chen et al. Microsoft. ECCV 2020.

Vision Transformer (ViT)

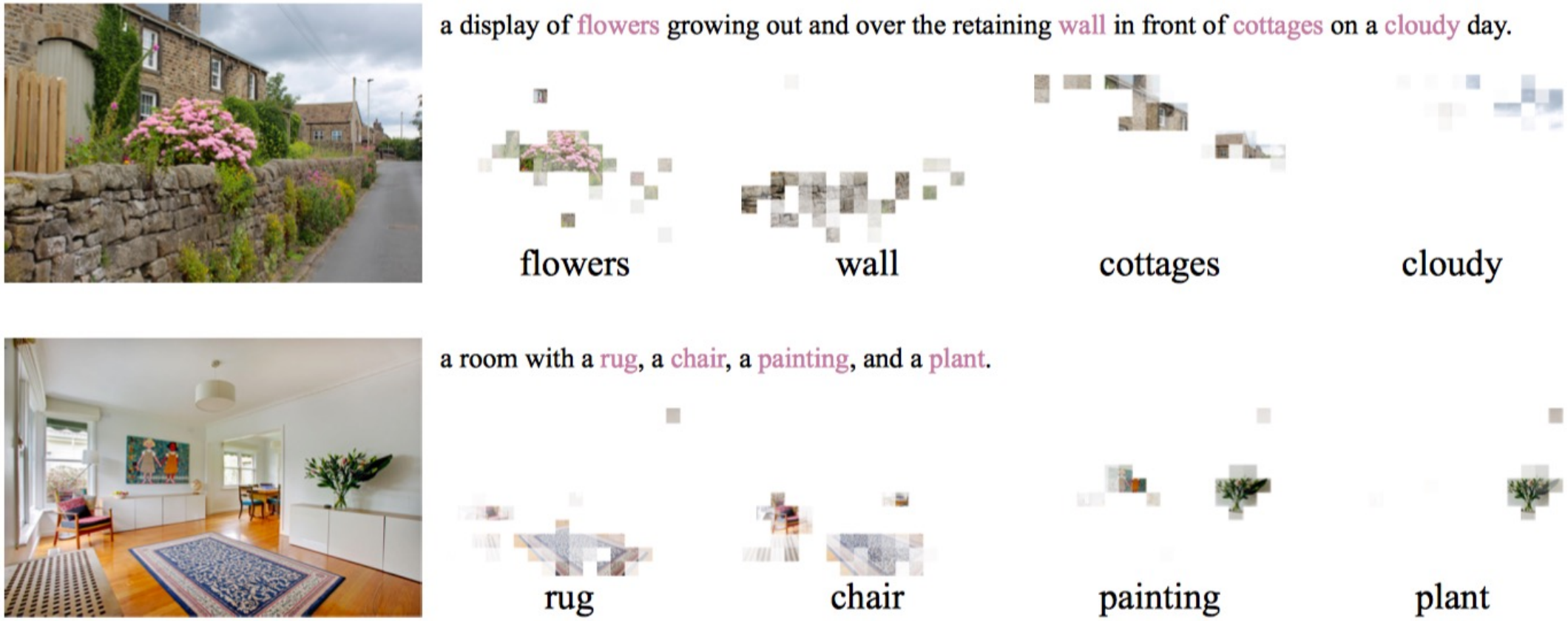
- Raw image-> patch token sequence : divide image into multiple patches -> patch token representation
- The success of Transformer architecture in the visual field, replacing the CNN



An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. Dosovitskiy et al. Google. ICLR 2021.

ViLT

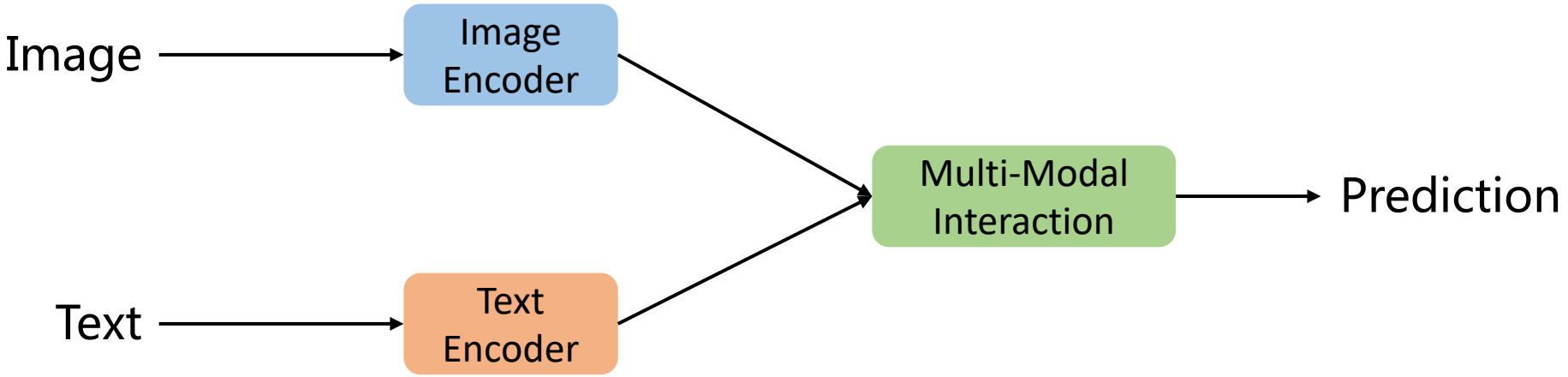
- ViLT can implicitly learn some semantic correspondence between word and image patch



ViLT: Vision-and-Language Transformer Without Convolution or Region Supervision. Kim et al. NAVER AL Lab. ICML 2021.

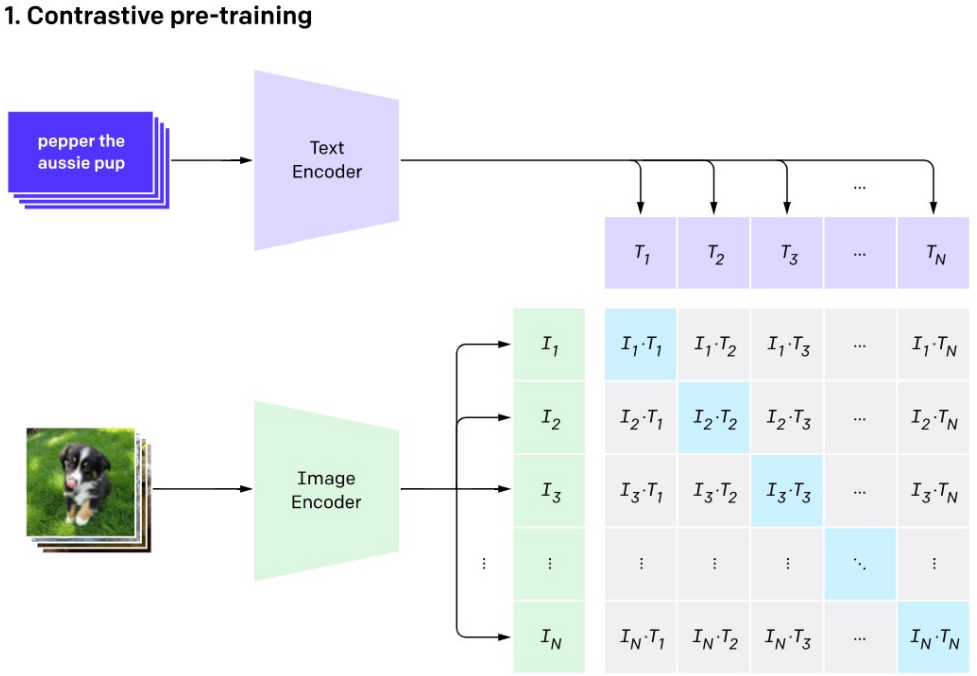
Dual-Encoder

- The image and text are encoded by two independent image encoders and text encoders respectively, and the modal interaction is carried out by calculating the similarity between the image features and text features respectively encoded



CLIP: Connecting Text and Images

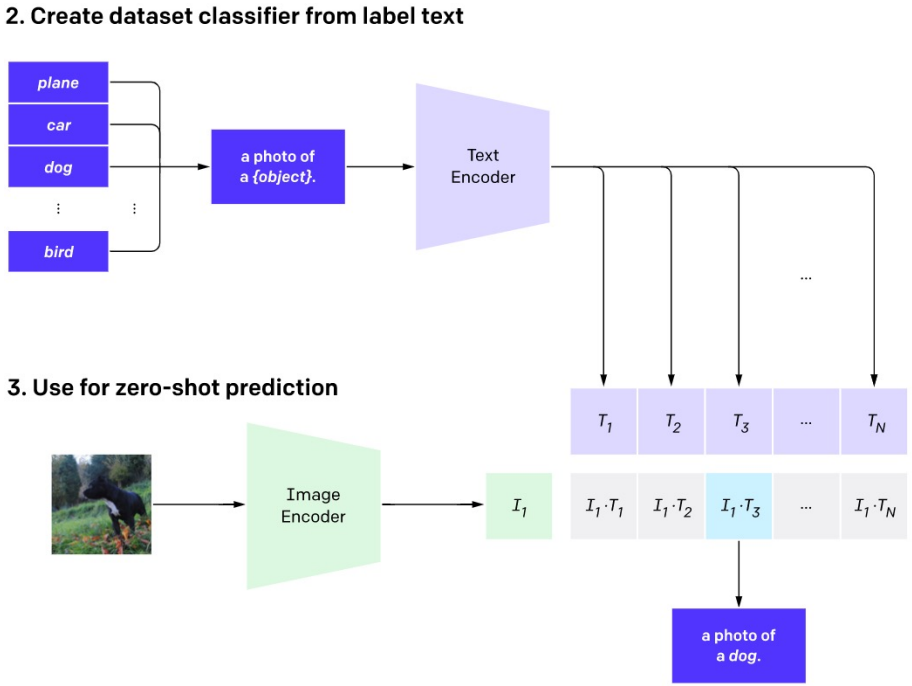
- Dual-Encoder : Text and image are encoded by two separate encoders, respectively
- Contrast learning is used for self-supervised learning: matched images and text are regarded as positive samples, and unmatched images and text are regarded as negative samples

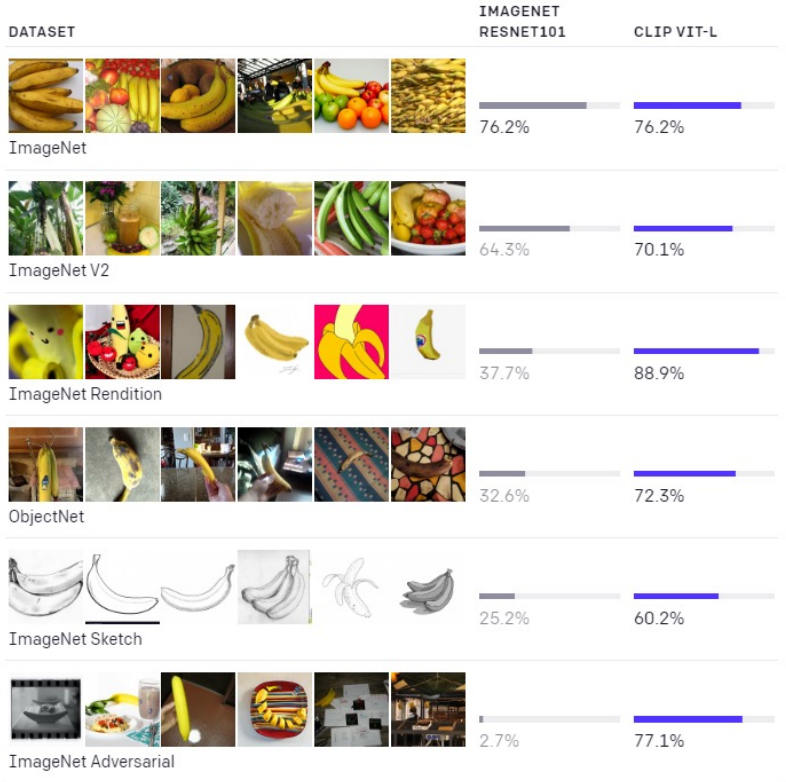


Learning Transferable Visual Models From Natural Language Supervision. Radford et al. OpenAI. ICML 2021.

CLIP: Connecting Text and Images

- CLIP trains 400 million data collected from the web, and because of the rich visual concepts contained in the massive text, CLIP can migrate zero samples to 30 different tasks (fine-grained object classification, action recognition, facial emotion recognition, optical character recognition, etc.).

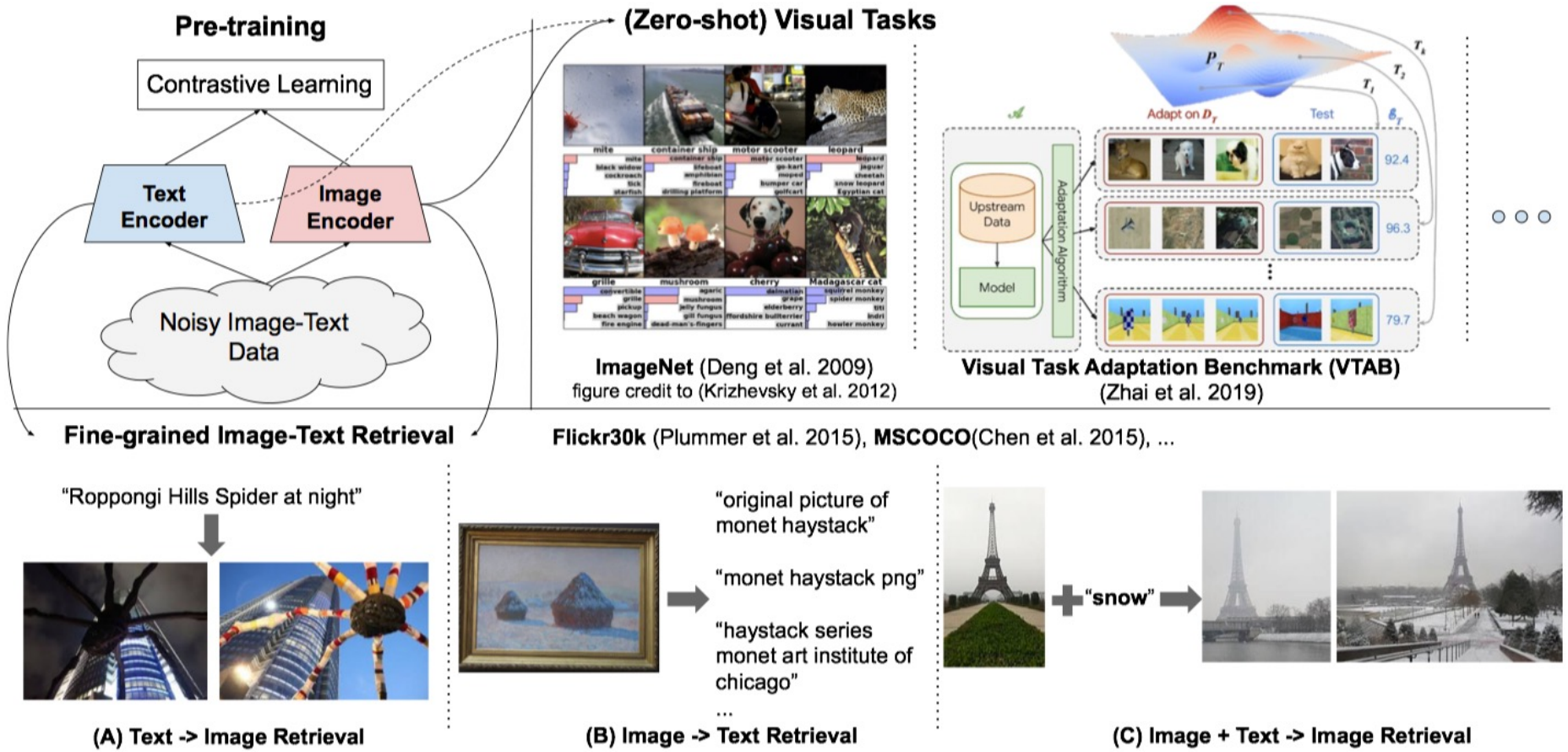




Learning Transferable Visual Models From Natural Language Supervision. Radford et al. OpenAI. ICML 2021.

ALIGN: A Large-scale ImaGe and Noisy-text embedding

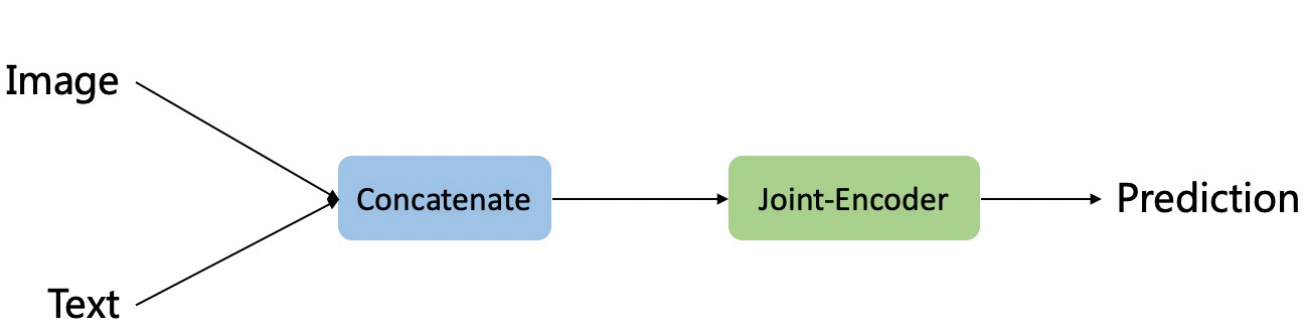
- Larger pre-training data: 1.2 billion (compare: CLIP 400 million)



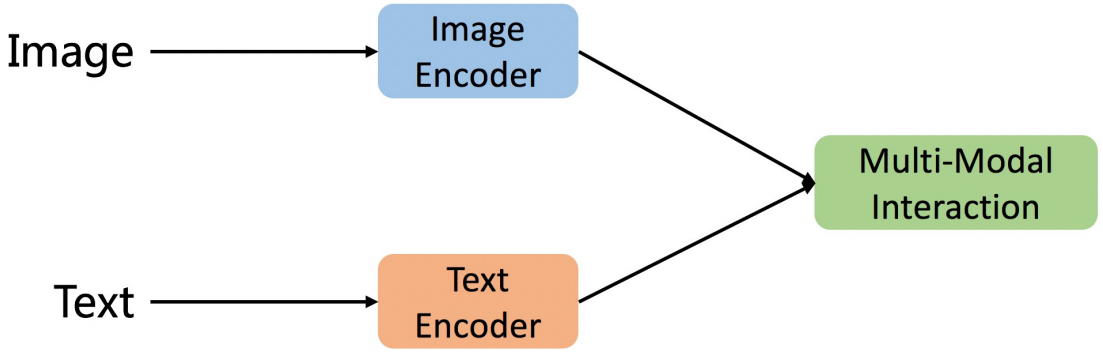
Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision. Jia et al. Salesforce Research. ICML 2021

Comparison of Joint-Encoder and Dual-Encoder

- **Joint-Encoder :**
  - Using the cross-modal attention mechanism, text and image features can be interacted and integrated more fully, so that complex cross-media reasoning can be accomplished better ;
  - Low retrieval efficiency ;
- **Dual-Encoder :**
  - High retrieval efficiency ;
  - Visual information and text information only interact through similarity calculation, and this shallow interaction is not suitable for complex cross-media reasoning tasks, such as Visual Question Answering.



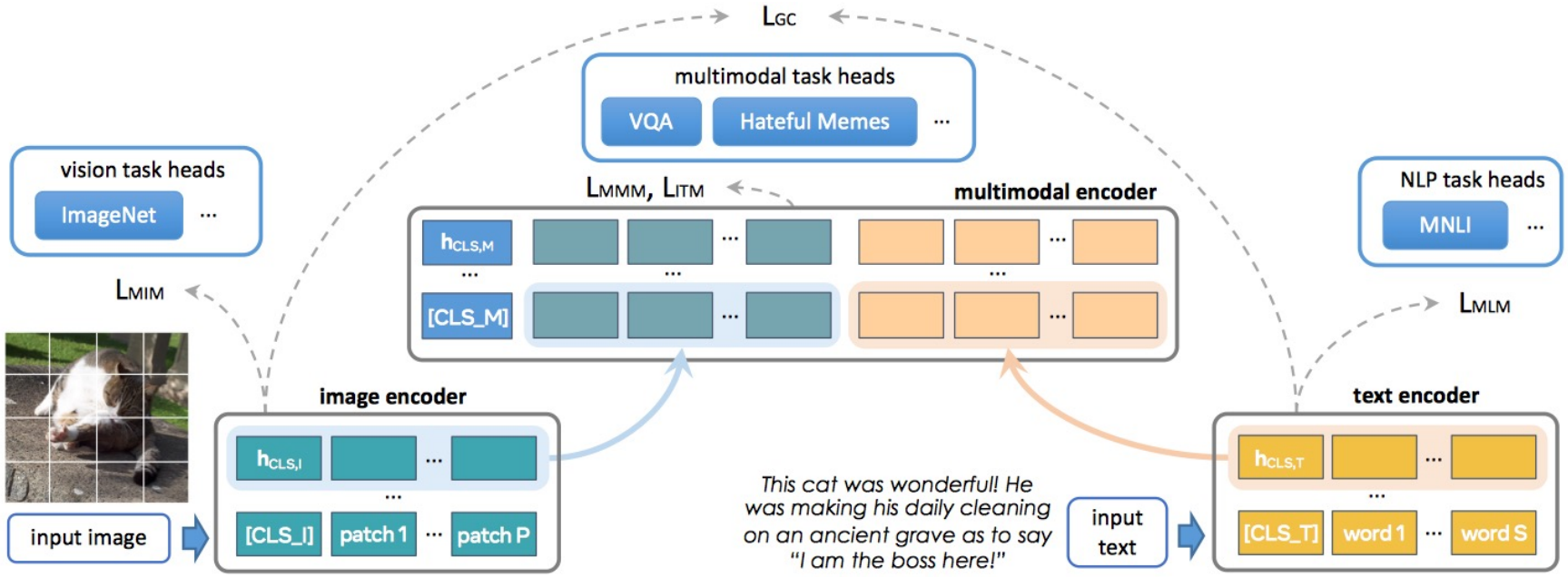
a. Joint-Encoder



b. Dual-Encoder

Hybrid Method - FLAVA

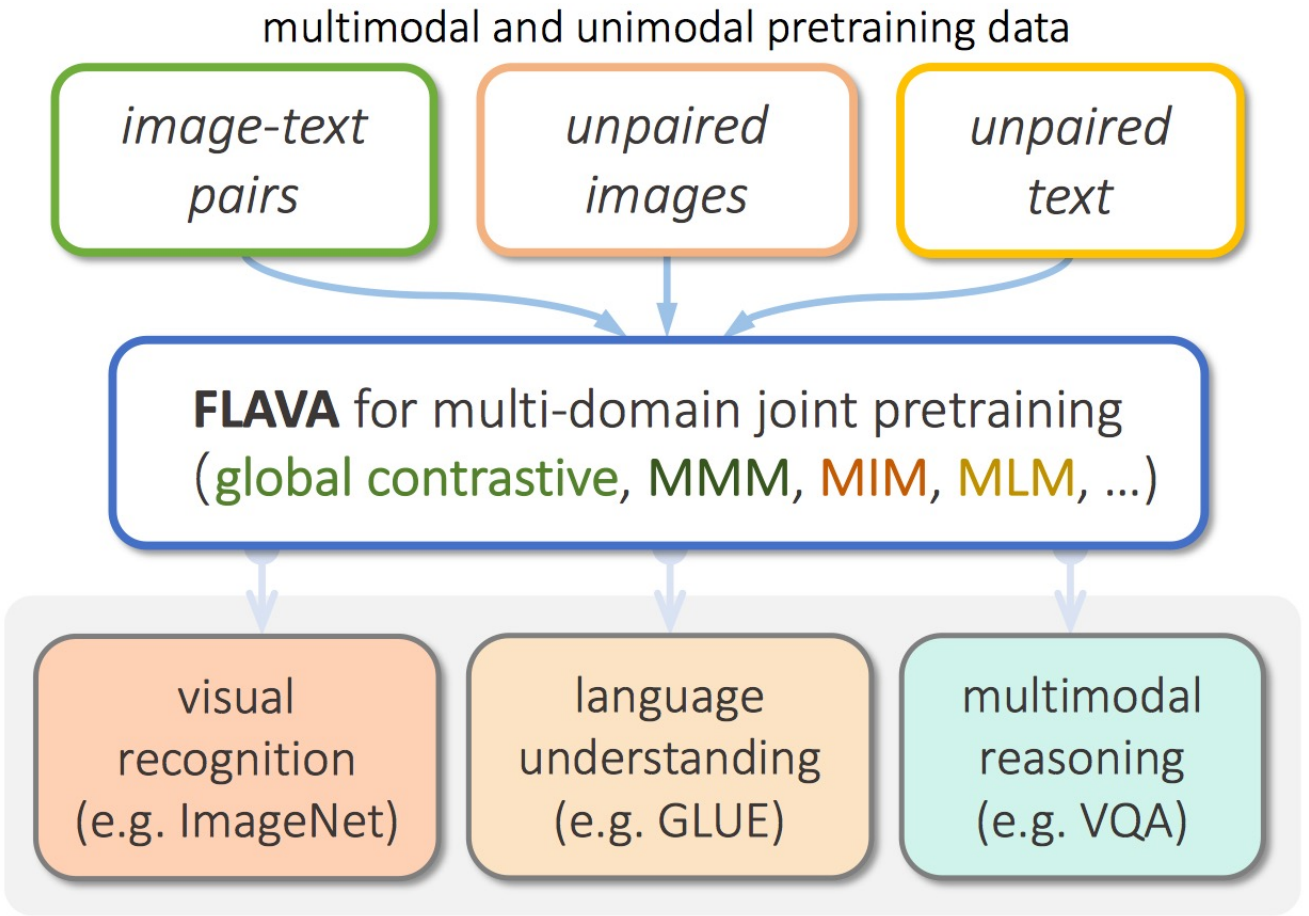
- For single-modal tasks (vision task, NLP task) and cross-modal retrieval tasks, complete with Dual-Encoder (two separate visual encoders and text encoders)
- For the cross-modal reasoning task, the image and text representations encoded by the Dual-Encoder structure are fed into the multimodal encoder (Joint-Encoder) for full cross-modal interaction and final prediction results



FLAVA: A Foundational Language And Vision Alignment Model. Singh et al. Facebook AI Research. CoRR 2021.

Hybrid Method - FLAVA

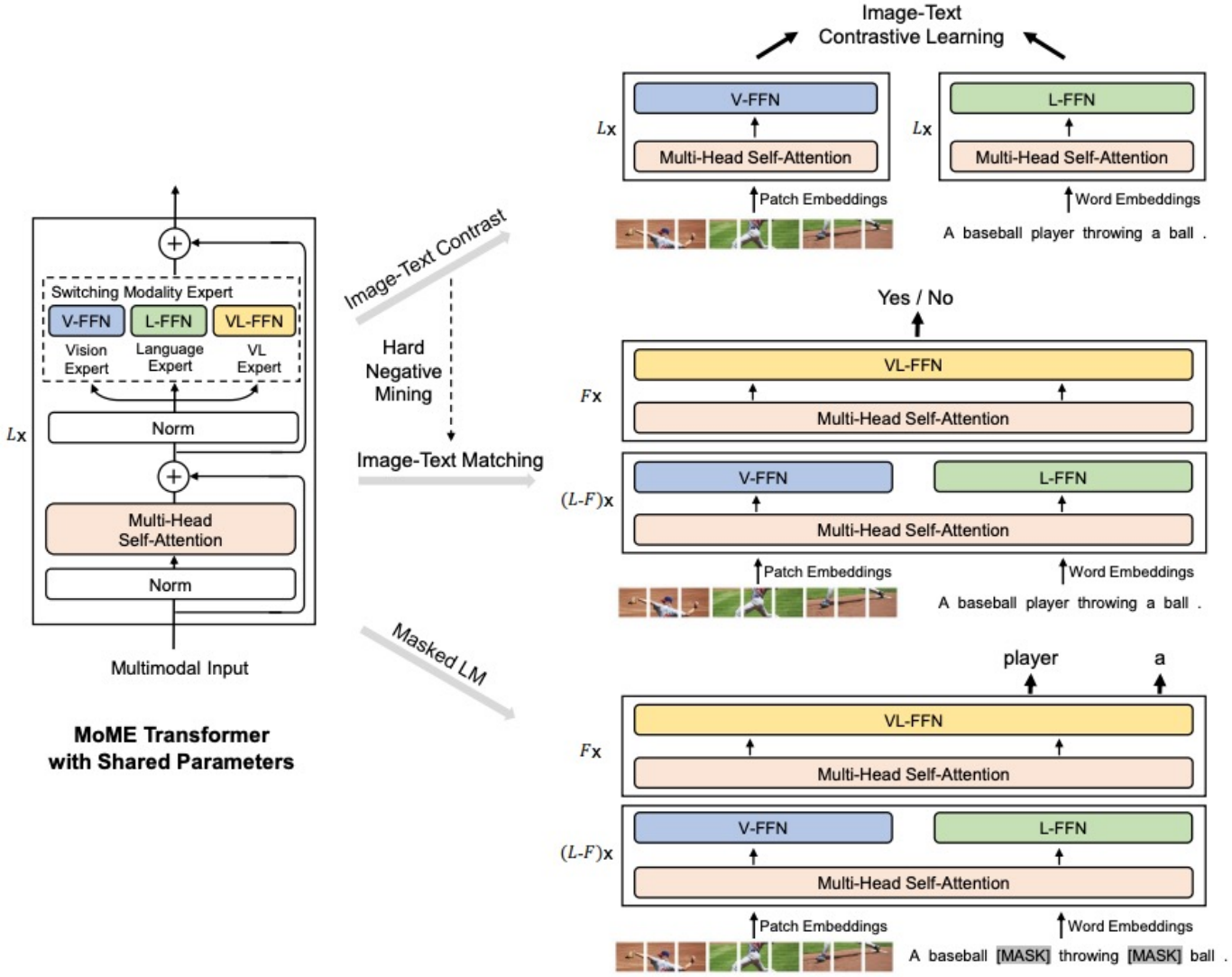
- After multi-modal pre-training, the model has achieved good performance on 35 downstream tasks (CV, NLP, MM)



FLAVA: A Foundational Language And Vision Alignment Model. Singh et al. Facebook AI Research. CoRR 2021.

VLMo: Unified Vision-Language Pre-Training

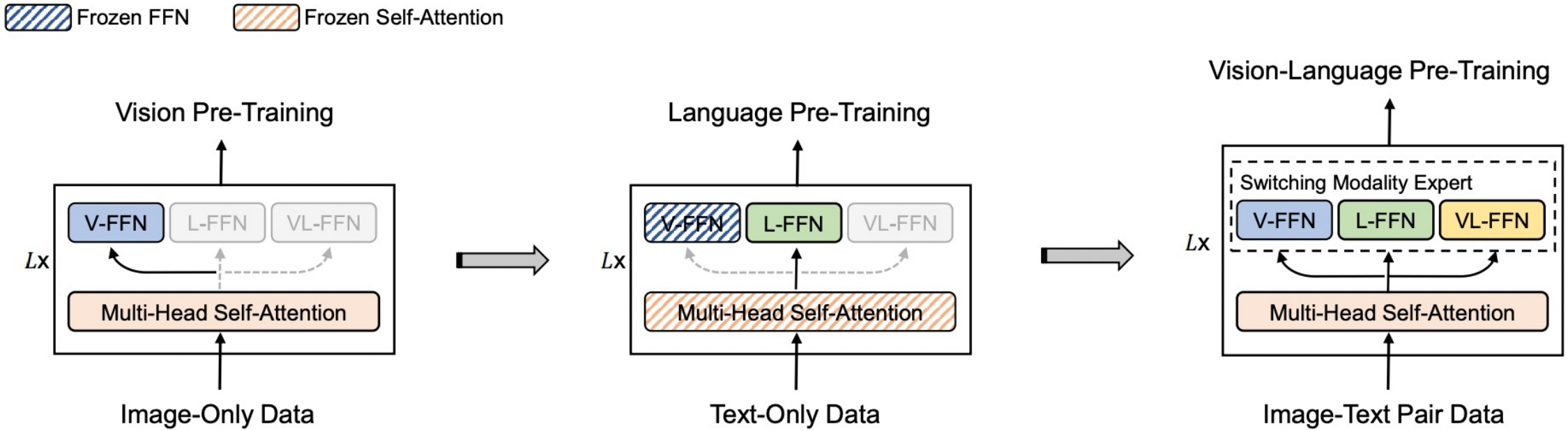
- Mixture-of-Modality-Experts (MoME) Transformer: Can switch between different "Modality Expert" according to different input types (image, text, image-text pairs)
- Replace the feed-forward network (FFN) in Transformer with three FFNS, each for handling different types of input
- Different types of inputs share parameters in other parts of Transformer (Multi-Head Self-Attention)



VLMO: Unified Vision-Language Pre-Training with Mixture-of-Modality-Experts. Wang et al. Microsoft Research. NeurIPS 2022.

VLMo: Unified Vision-Language Pre-Training

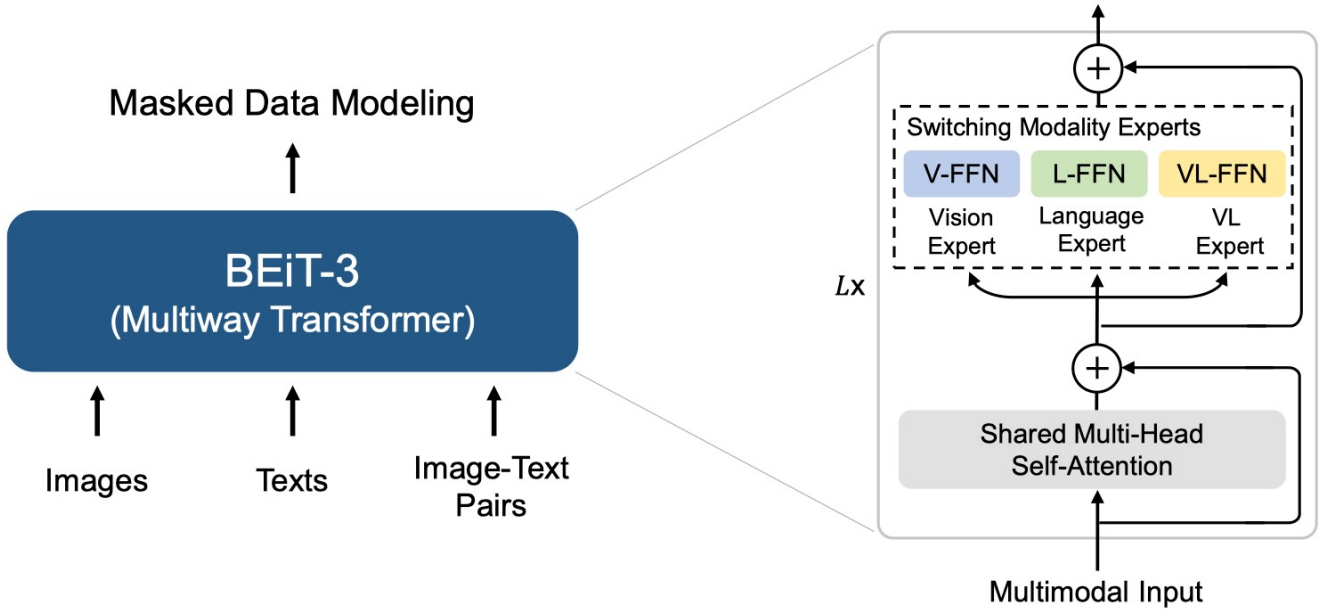
- By adapting MoME Transformer, a unified model is implemented to solve visual, linguistic, and cross-media tasks simultaneously
- Unified model can simultaneously use visual pre-training, language pre-training, multi-modal pre-training to train the model, so the model is more robust and universal
- Pre-training order:



VLMO: Unified Vision-Language Pre-Training with Mixture-of-Modality-Experts. Wang et al. Microsoft Research. NeurIPS 2022.

BEiT-3: Pre-training for All Vision and Vision-Language Tasks

- BEiT-3 uses MoME Transformer to build the unified model and uses a larger parameter count (1.9B) to enhance model characterization



Model	#Layers	Hidden Size	MLP Size	#Parameters				
				V-FFN	L-FFN	VL-FFN	Shared Attention	Total
BEIT-3	40	1408	6144	692M	692M	52M	317M	1.9B

Image as a Foreign Language: BEIT Pretraining for All Vision and Vision-Language Tasks. Wang et al. Microsoft. arXiv 2022.

BEiT-3: Pre-training for All Vision and Vision-Language Tasks

- Image is also likened to a kind of "language" (Imglish), so masked "language" modeling can be used as a unified pre-training task for three types of data: text, image, image-text (" parallel sentences ")
- Unified pre-training can make full use of various forms of data and further improve the generalization ability of the model.

Data	Source	Size
Image-Text Pair	CC12M, CC3M, SBU, COCO, VG	21M pairs
Image	ImageNet-21K	14M images
Text	English Wikipedia, BookCorpus, OpenWebText, CC-News, Stories	160GB documents

Image as a Foreign Language: BEIT Pretraining for All Vision and Vision-Language Tasks. Wang et al. Microsoft. arXiv 2022.

BEiT-3: Pre-training for All Vision and Vision-Language Tasks

- With a unified model, larger parameters, and joint pre-training, BEiT-3 achieves significant performance improvements on multiple tasks in the field of vision and multimodality

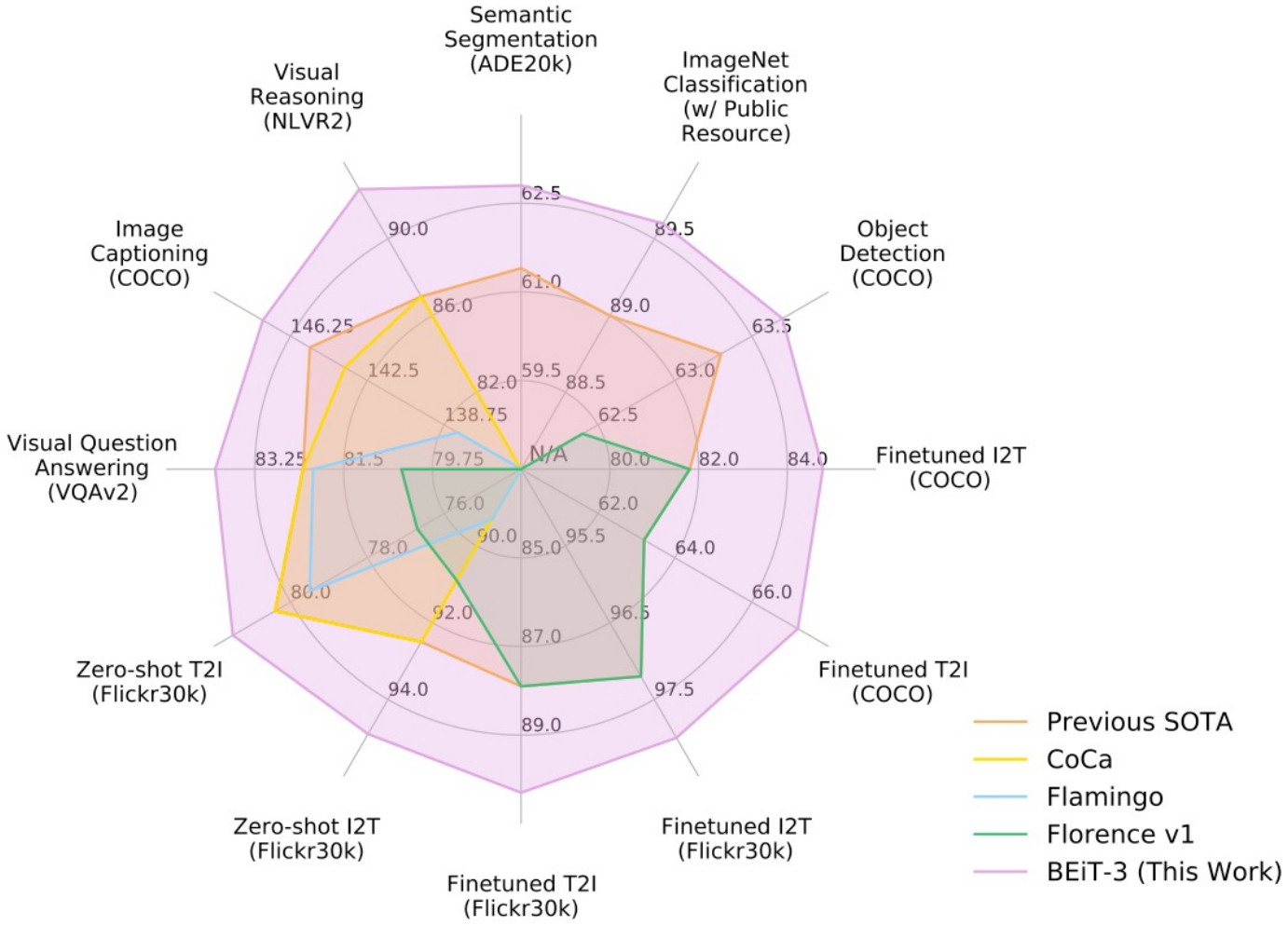


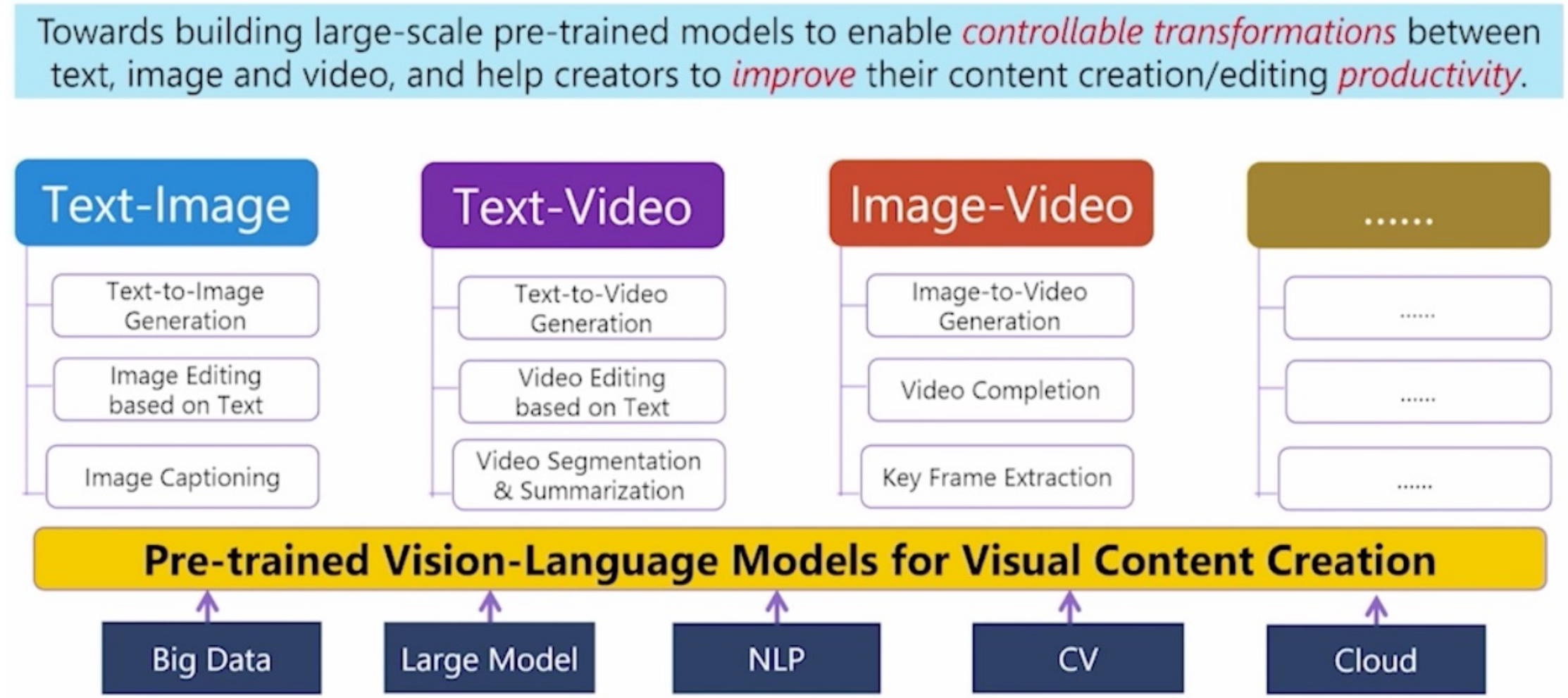
Image as a Foreign Language: BEIT Pretraining for All Vision and Vision-Language Tasks. Wang et al. Microsoft. arXiv 2022.

Large-scale Visual-language Pretrained Models

- Multi-modal Pretrained Models
  - Joint-Encoder
  - Dual-Encoder
  - Hybrid Method – Extremely Large Unified Models
  - Multi-modal Few-shot Learning

- Applications: Multi-modal Generation

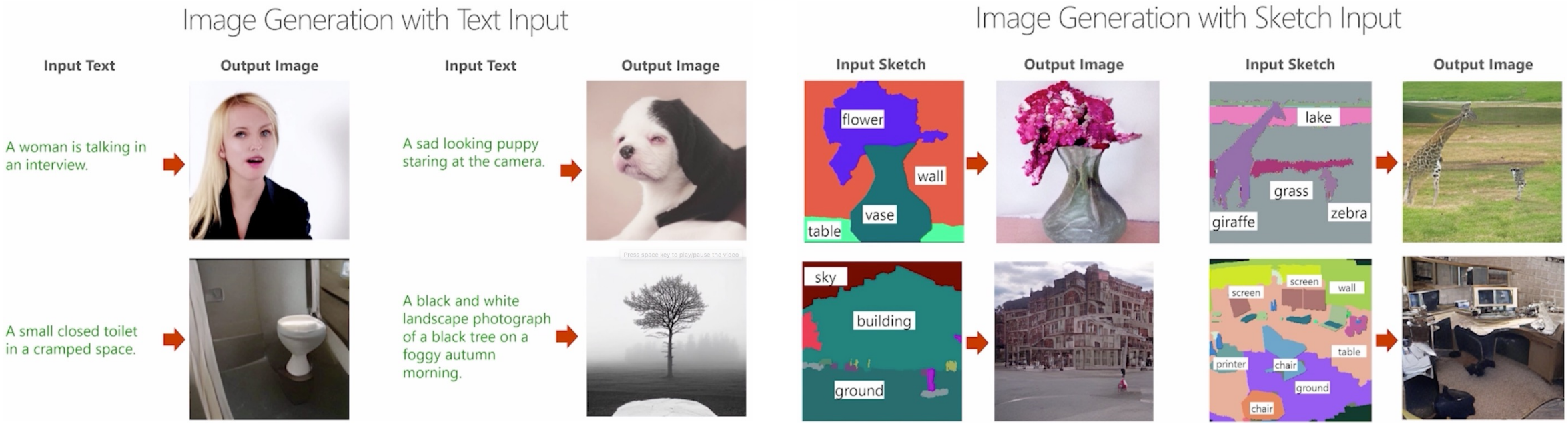
NUWA



NUWA: Neural visUal World creAtion with multimodal pre-training. Wu et al. Microsoft.

NUWA

- Generate images/videos based on text
- Generate images/videos based on sketches

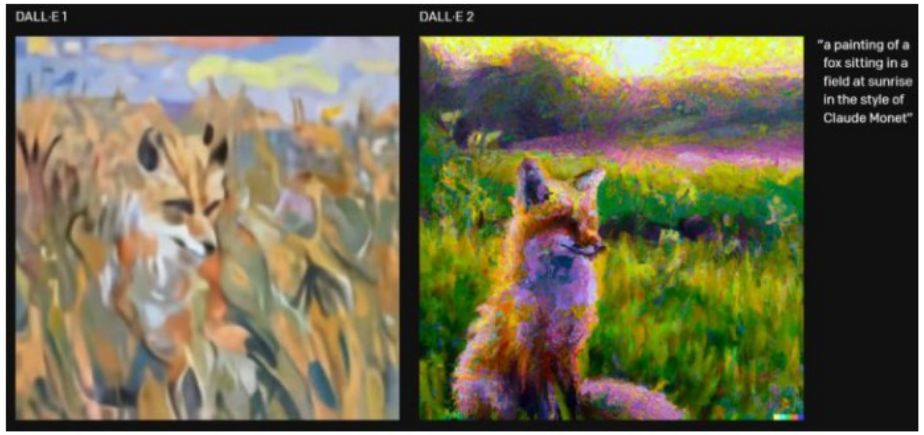


NUWA: Neural visUal World creAtion with multimodal pre-training. Wu et al. Microsoft.

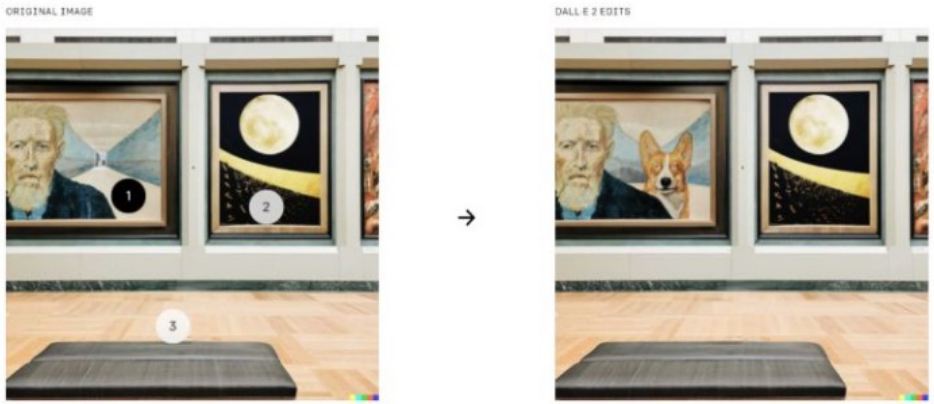
DALL·E2

DALL·E2 can generate surrealistic images, and has image conversion, image editing and other functions

Generate surrealistic images



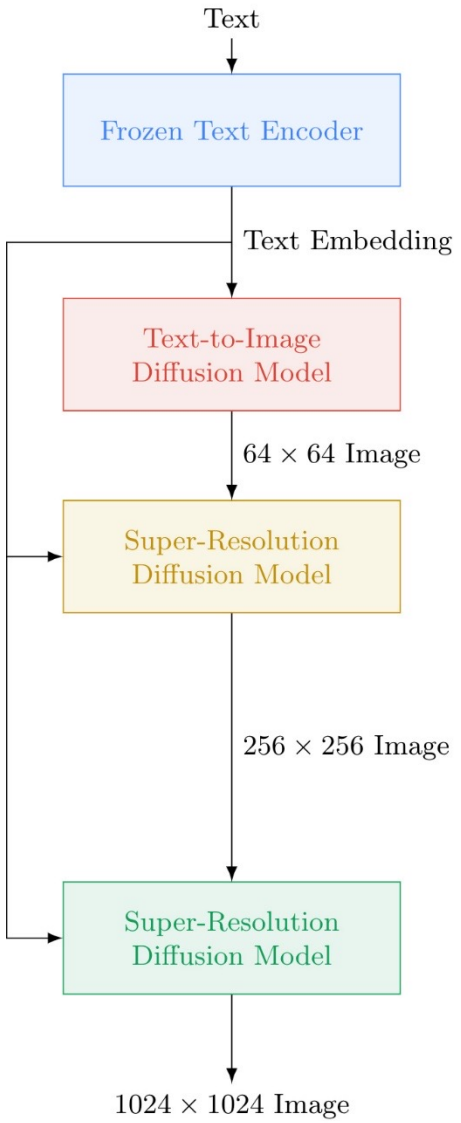
Style transfer (Monet)

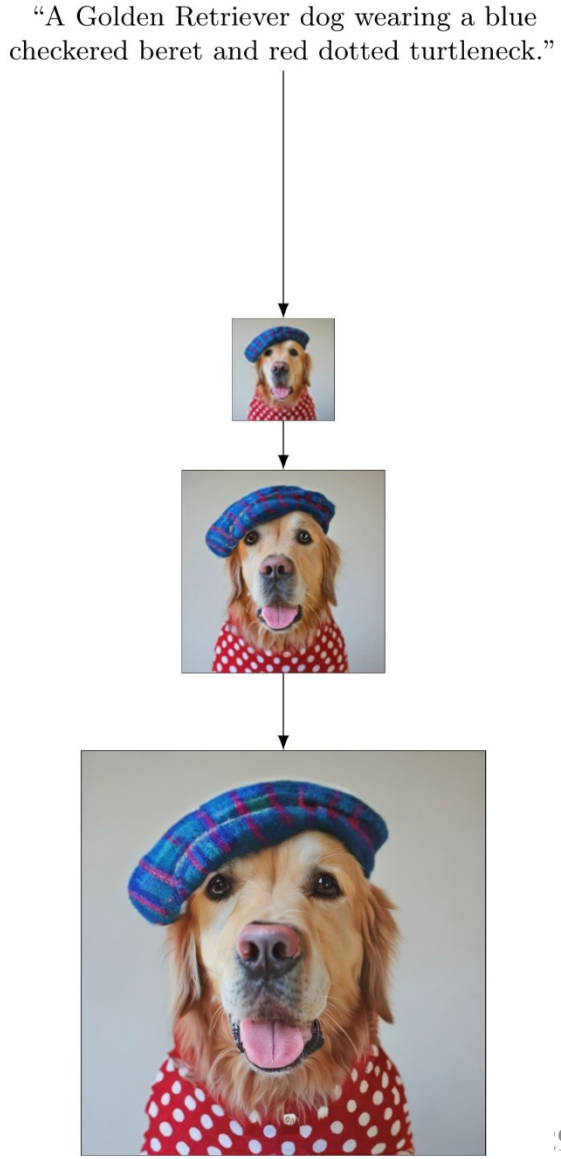


Edit the image by area

Hierarchical Text-Conditional Image Generation with CLIP Latent. Ramesh et al. OpenAI. 2022

Imagen: by Google Research





https://imagen.research.google

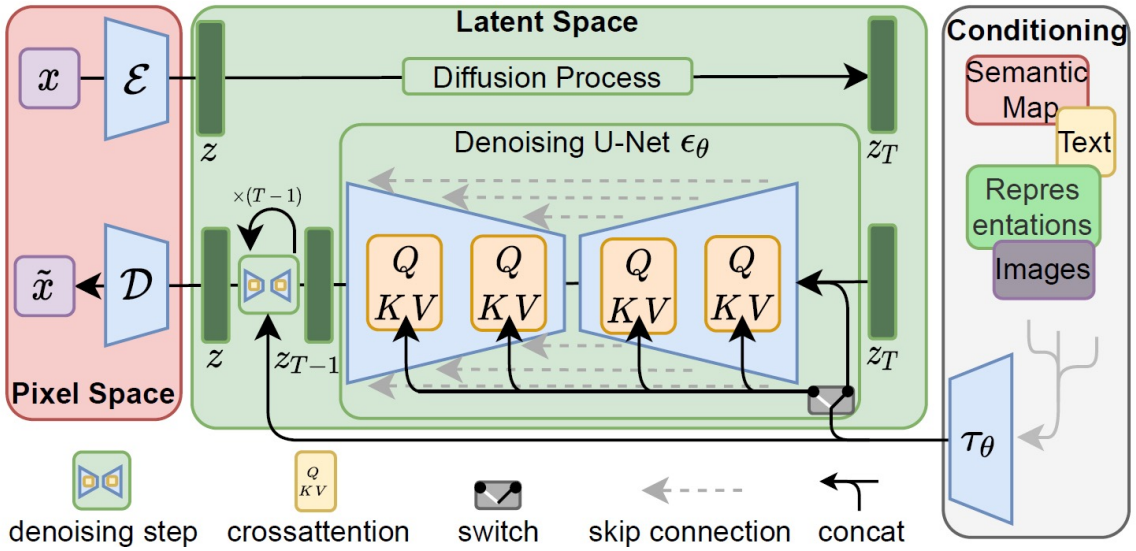
ImagenVideo: by Google Research

- Combining sequence model and upsampling model, text to video generation can be realized

Source：https://imagen.research.google/video/

Stable Diffusion

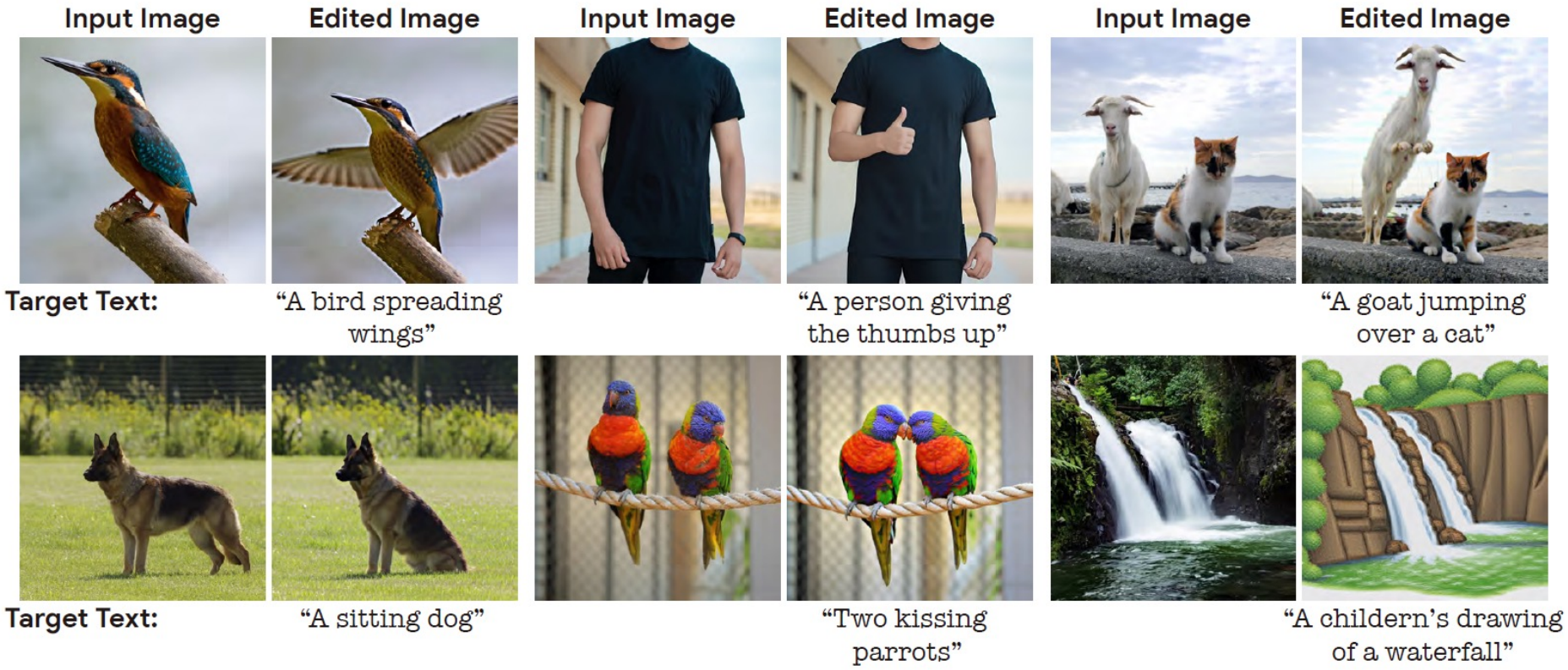
- Latent Diffusion Model (LDM) :
- The image is encoded into the hidden space, and the diffusion model is applied in the hidden space.

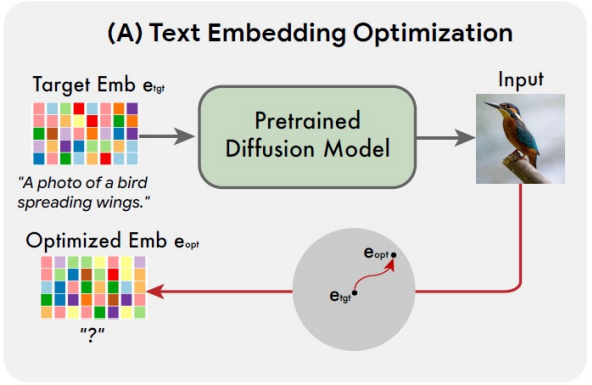


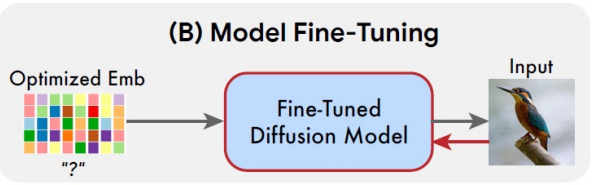
High-Resolution Image Synthesis with Latent Diffusion Models. Rombach et al. CVPR 2022.

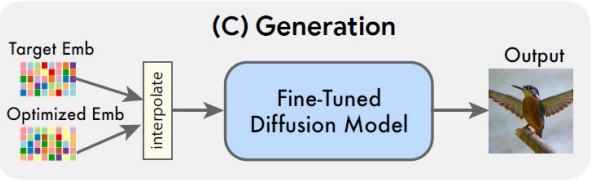
Imagic: Text-Based Real Image Editing with DM

- Based on diffusion model, edit images from text (conditional generation)









Imagic: Text-Based Real Image Editing with Diffusion Models. Kawar et al. arXiv 2022.

Thank you!