JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

Accelerating Vision-Language Pretraining with Free Language Modeling

Teng Wang, *Yixiao Ge, Feng Zheng, Ran Cheng, Ying Shan, Xiaohu Qie, Ping Luo*

Poster : THU-PM-244

Paper: https://arxiv.org/abs/2303.14038

Project: https://github.com/TencentARC/FLM

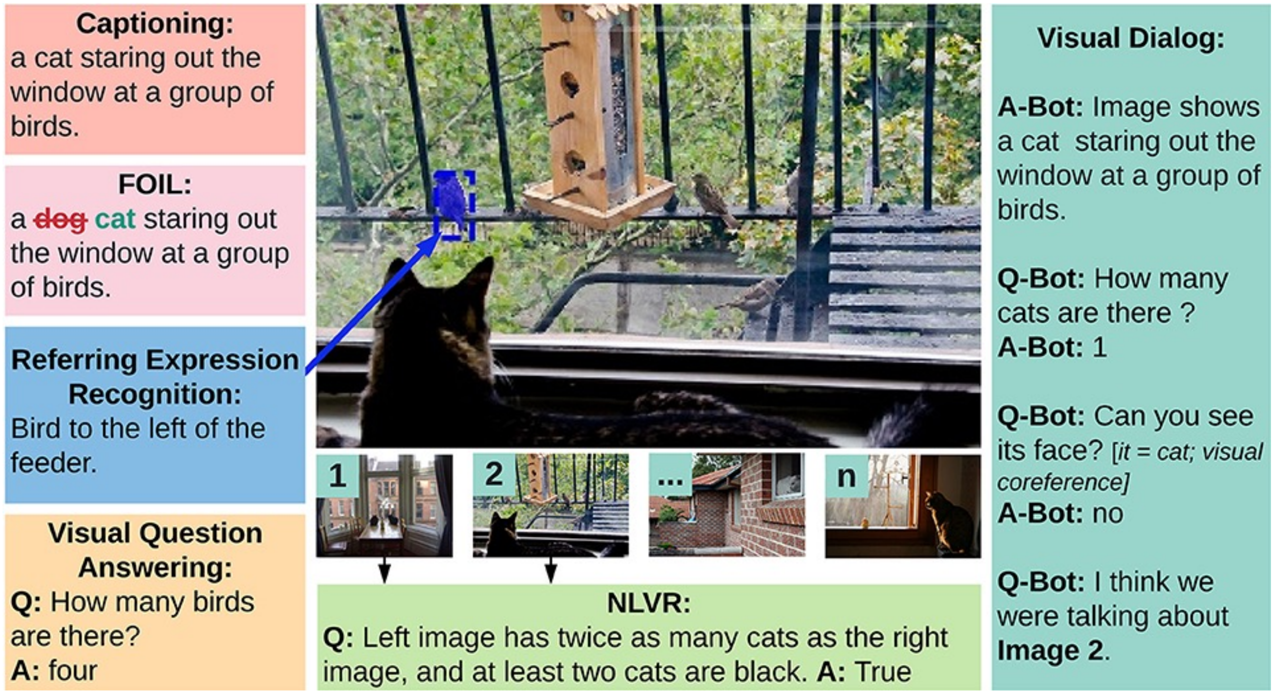
SUSTech Southern University of Science and Technology
香 港 大 學 THE UNIVERSITY OF HONG KONG
Tencent ARC Applied Research Center
鹏城实验室 PENGCHENG LABORATORY
上海人工智能实验室 Shanghai Artificial Intelligence Laboratory

Vision-Language Pretraining (VLP)

VLP aims to pretrain the model by mining multimodal associations from large-scale unlabeled image-text data, serving as an initial stage for subsequent finetuning.

Vision-Language Pretraining Model

Finetuing



Diverse Vision-Language Tasks [Kushal et al.]

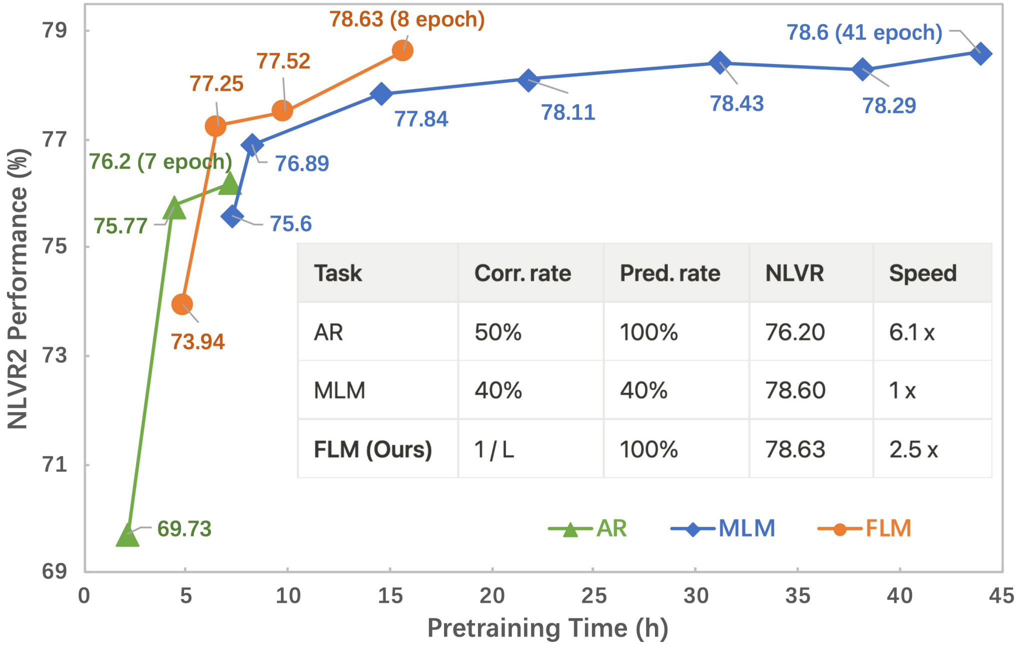
[1] Kafle, Kushal, et al. "Challenges and prospects in vision and language research." *Frontiers in Artificial Intelligence* 2 (2019): 28.

Overview

- We propose a new pretraining objective, i.e., Free language modeling (FLM), for accelerating vision-language pretraining.

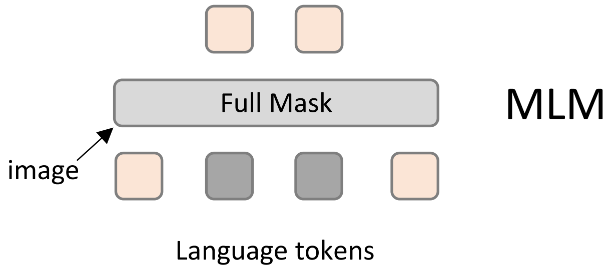
- FLM frees the prediction rate from the constraints of the corruption rate, enabling an appealing 100% prediction rate for better convergence.

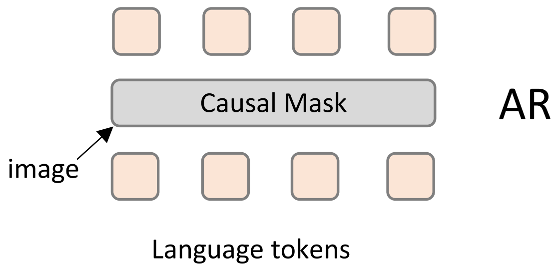
- With less than 50% pretraining time, FLM could achieve competitive performance on both vision-language understanding and generation tasks.

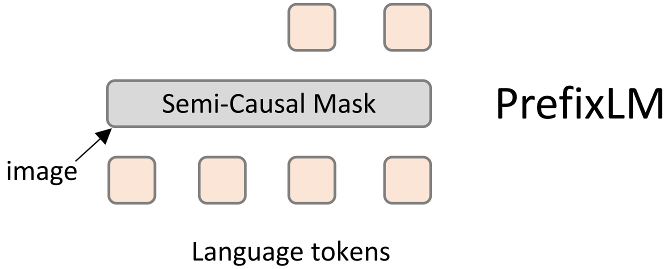


Task	Corr. rate	Pred. rate	NLVR	Speed
AR	50%	100%	76.20	6.1 x
MLM	40%	40%	78.60	1 x
FLM (Ours)	1 / L	100%	78.63	2.5 x

Related Work: Language Modeling



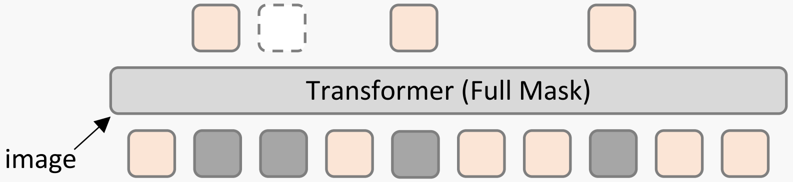




- Masked Language Modeling (MLM)
  - Mask ratio mainly in 15%-40%,
  - Large portion of output tokens is not utilized, impeding efficiency
  - E.g., VLBERT, VisualBERT, BEiT3
- Auto-regressive (AR)
  - 100% output tokens are utilized
  - Converge faster with high efficiency
  - Inferior performance on VL understanding tasks
  - E.g., Coca
- Others (mainly in NLP)
  - Prefix Language Modeling
  - Permuted Language Modeling
  - General Language Modeling

Can we **accelerate the convergence** of VLP by predicting 100% tokens like AR meanwhile achieving **competitive performance with MLM**?

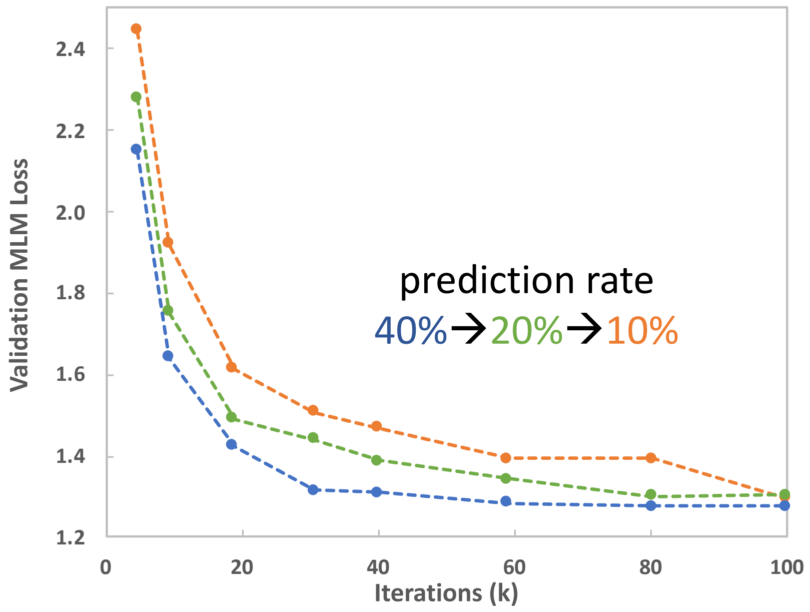
Prediction Rate in MLM



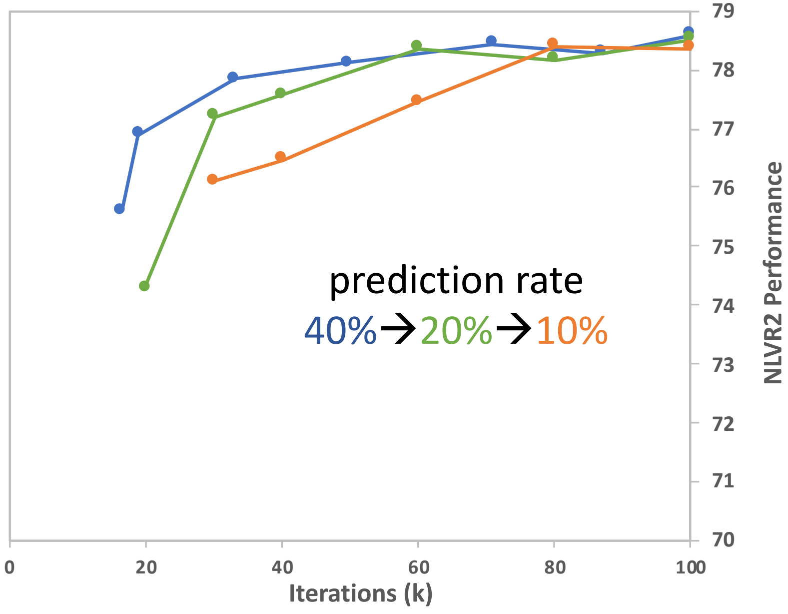
prediction rate: $r_{pred} = \frac{\#output\ tokens\ for\ loss\ calculation}{\#input\ koens}$

corruption rate: $r_{corr} = \frac{\#corrupted\ input\ tokens}{\#input\ koens}$

Validation loss

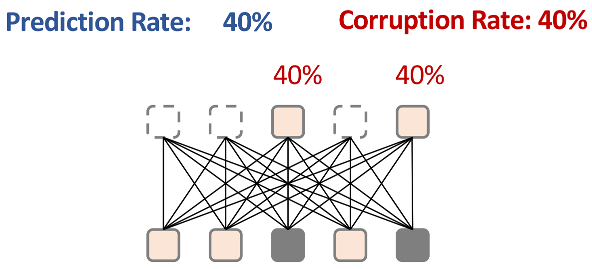


NLVR Accuracy

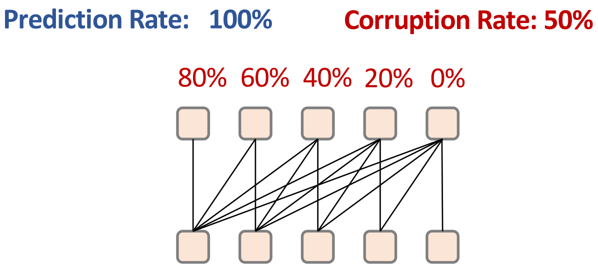


Increasing prediction rate helps convergence.

How to increase efficiency while keeping competitive performance?



MLM

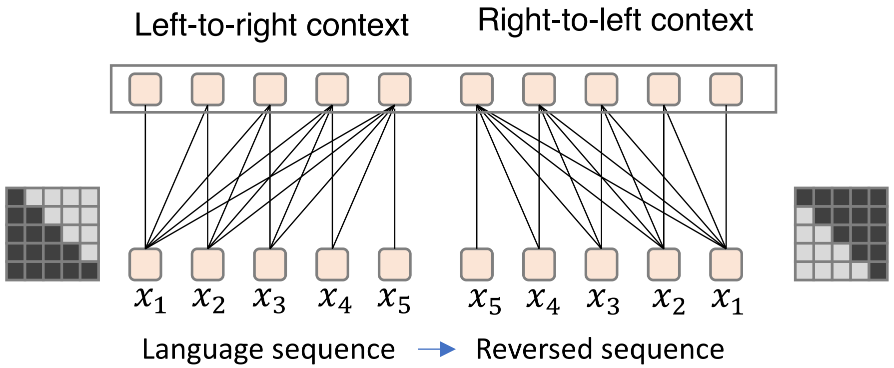


Auto-regressive LM

- **Efficiency:** Increase r_{pred}
  - Auto-regressive LM: $r_{pred} = 100\%$
  - MLM: r_{pred} can not be larger due to the coupling between r_{pred} and r_{corr}
  - **FLM (Ours):** $r_{pred} = 100\%$

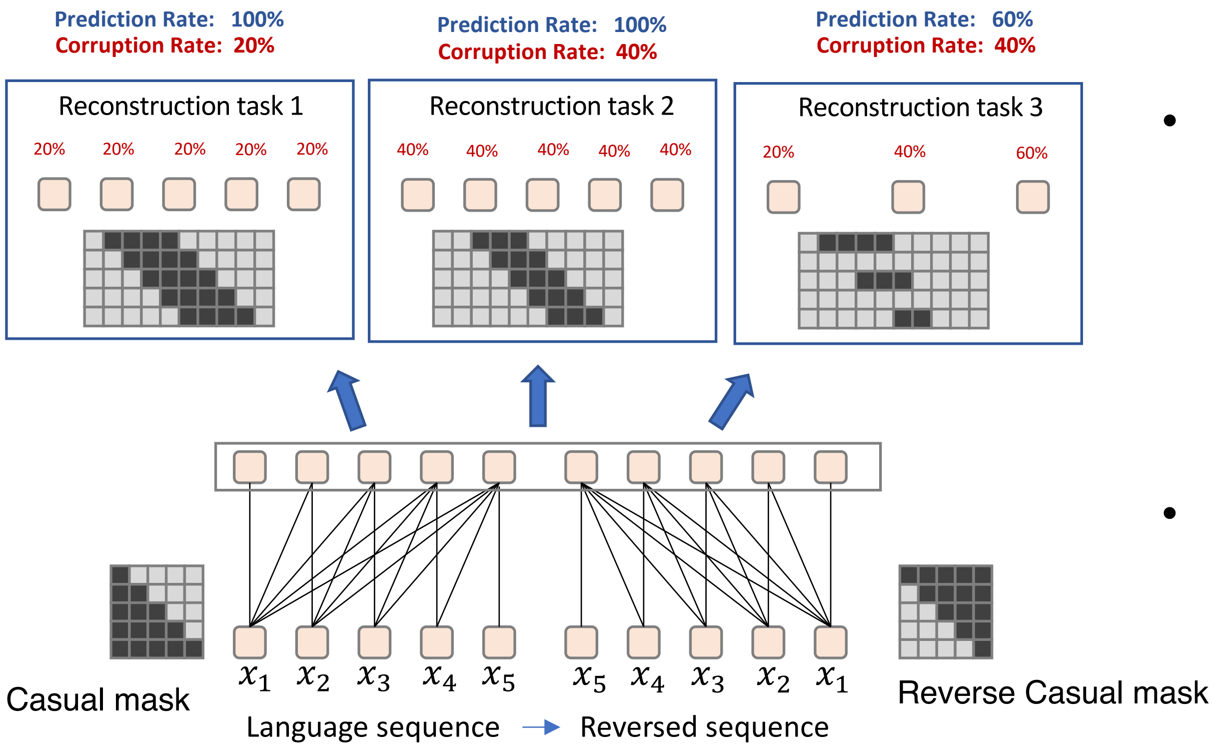
- **Performance:** find the best r_{corr} & Bidirectional context modeling
  - Auto-regressive: uni-directional context → unsatisfied performance
  - MLM: r_{pred} is coupled with and r_{corr}
  - **FLM (Ours): Decouple r_{pred} and r_{corr} , and use bidirectional context**

The proposed Free Language Modeling



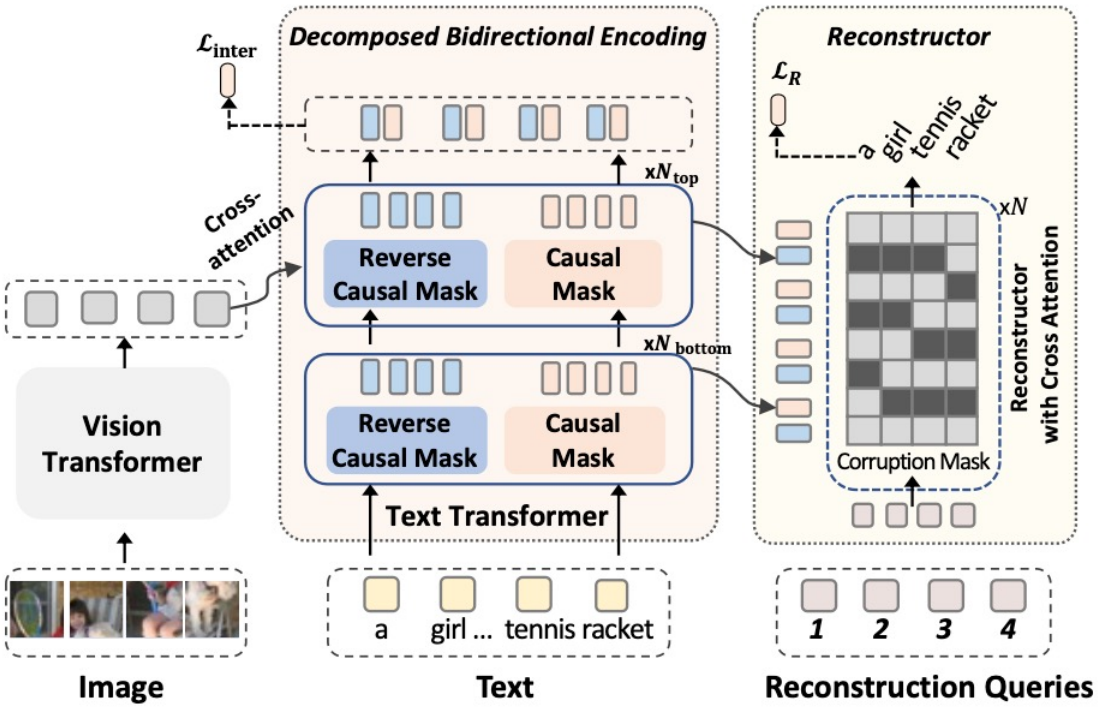
- **Decomposed bidirectional encoding**
  - Left-to-right
  - Right-to-left

The proposed Free Language Modeling



- **constructing reconstruct tasks via feature recombination**
  - Customized cross-attention masks
  - Flexible combination between r_{corr} and r_{pred}
- **Decomposed bidirectional encoding**
  - Left-to-right
  - Right-to-left

Vision-Language Pretraining with Free Language Modeling



Vision Transformer: encoding image into tokens
Text transformer: encoding text features, followed by image-text fusion (cross-attention)
Reconstructor: Feature recombination and solving reconstruction tasks

Training objectives: reconstruction loss L_R + intermediate loss L_{inter} (caption loss)

Experiments: Comparison with Language Modeling

Pretraining dataset: COCO+VG+SBU+CC3M (4M data)

Method	r_{corr}	r_{pred}	VQA test-dev	NLVR^2 dev	NLVR^2 test	Retrieval (Flickr30K) IR@1	Retrieval (Flickr30K) TR@1	COCO Captioning BLEU	COCO Captioning METER	COCO Captioning CIDEr	GPU Days (speed-up)
AR	50%	100%	72.85	75.79	76.29	66.59	84.10	<u>35.70</u>	<u>28.86</u>	<u>120.6</u>	9.6 (6.1×)
PrefixLM	25%	50%	72.64	75.73	76.17	66.21	82.70	35.50	28.79	119.4	10.0 (5.9×)
MLM	15%	15%	73.52	77.46	78.28	71.33	<u>88.40</u>	34.90	28.50	117.5	58.7 (1×)
MLM	40%	40%	**73.95**	<u>77.62</u>	<u>78.60</u>	**73.41**	**89.20**	35.50	28.79	120.3	58.7 (1×)
FLM (Ours)	$1/L$	100%	<u>73.85</u>	**77.99**	**78.63**	<u>72.81</u>	87.40	**36.68**	**29.17**	**123.0**	22.7 (2.5×)

- FLM achieves a 2.5× speed-up over MLM
- FLM keeps comparable performance on VL understanding tasks and superior performance on VL generation tasks.

Experiments: Corruption & Prediction Rate

corruption on features

corruption on inputs

Corruption	VQA	NLVR^2
span corruption ($1/L$)	73.85	78.63
span corruption (30%)	73.96	**78.83**
span corruption (40%)	**74.04**	78.82
span corruption (50%)	74.01	77.84
random corruption (15%)	73.93	78.38
random corruption (30%)	73.69	77.74

(d) Corruption Rate. FLM enables a flexible choice of the corruption rate.

Prediction rate	VQA	NLVR^2
50%	73.74	77.47
75%	73.89	77.65
90%	**74.00**	78.17
100%	73.85	**78.63**

- With a 100% prediction rate, FLM benefits from a span corruption of 40%.
- No obvious performance gap between Input-level corruption and feature-level corruption.

- FLM with a larger prediction rate improves performance.

Experiments: Comparison with SOTA

Model	Pretrain. Task	Pretrain. Time (GPU Days)	VQAv2 test-dev	VQAv2 test-std	NLVR^2 dev	NLVR^2 test	BLEU4	COCO Captioning METEOR	COCO Captioning CIDEr	COCO Captioning SPICE
Pre-trained with <10M images										
UNITER$_{\text{LARGE}}$ [6]	MLM, ITM, MVM, WRA	152 (V100)	73.82	74.02	79.12	79.98	-	-		
UNIMO$_{\text{LARGE}}$ [20]	MLM, MVM, ITC	640 (V100)	75.06	75.27	-	-	-	-		
OSCAR	MLM, ITM	220 (V100)	73.61	73.82	79.12	80.37	37.4	**30.7**	127.8	23.5
VinVL$_{\text{BASE}}$ [41]	MLM, ITM	320 (V100)	75.95	76.12	82.05	83.08	38.2	30.3	129.3	23.6
VinVL$_{\text{LARGE}}$ [41]	MLM, ITM	320 (V100)	76.52	76.60	82.67	**83.98**	38.5	30.4	130.8	23.4
PixelBERT [12]	MLM, ITM	-	74.45	74.55	76.5	77.2	-	-	-	
CLIP-ViL [30]	MLM, ITM, VQA	40 (A100)	76.48	76.70	-	-	40.2*	29.7*	134.2*	23.8*
ViLT [41]	MLM, ITM, WRA	192 (V100)	71.26	-	75.70	76.13	-	-		
ALBEF (4M) [17]	MLM, ITM	28 (A100)	71.40	-	-	77.51	-	-		
ALBEF (4M) [17]	MLM, ITM, ITC	28 (A100)	74.54	74.70	80.24	80.50	-	-		
METER$_{\text{BASE}}$ [9]	MLM, ITM	64 (A100)	77.68	77.64	82.33	83.05	38.8	30.0	128.2	23.0
Ours$_{\text{LARGE}}$ (4M)	FLM	18 (V100)	77.80	77.84	81.77	81.83	38.3	30.2	130.9	-
Pre-trained with 10M∼100M images										
ALBEF (14M) [17]	MLM, ITM, ITC	140 (A100)	75.84	76.04	82.55	83.14	-	-		
BLIP (14M) [16]	AR, ITM, ITC	112 (A100)	77.54	77.62	82.67	82.30	38.6	-	129.7	-
Ours$_{\text{LARGE}}$ (13M)	FLM	52 (V100)	**78.18**	**78.24**	**82.90**	83.86	**39.1**	30.3	**132.7**	-
Pre-trained with >100M images										
SimVLM$_{\text{BASE}}$ (1.8B) [36]	PrefixLM	-	77.87	78.14	81.72	81.77	39.0	32.9	134.8	24.0
SimVLM$_{\text{HUGE}}$ (1.8B) [36]	PrefixLM	-	80.03	80.34	84.53	85.15	40.6	33.7	143.3	25.4
LEMON (400M)	MLM	-	-	-	-	-	40.3	30.2	133.3	23.3

- Competitive performance on VQA, NLVR2, Image Captioning
- Training with FLM is more efficient than previous methods

JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

Thank you for your listening!

SUSTech
Southern University of Science and Technology
香 港 大 學
THE UNIVERSITY OF HONG KONG
Tencent
ARC Applied Research Center
PENGCHENG LABORATORY
上海人工智能实验室
Shanghai Artificial Intelligence Laboratory

Project Page

https://github.com/TencentARC/FLM