ICML | 2022

ICML
International Conference
On Machine Learning

VLMixer: Unpaired Vision-Language Pre-training via Cross-Modal CutMix

Teng Wang[12], Wenhao Jiang[3], Zhichao Lu[1], Feng Zheng[1], Ran Cheng[1], Chengguo Yin[3], Ping Luo[2]

[1] Southern University of Science and Technology [2] The University of Hong Kong

[3] Data Platform, Tencent

SUSTech
Southern University of Science and Technology

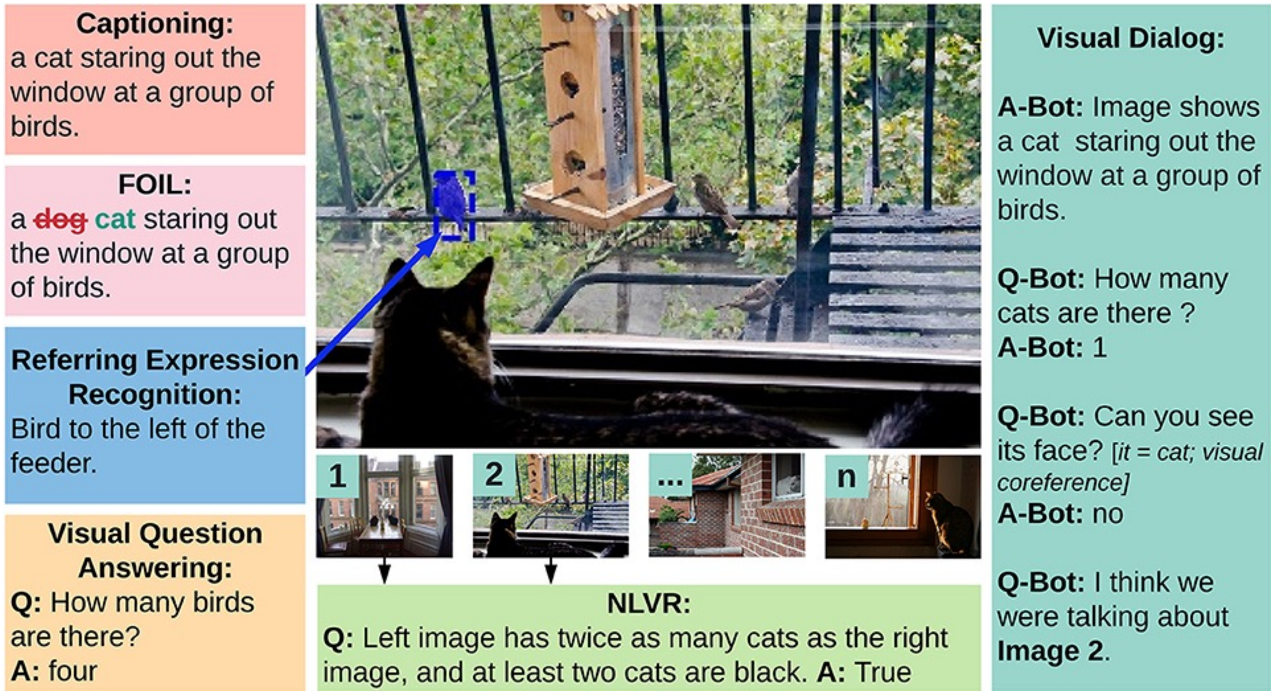
香港大學
THE UNIVERSITY OF HONG KONG

Tencent 腾讯

Vision-Language Pre-training (VLP)

Pretrained VL Model

Adaption



Captioning: a cat staring out the window at a group of birds.

FOIL: a ~~dog~~ cat staring out the window at a group of birds.

Referring Expression Recognition: Bird to the left of the feeder.

Visual Question Answering:
Q: How many birds are there?
A: four

NLVR:
Q: Left image has twice as many cats as the right image, and at least two cats are black. **A:** True

Visual Dialog:

A-Bot: Image shows a cat staring out the window at a group of birds.

Q-Bot: How many cats are there ?
A-Bot: 1

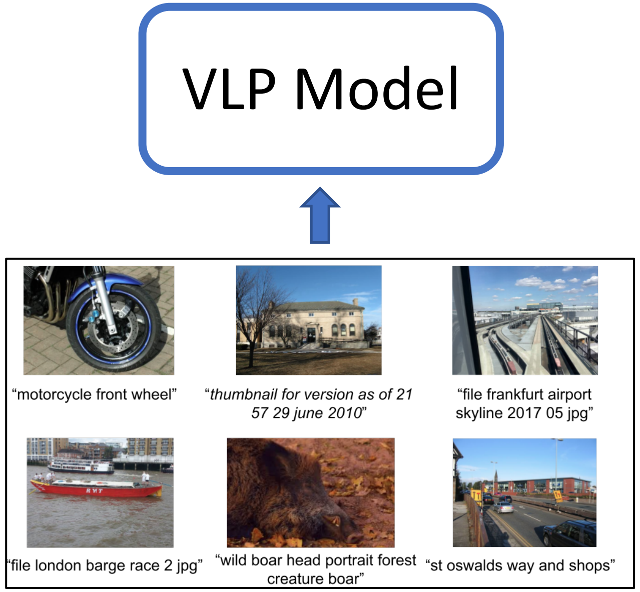
Q-Bot: Can you see its face? *[it = cat; visual coreference]*
A-Bot: no

Q-Bot: I think we were talking about **Image 2**.

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.

Paired VLP



VLP Model

"motorcycle front wheel"

"*thumbnail for version as of 21 57 29 june 2010*"

"file frankfurt airport skyline 2017 05 jpg"

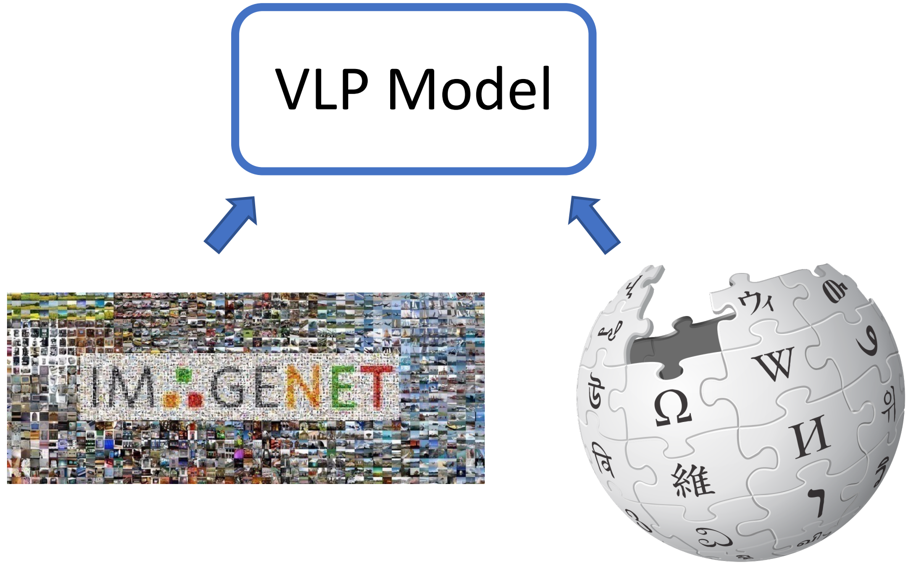
"file london barge race 2 jpg"

"wild boar head portrait forest creature boar"

"st oswalds way and shops"

Image-text pairs (eg., COCO, CC3M)

Unpaired VLP



VLP Model

Image datasets
(eg., ImageNet)

Text corpora
(eg. Wikipedia)

- **Human-annotated (COCO, Visual Genome)**
  - Hard to scale-up
  - Language bias
- **Auto-crawled from Internet (Conceptual Captions)**
  - Complicated data cleaning
  - Weak Alignment
  - Unfriendly to minority language

Stand-alone images and texts

- Easy to scaling-up
- Diverse visual/language patterns
- Less bias

Alignment Matters!

Paired VLP

- **Instance-level alignment**
  - Text-Image contrastive learning
  - Text-Image matching
- **Token-level alignment**
  - Masked language/image modeling (MLM/MIM)

Unpaired VLP

- **Instance-level alignment**
  - Contrast between a sentence and its multimodal view
- **Token-level alignment**
  - MLM on the multimodal sentence

Cross-modal CutMix

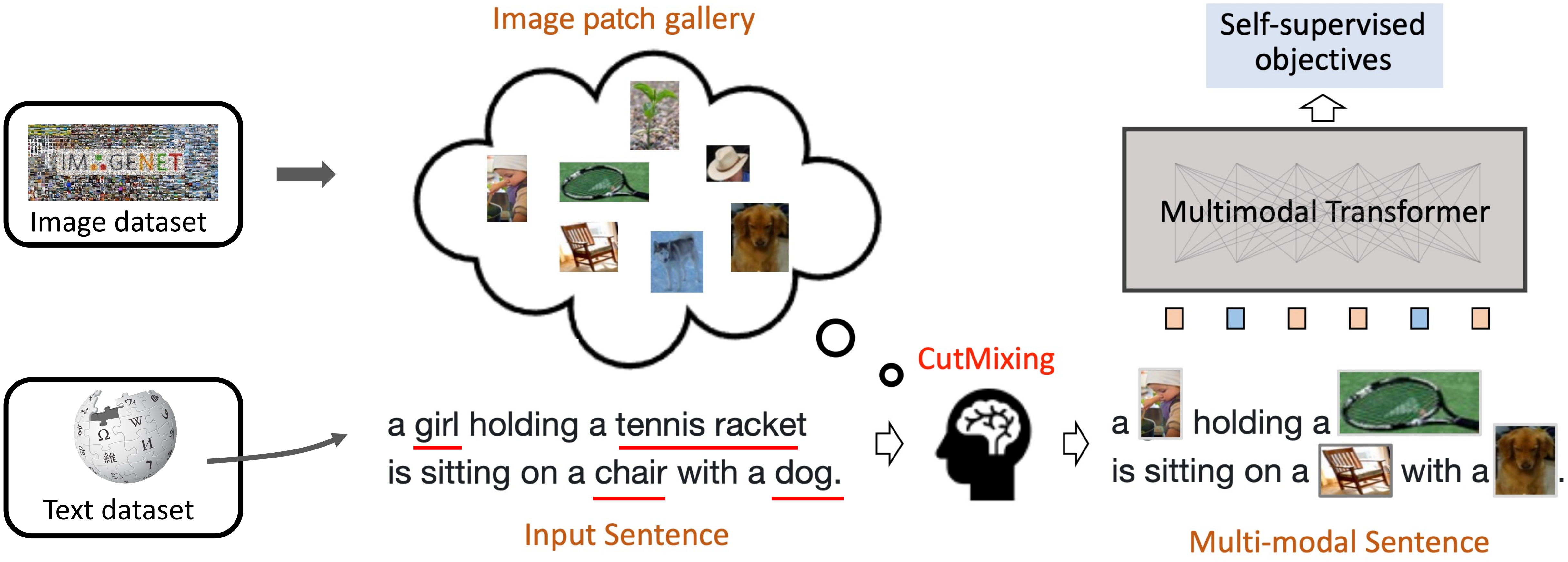


Image dataset

Text dataset

Image patch gallery

a girl holding a tennis racket is sitting on a chair with a dog.

Input Sentence

CutMixing

a holding a is sitting on a with a .

Multi-modal Sentence

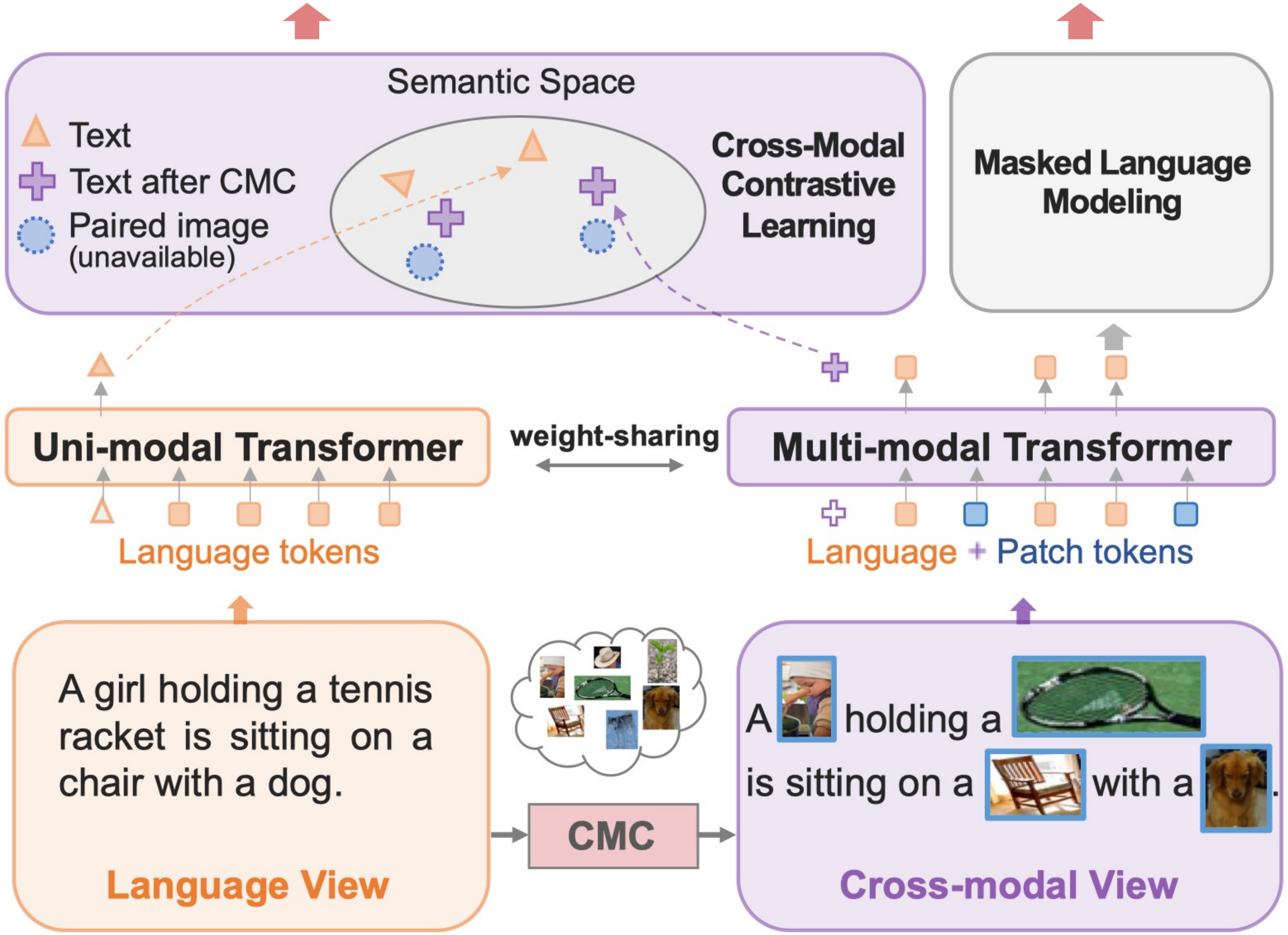
Multimodal Transformer

Self-supervised objectives

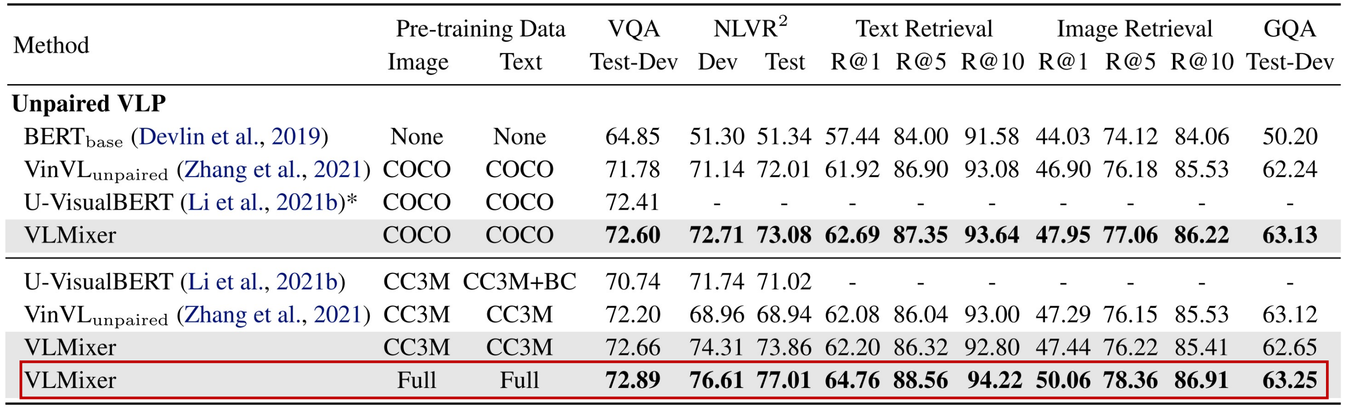
Self-supervised Pretraining Objectives

Instance-level alignment

Token-level alignment

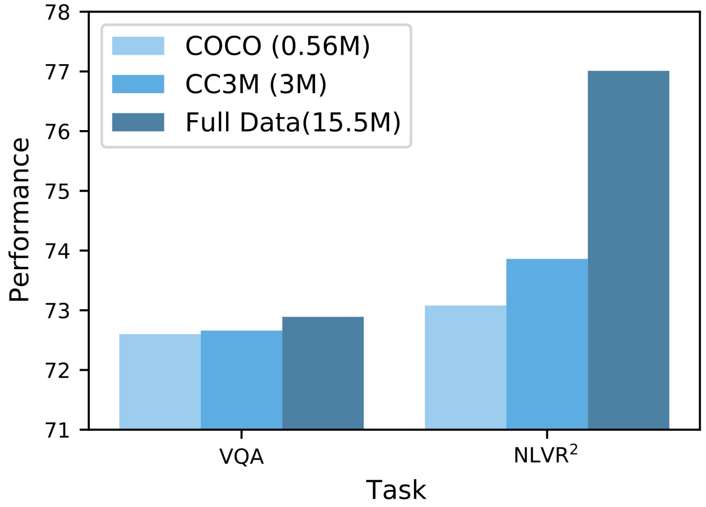


Experiments



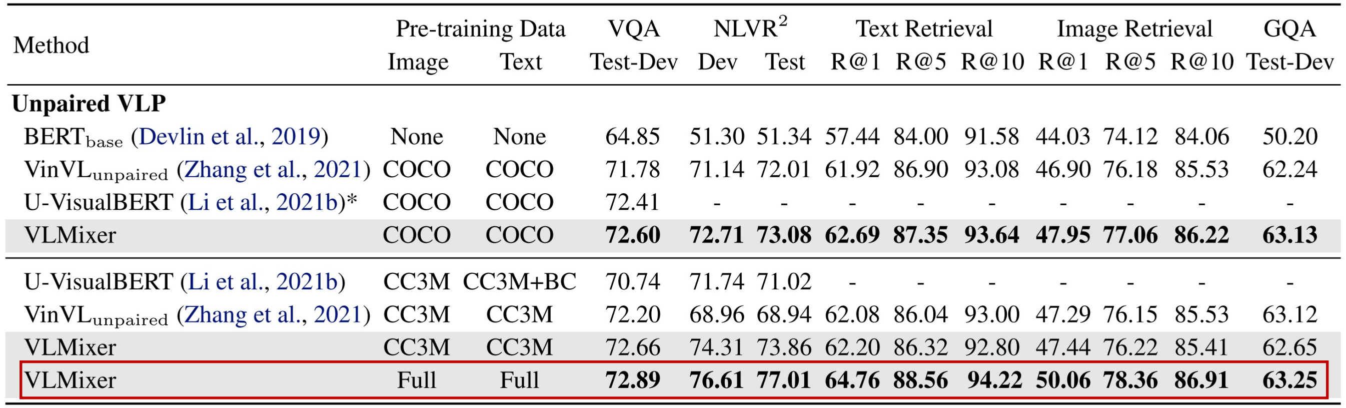
Method	Pre-training Data		VQA	NLVR^2		Text Retrieval			Image Retrieval			GQA
	Image	Text	Test-Dev	Dev	Test	R@1	R@5	R@10	R@1	R@5	R@10	Test-Dev
Unpaired VLP												
BERT$_{\text{base}}$ (Devlin et al., 2019)	None	None	64.85	51.30	51.34	57.44	84.00	91.58	44.03	74.12	84.06	50.20
VinVL$_{\text{unpaired}}$ (Zhang et al., 2021)	COCO	COCO	71.78	71.14	72.01	61.92	86.90	93.08	46.90	76.18	85.53	62.24
U-VisualBERT (Li et al., 2021b)*	COCO	COCO	72.41	-	-	-	-	-	-	-	-	-
VLMixer	COCO	COCO	**72.60**	**72.71**	**73.08**	**62.69**	**87.35**	**93.64**	**47.95**	**77.06**	**86.22**	**63.13**
U-VisualBERT (Li et al., 2021b)	CC3M	CC3M+BC	70.74	71.74	71.02	-	-	-	-	-	-	-
VinVL$_{\text{unpaired}}$ (Zhang et al., 2021)	CC3M	CC3M	72.20	68.96	68.94	62.08	86.04	93.00	47.29	76.15	85.53	63.12
VLMixer	CC3M	CC3M	72.66	74.31	73.86	62.20	86.32	92.80	47.44	76.22	85.41	62.65
VLMixer	Full	Full	**72.89**	**76.61**	**77.01**	**64.76**	**88.56**	**94.22**	**50.06**	**78.36**	**86.91**	**63.25**

Superior performance on five downstream tasks.



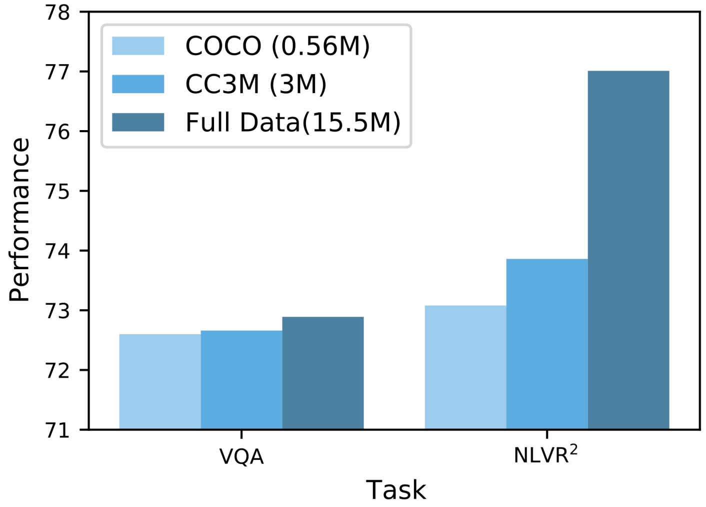
VLMixer benefits from the data scale.

Experiments

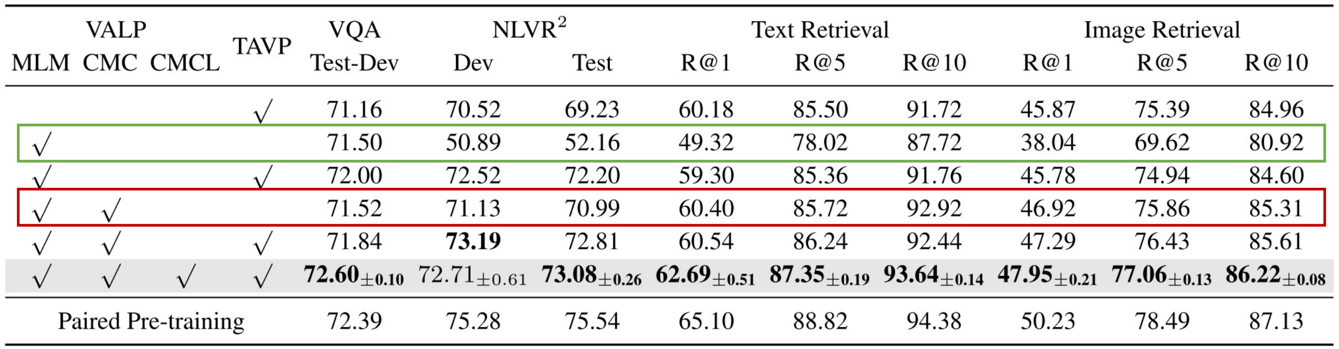


Method	Pre-training Data		VQA	NLVR^2		Text Retrieval			Image Retrieval			GQA
	Image	Text	Test-Dev	Dev	Test	R@1	R@5	R@10	R@1	R@5	R@10	Test-Dev
Unpaired VLP												
BERT$_{\text{base}}$ (Devlin et al., 2019)	None	None	64.85	51.30	51.34	57.44	84.00	91.58	44.03	74.12	84.06	50.20
VinVL$_{\text{unpaired}}$ (Zhang et al., 2021)	COCO	COCO	71.78	71.14	72.01	61.92	86.90	93.08	46.90	76.18	85.53	62.24
U-VisualBERT (Li et al., 2021b)*	COCO	COCO	72.41	-	-	-	-	-	-	-	-	-
VLMixer	COCO	COCO	**72.60**	**72.71**	**73.08**	**62.69**	**87.35**	**93.64**	**47.95**	**77.06**	**86.22**	**63.13**
U-VisualBERT (Li et al., 2021b)	CC3M	CC3M+BC	70.74	71.74	71.02	-	-	-	-	-	-	-
VinVL$_{\text{unpaired}}$ (Zhang et al., 2021)	CC3M	CC3M	72.20	68.96	68.94	62.08	86.04	93.00	47.29	76.15	85.53	63.12
VLMixer	CC3M	CC3M	72.66	74.31	73.86	62.20	86.32	92.80	47.44	76.22	85.41	62.65
VLMixer	Full	Full	**72.89**	**76.61**	**77.01**	**64.76**	**88.56**	**94.22**	**50.06**	**78.36**	**86.91**	**63.25**

Superior performance on five downstream tasks.

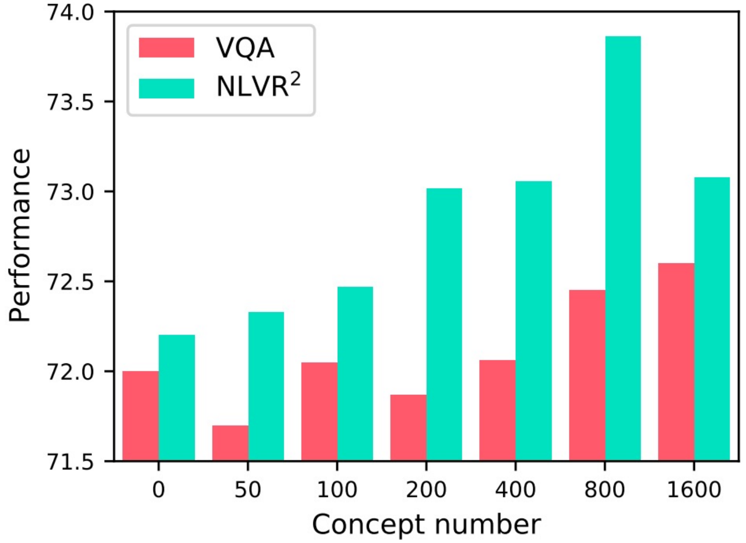


VLMixer benefits from the data scale.



VALP			TAVP	VQA	NLVR^2		Text Retrieval			Image Retrieval		
MLM	CMC	CMCL		Test-Dev	Dev	Test	R@1	R@5	R@10	R@1	R@5	R@10
			√	71.16	70.52	69.23	60.18	85.50	91.72	45.87	75.39	84.96
√				71.50	50.89	52.16	49.32	78.02	87.72	38.04	69.62	80.92
√			√	72.00	72.52	72.20	59.30	85.36	91.76	45.78	74.94	84.60
√	√			71.52	71.13	70.99	60.40	85.72	92.92	46.92	75.86	85.31
√	√		√	71.84	**73.19**	72.81	60.54	86.24	92.44	47.29	76.43	85.61
√	√	√	√	**72.60**$_{\pm0.10}$	72.71$_{\pm0.61}$	**73.08**$_{\pm0.26}$	**62.69**$_{\pm0.51}$	**87.35**$_{\pm0.19}$	**93.64**$_{\pm0.14}$	**47.95**$_{\pm0.21}$	**77.06**$_{\pm0.13}$	**86.22**$_{\pm0.08}$
Paired Pre-training				72.39	75.28	75.54	65.10	88.82	94.38	50.23	78.49	87.13

- CMC improves NLVR^2 and retrieval tasks.
- CMC + CMCL improve VQA.
- Unpaired model is slightly inferior to the paired counterpart.



Diverse patch gallery helps cross-modal alignment.

Thanks for your listening!