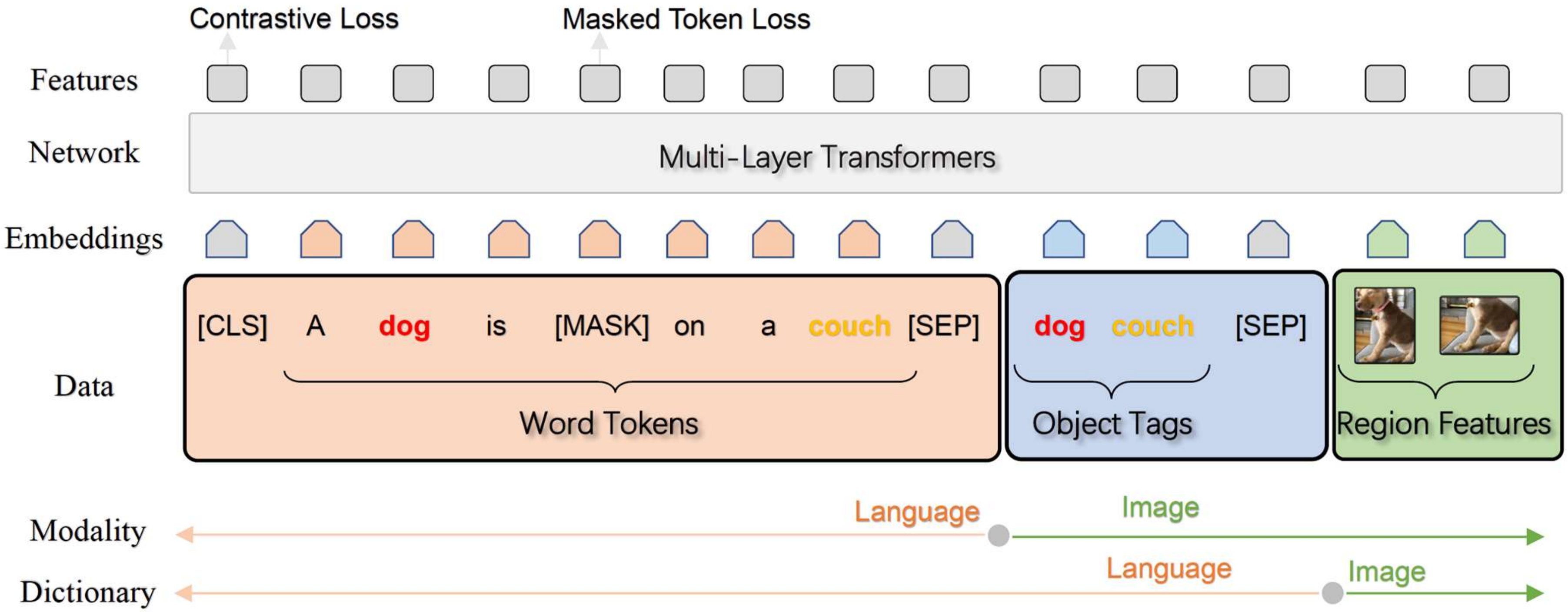
SimVLM: Simple Visual Language Model Pretraining with Weak Supervision

https://arxiv.org/abs/2108.10904

Background

- Vision-Language Pretraining
- Task:
  - VGA, NLVR2, Ranking, Img2Text
- Previous methods:
  - Need object detection features. LXMERT (Tan & Bansal, 2019), VLBERT (Su et al., 2020), VisualBERT (Li et al., 2019), UNITER (Chen et al., 2020b), Villa (Gan et al., 2020), Oscar (Li et al., 2020), ERNIE-ViL (Yu et al., 2021), UNIMO (Li et al., 2021), VinVL (Zhang et al., 2021), VIVO (Hu et al., 2021) VL-T5 (Cho et al., 2021)
  - Trained on human annotated data sets like Visual Genome (Krishna et al., 2016)
- Pre-Train Task:
  - image-text matching
  - object attribute prediction
  - word-region alignment
  - word-patch alignment
  - Mask LM

Oscar


Contrastive Loss
Masked Token Loss
Features
Network
Multi-Layer Transformers
Embeddings
Data
[CLS] A dog is [MASK] on a couch [SEP]
Word Tokens
dog couch [SEP]
Object Tags
Region Features
Modality
Language
Image
Dictionary
Language
Image

ViLT

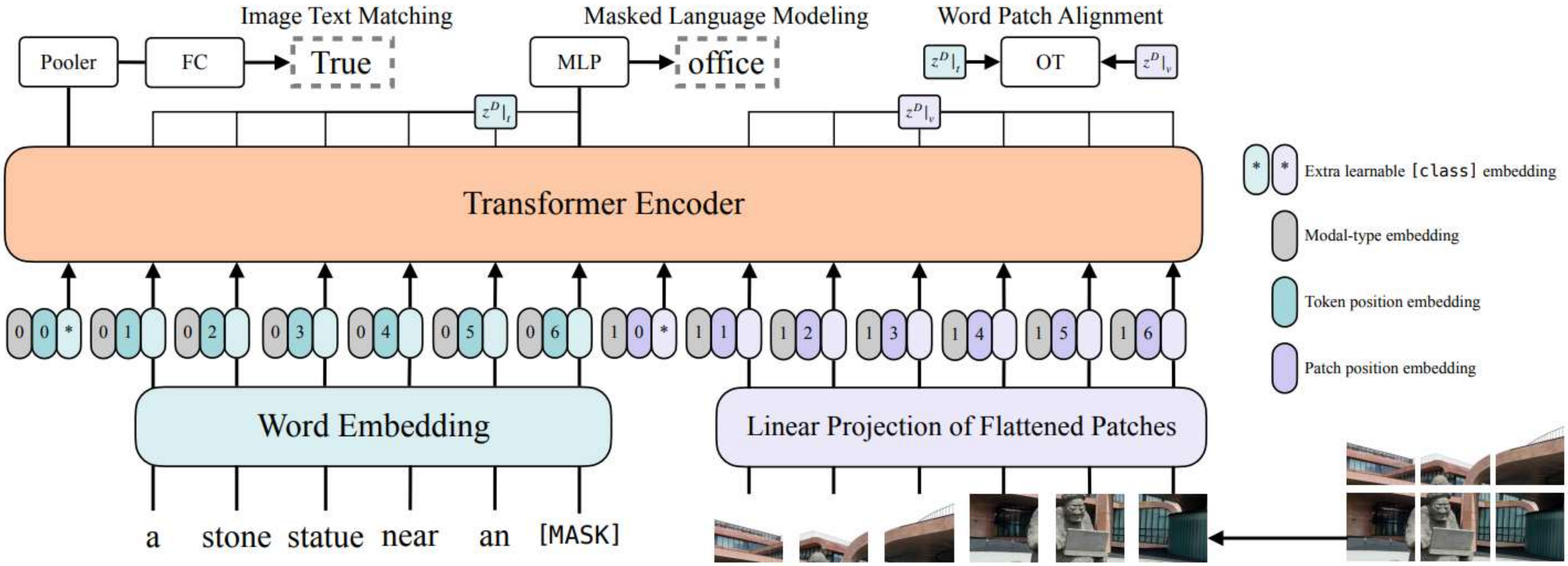
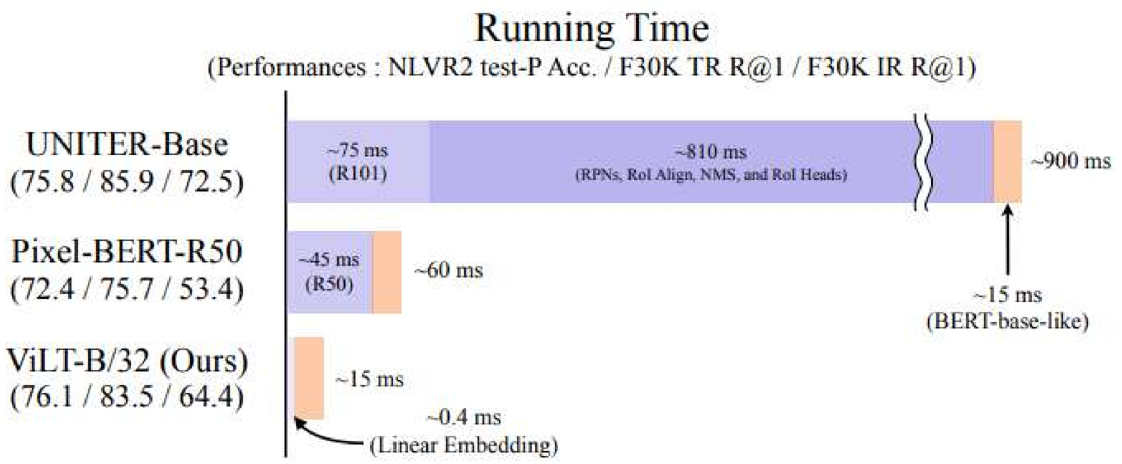


Image Text Matching
Masked Language Modeling
Word Patch Alignment
Pooler
FC
True
MLP
office
$z^D|_t$
OT
$z^D|_v$
$z^D|_t$
$z^D|_v$
Transformer Encoder
0 0 *
0 1
0 2
0 3
0 4
0 5
0 6
1 0 *
1 1
1 2
1 3
1 4
1 5
1 6
Word Embedding
Linear Projection of Flattened Patches
a
stone
statue
near
an
[MASK]
* * Extra learnable [class] embedding
Modal-type embedding
Token position embedding
Patch position embedding

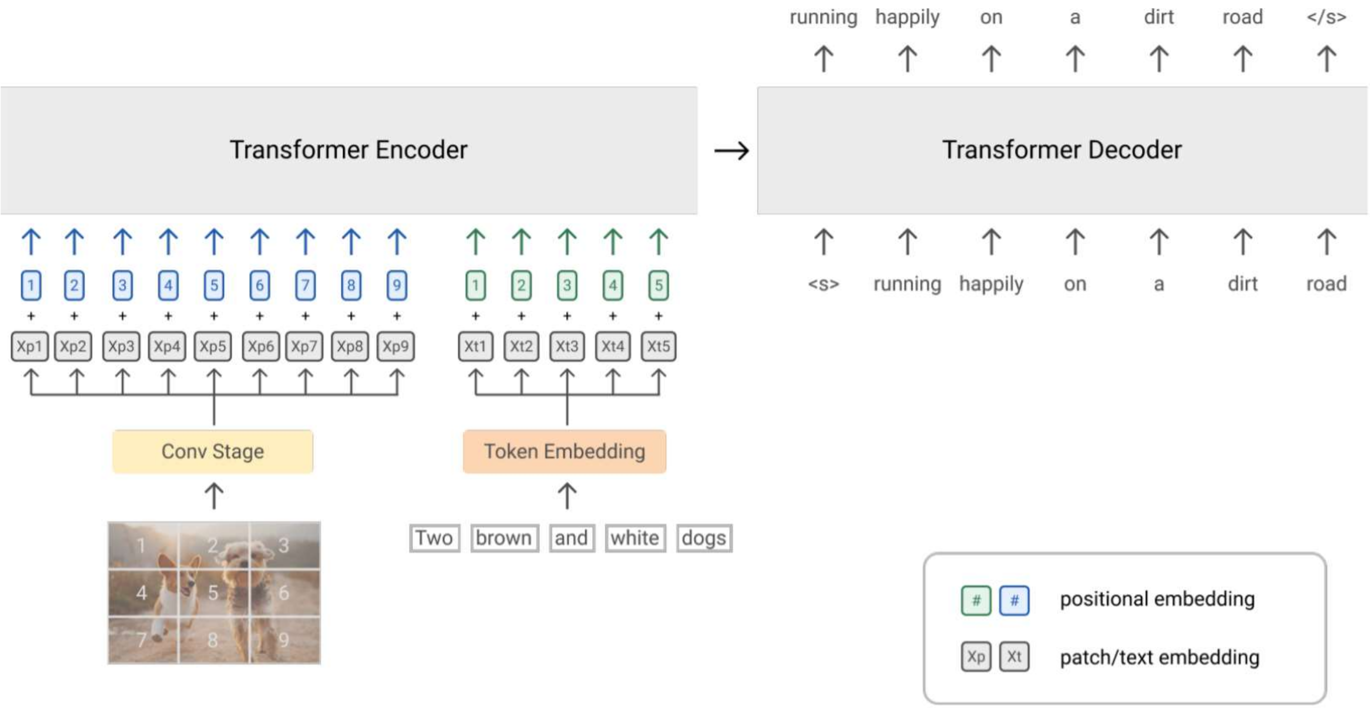
Background


Running Time
(Performances : NLVR2 test-P Acc. / F30K TR R@1 / F30K IR R@1)
UNITER-Base
(75.8 / 85.9 / 72.5)
~75 ms
(R101)
~810 ms
(RPNs, RoI Align, NMS, and RoI Heads)
~900 ms
Pixel-BERT-R50
(72.4 / 75.7 / 53.4)
~45 ms
(R50)
~60 ms
~15 ms
(BERT-base-like)
ViLT-B/32 (Ours)
(76.1 / 83.5 / 64.4)
~15 ms
~0.4 ms
(Linear Embedding)

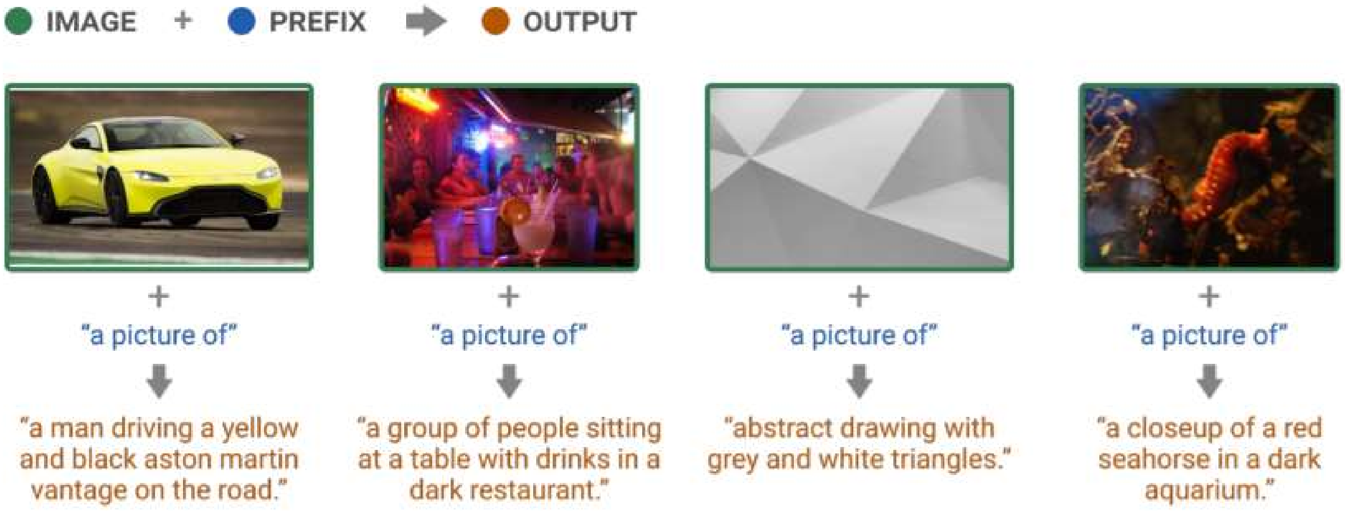
Introduction

- Vision-Language Pretraining
- No region-of-interest (ROI) features from images
- A single prefix language modeling objective
  - Encoder-decoder arch
    - Encoder (ViT + Conv)
      - Image(Bi) + Text(Uni), prefix language mask(like UniLM, also M6)
    - Decoder
      - Text
  - Task:
    - Image+Text2Text
    - Text2Text (T5)
- Trained on large-scale weak supervision dataset
- Zero-shot, transfer ability

Method


running
happily
on
a
dirt
road
</s>
Transformer Encoder
Transformer Decoder
<s>
running
happily
on
a
dirt
road
1
2
3
4
5
6
7
8
9
1
2
3
4
5
Xp1
Xp2
Xp3
Xp4
Xp5
Xp6
Xp7
Xp8
Xp9
Xt1
Xt2
Xt3
Xt4
Xt5
Conv Stage
Token Embedding
1
2
3
4
5
6
7
8
9
Two
brown
and
white
dogs
#
#
positional embedding
Xp
Xt
patch/text embedding

Method



● IMAGE + ● PREFIX → ● OUTPUT

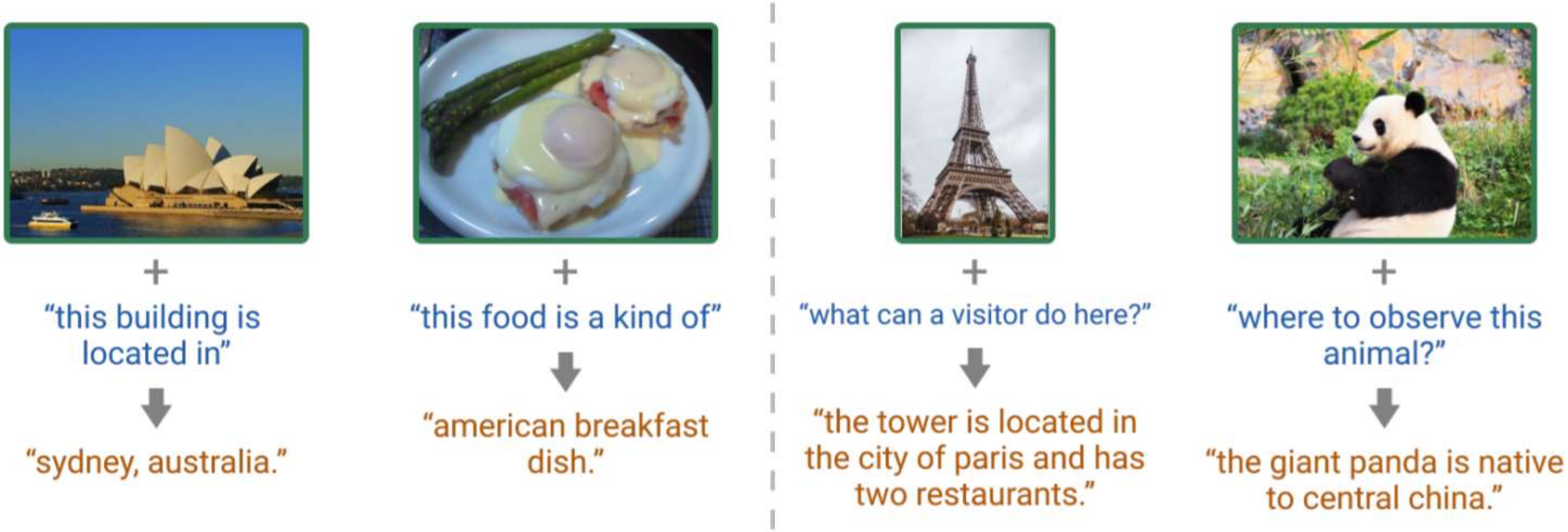
+ "a picture of" → "a man driving a yellow and black aston martin vantage on the road."

+ "a picture of" → "a group of people sitting at a table with drinks in a dark restaurant."

+ "a picture of" → "abstract drawing with grey and white triangles."

+ "a picture of" → "a closeup of a red seahorse in a dark aquarium."

Method



+

“this building is located in”

“sydney, australia.”

+

“this food is a kind of”

“american breakfast dish.”

+

“what can a visitor do here?”

“the tower is located in the city of paris and has two restaurants.”

+

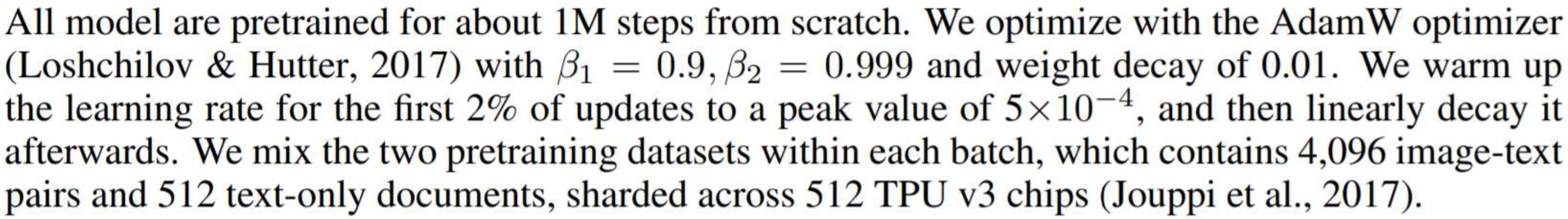
“where to observe this animal?”

“the giant panda is native to central china.”

Method

- Image:
  - Raw Image -> Conv -> Contextualized Patches -> Encoder
  - Patch size 16x16, 14 x 14 image patches
- Text:
  - Raw Text -> BEP -> Embedding -> Encoder
  - Max sequence Length 256 for encoder & decoder
- Details:
  - trainable 1D positional embeddings for image and text inputs separately
  - 2D relative attention for the image patches
  - No modality type embeddings(no gains)
  - Extra Conv for Image (pixel is low level feature, BPE is more abstract)
- Details:
  - Training from scratch using large-scale noisy image-text data (In House)
  - 1.8B image-text pairs and C4 (800GB) (CC 3M)
  - 4,096 image-text pairs and 512 text-only documents, sharded across 512 TPU v3 chips

Pre-training

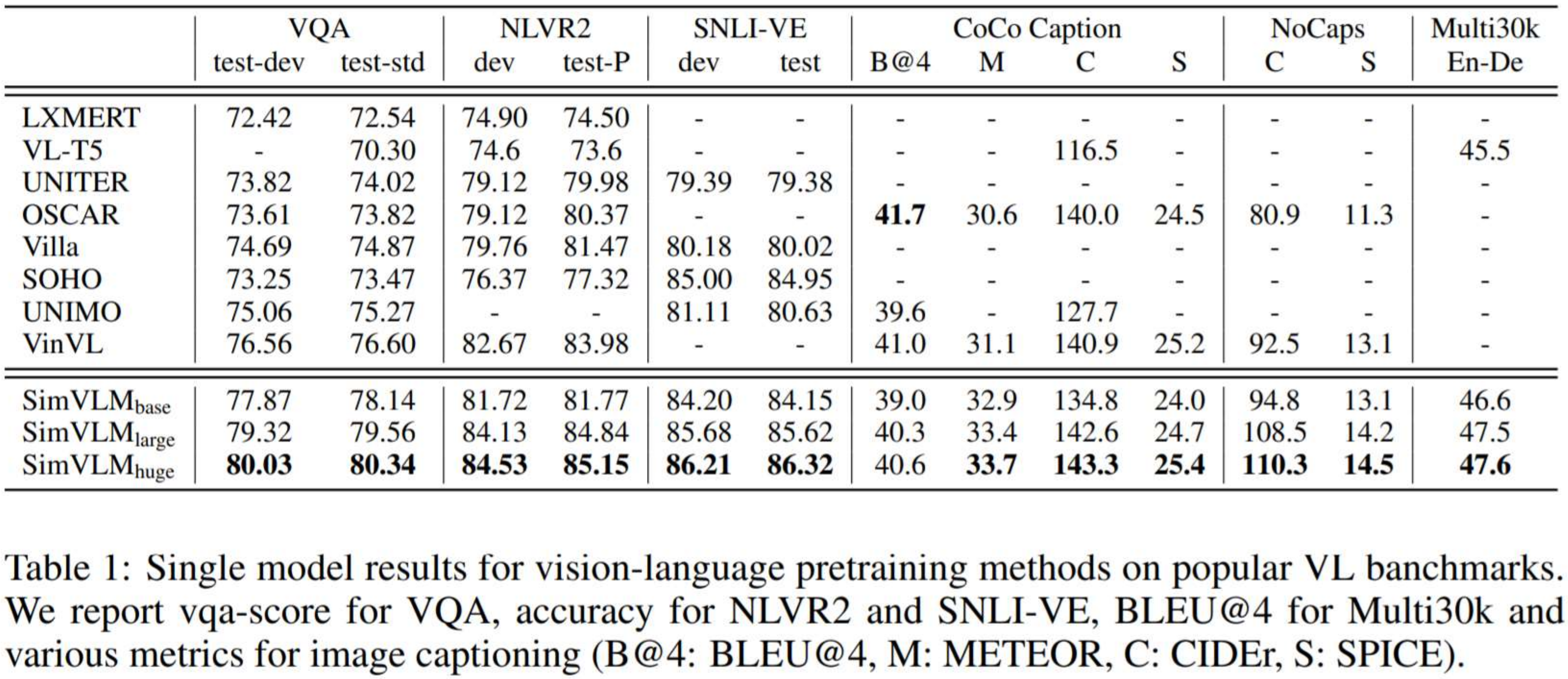


All model are pretrained for about 1M steps from scratch. We optimize with the AdamW optimizer (Loshchilov & Hutter, 2017) with $\beta_1 = 0.9, \beta_2 = 0.999$ and weight decay of 0.01. We warm up the learning rate for the first 2% of updates to a peak value of 5×10^{-4}, and then linearly decay it afterwards. We mix the two pretraining datasets within each batch, which contains 4,096 image-text pairs and 512 text-only documents, sharded across 512 TPU v3 chips (Jouppi et al., 2017).

Fine-tuning

- Visual question answering and Visual entailment:
  - Enc: image (resolution up to 480x480 by interpolation)
  - Dec: question, last token for prediction
- Visual reasoning:
  - Two Image and one text.
  - Get rep of two pair for prediction
- Image captioning:
  - No apply task-specific tricks such as CIDEr optimization
- Multimodal translation:
  - translate image descriptions in source language to target language
  - Enc: Image + Source, Dec: target

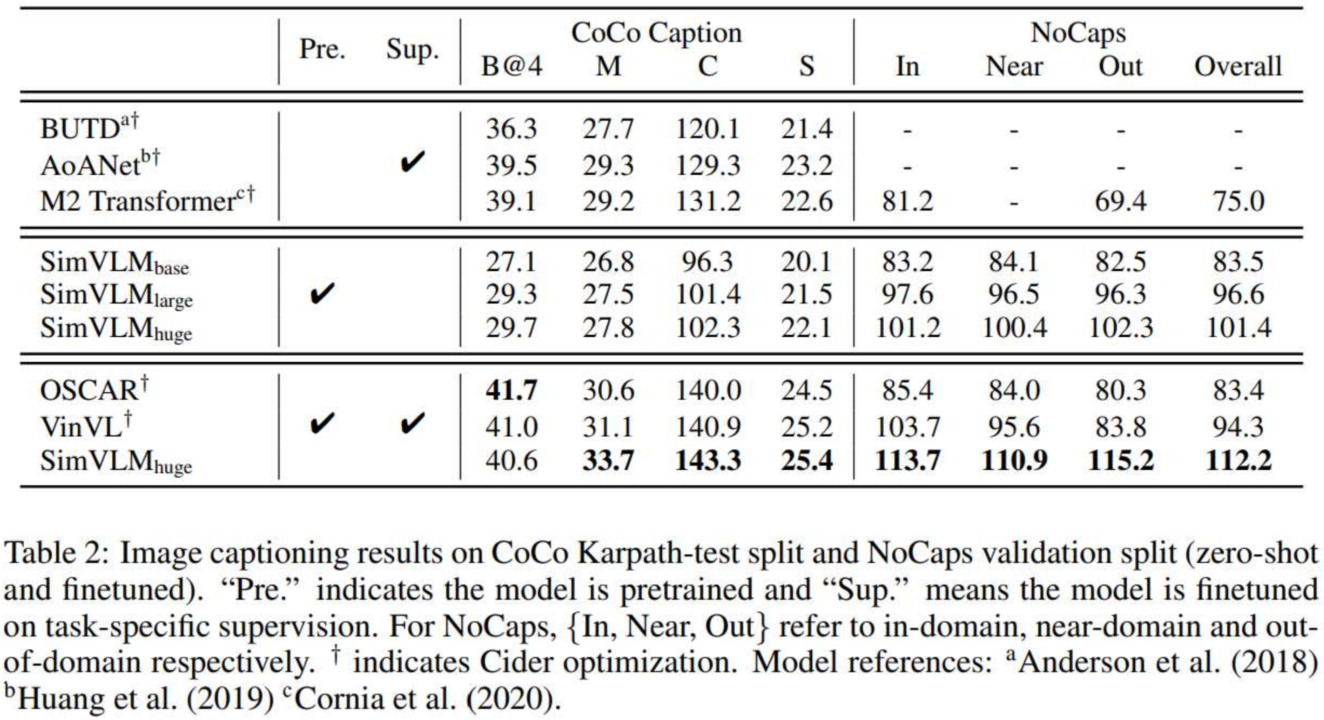
Fine-tuning



	VQA		NLVR2		SNLI-VE		CoCo Caption				NoCaps		Multi30k
	test-dev	test-std	dev	test-P	dev	test	B@4	M	C	S	C	S	En-De
LXMERT	72.42	72.54	74.90	74.50	-	-	-	-	-	-	-	-	-
VL-T5	-	70.30	74.6	73.6	-	-	-	-	116.5	-	-	-	45.5
UNITER	73.82	74.02	79.12	79.98	79.39	79.38	-	-	-	-	-	-	-
OSCAR	73.61	73.82	79.12	80.37	-	-	**41.7**	30.6	140.0	24.5	80.9	11.3	-
Villa	74.69	74.87	79.76	81.47	80.18	80.02	-	-	-	-	-	-	-
SOHO	73.25	73.47	76.37	77.32	85.00	84.95	-	-	-	-	-	-	-
UNIMO	75.06	75.27	-	-	81.11	80.63	39.6	-	127.7	-	-	-	-
VinVL	76.56	76.60	82.67	83.98	-	-	41.0	31.1	140.9	25.2	92.5	13.1	-
SimVLM$_{\text{base}}$	77.87	78.14	81.72	81.77	84.20	84.15	39.0	32.9	134.8	24.0	94.8	13.1	46.6
SimVLM$_{\text{large}}$	79.32	79.56	84.13	84.84	85.68	85.62	40.3	33.4	142.6	24.7	108.5	14.2	47.5
SimVLM$_{\text{huge}}$	**80.03**	**80.34**	**84.53**	**85.15**	**86.21**	**86.32**	40.6	**33.7**	**143.3**	**25.4**	**110.3**	**14.5**	**47.6**

Table 1: Single model results for vision-language pretraining methods on popular VL banchmarks. We report vqa-score for VQA, accuracy for NLVR2 and SNLI-VE, BLEU@4 for Multi30k and various metrics for image captioning (B@4: BLEU@4, M: METEOR, C: CIDEr, S: SPICE).

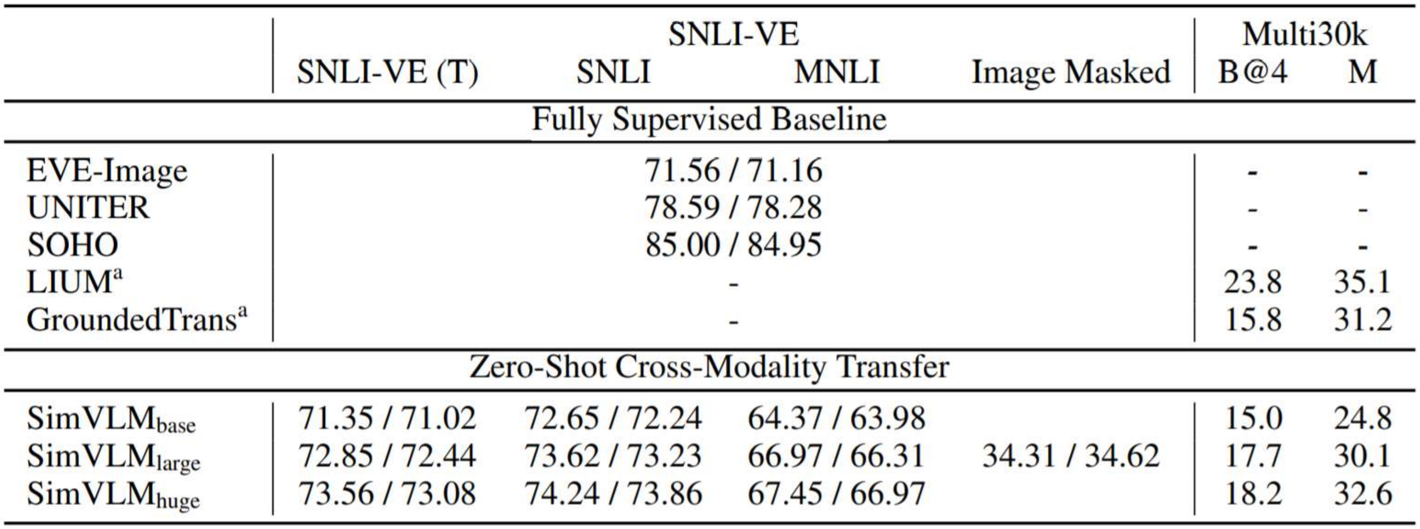
Image Caption



	Pre.	Sup.	CoCo Caption				NoCaps			
			B@4	M	C	S	In	Near	Out	Overall
BUTD[a]†		✔	36.3	27.7	120.1	21.4	-	-	-	-
AoANet[b]†			39.5	29.3	129.3	23.2	-	-	-	-
M2 Transformer[c]†			39.1	29.2	131.2	22.6	81.2	-	69.4	75.0
SimVLM$_{\text{base}}$	✔		27.1	26.8	96.3	20.1	83.2	84.1	82.5	83.5
SimVLM$_{\text{large}}$			29.3	27.5	101.4	21.5	97.6	96.5	96.3	96.6
SimVLM$_{\text{huge}}$			29.7	27.8	102.3	22.1	101.2	100.4	102.3	101.4
OSCAR†	✔	✔	**41.7**	30.6	140.0	24.5	85.4	84.0	80.3	83.4
VinVL†			41.0	31.1	140.9	25.2	103.7	95.6	83.8	94.3
SimVLM$_{\text{huge}}$			40.6	**33.7**	**143.3**	**25.4**	**113.7**	**110.9**	**115.2**	**112.2**

Table 2: Image captioning results on CoCo Karpath-test split and NoCaps validation split (zero-shot and finetuned). "Pre." indicates the model is pretrained and "Sup." means the model is finetuned on task-specific supervision. For NoCaps, {In, Near, Out} refer to in-domain, near-domain and out-of-domain respectively. † indicates Cider optimization. Model references: [a]Anderson et al. (2018) [b]Huang et al. (2019) [c]Cornia et al. (2020).

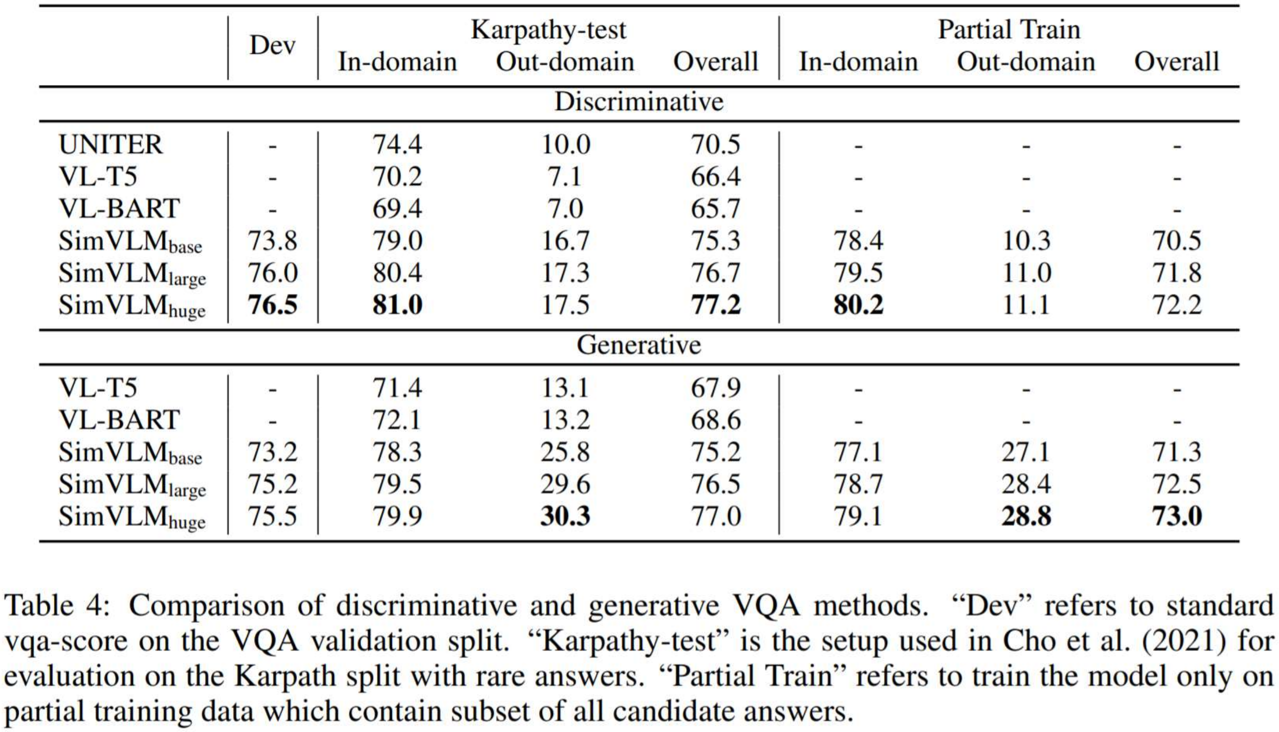
Zero-shot



	SNLI-VE				Multi30k	
	SNLI-VE (T)	SNLI	MNLI	Image Masked	B@4	M
Fully Supervised Baseline						
EVE-Image	71.56 / 71.16				-	-
UNITER	78.59 / 78.28				-	-
SOHO	85.00 / 84.95				-	-
LIUM[a]	-				23.8	35.1
GroundedTrans[a]	-				15.8	31.2
Zero-Shot Cross-Modality Transfer						
SimVLM$_{\text{base}}$	71.35 / 71.02	72.65 / 72.24	64.37 / 63.98		15.0	24.8
SimVLM$_{\text{large}}$	72.85 / 72.44	73.62 / 73.23	66.97 / 66.31	34.31 / 34.62	17.7	30.1
SimVLM$_{\text{huge}}$	73.56 / 73.08	74.24 / 73.86	67.45 / 66.97		18.2	32.6

Table 3: Zero-shot cross-modality transfer results on SNLI-VE and Multi30k. For SNLI-VE, the zero-shot model is finetuned on three source datasets: text-only SNLI-VE (Xie et al., 2019), SNLI (Bowman et al., 2015), and MNLI (Williams et al., 2017). Model reference: [a](Specia et al., 2016).

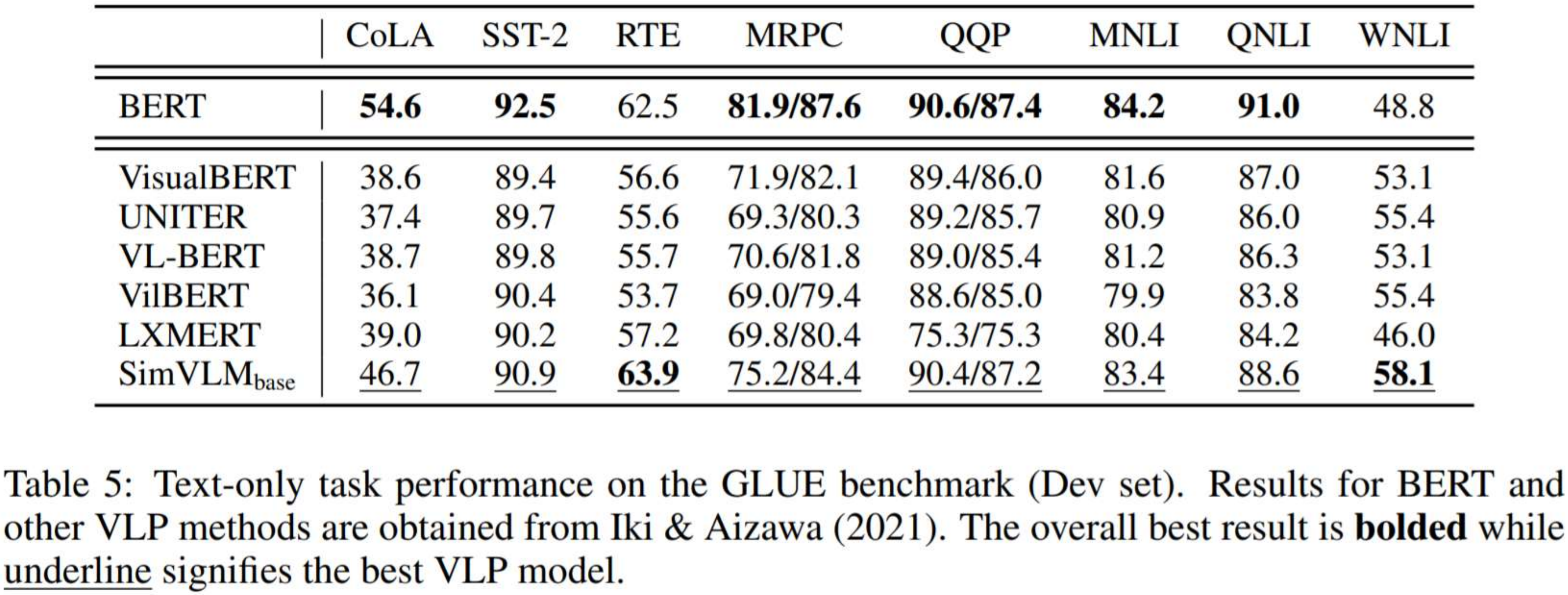
Discriminative vs Generative



	Dev	Karpathy-test			Partial Train		
		In-domain	Out-domain	Overall	In-domain	Out-domain	Overall
Discriminative							
UNITER	-	74.4	10.0	70.5	-	-	-
VL-T5	-	70.2	7.1	66.4	-	-	-
VL-BART	-	69.4	7.0	65.7	-	-	-
SimVLM$_{base}$	73.8	79.0	16.7	75.3	78.4	10.3	70.5
SimVLM$_{large}$	76.0	80.4	17.3	76.7	79.5	11.0	71.8
SimVLM$_{huge}$	**76.5**	**81.0**	17.5	**77.2**	**80.2**	11.1	72.2
Generative							
VL-T5	-	71.4	13.1	67.9	-	-	-
VL-BART	-	72.1	13.2	68.6	-	-	-
SimVLM$_{base}$	73.2	78.3	25.8	75.2	77.1	27.1	71.3
SimVLM$_{large}$	75.2	79.5	29.6	76.5	78.7	28.4	72.5
SimVLM$_{huge}$	75.5	79.9	**30.3**	77.0	79.1	**28.8**	**73.0**

Table 4: Comparison of discriminative and generative VQA methods. “Dev” refers to standard vqa-score on the VQA validation split. “Karpathy-test” is the setup used in Cho et al. (2021) for evaluation on the Karpath split with rare answers. “Partial Train” refers to train the model only on partial training data which contain subset of all candidate answers.

NLU benchmark

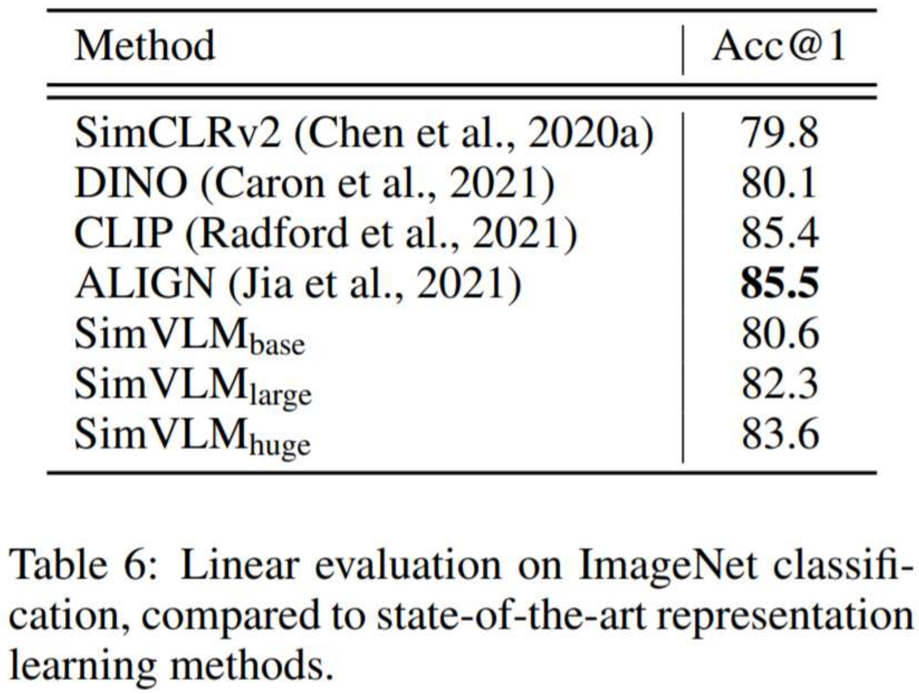


	CoLA	SST-2	RTE	MRPC	QQP	MNLI	QNLI	WNLI
BERT	**54.6**	**92.5**	62.5	**81.9/87.6**	**90.6/87.4**	**84.2**	**91.0**	48.8
VisualBERT	38.6	89.4	56.6	71.9/82.1	89.4/86.0	81.6	87.0	53.1
UNITER	37.4	89.7	55.6	69.3/80.3	89.2/85.7	80.9	86.0	55.4
VL-BERT	38.7	89.8	55.7	70.6/81.8	89.0/85.4	81.2	86.3	53.1
VilBERT	36.1	90.4	53.7	69.0/79.4	88.6/85.0	79.9	83.8	55.4
LXMERT	39.0	90.2	57.2	69.8/80.4	75.3/75.3	80.4	84.2	46.0
SimVLM$_{base}$	46.7	90.9	**63.9**	75.2/84.4	90.4/87.2	83.4	88.6	**58.1**

Table 5: Text-only task performance on the GLUE benchmark (Dev set). Results for BERT and other VLP methods are obtained from Iki & Aizawa (2021). The overall best result is **bolded** while underline signifies the best VLP model.

(SimVLM is Joint training with T5)

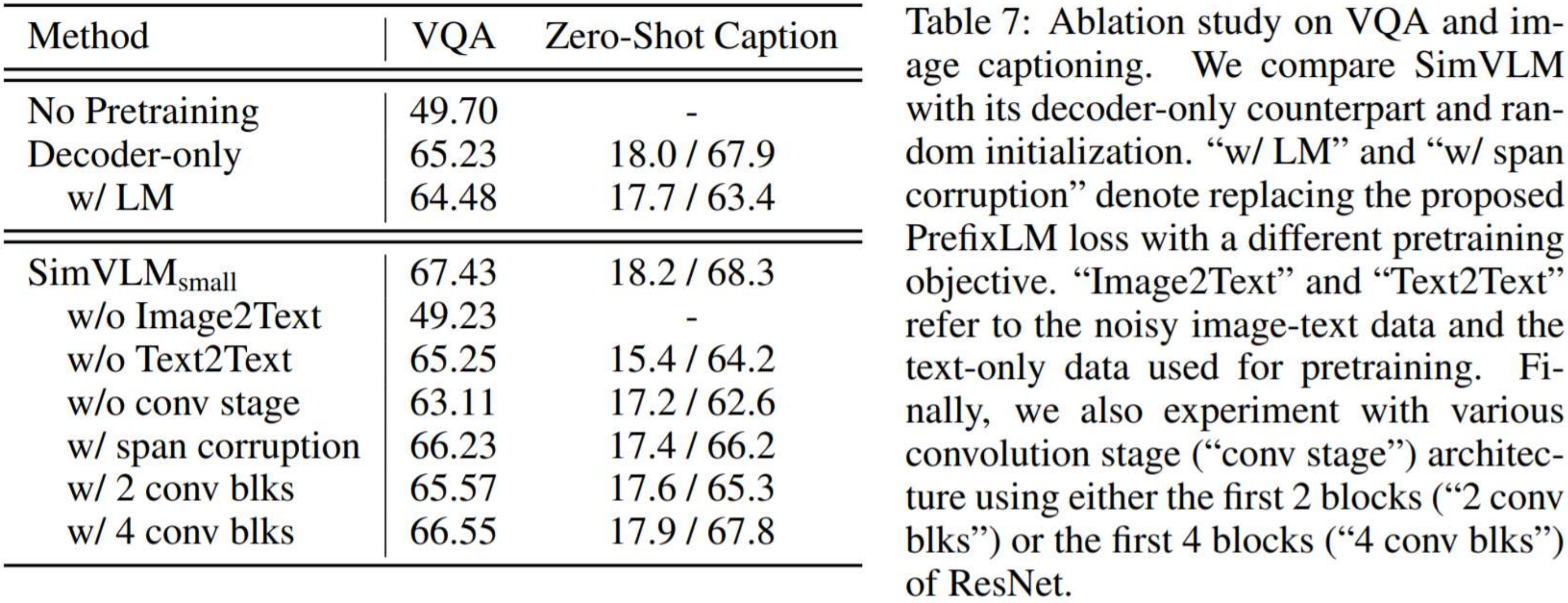
Linear probe



Method	Acc@1
SimCLRv2 (Chen et al., 2020a)	79.8
DINO (Caron et al., 2021)	80.1
CLIP (Radford et al., 2021)	85.4
ALIGN (Jia et al., 2021)	**85.5**
SimVLM$_{\text{base}}$	80.6
SimVLM$_{\text{large}}$	82.3
SimVLM$_{\text{huge}}$	83.6

Table 6: Linear evaluation on ImageNet classification, compared to state-of-the-art representation learning methods.

Ablation



Method	VQA	Zero-Shot Caption
No Pretraining	49.70	-
Decoder-only	65.23	18.0 / 67.9
w/ LM	64.48	17.7 / 63.4
$\text{SimVLM}_{\text{small}}$	67.43	18.2 / 68.3
w/o Image2Text	49.23	-
w/o Text2Text	65.25	15.4 / 64.2
w/o conv stage	63.11	17.2 / 62.6
w/ span corruption	66.23	17.4 / 66.2
w/ 2 conv blks	65.57	17.6 / 65.3
w/ 4 conv blks	66.55	17.9 / 67.8

Table 7: Ablation study on VQA and image captioning. We compare SimVLM with its decoder-only counterpart and random initialization. “w/ LM” and “w/ span corruption” denote replacing the proposed PrefixLM loss with a different pretraining objective. “Image2Text” and “Text2Text” refer to the noisy image-text data and the text-only data used for pretraining. Finally, we also experiment with various convolution stage (“conv stage”) architecture using either the first 2 blocks (“2 conv blks”) or the first 4 blocks (“4 conv blks”) of ResNet.

End

- Explore the transfer ability.
- Simple training objective and good performance (with huge weakly supervised pair data)