

GRIT: Faster and Better Image captioning Transformer Using Dual Visual Features

Van-Quang Nguyen¹, Masanori Suganuma^{1,2}, Takayuki Okatani^{1,2}
¹GSIS, Tohoku University, ²RIKEN Center for AIP

Abstract

We propose Transformer-only neural architecture named **GRIT** that effectively utilizes dual visual features to generate better captions:

- GRIT replaces the CNN-based detector employed in previous methods with a DETR-based one, making it computationally faster.
- Its monolithic design consisting only of Transformers enables end-to-end training of the model.

Question: How do we extract and fuse good visual representations?

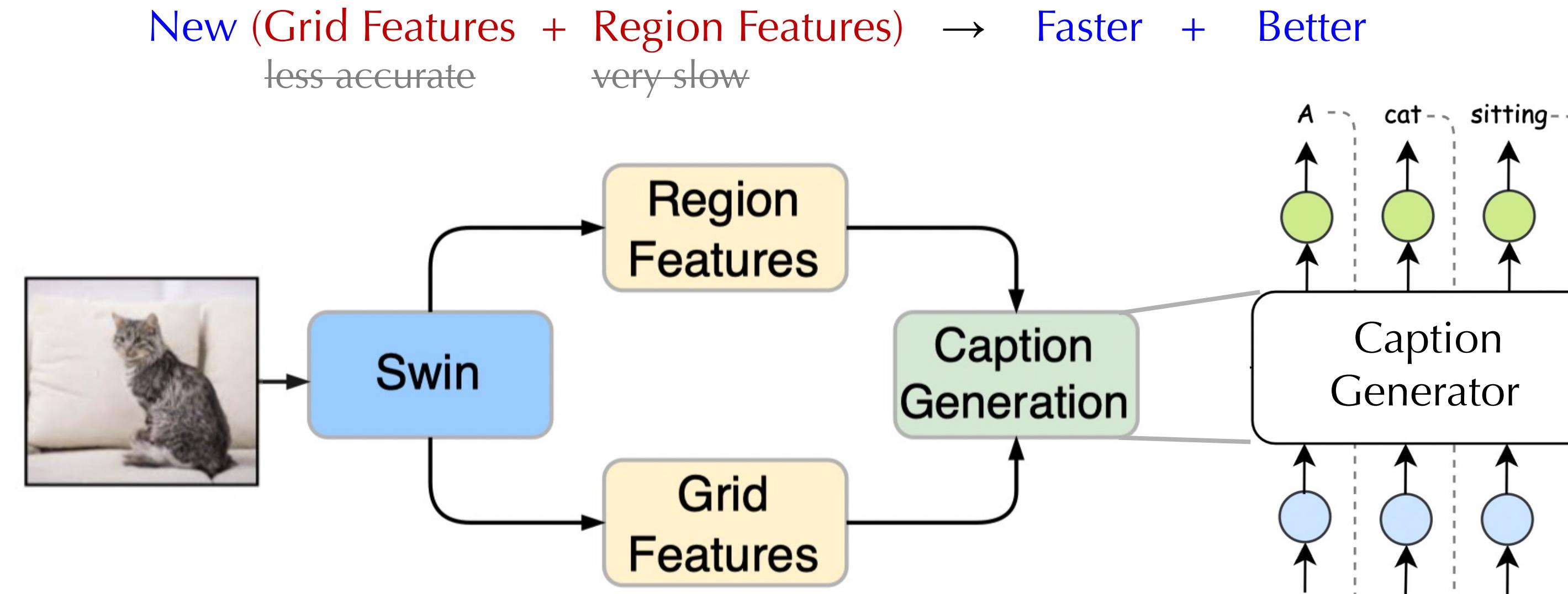
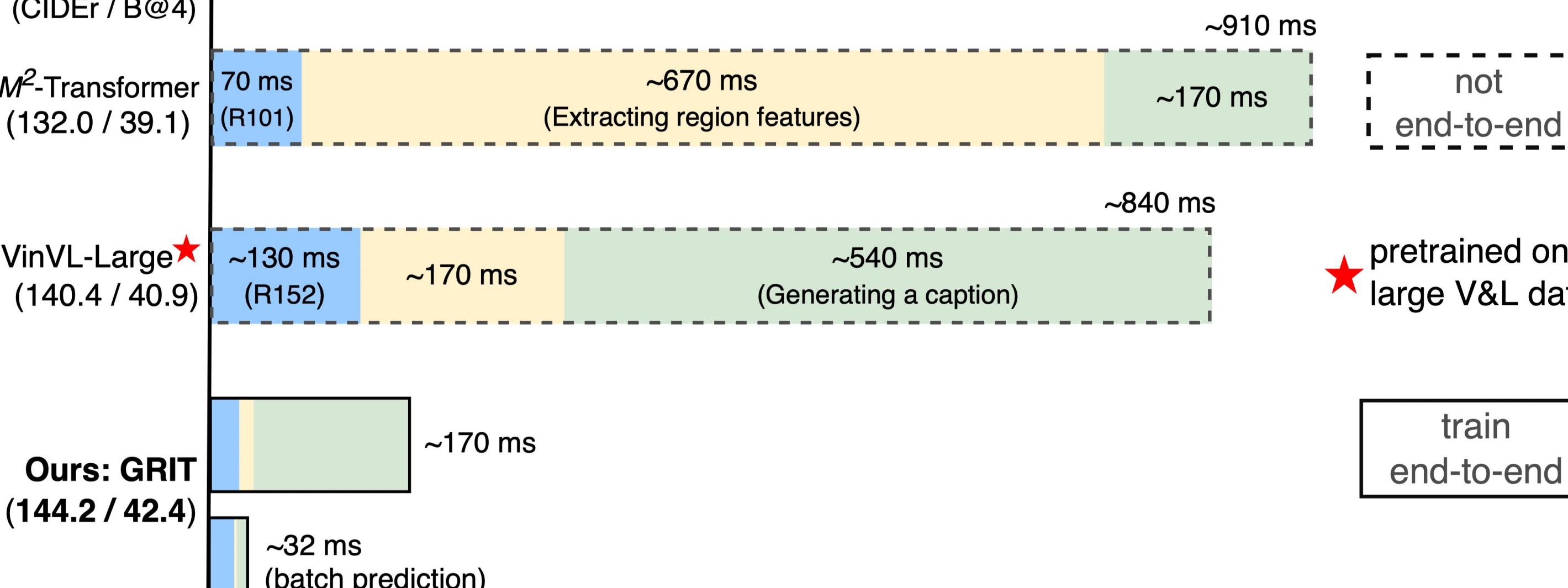
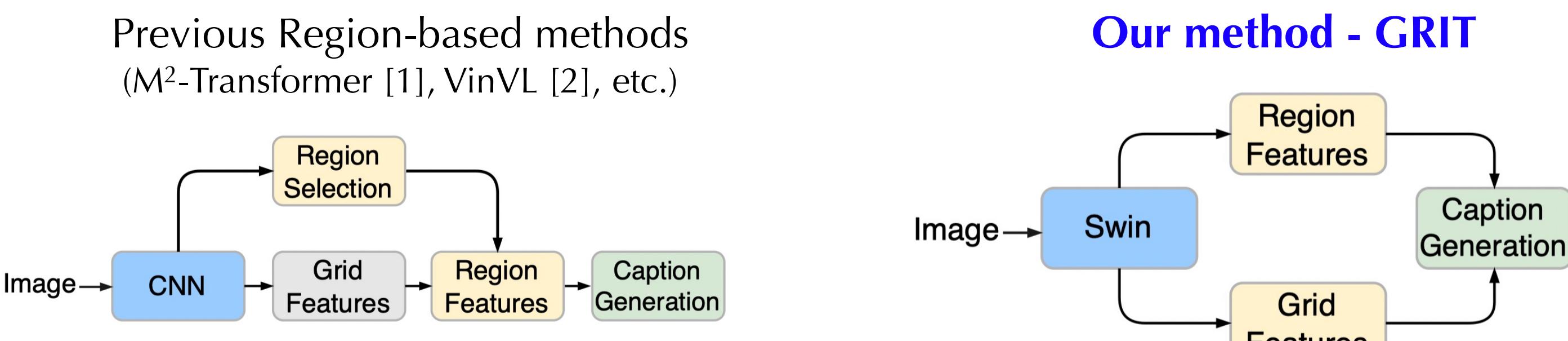


Figure: GRIT: Grid- and Region-based Image captioning Transformer. It utilizes both region and grid features, making more accurate. It uses DETR-based detector, making it faster.

Speed and Accuracy Comparison



Previous Region-based methods
(M²-Transformer [1], VinVL [2], etc.)



Our method - GRIT
Image → Swin → Region Selection → Region Features → Caption Generation
Image → CNN → Grid Features → Region Features → Caption Generation

Our Method

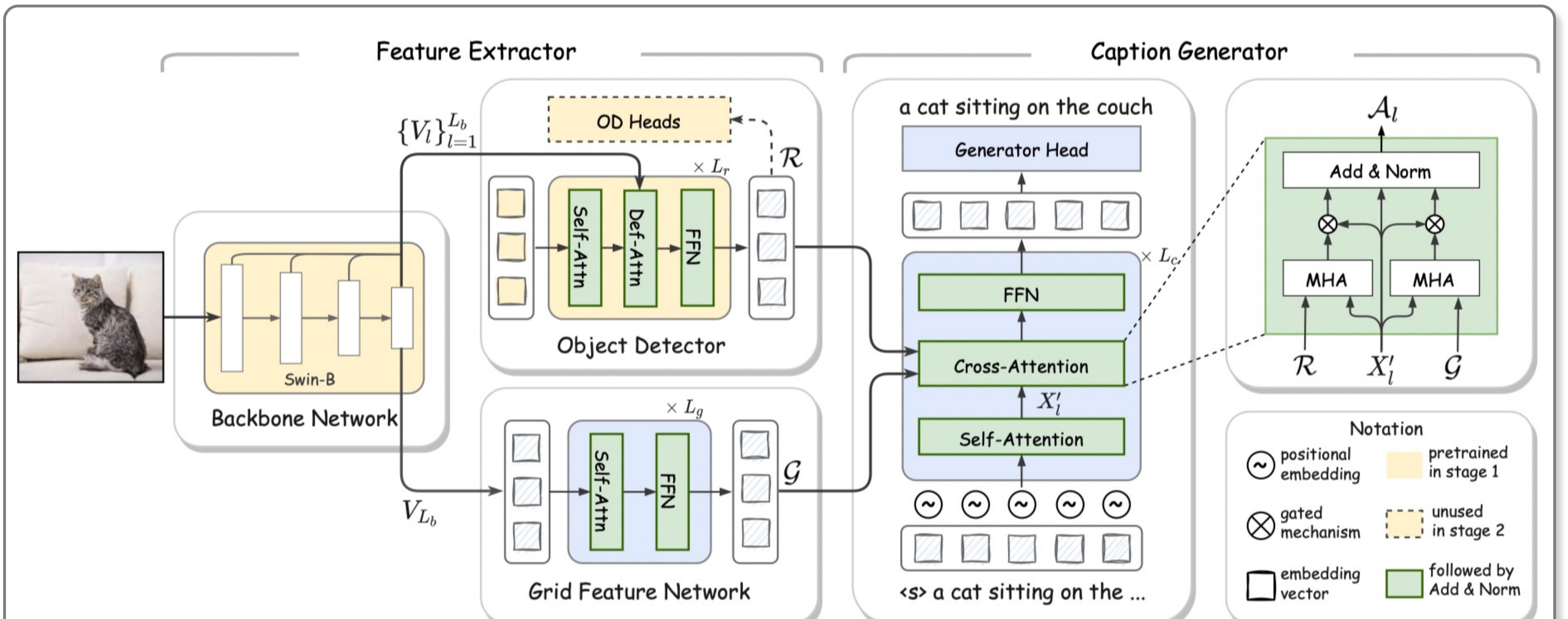


Figure: (1) Backbone Network to extract initial multi-level feature maps
(2) Object Detector to extract region features
(3) Grid Feature Network to extract grid features
(4) Caption Generator to generate a caption.

Object Detector

To extract region features

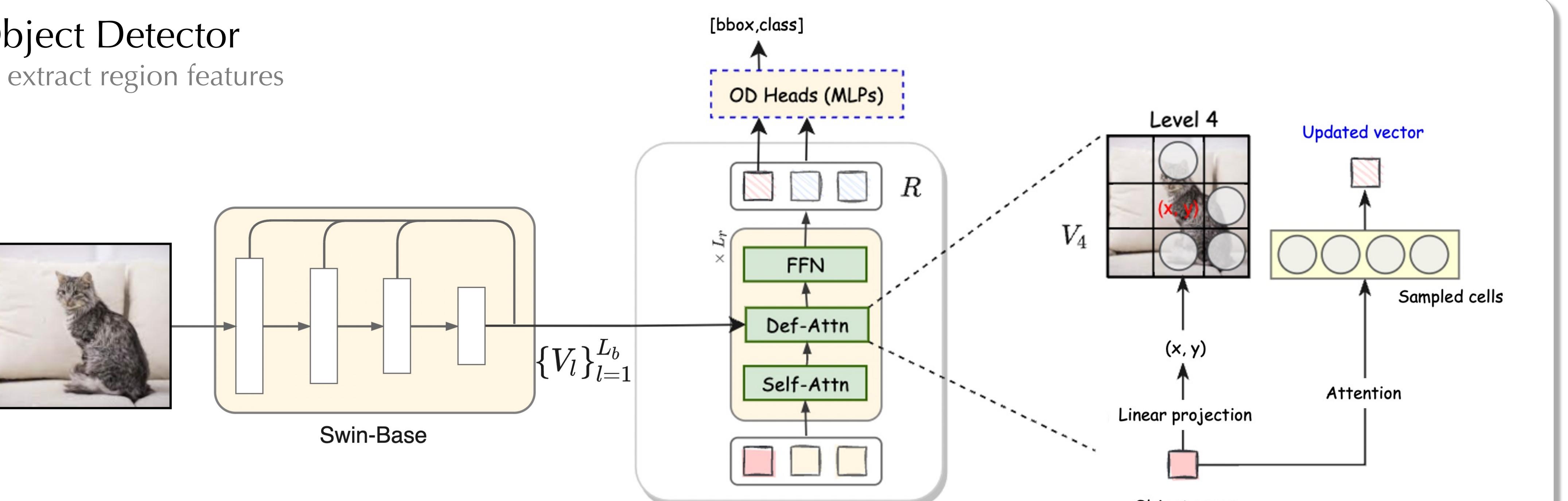


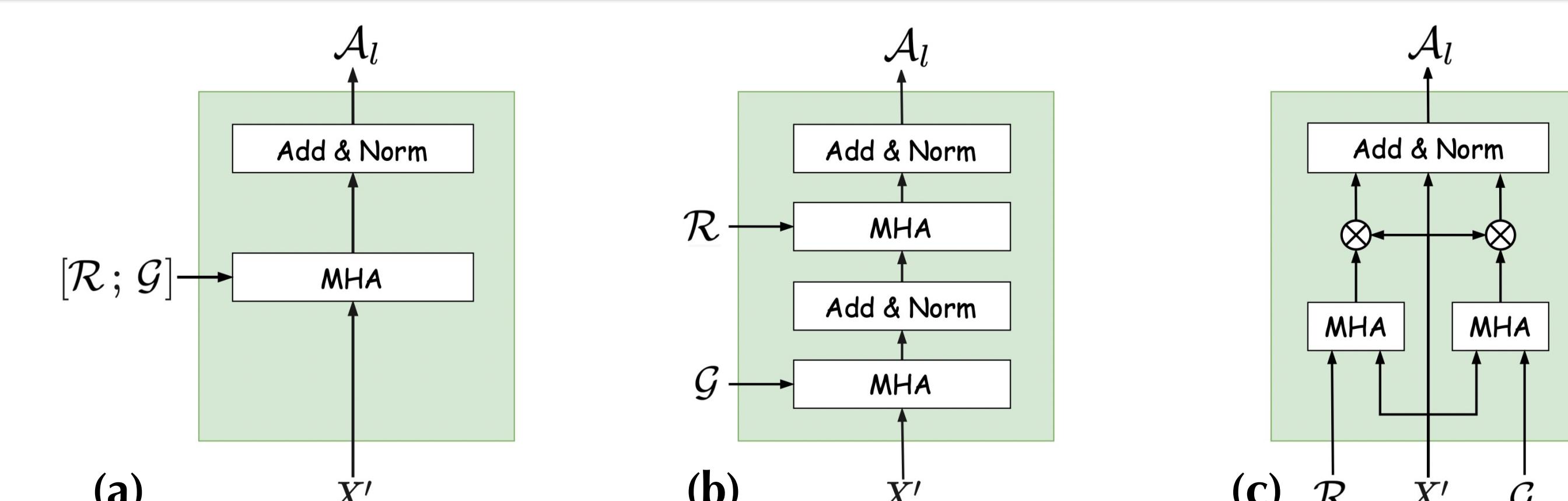
Figure: Use Swin-Base to extract multi-level feature maps as input to the object detector. Build object detector by stacking Deformable Transformer Layer which perform attention on the sampled input keys; The region features (R) are extracted from the object detector.

Cross-Attention Mechanism

The mechanism to use the dual visual features

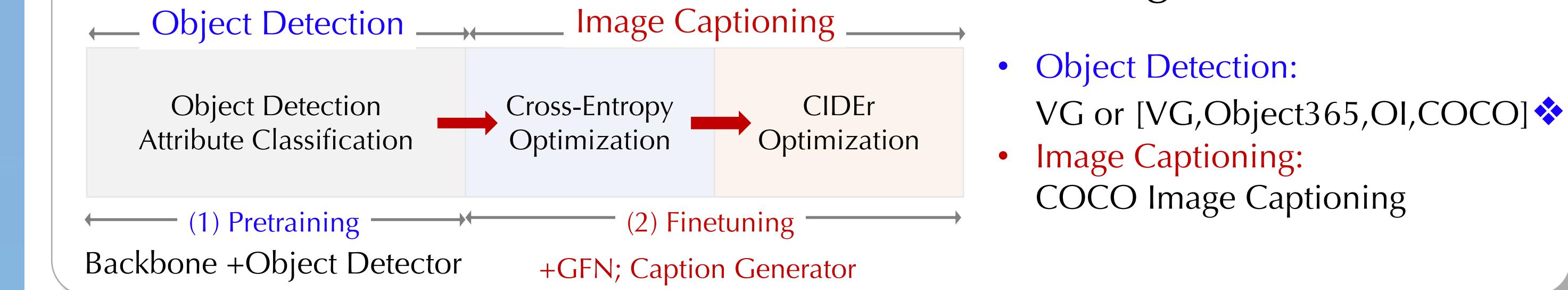
- (a) Concatenated Cross-Attention
- (b) Sequential Cross-Attention
- (c) Parallel Cross-Attention

-> Parallel Cross-Attention yields the best!



Experimental Results

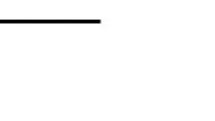
End-to-End Training



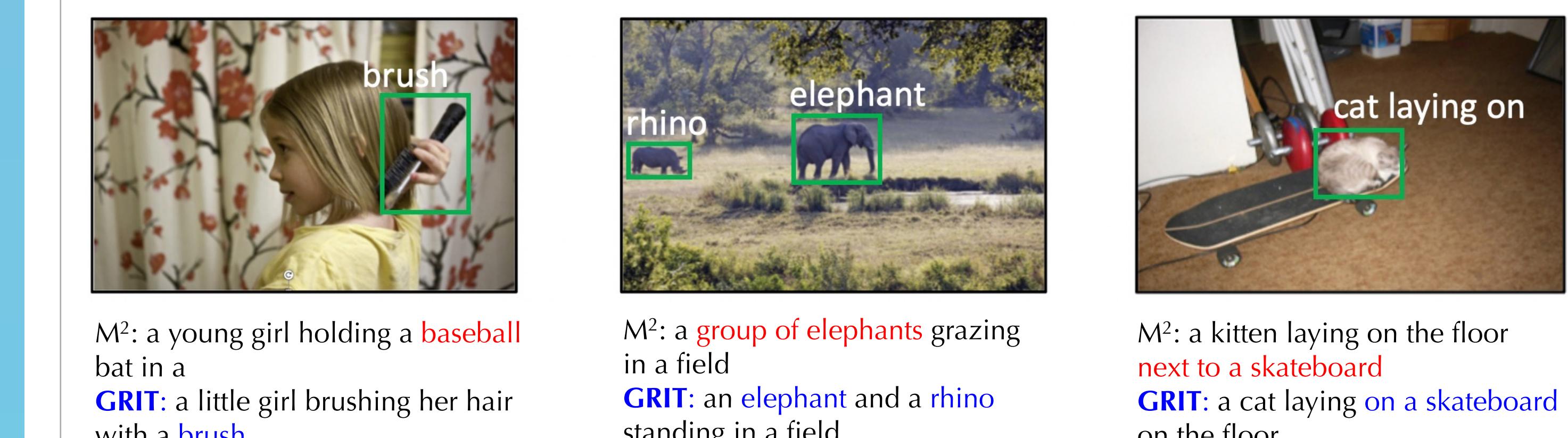
Training Datasets

- Object Detection:
VG or [VG, Object365, OI, COCO] 
- Image Captioning:
COCO Image Captioning

Results on the Karpathy test split of COCO

Method	V. E. Type	# VL Data	Performance Metrics				
			B@1	B@4	M	R	C
w/ VL pretraining							
UVLP	R	3.0M	-	39.5	29.3	-	129.3
Oscar _{base}	R	6.5M	-	40.5	29.7	-	137.6
VinVL [†] _{large}	R	8.9M	-	41.0	31.1	-	140.9
SimVLM _{huge}	G	1.8B	-	40.6	33.7	-	143.3
w/o VL pretraining							
SAT	G	-	-	31.9	25.5	54.3	106.3
RSTNet	G	-	81.8	40.1	29.8	59.5	135.6
Up-Down	R	-	79.8	36.3	27.7	56.9	120.1
AoA	R	-	80.2	38.9	29.2	58.8	129.8
\mathcal{M}^2 Transformer	R	-	80.8	39.1	29.2	58.6	131.2
TCIC	R	-	81.8	40.8	29.5	59.2	135.4
Dual Global	R+G	-	81.3	40.3	29.2	59.4	132.4
DLCT	R+G	-	81.4	39.8	29.5	59.1	133.8
GRIT	R+G	-	83.5	41.9	30.5	60.5	142.2
GRIT ( 4 datasets)	R+G	-	84.2	42.4	30.6	60.7	144.2
GRIT	R+G	-	84.2	42.4	30.6	60.7	144.2
GRIT	R+G	-	84.2	42.4	30.6	60.7	144.2

Qualitative Examples



References & Our Code

- [1] Cornia, M., Stefanini, M., Baraldi, L., Cucchiara, R.: Meshed-memory transformer for image captioning. In CVPR, 2020.
- [2] Zhang, P., Li, X., Hu, X., Yang, J., Zhang, L., Wang, L., Choi, Y., Gao, J.: Vinvl: Revisiting visual representations in vision-language models. In CVPR, 2021.

Code and pretrained models given at: <https://www.github.com/davidnvq/grit>

