

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

Van-Quang Nguyen<sup>1</sup>, Masanori Suganuma<sup>2,1</sup>, and Takayuki Okatani<sup>1,2</sup>

<sup>1</sup> Graduate School of Information Sciences, Tohoku University

<sup>2</sup> RIKEN Center for AIP

{quang,suganuma,okatani}@vision.is.tohoku.ac.jp

**Abstract.** Current state-of-the-art methods for image captioning employ region-based features, as they provide object-level information that is essential to describe the content of images; they are usually extracted by an object detector such as Faster R-CNN. However, they have several issues, such as lack of contextual information, the risk of inaccurate detection, and the high computational cost. The first two could be resolved by additionally using grid-based features. However, how to extract and fuse these two types of features is uncharted. This paper proposes a Transformer-only neural architecture, dubbed GRIT (Grid- and Region-based Image captioning Transformer), that effectively utilizes the two 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. Moreover, its monolithic design consisting only of Transformers enables end-to-end training of the model. This innovative design and the integration of the dual visual features bring about significant performance improvement. The experimental results on several image captioning benchmarks show that GRIT outperforms previous methods in inference accuracy and speed.

**Keywords:** Image Captioning, Grid Features, Region Features

## 1 Introduction

Image captioning is the task of generating a semantic description of a scene in natural language, given its image. It requires a comprehensive understanding of the scene and its description reflecting the understanding. Therefore, most existing methods solve the task in two corresponding steps; they first extract visual features from the input image and then use them to generate a scene’s description. The key to success lies in the problem of how we can extract good features.

Researchers have considered several approaches to the problem. There are two primary methods, referred to as grid features [51, 41, 32] and region features [4]. Grid features are local image features extracted at the regular grid points, often obtained directly from a higher layer feature map(s) of CNNs/ViTs. Region features are a set of local image features of the regions (i.e., bounding boxes)

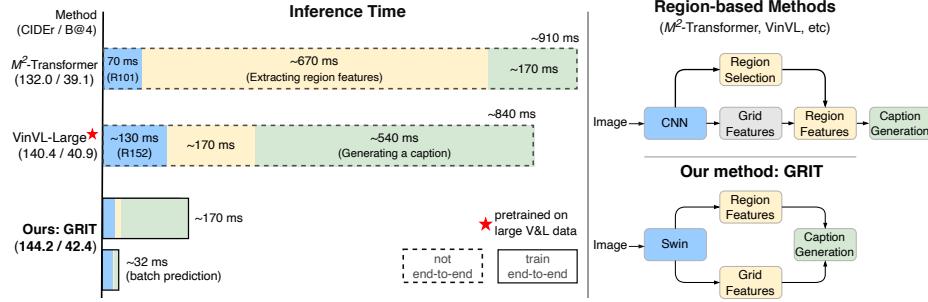


Fig. 1: Comparison of GRIT and other region-based methods for image captioning. Left: Running time per image of performing inference with beam size of five and the maximum length of 20 on a V100 GPU. Right: Their architectures

detected by an object detector. The current state-of-the-art methods employ the region features since they encode detected object regions directly. Identifying objects and their relations in an image will be useful to correctly describing the image. However, the region features have several issues. First, they do not convey contextual information such as objects' relation since the regions do not cover the areas between objects. Second, there is a risk of erroneous detection of objects; important objects could be overlooked, etc. Third, computing the region feature is computationally costly, which is especially true when using a high-performance CNN-based detector, such as Faster R-CNN [40].

The grid features are extracted from the entire image, typically a high-layer feature map of a backbone network. While they do not convey object-level information, they are free from the first two issues with the region features. They may represent contextual information such as objects' relations in images, and they are free from the risk of erroneous object detection.

In this study, we consider using such region and grid features in an integrated manner, aiming to build a better model for image captioning. The underlying idea is that properly integrating the two types of features will provide a better representation of input images since they are complementary, as explained above. While a few recent studies consider their integration [34, 49], it is still unclear what the best way is. In this study, we reconsider how to extract each from input images and then consider how to integrate them.

There is yet another issue with the region features, usually obtained by a CNN-based detector. At the last stage of its computation, CNN-based detectors employ non-maximum suppression (NMS) to eliminate redundant bounding boxes. This makes the end-to-end training of the entire model hard, i.e., jointly training the decoder part of the image captioning model and the detector by minimizing a single loss. Recent studies detach the two parts in training; they first train a detector on the object detection task and then train only the

decoder part on image captioning. This could be a drag on achieving optimal performance of image captioning.

To overcome this limitation of CNN-based detectors and also cope with their high-computational cost, we employ the framework of DETR [7], which does not need NMS. We choose Deformable DETR [60], an improved variant, for its high performance, and also replace a CNN backbone used in the original design with Swin Transformer [31] to extract initial features from the input image. We also obtain the grid features from the same Swin Transformer. We input its last layer features into a simple self-attention Transformer and update them to obtain our grid features. This aims to model spatial interaction between the grid features, retrieving contextual information absent in our region features.

The extracted two types of features are fed into the second half of the model, the caption generator. We design it as a lightweight Transformer generating a caption sentence in an autoregressive manner. It is equipped with a unique cross-attention mechanism that computes and applies attention from the two types of visual features to caption sentence words.

These components form a Transformer-only neural architecture, dubbed GRIT (Grid- and Region-based Image captioning Transformer). Our experimental results show that GRIT has established a new state-of-the-art on the standard image captioning benchmark of COCO [30]. Specifically, in the offline evaluation using the Karpathy test split, GRIT outperforms all the existing methods without vision and language (V&L) pretraining. It also performs at least on a par with SimVLM<sub>huge</sub> [48] leveraging V&L pretraining on 1.8B image-text pairs.

## 2 Related Work

### 2.1 Visual Representations for Image Captioning

Recent image captioning methods typically employ an encoder-decoder architecture. Specifically, given an image, the encoder extracts visual features; the decoder receives the visual features as inputs and generates a sequence of words. Early methods use a CNN to extract a global feature as a holistic representation of the input image [46,21]. Although it is simple and compact, this holistic representation suffers from information loss and insufficient granularity. To cope with this, several studies [51,41,32] employed more fine-grained grid-based features to represent input images and also used attention mechanisms to utilize the granularity for better caption generation. Later, Anderson et al. [4] introduced the method of using an object detector, such as Faster R-CNN, to extract object-oriented features, called region features, showing that this leads to performance improvement in many V&L tasks, including image captioning and visual question answering. Since then, region features have become the de facto choice of visual representation for image captioning. Pointing out the high computational cost of the region features, Jiang et al. [19] showed that the grid features extracted by an object detector perform well on the VQA task. RSTNet [58] has recently applied these grid features to image captioning.

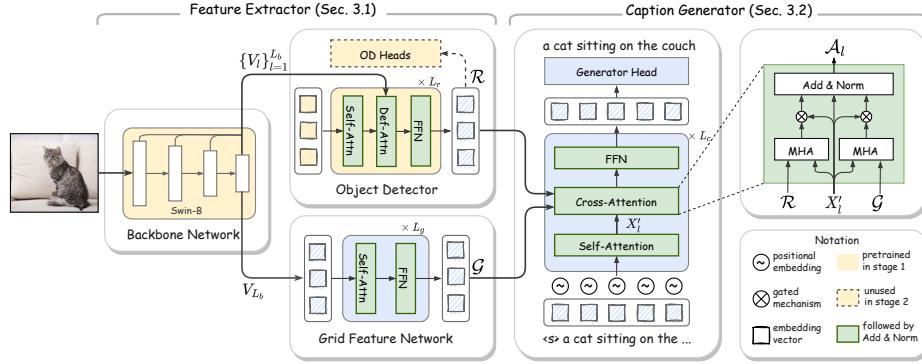


Fig. 2: Overview of the architecture of GRIT

## 2.2 Application of Transformer in Vision/Language Tasks

Transformer has long been a standard neural architecture in natural language processing [44,9,39], and started to be extended to computer vision tasks. Besides ViT [10] for image classification, it was also applied to object detection, leading to DETR [7], followed by several variants [60,12,43]. A recent study [50] applied the framework of DETR to pretraining for various V&L tasks, where they did not use it to obtain the region features.

Transformer has been applied to image captioning, where it is used as an encoder for extracting and encoding visual features and a decoder for generating captions. Specifically, Yang et al. [53] proposed to use the self-attention mechanism to encode visual features. Li et al. [27] used Transformer for obtaining the region features in combination with a semantic encoder that exploits knowledge from an external tagger. Several following studies proposed several variants of Transformer tailored to image captioning, such as Attention on Attention [17], X-Linear Attention [36], Memory-augmented Attention [8], etc. Transformer is naturally employed also as a caption decoder [14,13,34,48].

## 3 Grid- and Region-based Image captioning Transformer

This section describes the architecture of GRIT (Grid- and Region-based Image captioning Transformer). It consists of two parts, one for extracting the dual visual features from an input image (Sec. 3.1) and the other for generating a caption sentence from the extracted features (Sec. 3.2).

### 3.1 Extracting Visual Features from Images

**Backbone Network for Extracting Initial Features** A lot of efforts have been made to apply the Transformer architecture to various computer vision

tasks since ViT [10] applied it to image classification. ViT divides an input image into small patches and computes global attention over them. This is not suitable for tasks requiring spatially dense prediction, e.g., object detection since the computational complexity increases quadratically with the image resolution.

Swin Transformer [31] mitigates this issue to a great extent by incorporating operations such as patch reduction and shifted windows that support local attention. It is currently a de facto standard as a backbone network for various computer vision tasks. We employ it to extract initial visual features from the input image in our model.

We briefly summarize its structure, explaining how we extract features from the input image and send them to the components following the backbone. Given an input image of resolution  $H \times W$ , Swin Transformer computes and updates feature maps through multiple stages; it uses the patch merging layer after every stage (but the last stage) to downsample feature maps in their spatial dimension by the factor of 2. We apply another patch merging layer to downsample the last layer’s feature map. We then collect the feature maps from all the stages, obtaining four multi-scale feature maps, i.e.,  $\{V_l\}_{l=1}^{L_b}$  where  $L_b = 4$ , which have the resolution from  $H/8 \times W/8$  to  $H/64 \times W/64$ . These are inputted to the subsequent modules, i.e., the object detector and the network for generating grid features.

**Generating Region Features** As in previous image captioning methods, ours also rely on an object detector to create region features. However, we employ a Transformer-based decoder framework, i.e., DETR [7] instead of CNN-based detectors, such as Faster R-CNN, which is widely employed by the SOTA image captioning models [4]. DETR formulates object detection as a direct set prediction problem, which makes the model free of the unideal computation for us, i.e., NMS and ROI alignment. This enables the end-to-end training of the entire model from the input image to the final output, i.e., a generated caption, and also leads to a significant reduction in computational time while maintaining the model’s performance on image captioning compared with the SOTA models.

Specifically, we employ Deformable DETR [60], a variant of DETR. Deformable DETR extracts multi-scale features from an input image with its encoder part, which are fed to the decoder part. We use only the decoder part, to which we input the multi-scale features from the Swin Transformer backbone. This leads to further reduction in computational time. We will refer this decoder part as “object detector” in what follows; see Fig. 2.

The object detector receives two inputs: the multi-scale feature maps generated by the backbone, and  $N$  learnable object queries  $R_0 = \{r_i\}_{i=1}^N$ , in which  $r_i \in \mathbb{R}^d$ . Before forwarding them into the object detector, we apply linear transformation to the multi-scale feature maps, mapping them into  $d$ -dimensional vectors as  $V_l \leftarrow W_l^r V_l$ , where  $\{W_l^r\}_{l=1}^{L_b}$  is a learnable projection matrix.

Receiving these two inputs, the object detector updates the object queries through a stack of  $L_r$  deformable layers, yielding  $R_{L_r} \in \mathbb{R}^{N \times d}$  from the last

layer; see [60] for details. We use  $R_{L_r} \in \mathbb{R}^{N \times d}$  as our region features  $\mathcal{R}$ . We forward this to the caption generator.

Although we train it as a part of our entire model, we pretrain our ‘‘object detector’’ including the vision backbone on object detection before the training of image captioning. For the pretraining, we follow the procedure of Deformable DETR; placing a three-layer MLP and a linear layer on its top to predict box coordinates and class category, respectively. We then minimize a set-based global loss that forces unique predictions via bipartite matching.

Following [4,57], we pretrain the model (i.e., our object detector including the vision backbone) in two steps. We first train it on object detection following the training method of Deformable DETR. We then fine-tune it on a joint task of object detection and object attribute prediction, aiming to make it learn fine-grained visual semantics with the following loss:

$$\mathcal{L}_v(y, \hat{y}) = \sum_{i=1}^N \underbrace{[-\log \hat{p}_{\hat{\sigma}(i)}(c_i) + \mathbf{1}_{c_i \neq \emptyset} \mathcal{L}_{box}(b_i, \hat{b}_{\hat{\sigma}(i)})]}_{\text{object detection}} \underbrace{[-\log \hat{p}_{\hat{\sigma}(i)}(a_i)]}_{\text{attribute prediction}}, \quad (1)$$

where  $\hat{p}_{\hat{\sigma}(i)}(a_i)$  and  $\hat{p}_{\hat{\sigma}(i)}(c_i)$  are the attribute and class probabilities,  $\mathcal{L}_{box}(b_i, \hat{b}_{\hat{\sigma}(i)})$  is the loss for normalized bounding box regression for object  $i$  [60].

**Grid Feature Network** This network receives the last one of the multi-scale feature maps from the Swin Transformer backbone, i.e.,  $V_{L_b} \in \mathbb{R}^{M \times d_{L_b}}$ , where  $M = H/64 \times W/64$ . As with the input to the object detector, we apply a linear transformation with a learnable matrix  $W^g \in \mathbb{R}^{d \times d_{L_b}}$  to  $V_{L_b}$ , obtaining  $G_0 = W^g V_{L_b}$ . We employ the standard self-attention Transformer having  $L_g$  layers. This network updates  $V_{L_b}$  through these layers, yielding our grid features  $\mathcal{G}$  represented as a  $M \times d$  matrix. We intend to extract contextual information hidden in the input image by modeling the spatial interaction between the grid features.

### 3.2 Caption Generation Using Dual Visual Features

**Overall Design of Caption Generator** The caption generator receives the two types of visual features, the region features  $\mathcal{R} \in \mathbb{R}^{N \times d}$  and the grid features  $\mathcal{G} \in \mathbb{R}^{M \times d}$ , as inputs. Apart from this, we employ the basic design employed in previous studies [44,14] that is based on the Transformer architecture. It generates a caption sentence in an autoregressive manner; receiving the sequence of predicted words (rigorously their embeddings) at time  $t - 1$ , it predicts the next word at time  $t$ . We employ the sinusoidal positional embedding of time step  $t$  [44]; we add it to the word embedding to obtain the input  $x_0^t \in \mathbb{R}^d$  at  $t$ .

The caption generator consists of a stack of  $L_c$  identical layers. The initial layer receives the sequence of predicted words and the output from the last layer is input to a linear layer whose output dimension equals the vocabulary size to predict the next word.

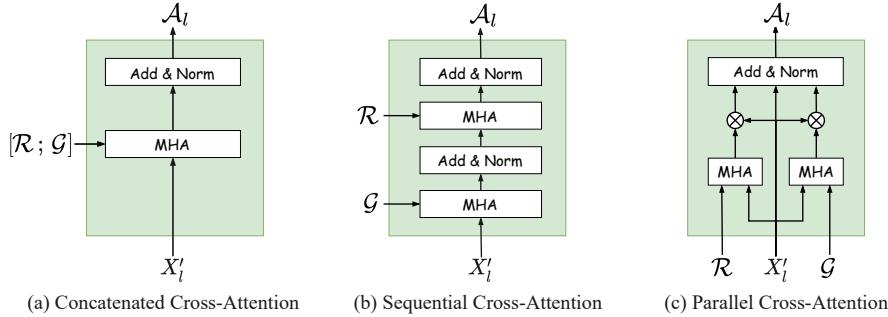


Fig. 3: Three designs of cross-attention mechanism to use dual visual features

Each transformer layer has a sub-layer of masked self-attention over the sentence words and a sub-layer(s) of cross-attention between them and the visual features in this order, followed by a feedforward network (FFN) sub-layer. The masked self-attention sub-layer at the  $l$ -th layer receives an input sequence  $\{x_i^{l-1}\}_{i=0}^t$  at time step  $t$ , and computes and applies self-attention over the sequence to update the tokens with the attention mask to prevent the interaction from the future words during training.

The cross-attention sub-layer in the layer  $l$ , located after the self-attention sub-layer, fuses its output with the dual visual features by cross-attention between them, yielding  $A_l$ . We consider the three design choices shown in Fig. 3 and described below. We examine their performance through experiments.

**Cross-attention between Caption Word and Dual Visual Features** We show three designs of cross-attention between the word features and the dual visual features (i.e., the region features  $\mathcal{R}$  and the grid features  $\mathcal{G}$ ) as below.

*Concatenated Cross-Attention* The simplest approach is to concatenate the two visual features and use the resultant features as keys and values in the standard multi-head attention sub-layer, where the words serve as queries; see Fig. 3(a).

*Sequential Cross-Attention* Another approach is to perform cross-attention computation separately for the two visual features. The corresponding design is to place two independent multi-head attention sub-layers in a sequential fashion, and uses one for the grid features and the other for the region features (or the opposite combination); see Fig. 3(b). Note that their order could affect the performance.

*Parallel Cross-Attention* The third approach is to perform multi-head attention computation on the two visual features in parallel. To do so, we use two multi-head attention mechanisms with independent learnable parameters. The detailed design is as follows. Let  $X_{l-1} = \{x_i^{l-1}\}$  be the word features inputted to the

meta-layer  $l$  containing this cross attention sub-layer. As shown in Fig. 2, they are first input to the self-attention sub-layer, converted into  $X'_l = \{x'_i\}$  (layer index  $l$  omitted for brevity) and then input to this cross attention sub-layer. In this sub-layer, multi-head attention (MHA) is computed with  $\{x'_i\}$  as queries and the region features  $\mathcal{R}$  as keys and values, yielding attended features  $\{a_i^r\}$ . The same computation is performed in parallel with the grid features  $\mathcal{G}$  as keys and values, yielding  $\{a_i^g\}$ . Next, we concatenate them with  $x'_i$  as  $[a_i^r; x'_i]$  and  $[a_i^g; x'_i]$ , projecting them back to  $d$ -dimensional vector using learnable affine projections. Normalizing them with sigmoid into probabilities  $\{c_i^r\}$  and  $\{c_i^g\}$ , respectively, we have

$$c_i^g = \text{sigmoid}(W^g[a_i^g; x'_i] + b^g), \quad (2)$$

$$c_i^r = \text{sigmoid}(W^r[a_i^r; x'_i] + b^r). \quad (3)$$

We then multiply them with  $\{a_i^r\}$  and  $\{a_i^g\}$ , add the resultant vectors to  $\{x'_i\}$ , and finally feed to layer normalization, obtaining  $\mathcal{A}_l = \{a_i^{(l)}\}$  as follows:

$$a_i^{(l)} = \text{LN}(c_i^g \otimes a_i^g + c_i^r \otimes a_i^r + x'_i). \quad (4)$$

**Caption Generator Losses** Following a standard practice of image captioning studies, we pre-train our model with a cross-entropy loss (XE) and finetune it using the CIDEr-D optimization with self-critical sequence training strategy [41]. Specifically, the model is first trained to predict the next word  $x_t^*$  at  $t = 1..T$ , given the ground-truth sentence  $x_{1:T}^*$ . This is equal to minimize the following XE loss with respect to the model's parameter  $\theta$ :

$$\mathcal{L}_{XE}(\theta) = - \sum_{t=1}^T \log(p_\theta(x_t^* | x_{0:t-1}^*)). \quad (5)$$

We then finetune the model with the CIDEr-D optimization, where we use the CIDEr score as the reward and the mean of the rewards as the reward baseline, following [8]. The loss for self-critical sequence training is given by

$$\mathcal{L}_{RL}(\theta) = -\frac{1}{k} \sum_{i=1}^k (r(\mathbf{w}^i) - b) \log p(\mathbf{w}^i), \quad (6)$$

where  $\mathbf{w}^i$  is the  $i$ -th sentence in the beam;  $r(\cdot)$  is the reward function; and  $b$  is the reward baseline; and  $k$  is the number of samples in the batch.

## 4 Experiments

### 4.1 Datasets

**Object Detection** As mentioned earlier, we train our object detector (including the backbone) in two steps. In the first step, we train it on object detection

using either Visual Genome [25] or a combination [57] of four datasets: COCO [30], Visual Genome, Open Images [26], and Object365 [42], depending on what previous methods we experimentally compare. In the second step, we train the model on object detection plus attribute prediction using Visual Genome. Note that following the standard practice, we exclude the duplicated samples appearing in the testing and validation splits of the COCO and nocaps [2] datasets to remove data contamination. See the supplementary material for more details.

**Image Captioning** We conduct our experiments on the COCO dataset, the standard for the research of image captioning [30]. The dataset contains 123,287 images, each annotated with five different captions. For offline evaluation, we follow the widely adopted Karpathy split [20], where 113,287, 5,000, and 5,000 images are used for training, validation, and testing respectively.

To test our method’s effectiveness on other image captioning datasets, we also report the performances on the nocaps dataset and the Artemis dataset [1]. See the supplementary material for more details.

## 4.2 Implementation Details

**Evaluation Metrics** We employ the standard evaluation protocol for the evaluation of methods. Specifically, we use the full set of captioning metrics: BLEU@N [37], METEOR [5], ROUGE-L [29], CIDEr [45], and SPICE [3]. We will use the abbreviations, B@N, M, R, C, and S, to denote BLEU@N, METEOR, ROUGE-L, CIDEr, and SPICE, respectively.

**Hyperparameters Settings** In our model, we set the dimension  $d$  of each layer to 512, the number of heads to eight. We employ dropout with the dropout rate of 0.2 on the output of each MHA and FFN sub-layer following [44]. We set the number of layers as  $L_r = 6$  for the object detector, as  $L_g = 3$  for the grid feature network, and as  $L_c = 3$  for the caption generator. Following previous studies, we convert all the captions to lower-case, remove punctuation characters, and perform tokenization with the SpaCy toolkit [16]. We build the vocabularies, excluding the words which appear less than five times in the training and validation splits.

## 4.3 Training Details

**First Stage** In the first stage, we pretrain the object detector with the backbone. We consider several existing region-based methods for comparison, which employ similar pretraining of an object detector but use different datasets. For a fair comparison, we consider two settings. One uses Visual Genome for training, following most previous methods. We train our detector for 150,000 iterations with a batch size of 32. The other (results indicated with † in what follows) uses the four datasets mentioned above, following [57]. We train the detector for 125,000 iterations with a batch size of 256. In both settings, the input image is

Table 1: Results of ablation tests on the COCO test split. All the models are trained with the XE loss and finetuned by the CIDEr optimization.

(a)			(b)		
Factor	Choice	CIDEr B@4	Cross Attention	Choice	CIDEr B@4
(1) Backbone Network - Training data	ImageNet	135.5 41.5	(1) Concatenated - Visual features	$\mathcal{G}$	142.1 41.7
	VG	142.3 41.9		$\mathcal{R}$	142.9 41.9
	4DS	<b>144.2 42.4</b>		$[\mathcal{G}; \mathcal{R}]$	143.1 41.9
(2) Region features - Number of vectors (trained on VG)	50	141.4 41.9	(2) Sequential - Sequential order	$\mathcal{G} \rightarrow \mathcal{R}$	144.0 42.1
	100	141.8 41.5		$\mathcal{R} \rightarrow \mathcal{G}$	143.6 42.1
	150	<b>142.3 41.9</b>			
(3) Training strategy - End-to-end training	Yes	<b>144.2 42.4</b>	(3) Parallel - Gated activation	Sigmoid	<b>144.2 42.4</b>
	No	139.6 <b>42.7</b>		Identity	143.9 41.6

resized so that the maximum for the shorter side is 800 and for the longer side is 1333. We use Adam optimizer [24] with a learning rate of  $10^{-4}$ , decreased by 10 at iteration 120,000 and 100,000 in the first and second settings, respectively. We follow [60] for other training procedures. After this, we finetune the models on object detection plus attribute prediction using Visual Genome for additional five epochs with a learning rate of  $10^{-5}$ , following [4,57]. The supplementary material presents the details of implementation and experimental results on object detection.

**Second Stage** We train the entire model for the image captioning task in the second stage. We employ the standard method for word representation, i.e., linear projections of one-hot vectors to vectors of dimension  $d = 512$ . In this stage, we resize all the input images so that the maximum dimensions for the shorter side and longer side are 384 and 640, respectively. We train models, as explained earlier. Specifically, we train models with the cross-entropy loss  $\mathcal{L}_{XE}$  for ten epochs, in which we warm up the learning rates for the grid feature network and the caption generator from  $10^{-5}$  to  $10^{-4}$  in the first epoch, while we fix those for the backbone network and the object detector at  $10^{-5}$ . Then, we finetune the model based on the CIDEr-D optimization for ten epochs, where we set the fixed learning rate to  $5 \times 10^{-6}$  for the entire model. We use the Adam optimizer [24] with a batch size of 128. For the CIDEr-D optimization, we use beam search with a beam size of 5 and a maximum length of 20.

#### 4.4 Performance of Different Configurations

Our method has several design choices. We conduct experiments to examine which configuration is the best. The results are shown in Table 1. We used an identical configuration unless otherwise noted. Specifically, we use the feature extractor pretrained on the four datasets and parallel cross-attention for fusing the region and grid features.

The first block of Table 1(a) shows the effects of different (pre)training strategies of the visual backbone on image captioning performance. The ‘ImageNet’ column shows the result of the model using a Swin Transformer backbone pre-trained on ImageNet21K and the grid features alone; ‘VG’ and ‘4DS’ indicate the models with a detector pretrained on Visual Genome and the four datasets, respectively. They show that using more datasets leads to better performance.

The second block of Table 1(a) shows the effects of the number of object queries, or equivalently region features. The performance increases as they vary as 50, 100, and 150. We also confirmed that the performance is saturated for more region features, while the computational cost and false detection increase.

The third block shows the effect of the end-to-end training of the entire model. ‘Yes’ indicates the end-to-end training of the entire model and ‘No’ indicates training the model but the vision backbone. The results show that the end-to-end training considerably improves CIDEr score (from 139.6 to 144.3) with little sacrifice of B@4. This validates our expectation about the effectiveness of the end-to-end training; it arguably helps reduce the domain gap between object detection and image captioning.

The first block of Table 1(b) shows the performances of the model employing the concatenated cross-attention and its two variants using the grid features alone or the region features alone. They show that the region features alone work better than the grid features alone, and their fusion achieves the highest performance.

The three blocks of Table 1(b) show the performances of the three cross-attention architectures explained in Sec. 3.2. The second block shows the two variants of the sequential cross-attention, and the third block shows the two variants of the parallel cross-attention with different gated activation functions, i.e., sigmoid and identity. By identity activation, we mean setting all the values of  $c_l^q$  and  $c_l^r$  in Eq.(4) to one. These results show that the parallel cross-attention with sigmoid activation function performs the best; the sequential cross-attention in the order  $\mathcal{G} \rightarrow \mathcal{R}$  attains the second best result.

#### 4.5 Results on the COCO Dataset

We next show complete results on the COCO dataset by the offline and online evaluations. We present example results in the supplementary material.

**Offline Evaluation** Table 2 shows the performances of our method and the current state-of-the-art methods on the offline Karpathy test split. The compared methods are as follows: grid-based methods [46,41,56,58], region-based methods [4,22,55,38,52,17,17,13,18,14,27,8,36,11], the methods employing both grid and region features [49,34], and also the methods relying on large-scale pretraining on vision and language (V&L) tasks using a large image-text corpus [59,28,57], including SimVLM<sub>huge</sub>, a model pretrained on an extremely large dataset (i.e., 1.8 billion image-caption pairs) [48].

For fair comparison with the region-based methods, we report the results of two variants of our model, one with the object detector pretrained on Visual

Table 2: Offline results evaluated on the COCO Karpathy test split. ‘V. E. type’ indicates the type of visual features; ‘# VL Data’ is the number of image-text pairs used for vision-language pretraining.

Method	V. E.	# VL	Performance Metrics					
	Type		B@1	B@4	M	R	C	S
<b>w/ VL pretraining</b>								
UVLP [59]	$\mathcal{R}$	3.0M	-	39.5	29.3	-	129.3	23.2
Oscar <sub>base</sub> [28]	$\mathcal{R}$	6.5M	-	40.5	29.7	-	137.6	22.8
VinVL <sup>†</sup> <sub>large</sub> [57]	$\mathcal{R}$	8.9M	-	<b>41.0</b>	31.1	-	140.9	25.2
SimVLM <sub>huge</sub> [48]	$\mathcal{G}$	1.8B	-	40.6	<b>33.7</b>	-	<b>143.3</b>	<b>25.4</b>
<b>w/o VL pretraining</b>								
SAT [46]	$\mathcal{G}$	-	-	31.9	25.5	54.3	106.3	-
SCST [41]	$\mathcal{G}$	-	-	34.2	26.7	55.7	114.0	-
RSTNet [58]	$\mathcal{G}$	-	81.8	40.1	29.8	59.5	135.6	23.0
Up-Down [4]	$\mathcal{R}$	-	79.8	36.3	27.7	56.9	120.1	21.4
RFNet [22]	$\mathcal{R}$	-	79.1	36.5	27.7	57.3	121.9	21.2
GCN-LSTM [55]	$\mathcal{R}$	-	80.5	38.2	28.5	58.3	127.6	22.0
LBPF [38]	$\mathcal{R}$	-	80.5	38.3	28.5	58.4	127.6	22.0
SGAE [52]	$\mathcal{R}$	-	80.8	38.4	28.4	58.6	127.8	22.1
AoA [17]	$\mathcal{R}$	-	80.2	38.9	29.2	58.8	129.8	22.4
NG-SAN [13]	$\mathcal{R}$	-	-	39.9	29.3	59.2	132.1	23.3
GET [18]	$\mathcal{R}$	-	81.5	39.5	29.3	58.9	131.6	22.8
ORT [14]	$\mathcal{R}$	-	80.5	38.6	28.7	58.4	128.3	22.6
ETA [27]	$\mathcal{R}$	-	81.5	39.3	28.8	58.9	126.6	22.6
$\mathcal{M}^2$ Transformer [8]	$\mathcal{R}$	-	80.8	39.1	29.2	58.6	131.2	22.6
X-LAN [36]	$\mathcal{R}$	-	80.8	39.5	29.5	59.2	132.0	23.4
TCIC [11]	$\mathcal{R}$	-	81.8	40.8	29.5	59.2	135.4	22.5
Dual Global [49]	$\mathcal{R+G}$	-	81.3	40.3	29.2	59.4	132.4	23.3
DLCT [34]	$\mathcal{R+G}$	-	81.4	39.8	29.5	59.1	133.8	23.0
GRIT	$\mathcal{R+G}$	-	83.5	41.9	30.5	60.5	142.2	24.2
GRIT <sup>†</sup>	$\mathcal{R+G}$	-	<b>84.2</b>	<b>42.4</b>	<b>30.6</b>	<b>60.7</b>	<b>144.2</b>	<b>24.3</b>

Genome alone and the other (marked with <sup>†</sup>) with the object detector pre-trained on the four datasets, as explained earlier. It is seen from Table 2 that our models, regardless of the datasets used for the detector’s pre-training, outperform all the methods that do not use large-scale pre-training of vision and language tasks (i.e., the methods in the second block entitled ‘w/o VL pre-training’). Moreover, our model with the detector pretrained solely on Visual Genome (i.e., ‘GRIT’) performs better than those relying on large-scale V&L pre-training but SimVLM<sub>huge</sub>. Finally, our model with the pretrained detector on multiple datasets (i.e., ‘GRIT<sup>†</sup>’) outperforms SimVLM<sub>huge</sub> leveraging large-scale V&L pre-training in CIDEr score (i.e., 144.2 vs 143.3).

**Online Evaluation** We also evaluate our models (i.e., a single model and an ensemble of six models) on the 40K testing images by submitting their results on the official evaluation server. Table 3 shows the results and those of all the published methods on the leaderboard. Table 3 presents the metric scores based

Table 3: Online evaluation results on the COCO image captioning dataset

Method	Ensemble	B-1		B-2		B-3		B-4		M		R		C	
		c5	c40	c5	c40										
w/ VL pretraining															
VinVL <sup>large</sup> [57]	<b>x</b>	81.9	96.9	66.9	92.4	52.6	84.7	40.4	74.9	30.6	40.8	60.4	76.8	134.7	138.7
w/o VL pretraining															
SCST [41]	✓	78.1	93.7	61.9	86.0	47.0	75.9	35.2	64.5	27.0	35.5	56.3	70.7	114.7	116.7
Up-Down [4]	✓	80.2	95.2	64.1	88.8	49.1	79.4	36.9	68.5	27.6	36.7	57.1	72.4	117.9	120.5
HAN [47]	✓	80.4	94.5	63.8	87.7	48.8	78.0	36.5	66.8	27.4	36.1	57.3	71.9	115.2	118.2
GCN-LSTM [55]	✓	80.8	95.2	65.5	89.3	50.8	80.3	38.7	69.7	28.5	37.6	58.5	73.4	125.3	126.5
SGAE [52]	✓	81.0	95.3	65.6	89.5	50.7	80.4	38.5	69.7	28.2	37.2	58.6	73.6	123.8	126.5
AoA [17]	✓	81.0	95.0	65.8	89.6	51.4	81.3	39.4	71.2	29.1	38.5	58.9	74.5	126.9	129.6
HIP [54]	<b>x</b>	81.6	95.9	66.2	90.4	51.5	81.6	39.3	71.0	28.8	38.1	59.0	74.1	127.9	130.2
$\mathcal{M}^2$ Trans. [8]	✓	81.6	96.0	66.4	90.8	51.8	82.7	39.7	72.8	29.4	39.0	59.2	74.8	129.3	132.1
X-LAN [36]	✓	81.9	95.7	66.9	90.5	52.4	82.5	40.3	72.4	29.6	39.2	59.5	75.0	131.1	133.5
Dual Global [49]	<b>x</b>	80.8	95.1	65.6	81.3	51.1	81.3	39.1	71.2	28.9	38.4	58.9	74.4	126.3	129.2
DLCT [34]	✓	82.4	96.6	67.4	91.7	52.8	83.8	40.6	74.0	29.8	39.6	59.8	75.3	133.3	135.4
GRIT <sup>†</sup>	<b>x</b>	83.7	97.4	68.5	92.8	53.9	85.3	41.5	75.6	30.3	40.2	60.2	75.9	138.3	141.8
GRIT <sup>†</sup>	✓	<b>84.1</b>	<b>97.6</b>	<b>69.4</b>	<b>93.5</b>	<b>54.9</b>	<b>86.3</b>	<b>42.5</b>	<b>76.8</b>	<b>30.9</b>	<b>41.0</b>	<b>61.2</b>	<b>77.1</b>	<b>141.3</b>	<b>143.8</b>

on five (c5) and 40 reference captions (c40) per image. We can see that our method achieves the best scores for all the metrics. Note that even our single model outperforms all the published methods that use ensembles.

#### 4.6 Results on the ArtEmis and nocaps Datasets

As explained above, we evaluate our method on the ArtEmis and nocaps datasets. For nocaps, we evaluate zero-shot inference performance, i.e., the performance of the model trained on COCO. For ArtEmis, we train the model in the same way as COCO except for the number of training epochs, precisely, five epochs each for the training with the XE loss and that with the CIDEr-D optimization.

Table 4(a) shows the results of our method on the test split of ArtEmis [1]. It also shows the results of existing methods reported in [1], which are grid-based [35,46], region-based [8], and a nearest neighbor method using a holistic vector to encode images (denoted as  $\mathcal{H}$ ). Our method outperforms all these methods by a large margin.

Table 4 shows the results on the nocaps dataset, including the baseline methods reported in [2,8]. All the models are trained on the training split of the COCO datasets and tested on the validation split of nocaps, which consists of images with novel objects and captions with unseen vocabularies. Our method surpasses all the other methods including region-based methods [33,4,8] in both in-domain and out-of-domain images. See the supplementary material for the full results.

#### 4.7 Computational Efficiency

We measured the inference time of GRIT and two representative region-based methods, VinVL [57] and  $\mathcal{M}^2$  Transformer [8]. It is the computational time per image from image input to caption generation. Specifically, we measured the time

Table 4: Performance on the ArtEmis and nocaps datasets

Method	a) Performance on the ArtEmis test split						b) Performance on the nocaps validation split								
	Type	Performance Metrics					Method	In-Domain					Overall		
		B@1	B@2	B@3	B@4	M		C	S	C	S	C			
NN [1]	$\mathcal{H}$	36.4	13.9	5.4	2.2	10.2	21.0	NBT [2]	$\mathcal{R}$	62.7	10.1	54.0	8.6	53.9	9.2
ANP [1]	$\mathcal{G}$	39.6	13.4	4.2	1.4	8.8	20.2	Up-down [2]	$\mathcal{R}$	78.1	11.6	31.3	8.3	55.3	10.1
SAT [1]	$\mathcal{G}$	53.6	29.0	15.5	8.7	14.2	29.7	Trans. [8]	$\mathcal{R}$	78.0	11.0	29.7	7.8	54.7	9.8
$\mathcal{M}^2$ Trans. [1]	$\mathcal{R}$	50.7	28.2	15.9	9.5	13.7	28.0	$\mathcal{M}^2$ Trans. [8]	$\mathcal{R}$	85.7	12.1	38.9	8.9	64.5	11.1
GRIT <sup>†</sup>	$\mathcal{R} + \mathcal{G}$	<b>70.1</b>	<b>40.1</b>	<b>20.9</b>	<b>11.3</b>	<b>16.8</b>	<b>33.3</b>	GRIT <sup>†</sup>	$\mathcal{R} + \mathcal{G}$	<b>105.9</b>	<b>13.6</b>	<b>72.6</b>	<b>11.1</b>	<b>90.2</b>	<b>12.8</b>

to generate a caption of length 20 with a beam size of five on a V100 GPU. The input image resolution was set to  $800 \times 1333$  for VinVL and  $\mathcal{M}^2$  Transformer as reported in [4,57]. We set it to  $384 \times 640$  for GRIT since it already achieves higher accuracy. Figure 1 shows the breakdown of the inference time for the three methods. GRIT reduces the time for feature extraction by a factor of 10 compared with the others. Similar to  $\mathcal{M}^2$  Transformer, GRIT has a lightweight caption generator and thus spends much less time than VinVL for generating a caption after receiving the visual features. GRIT can run with minibatch size up to 64 on a single V100 GPU, while others cannot afford large minibatch. With minibatch size  $\geq 32$ , the per-image inference time decreases to about 32ms. More details are given in the supplementary material.

## 5 Summary and Conclusion

In this paper, we have proposed a Transformer-based architecture for image captioning named GRIT. It integrates the region features and the grid features extracted from an input image to extract richer visual information from input images. Previous SOTA methods employ a CNN-based detector to extract region features, which prevents the end-to-end training of the entire model and makes to high computational costs. Using the Swin Transformer for a backbone extracting the initial visual feature, GRIT resolves these two issues by employing a DETR-based detector. Furthermore, GRIT obtains grid features by updating the feature from the same backbone using a self-attention Transformer, aiming to extract richer context information complementing the region feature. These two features are fed to the caption generator equipped with a unique cross-attention mechanism, which computes and applies attention from the dual features on the generated caption sentence. The integration of all these components led to significant performance improvement. The experimental results validated our approach, showing that GRIT outperforms all published methods by a large margin in inference accuracy and speed.

**Acknowledgments** This work was supported by JST [Moonshot Research and Development], Grant Number [JPMJMS2032] and by JSPS KAKENHI Grant Number 20H05952 and 19H01110.

## A Additional Details for Object Detection

### A.1 Object Detection Datasets

When pretraining our model on the four datasets (i.e., Visual Genome (VG), COCO, OpenImages, and Objects365), we follow [57] to build a unified training corpus with the statistics shown in Table 5 except that we do not use the annotations from COCO stuff [6]. The resultant corpus has images with 1848 categories.

Table 5: Statistics of the pretraining datasets for object detection.

Source	VG	COCO	Objects365	OpenImages
Images	97k	111k	609k	1.67M
Categories	1594	80	365	500
Sampling	$\times 8$	$\times 8$	$\times 2$	$\times 1$

### A.2 Implementation Details

For the object detector, we set the number of queries  $N = 150$ , the number of sampling points equal to 4, and the hidden dimension  $d = 512$ . The backbone network weights are initialized by the weights of Swin-Base ( $384 \times 384$ ) pretrained on ImageNet21K [31]. Following [60], the loss for normalized bounding box regression for object  $i$ ,  $\mathcal{L}_{box}(b_i, \hat{b}_{\hat{\sigma}(i)})$  is computed as the weighted summation of a box distance  $\mathcal{L}_{l_1}$  and a GIoU loss  $\mathcal{L}_{iou}$ :

$$\mathcal{L}_{l_1}(b_i, \hat{b}_{\hat{\sigma}(i)}) = \|b_i - \hat{b}_{\hat{\sigma}(i)}\|_1, \quad (7)$$

$$\mathcal{L}_{iou}(b_i, \hat{b}_{\hat{\sigma}(i)}) = 1 - \left( \frac{|b_i \cap \hat{b}_{\hat{\sigma}(i)}|}{|b_i \cup \hat{b}_{\hat{\sigma}(i)}|} - \frac{|\mathbf{B}(b_i, \hat{b}_{\hat{\sigma}(i)}) \setminus b_i \cup \hat{b}_{\hat{\sigma}(i)}|}{|\mathbf{B}(b_i, \hat{b}_{\hat{\sigma}(i)})|} \right), \quad (8)$$

$$\mathcal{L}_{box}(b_i, \hat{b}_{\hat{\sigma}(i)}) = \alpha_{l_1} \mathcal{L}_{l_1}(b_i, \hat{b}_{\hat{\sigma}(i)}) + \alpha_{iou} \mathcal{L}_{iou}(b_i, \hat{b}_{\hat{\sigma}(i)}), \quad (9)$$

where  $\alpha_{l_1} = 5$ ,  $\alpha_{iou} = 2$ , and  $\mathbf{B}$  outputs the largest box covering  $b_i$  and  $\hat{b}_{\hat{\sigma}(i)}$ . We also employ two training strategies, i.e., iterative bounding box refinement and auxiliary losses; see [60] and our configuration files for details.

### A.3 Object Detection Results

Table 6 shows the performance on the COCO validation split and the Visual Genome test split of our object detector compared with VinVL and BUTD [4]. It is seen that the object detector of GRIT attains comparable or higher performance on the two datasets as compared with BUTD and VinVL when pretrained on the similar datasets.

Table 6: Performance of object detection on the COCO and Visual Genome datasets. ‘4DS’ denotes the four object detection datasets.

Model	Training Data	mAP (COCO)	mAP <sup>50</sup> (VG)
BUTD [4]	VG	-	10.2
VinVL [57]	4DS	50.5	13.8
GRIT	VG	33.6	14.2
GRIT <sup>†</sup>	4DS	50.8	15.1

## B Additional Details for Image Captioning

**Class Token** We prepend a class token embedding  $g_{\langle \text{cls} \rangle} \in \mathbb{R}^d$  to  $G_0$  before forwarding them to the grid feature network. We use this class token embedding to predict the emotion category of the input image when training an emotion-grounded model on the ArtEmis dataset; see Sec. B.2.

**Boundary Tokens** Following previous studies, we prepend a special token  $\langle \text{sos} \rangle$  to the beginning of captions, and append another special token  $\langle \text{eos} \rangle$  to the end of captions during training. During inference, we start the generation by setting the first token to  $\langle \text{sos} \rangle$ .

Table 7: Breakdown of SPICE F-scores over various sub-categories and the CLIP scores.

Method	SPICE	Object	Attr.	Relation	Color	Count	Size	CLIP
Up-Down [4]	21.4	39.1	10.0	6.5	11.4	18.4	3.2	-
Transformer [8]	21.1	38.6	9.6	6.3	9.2	17.5	2.0	-
$\mathcal{M}^2$ Trans. [8]	22.6	40.0	11.6	6.9	12.9	20.4	3.5	73.4
GRIT <sup>†</sup>	24.3	42.7	13.5	7.7	14.7	29.3	4.5	77.2

### B.1 Image Captioning on the COCO dataset

**SPICE Sub-category and CLIPscore Metrics** Table 7 reports a breakdown of SPICE F-scores over various sub-categories on the “Karpathy” test split, in comparison with the region-based methods: Up-Down [4], vanilla Transformer [8], and  $\mathcal{M}^2$  Transformer [8]. These scores give a quantitative assessment of performance on different aspects when describing the content of images. As seen in Table 7, our method attains better scores over all sub-categories, showing significant improvement on identifying and counting objects, attributes, and relationships between objects. The table also reports the CLIP scores [15] of the

two methods, showing consistent improvement of our method over the compared method.

## B.2 Image Captioning on the ArtEmis dataset

**ArtEmis Dataset** This dataset consists of 80,031 unique images divided into the training, validation, and test splits with the ratios of 85%, 5%, and 10%, respectively. Each caption of a given image is annotated with an emotion label. In total, there are 454,684 captions along with 8 unique emotion categories; see [1] for details.

**Emotion Grounded Model** Following [1], we also trained an emotion grounded model, which predicts the emotion associated with the caption. Specifically, we mapped the updated class embedding  $g_{\langle \text{cls} \rangle}$  into an 8-dimensional vector using a linear projection. During training, we minimized the summation of the two losses, i.e., emotion prediction and caption generation.

**Full Results** Table 8 shows the full results of different models on the test split of the Artemis dataset including the emotion grounded models. It is noted that the ground truth emotion labels are not provided during inference.

Table 8: Performance on the ArtEmis test split.

Method	Emotion	V. E.	Performance Metrics					
	Grounded	Type	B@1	B@2	B@3	B@4	M	R
NN [1]	No	$\mathcal{H}$	36.4	13.9	5.4	2.2	10.2	21.0
ANP [1]	No	$\mathcal{G}$	39.6	13.4	4.2	1.4	8.8	20.2
$\mathcal{M}^2$ Trans. [1]	Yes	$\mathcal{R}$	51.1	28.2	15.4	9.0	13.7	28.6
$\mathcal{M}^2$ Trans. [1]	No	$\mathcal{R}$	50.7	28.2	15.9	9.5	14.0	28.0
SAT [1]	Yes	$\mathcal{G}$	52.0	28.0	14.6	7.9	13.4	29.4
SAT [1]	No	$\mathcal{G}$	53.6	29.0	15.5	8.7	14.2	29.7
GRIT <sup>†</sup>	Yes	$\mathcal{R} + \mathcal{G}$	69.3	39.4	19.2	11.1	16.5	33.0
GRIT <sup>†</sup>	No	$\mathcal{R} + \mathcal{G}$	<b>70.1</b>	<b>40.1</b>	<b>20.9</b>	<b>11.3</b>	<b>16.8</b>	<b>33.3</b>

## B.3 Image Captioning on the nocaps Dataset

**Full results** We report the full results on the validation split of the nocaps dataset for different domains, i.e., in-domain, near-domain, and out-of-domain, in Table 9.

Table 9: Performance on the nocaps validation split.

Method	V.E Type	in-domain		near-domain		out-domain		Overall	
		C	S	C	S	C	S	C	S
NBT [2]	$\mathcal{R}$	62.7	10.1	51.9	9.2	54.0	8.6	53.9	9.2
Up-down [2]	$\mathcal{R}$	78.1	11.6	57.7	10.3	31.3	8.3	55.3	10.1
Trans. [8]	$\mathcal{R}$	78.0	11.0	-	-	29.7	7.8	54.7	9.8
$\mathcal{M}^2$ Trans. [8]	$\mathcal{R}$	85.7	12.1	-	-	38.9	8.9	64.5	11.1
GRIT <sup>†</sup>	$\mathcal{R} + \mathcal{G}$	<b>105.9</b>	<b>13.6</b>	<b>92.16</b>	<b>13.05</b>	<b>72.6</b>	<b>11.1</b>	<b>90.2</b>	<b>12.8</b>

#### B.4 Computational Efficiency

We measured the inference time of GRIT and two representative region-based methods, VinVL [57] and  $\mathcal{M}^2$  Transformer [8], on the same machine having a Tesla V100-SXM2 of 16GB memory with CUDA version 10.0 and Driver version 410.104. It has Intel(R) Xeon(R) Gold 6148 CPU. The comparison was conducted following [19,23]. Specifically, we excluded the time of preprocessing the image and loading it to the GPU device. Also, the images are rescaled to the resolutions such that all the compared methods achieve its highest performance for image captioning. For the compared methods, we used the official implementations of  $\mathcal{M}^2$  Transformer<sup>3</sup> and VinVL<sup>4</sup>.

Regarding feature extraction, we extracted the region features from Faster R-CNN using the original implementation<sup>5</sup> used by  $\mathcal{M}^2$  Transformer and another implementation<sup>6</sup> used by VinVL. It is seen that VinVL and  $\mathcal{M}^2$  Transformer spend considerable time on feature extraction due to the forward pass through the CNN backbone with high resolution inputs and the computationally expensive regional operations. It is also noted that VinVL introduced class-agnostic NMS operations, which reduce a great amount of time consumed by class-aware NMS operations in the standard Faster R-CNN. On the other hand, we employ a Deformable DETR-based detector to extract region features without using all such operations. Table 10 shows the comparison on feature extraction.

Regarding caption generation, all the methods use beam search as the decoding strategy, with beam size of 5 and the maximum caption length of 20. Both  $\mathcal{M}^2$  Transformer and GRIT employ a lightweight caption generator (caption decoder) having only 3 transformer layers with hidden dimension of 512 while VinVL<sub>large</sub> has 24 transformer layers with hidden dimension of 1024; see Table 11. Thus, with the visual features as inputs,  $\mathcal{M}^2$  Transformer and GRIT spend less inference time generating words than VinVL<sub>large</sub> in the autoregressive manner.

<sup>3</sup> <https://github.com/aimagelab/meshed-memory-transformer>

<sup>4</sup> <https://github.com/pzzhang/VinVL>

<sup>5</sup> <https://github.com/peteanderson80/bottom-up-attention>

<sup>6</sup> [https://github.com/microsoft/scene\\_graph\\_benchmark](https://github.com/microsoft/scene_graph_benchmark)

Table 10: The inference time on feature extraction of different methods.

Method	Backbone	Detector	Regional Operations	Inference Time
VinVL <sub>large</sub> [57]	ResNeXt-152	Faster R-CNN	Class-Agnostic NMS RoI Align, etc	304 ms
$\mathcal{M}^2$ Trans. [8]	ResNet-101	Faster R-CNN	Class-Aware NMS RoI Align, etc	736 ms
GRIT	Swin-Base	DETR-based	-	31 ms

Table 11: The inference time on caption generation of different methods.

Method	No. of Layers	Hidden Dim.	Inference Time
VinVL <sub>large</sub> [57]	24	1024	542 ms
$\mathcal{M}^2$ Transformer [8]	3	512	174 ms
GRIT	3	512	138 ms

## B.5 Qualitative Examples

Figure 4, 5, 6, and 7 show some examples of the captions generated by our proposed method (GRIT) and another region-based method ( $\mathcal{M}^2$  Transformer) given the same input images from the COCO test split. It is observed that the generated captions from GRIT are qualitatively better than those generated by the baseline method in terms of detecting and counting objects as well as describing their relationships in the given images. The inaccuracy of the captions generated by the baseline method might be due to the drawbacks of the region features extracted by a frozen pretrained object detector which produces wrong detection and lacks of contextual information.

## References

1. Achlioptas, P., Ovsjanikov, M., Haydarov, K., Elhoseiny, M., Guibas, L.J.: Artemis: Affective language for visual art. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 11569–11579 (2021)
2. Agrawal, H., Desai, K., Wang, Y., Chen, X., Jain, R., Johnson, M., Batra, D., Parikh, D., Lee, S., Anderson, P.: nocaps: novel object captioning at scale. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 8948–8957 (2019)
3. Anderson, P., Fernando, B., Johnson, M., Gould, S.: Spice: Semantic propositional image caption evaluation. In: Proceedings of European Conference on Computer Vision. pp. 382–398 (2016)
4. Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., Zhang, L.: Bottom-up and top-down attention for image captioning and visual question answering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6077–6086 (2018)

5. Banerjee, S., Lavie, A.: Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In: Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization. pp. 65–72 (2005)
6. Caesar, H., Uijlings, J., Ferrari, V.: Coco-stuff: Thing and stuff classes in context. In: Computer vision and pattern recognition (CVPR), 2018 IEEE conference on. IEEE (2018)
7. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: Proceedings of European Conference on Computer Vision. pp. 213–229 (2020)
8. Cornia, M., Stefanini, M., Baraldi, L., Cucchiara, R.: Meshed-memory transformer for image captioning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 10578–10587 (2020)
9. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv:1810.04805 (2018)
10. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv:2010.11929 (2020)
11. Fan, Z., Wei, Z., Wang, S., Wang, R., Li, Z., Shan, H., Huang, X.: Tcic: Theme concepts learning cross language and vision for image captioning. arXiv:2106.10936 (2021)
12. Fang, Y., Liao, B., Wang, X., Fang, J., Qi, J., Wu, R., Niu, J., Liu, W.: You only look at one sequence: Rethinking transformer in vision through object detection. In: Proceedings of Advances in Neural Information Processing Systems (2021)
13. Guo, L., Liu, J., Zhu, X., Yao, P., Lu, S., Lu, H.: Normalized and geometry-aware self-attention network for image captioning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 10327–10336 (2020)
14. Herdade, S., Kappeler, A., Boakye, K., Soares, J.: Image captioning: Transforming objects into words. In: Proceedings of Advances in Neural Information Processing Systems (2019)
15. Hessel, J., Holtzman, A., Forbes, M., Bras, R.L., Choi, Y.: Clipscore: A reference-free evaluation metric for image captioning. arXiv preprint arXiv:2104.08718 (2021)
16. Honnibal, M., Montani, I.: spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing (2017), to appear
17. Huang, L., Wang, W., Chen, J., Wei, X.Y.: Attention on attention for image captioning. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 4634–4643 (2019)
18. Ji, J., Luo, Y., Sun, X., Chen, F., Luo, G., Wu, Y., Gao, Y., Ji, R.: Improving image captioning by leveraging intra-and inter-layer global representation in transformer network. In: Proceedings of the AAAI Conference on Artificial Intelligence. pp. 1655–1663 (2021)
19. Jiang, H., Misra, I., Rohrbach, M., Learned-Miller, E., Chen, X.: In defense of grid features for visual question answering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 10267–10276 (2020)
20. Karpathy: Karpathy/neuraltalk: Neuraltalk is a python+numpy project for learning multimodal recurrent neural networks that describe images with sentences., <https://github.com/karpathy/neuraltalk>
21. Karpathy, A., Fei-Fei, L.: Deep visual-semantic alignments for generating image descriptions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3128–3137 (2015)

22. Ke, L., Pei, W., Li, R., Shen, X., Tai, Y.W.: Reflective decoding network for image captioning. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 8888–8897 (2019)
23. Kim, W., Bokyung, S., Ildoo, K., Kim, W.: Vilt: Vision-and-language transformer without convolution or region supervision. In: International Conference on Machine Learning (2021)
24. Kingma, D.P., Ba, J.: Adam: A Method for Stochastic Optimization. In: Proceedings of International Conference on Representation Learning (2015)
25. Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.J., Shamma, D.A., Bernstein, M., Fei-Fei, L.: Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. International Journal of Computer Vision **123**(1), 32–73 (2017)
26. Kuznetsova, A., Rom, H., Alldrin, N., Uijlings, J., Krasin, I., Pont-Tuset, J., Kamali, S., Popov, S., Malloci, M., Kolesnikov, A., Duerig, T., Ferrari, V.: The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. International Journal of Computer Vision **128**, 1956–1981 (2020)
27. Li, G., Zhu, L., Liu, P., Yang, Y.: Entangled transformer for image captioning. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 8928–8937 (2019)
28. Li, X., Yin, X., Li, C., Zhang, P., Hu, X., Zhang, L., Wang, L., Hu, H., Dong, L., Wei, F., et al.: Oscar: Object-semantics aligned pre-training for vision-language tasks. In: Proceedings of European Conference on Computer Vision. pp. 121–137 (2020)
29. Lin, C.Y.: Rouge: A package for automatic evaluation of summaries. In: Text Summarization Branches Out. pp. 74–81 (2004)
30. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Proceedings of European Conference on Computer Vision. pp. 740–755 (2014)
31. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 10012–10022 (2021)
32. Lu, J., Xiong, C., Parikh, D., Socher, R.: Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 375–383 (2017)
33. Lu, J., Yang, J., Batra, D., Parikh, D.: Neural baby talk. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7219–7228 (2018)
34. Luo, Y., Ji, J., Sun, X., Cao, L., Wu, Y., Huang, F., Lin, C.W., Ji, R.: Dual-level collaborative transformer for image captioning. In: Proceedings of the AAAI Conference on Artificial Intelligence. pp. 2286–2293 (2021)
35. Mathews, A., Xie, L., He, X.: Senticap: Generating image descriptions with sentiments. In: Proceedings of the AAAI conference on artificial intelligence. pp. 3574–3580 (2016)
36. Pan, Y., Yao, T., Li, Y., Mei, T.: X-linear attention networks for image captioning. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 10971–10980 (2020)
37. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics. pp. 311–318 (2002)

38. Qin, Y., Du, J., Zhang, Y., Lu, H.: Look back and predict forward in image captioning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 8367–8375 (2019)
39. Radford, A., Narasimhan, K., Salimans, T., Sutskever, I.: Improving language understanding by generative pre-training. In: Technical report. OpenAI (2018)
40. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. In: Proceedings of Advances in Neural Information Processing Systems. pp. 91–99 (2015)
41. Rennie, S.J., Marcheret, E., Mroueh, Y., Ross, J., Goel, V.: Self-critical sequence training for image captioning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7008–7024 (2017)
42. Shao, S., Li, Z., Zhang, T., Peng, C., Yu, G., Zhang, X., Li, J., Sun, J.: Objects365: A large-scale, high-quality dataset for object detection. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 8430–8439 (2019)
43. Song, H., Sun, D., Chun, S., Jampani, V., Han, D., Heo, B., Kim, W., Yang, M.H.: Vidit: An efficient and effective fully transformer-based object detector. arXiv:2110.03921 (2021)
44. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. arXiv:1706.03762 (2017)
45. Vedantam, R., Lawrence Zitnick, C., Parikh, D.: Cider: Consensus-based image description evaluation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4566–4575 (2015)
46. Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Show and tell: A neural image caption generator. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3156–3164 (2015)
47. Wang, W., Chen, Z., Hu, H.: Hierarchical attention network for image captioning. In: Proceedings of the AAAI Conference on Artificial Intelligence. pp. 8957–8964 (2019)
48. Wang, Z., Yu, J., Yu, A.W., Dai, Z., Tsvetkov, Y., Cao, Y.: Simvlm: Simple visual language model pretraining with weak supervision. arXiv:2108.10904 (2021)
49. Xian, T., Li, Z., Zhang, C., Ma, H.: Dual global enhanced transformer for image captioning. Neural Networks **148**, 129–141 (2022)
50. Xu, H., Yan, M., Li, C., Bi, B., Huang, S., Xiao, W., Huang, F.: E2e-vlp: End-to-end vision-language pre-training enhanced by visual learning. arXiv:2106.01804 (2021)
51. Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., Bengio, Y.: Show, attend and tell: Neural image caption generation with visual attention. In: Proceedings of International Conference on Machine Learning. pp. 2048–2057 (2015)
52. Yang, X., Tang, K., Zhang, H., Cai, J.: Auto-encoding scene graphs for image captioning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 10685–10694 (2019)
53. Yang, X., Zhang, H., Cai, J.: Learning to collocate neural modules for image captioning. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 4250–4260 (2019)
54. Yao, T., Pan, Y., Li, Y., Mei, T.: Hierarchy parsing for image captioning. Proceedings of International Conference on Computer Vision pp. 2621–2629 (2019)
55. Yao, T., Pan, Y., Li, Y., Mei, T.: Exploring visual relationship for image captioning. In: Proceedings of European Conference on Computer Vision. pp. 684–699 (2018)

56. Yao, T., Pan, Y., Li, Y., Qiu, Z., Mei, T.: Boosting image captioning with attributes. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 4894–4902 (2017)
57. 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: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5579–5588 (2021)
58. Zhang, X., Sun, X., Luo, Y., Ji, J., Zhou, Y., Wu, Y., Huang, F., Ji, R.: Rstnet: Captioning with adaptive attention on visual and non-visual words. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 15465–15474 (2021)
59. Zhou, L., Palangi, H., Zhang, L., Hu, H., Corso, J., Gao, J.: Unified vision-language pre-training for image captioning and vqa. In: Proceedings of the AAAI Conference on Artificial Intelligence. pp. 13041–13049 (2020)
60. Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J.: Deformable DETR: Deformable transformers for end-to-end object detection. In: Proceedings of International Conference of Learning Representations (2021)



**GT-1:** a child is brushing her hair in the mirror  
**GT-2:** a little girl is brushing her hair in a bathroom  
**M<sup>2</sup>:** a young girl holding a baseball bat in a  
**GRIT:** a little girl brushing her hair with a brush



**GT-1:** an elephant walking not to far from a rhino in a forest  
**GT-2:** an elephant and a rhino share a field with a pond  
**M<sup>2</sup>:** a group of elephants grazing in a field  
**GRIT:** an elephant and a rhino standing in a field



**GT-1:** a bike is parked alongside the lake shore  
**GT-2:** a bike is parked on the grass in front of the lake  
**M<sup>2</sup>:** a bicycle leaning against a bridge over the water  
**GRIT:** a bike parked next to a bridge on the water



**GT-1:** 2 female tennis players standing with their rackets  
**GT-2:** a pair of young women hold tennis balls and rackets  
**M<sup>2</sup>:** a woman hitting a tennis ball with a tennis racket  
**GRIT:** 2 people hold tennis rackets and balls on a court



**GT-1:** a cat holding a toothbrush in its mouth  
**GT-2:** a cat chewing on a packaged pink toothbrush  
**M<sup>2</sup>:** a cat laying on top of a pair of scissors  
**GRIT:** a cat with a toothbrush in its mouth on



**GT-1:** the boy is playing video games in his bedroom  
**GT-2:** a young man is sitting in a chair playing a video game  
**M<sup>2</sup>:** a young man sitting in a chair holding a wii remote  
**GRIT:** a man sitting in a chair playing a video game



**GT-1:** a woman is taking a turkey out of the oven  
**GT-2:** a woman is taking the cooked turkey out of the oven.  
**M<sup>2</sup>:** a woman taking a pizza out of an oven with a  
**GRIT:** a woman taking a turkey out of an oven with



**GT-1:** a giraffe standing outside of a building next to a tree.  
**GT-2:** a giraffe standing in a small piece of shade.  
**M<sup>2</sup>:** two giraffes are standing in a zoo enclosure  
**GRIT:** a giraffe standing in the dirt next to a building



**GT-1:** bowls on a table with meat and vegetables.  
**GT-2:** four plates of different kind of food sitting on a table  
**M<sup>2</sup>:** three plates of food on a wooden table with a  
**GRIT:** four bowls of food and a spoon on a table

Fig. 4: Qualitative examples from our method (GRIT) and a region-based method ( $M^2$  Transformer) on the COCO test images. Zoom in for better view.



**GT-1:** a white cat is laying on a black skateboard  
**GT-2:** A cat is sleeping on a skateboard.  
**M<sup>2</sup>:** a kitten laying on the floor next to a skateboard  
**GRIT:** a cat laying on a skateboard on the floor



**GT-1:** A baby elephant looking at a white duck  
**GT-2:** A small elephant standing next to a white bird  
**M<sup>2</sup>:** an elephant in a field with two birds in the  
**GRIT:** a baby elephant walking in a field of grass



**GT-1:** Two children wrapped in blankets reading on a bed.  
**GT-2:** Two children reading while lying in their bed  
**M<sup>2</sup>:** two people laying in a bed with a  
**GRIT:** two young boys sitting on a bed reading a book



**GT-1:** a kitchen with a refrigerator next to a sink.  
**GT-2:** a red bucket sits in a sink next to an open refrigerator  
**M<sup>2</sup>:** an open refrigerator with the door open in a kitchen  
**GRIT:** a kitchen with a sink and an open refrigerator



**GT-1:** a woman pulling her luggage past a fire hydrant.  
**GT-2:** a woman pulls a wheeled suitcase past a fire hydrant  
**M<sup>2</sup>:** a person riding a skateboard down a street with a  
**GRIT:** a person pulling a suitcase next to a fire hydrant



**GT-1:** two zebras in an animal park behind a wire fence  
**GT-2:** two zebras in a zoo, behind a wire fence  
**M<sup>2</sup>:** a zebra standing next to a fence in a  
**GRIT:** two zebras standing behind a fence in a zoo



**GT-1:** a small teddy bear is wedged into an opening in a car dashboard  
**GT-2:** little teddy bear attached to the dashboard of the car  
**M<sup>2</sup>:** a stuffed teddy bear sitting in the back of a car  
**GRIT:** a teddy bear sitting on the dashboard of a car



**GT-1:** horses racing on a race track with jockeys  
**GT-2:** a group of jockeys ride horses on a track  
**M<sup>2</sup>:** a group of people riding horses in a  
**GRIT:** a group of jockeys riding horses on a track



**GT-1:** two birds going up the back of a giraffe.  
**GT-2:** two birds sitting on the back of a giraffe.  
**M<sup>2</sup>:** a bird on the neck of a giraffe with a  
**GRIT:** two birds sitting on the back of a giraffe

Fig. 5: Qualitative examples from our method (GRIT) and a region-based method ( $\mathcal{M}^2$  Transformer) on the COCO test images. Zoom in for better view.

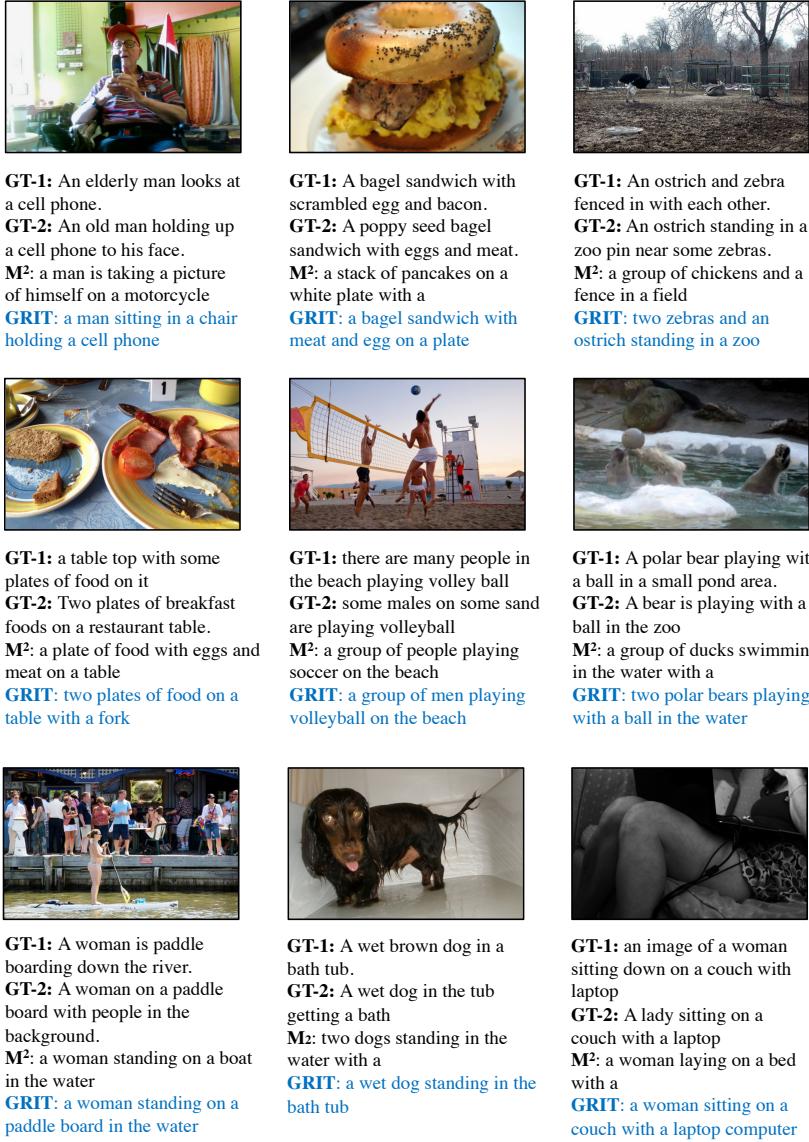


Fig. 6: Qualitative examples from our method (GRIT) and a region-based method ( $\mathcal{M}^2$  Transformer) on the COCO test images. Zoom in for better view.



**GT-1:** A dried black flower in a long, tall black & white vase.  
**GT-2:** Thin black and white vase with black flowers.  
**M<sup>2</sup>:** two white vases with a flower in them on a  
**GRIT:** a black and white vase with a flower in it



**GT-1:** The bushels of bananas on display are purple  
**GT-2:** A pile of black bananas and other fruit  
**M<sup>2</sup>:** a bunch of fruits and vegetables in a basket  
**GRIT:** a pile of bananas and other fruit on display



**GT-1:** A doll sitting at a table with fake food  
**GT-2:** The doll is posed at the table eating a meal  
**M<sup>2</sup>:** a young child sitting at a table with a plate of food  
**GRIT:** a doll sitting at a table with a plate of food



**GT-1:** A woman throwing a frisbee outside at a park  
**GT-2:** a woman is throwing a disk outside in the sun  
**M<sup>2</sup>:** a woman holding a blue umbrella in the street  
**GRIT:** a woman is throwing a frisbee in the street



**GT-1:** Two frisbees laying on the ground next to a sports water bottle.  
**GT-2:** Two flying disks on the ground next to a water bottle  
**M<sup>2</sup>:** a knife and a knife on a table with a  
**GRIT:** two frisbees laying on the ground next to a bottle



**GT-1:** Two knives are lying on a dark red surface.  
**GT-2:** Two knives placed on a dining table  
**M<sup>2</sup>:** a close up of a red tie with a  
**GRIT:** two knives are on a red table with



**GT-1:** A woman laying in bed reading a book while wearing purple socks  
**GT-2:** A woman is laying in bed reading a book  
**M<sup>2</sup>:** a dog is looking at a person on a bed  
**GRIT:** a woman laying on a bed with a book



**GT-1:** A zombie walking down a street covered in blood  
**GT-2:** A man dressed like a zombie with other zombies around him.  
**M<sup>2</sup>:** a man in a suit and tie walking with a group of people  
**GRIT:** a man dressed as zombies walking down a street



**GT-1:** A person is standing near a ski-lift with a view of mountains  
**GT-2:** A man stands beside a ski lift on a mountain  
**M<sup>2</sup>:** a person riding a snowboard down a snow covered slope  
**GRIT:** a person on a ski lift on a snowy mountain

Fig. 7: Qualitative examples from our method (GRIT) and a region-based method ( $\mathcal{M}^2$  Transformer) on the COCO test images. Zoom in for better view.