GRILL: Grounded Vision-language Pre-training via Aligning Text and Image Regions

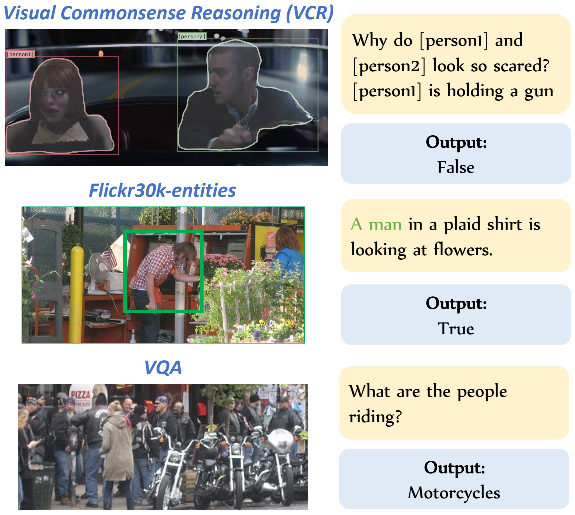
Anonymous ACL submission

Abstract

Cross-task generalization is an important ability for few-shot learners to achieve better zero-/few-shot performance on diverse tasks. However, such generalization to vision-language tasks including grounding and generation tasks has been under-explored. Furthermore, existing few-shot VL models struggle to handle tasks that involve object grounding and multiple images such as visual commonsense reasoning (Zellers et al., 2019) or NLVR2 (Suhr et al., 2019). In this paper, we introduce GRILL, **GR**ounded v**I**sion **L**anguage a**L**igning, a novel VL model that learns object grounding and localization in pre-training and can adapt to diverse grounding tasks with no or very few training instances. Specifically, GRILL exploits object-text alignments and learns to ground objects in pre-training, which enables it to transfer to tasks such as referring expression comprehension (Mao et al., 2016) and visual commonsense reasoning (Zellers et al., 2019) in a zero-/few-shot fashion. We evaluate our model on various zero-/few-shot VL tasks and show that it consistently surpasses the state-of-the-art few-shot methods.

1 Introduction

Cross-task generalization has been explored and investigated on zero-/few-shot NLP tasks by performing multi-task learning and generalizing unseen tasks with task-specific prompts (Sanh et al., 2021) or pre-training huge language models on a massive dataset and using a few examples as demonstrations for generalization (Brown et al., 2020). Similarly, few-shot learning methods aim to leverage the pre-trained language models and their powerful generalization abilities to adapt to vision-language (VL) domains and learn new tasks from zero or a few examples (Tsimpoukelli et al., 2021; Radford et al., 2021; Jin et al., 2021; Alayrac et al., 2022).



Why do [person1] and [person2] look so scared? [person1] is holding a gun

Output: False

A man in a plaid shirt is looking at flowers.

Output: True

What are the people riding?

Output: Motorcycles

Figure 1: **Examples of inputs and outputs of our task setup.** GRILL can generalize to diverse VL tasks including grounding tasks in a zero-/few-shot manner.

While few-shot learning methods can overcome the challenges of data-hungry supervised learning and avoid the need for task-specific fine-tuning, existing few-shot VL learners *do not address the challenge of grounding tasks* that require not only understanding the image and the language, but also locating and identifying relevant regions or objects in images, such as visual commonsense reasoning (VCR) (Zellers et al., 2019), where the model has to reason about the actions, intentions, and emotions of agents in the image; or Flickr30k-entities (Plummer et al., 2015), where the model has to align the mentions of entities in the captions with their corresponding regions in the image. These tasks are essential for VL models to achieve human-like reasoning and understanding. However existing few-shot methods lack the skills to address the challenge, as they *do not explicitly model the spatial and visual information of the regions or objects*. On the other hand, existing fine-tuning methods for grounding tasks rely on special representations for regions or objects, such as special tokens that mark the regions or objects in the captions and the images (Cho et al., 2021); object features extracted from a pre-trained object detector (Su et al., 2020;

Chen et al., 2019) and concatenated with the image features, etc. These methods achieve good results with fine-tuning, but they are not compatible with zero-/few-shot learning, due to the different designs of object representation for each task and the dependence on external object detectors that may not cover all the relevant concepts.

In this paper, we introduce GRILL, **GR**ounded v**I**sion **L**anguage a**L**igning, a new VL model that can learn object grounding and localization during pre-training and generalize to a wide range of VL tasks including grounding tasks in a zero-/few-shot fashion. Our model is a sequence-to-sequence transformer model (Vaswani et al., 2017) that uses a vision transformer (ViT) (Dosovitskiy et al., 2021; Liu et al., 2021) to process images with patch embeddings, where each patch represents a fixed-size region of the image. We represent a visual concept (object or region) that corresponds to a group of patches by aggregating information across the patches. This enables our model to generate better representations for any kind of regions or images without relying on pre-trained object detectors, which may be noisy, incomplete, or biased. We construct our pre-training dataset from MS-COCO (Lin et al., 2014; Chen et al., 2015) and Visual Genome (Krishna et al., 2017), where each caption contains images or bounding boxes within them, which provide rich and diverse information for the model to learn object grounding and localization. Given the dataset, we pre-train our model with prefix language modeling (PrefixLM) and masked language modeling (MaskedLM) objectives, which encourage the model to generate natural language from images and fill in the missing words in captions, respectively; and a discriminative objective, which encourages the model to distinguish between correct and incorrect captions for the same image.

We test our GRILL on 7 zero-/few-shot vision-language tasks including Visual Commonsense Reasoning (VCR) (Zellers et al., 2019), RefCOCOg (Mao et al., 2016), Flickr30k-entities (Plummer et al., 2015), NLVR2 (Suhr et al., 2019), SNLI-VE (Xie et al., 2019), visual question answering (Goyal et al., 2017), and Flickr30k captioning (Young et al., 2014). We observe that our model demonstrates better zero-/few-shot generalization on diverse tasks compared to baselines. We also notice that the discriminative objective and hybrid sequences in pre-training are vital for better zero-/few-shot performance.

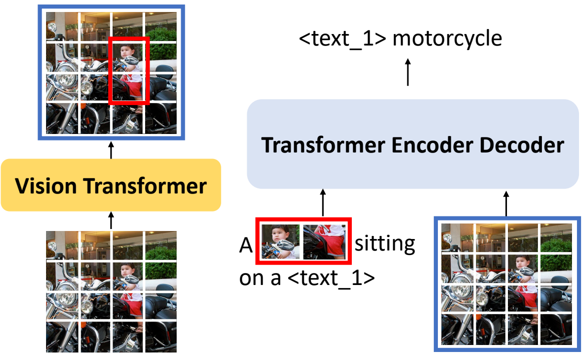


Figure 2: **Illustration of GRILL.** Our model is a sequence-to-sequence transformer that uses a vision transformer (ViT) (Dosovitskiy et al., 2021; Liu et al., 2021) to process images with patch embeddings, where each patch represents a fixed-size region of the image. We replace the referring words with the corresponding visual patches.

2 Generalization to Diverse Grounded Visual-language Tasks

Various VL tasks require phrase and object grounding and their task formats are different, which makes few-shot models challenging to generalize. In this work, we introduce a model that can generalize to grounded VL tasks with *no* or *a few* labeled examples. We first introduce the background, formal problem definition, and challenges.

2.1 Background: Visual Grounding

Visual grounding refers to the ability to link linguistic concepts (sentences, phrases, or words) to visual concepts (images and regions) (Chandu et al., 2021). Here we consider two types of visual grounding: image grounding and object grounding.

Image grounding refers to the linking of textual concepts to image concepts (Chandu et al., 2021). In this work, we consider image grounding as linking any type of text including sentences, phrases, and words to an entire image (e.g., image captioning, and image retrieval). Given an image and a corresponding caption, *object grounding* aims to localize objects in the image as mentioned by a noun phrase in the caption (or the entire caption sentence). Such object grounding occurs at word, phrase, and sentence levels in the language modality. Many VL tasks require object grounding implicitly or explicitly and we consider tasks that explicitly require localization as object grounding tasks. Referring expression comprehension (RefCOCOg (Mao et al., 2016)), phrase grounding (Flickr30k-entities (Plummer et al., 2015)), and visual commonsense reasoning (Zellers et al., 2019) are examples of localization tasks

2.2 Problem Formulation

In this work, we re-formulate the widely used pre-training task for image-caption datasets such that each caption may have one or more images including bounding boxes or regions in itself as a part of the text, denoted by $(T, \{V_j\}^N)$, in addition to the associated images. Note that some captions may not have images in themselves, $N = 0$. We refer to learning on the captions with images *grounded* learning. For pre-training, a VL model is pre-trained on image-caption datasets where captions include images or bounding boxes. For zero-shot tasks, the pre-trained model $\mathcal{L}$ cannot access training data $\mathcal{D}_{train}$ and validation data $\mathcal{D}_{val}$. We directly evaluate the model on the test data $\mathcal{D}_{test}$. For few-shot tasks, the model has access to K instances of training data for fine-tuning. For hyper-parameter tuning and model selection, we assume validation data $\mathcal{D}_{val}$ which has an equal number of instances to $\mathcal{D}_{train}$ to simulate a real-world low-resource environment and compose the validation data from training data. The sizes of $\mathcal{D}_{train}$ and $\mathcal{D}_{val}$ are 32 in our study.

Challenges Our goal is to pre-train a VL model that seamlessly transfers to various tasks not limited to visual commonsense reasoning, referring expression comprehension, and phrase grounding in a zero-shot or few-shot manner. Different tasks, especially grounding tasks, have different input and output formats, and thus the main challenge of this work is to generalize the zero-/few-shot ability to diverse tasks. Existing works on grounding tasks introduce special representations to depict regions such as special tokens (Cho et al., 2021) or object representations by an object detector (Su et al., 2020; Chen et al., 2019). While these works perform well on grounding tasks via expensive fine-tuning on labeled data, they have to design different object representations for different task formats. This makes it difficult to generalize to new tasks in a zero-shot fashion. For example, the object representations from an object detector are difficult to transfer to a task that refers to multiple images such as NLVR2 (Suhr et al., 2019). In this work, we tackle these challenges by introducing patch embeddings to represent objects, regions, and images; and pre-training our model with grounded sequences that contain captions and multiple images per caption.

3 Pre-training for Better Cross-Task Generalization

In this section, we introduce GRILL, a few-shot VL model for jointly learning contextualized representations from vision and language tasks. We first present an overview of GRILL (§3.1), our model architecture (§3.2), pre-training objectives (§3.3), and pre-training data (§3.4) in this section.

3.1 Overview

We propose GRILL, a new VL model that can learn object grounding and localization in pre-training and generalize to a wide range of VL tasks including grounding tasks in a zero-/few-shot manner. Our model is a sequence-to-sequence transformer (Vaswani et al., 2017) and takes a *hybrid sequence*, denoted by $(I, T, \{V_j\}^N)$, consisting of text T, an image I and visual concepts or regions $\{V_j\}^N$ as input and the output is a text sequence. We represent an input image with image patches by vision transformer (Dosovitskiy et al., 2021; Liu et al., 2021) and represent a region that corresponds to a set of patches by aggregating information among the patches. Given sequences with paired text outputs, we pre-train our model with prefix language modeling, masked language modeling, and a discriminative objective. We discuss how we create the hybrid sequences from image-caption datasets in §3.4.

3.2 Model Architecture

We adopt an encoder-decoder architecture (Vaswani et al., 2017) to encode visual and text inputs to generate target text. We represent an input image with a sequence of image patches by a vision transformer (Dosovitskiy et al., 2021; Liu et al., 2021). We adopt Swin transformer (Liu et al., 2021) as our vision transformer. It first splits an image into non-overlapping patches and linearly embeds all patches. Then, these patches are passed to the transformer layers, yielding $\{v_1, ..., v_m\}$. For an image of resolution of 224×224 and patch size of 32×32, we have $m = 49$. We assume that v_i encodes the information of the corresponding patch p_i. Therefore, we represent a visual concept (object or region) V_j that corresponds to a set of patches by aggregating information among the patches as shown in Figure 2. In addition, the entire patch representations are fed into the encoder by appending them to the text to encode the whole image. We train the model parameters θ

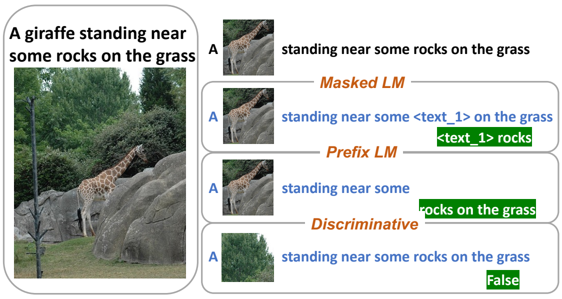

A giraffe standing near some rocks on the grass
A
standing near some rocks on the grass
Masked LM
A
standing near some <text_1> on the grass
<text_1> rocks
Prefix LM
A
standing near some
rocks on the grass
Discriminative
A
standing near some rocks on the grass
False

Figure 3: **Pre-training objectives.** We illustrate our pre-training objectives. We include masked language modeling, prefix language modeling, and the discriminative objective as our pre-training objectives.

by minimizing the negative log-likelihood of target text y tokens given input text x and image v:

$$L_\theta = -\sum_{i=1}^{|y|} \log P_\theta(y_i|y_{<i}, x, v). \quad (1)$$

3.3 Pre-training Objectives

Given images and captions with regions, we pre-train the models with prefix language modeling (PrefixLM), masked language modeling (MaskedLM), and a discriminative objective. We aim to learn grounding and localization through pre-training. Note that our model takes an image and text with regions as inputs and generates target text. In addition to the hybrid sequences, we also include raw text and raw images as our pre-training data. Fig. 3 illustrates the pre-training objectives.

Prefix language modeling. We include prefix language modeling (PrefixLM) following (Raffel et al., 2020; Jin et al., 2021). The objective randomly splits the text with regions input into two separate sequences. The first part may contain regions and is used as an input with an image to the encoder, and the second part does not contain regions and is used as target text to be generated by the decoder. The target text is not allowed to have region representations since our model generates text only

Masked language modeling. Masked language modeling (Cho et al., 2021; Jin et al., 2021) is to mask out random spans with numbered sentinel tokens, e.g., `<text_1>`, and then the masked sequence is fed into the encoder. Then the decoder generates the masked spans as target text. We randomly mask 15% of input text tokens and replace them with sentinel tokens. Note that the input sequence may include region representations in addition to a paired image and the region representations are not allowed to be masked.

Discriminative objective. This discriminative objective is important so that our model can do classification tasks where it has to determine whether the given sequence is correct or not. Thus, we pre-train GRILL with the discriminative objective and the model generates target texts, "true" for positive pairs and "false" for negative pairs. We consider an image and its captions with associated regions (if any) as positive pairs. With a probability of 50%, we replace the referring words with random region representations from the given image and treat this as a negative pair. The negative samples let the model learn the correct bindings of referring words and corresponding regions. For the raw text and raw image pairs, we randomly sample another training caption to create a negative pair.

3.4 Pre-training Data

To pre-train GRILL, we collect image-caption data from MS COCO (Lin et al., 2014; Chen et al., 2015) and Visual Genome (VG) (Krishna et al., 2017). The pre-training datasets contain 9.18M image-text pairs and 180K distinct images. From the image-caption pairs, we create our dataset for grounded image-caption pre-training, where each caption may have one or more region representations. We also include raw text and raw images as our pre-training data. To obtain captions with region representations, we introduce object-word alignments representing correspondence between words and objects. Given the object-word alignments, we replace words in a caption with a corresponding region representation so that the caption has a region representation as a substitute for the aligned word. With the object-word alignments, we prepare our dataset for pre-training. In addition, we include region descriptions and the aligned regions as hybrid sequences in Visual Genome for our pre-training.

3.4.1 Object-word Alignments

Given image-caption pairs, the process of getting object-word alignments consists of three steps: (1) object detection on images, (2) object tag-word matching, and (3) object-word alignments. We illustrate the process in Fig. 4.

Object detection on images. The first step is to detect objects in images and tags for the objects. We use the state-of-the-art object detector (Zhang et al., 2021) to get object bounding boxes and tags, yielding $\{(V_1, l_1), ..., (V_m, l_m)\}$ where V_i is a bounding box and l_i is a tag for the box. Given the set of tags $\{l_1, ..., l_m\}$, we find correspondence between

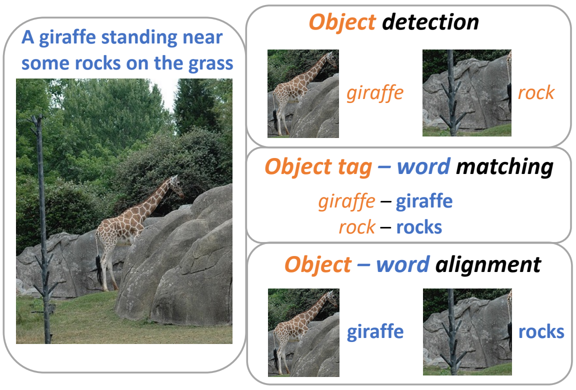


Figure 4: **Object-word alignments.**To create pre-training data, we use object-word alignments which is to replace referring words with corresponding bounding boxes.

the tags and words $\{w_1, ..., w_n\}$ in a caption in the next step.

Object tag-word matching. The second step is basically to find similar words between tags $\{l_1, ..., l_m\}$ and words $\{w_1, ..., w_n\}$. We define rules to find similar words as follows:

- Exact token matching
- Plural - Singular exact token matching
- Word vector similarity (Mikolov et al., 2013)
- WordNet Synonyms (Miller, 1995)

If one of the rules is satisfied, then we mark them as aligned tags and words $\{(l_i, w_j)\}$. Note that a word can be matched to multiple tags.

Object-word alignments. In the last step, we find alignments between object bounding boxes and words $\{(o_i, w_j)\}$ given the alignments between tags and words $\{(l_i, w_j)\}$ and an object list $\{(o_1, l_1), ..., (o_m, l_m)\}$. We simply find the object-word alignments since each tag is mapped to each bounding box, yielding $\{(o_i, l_i, w_j)\}$. However, note that some object bounding boxes share the same object tag; thus the alignments can include noisy correspondence between object boxes and words. We run CLIP (Radford et al., 2021) between aligned words and objects to find the most plausible alignment.

4 Experiments

4.1 Experiment Details

For pre-training, we use 1,280 batch size for GRILL and set learning rate 1e-4 with 5% linear warmup in pre-training. For the few-shot setting, we train models with 100 epochs and learning rate 1e-4, and choose the best checkpoint on the valid set. The model size of GRILL is 310M parameters. For baselines, we use their official codes to get zero-shot and few-shot performance.

4.2 Evaluation Setup

To evaluate few-shot performance, we randomly sample 5 different training and dev splits and measure the average performance on the 5 splits. We fine-tune the vision-language models with 100 epochs for the few-shot setup and choose the best checkpoint on the dev set. We report the model performance on the test set for RefCOCOg, NLVR2, Flickr30k-entities, SNLI-VE, and Flickr30k captioning (Karpathy split (Karpathy and Li, 2015)), and the validation set for VCR and VQAv2. We adopt accuracy for VCR, RefCOCOg, SNLI-VE, NLVR2, and VQA datasets; Recall@1,5,10 for Flickr30k-entities; and CIDEr (Vedantam et al., 2015) as evaluation metrics for captioning.

4.3 Baselines

For baselines, we include existing VL models: UNITER$_{large}$ (Chen et al., 2019), VL-T5 (Cho et al., 2021), GLIP-L (Li et al., 2022; Zhang et al., 2022), MDETR-ENB3 (Kamath et al., 2021); and few-shot VL models: FewVLM (Jin et al., 2021), Flamingo (Alayrac et al., 2022), and CPT (Yao et al., 2021). We exclude VQA datasets for VL-T5 for fair comparisons and pre-train the model using their code. Parameter sizes of each model are 303M for UNITER$_{large}$, 224M for VL-T5, 231M for GLIP-L, 152M for MDETR, 224M and 740M for FewVLM$_{base}$ and FewVLM$_{large}$, 3B and 80B for Flamingo, and 113M for CPT.

4.4 Downstream Tasks and Datasets

In this section, we compare our GRILL on a diverse set of 7 downstream tasks. We mainly focus on tasks that require phrase/object grounding: Visual Commonsense Reasoning, referring expression comprehension, and phrase grounding. Additionally, we evaluate our model on SNLI-VE, VQA, and captioning. VQA and captioning require generation for our method, while other datasets are classification tasks.

Visual Commonsense Reasoning (VCR) Visual Commonsense Reasoning (VCR) (Zellers et al., 2019) is a multiple-choice question-answering task that requires commonsense reasoning between objects in images. The task is decomposed into two sub-tasks, question answering (Q → A) and rationale prediction (QA → R). In the holistic setting (Q → AR), models have to predict answers and rationales. Following VL-T5 (Cho et al., 2021), we rank the choices with $P(\text{true})/(P(\text{true}) + P(\text{false}))$.

Method	Size	VCR			RefCOCOg	Flickr30k-entities			NLVR2	SNLI-VE	VQAv2	Flickr30k
		Q → A	QA → R	Q → AR	Acc	R@1	R@5	R@10	Acc	Acc	Acc	CIDEr
Random	-	25.0	25.0	6.3	19.0	6.5	27.7	47.8	50.0	33.3	0.0	-
UNITER$_{large}$	303M	32.6	26.1	8.7	10.0	-	-	-	49.1	17.9	0.0	-
VL-T5	224M	28.2	27.5	8.2	0.0	0.0	0.0	1.1	48.7	-	13.5	4.4
FewVLM$_{base}$	224M	25.9	25.4	6.5	0.0	0.0	0.0	0.0	50.6	-	43.4	31.0
FewVLM$_{large}$	740M	27.0	26.1	7.4	0.0	0.0	0.0	0.0	51.2	-	**47.7**	**36.5**
GRILL	310M	**40.6**	**39.3**	**16.2**	**47.5**	**18.9**	**53.4**	**70.3**	**56.1**	**46.9**	42.3	25.6

Table 1: **Zero-shot results.** We report performance on downstream tasks without any training data. Our model surpasses all baselines on classification tasks.

Method	Size	VCR			RefCOCOg	Flickr30k-entities			NLVR2	SNLI-VE	VQAv2	Flickr30k
		Q → A	QA → R	Q → AR	Acc	R@1	R@5	R@10	Acc	Acc	Acc	CIDEr
Random	-	25.0	25.0	6.3	19.0	6.5	27.7	47.8	50.0	33.3	0.0	-
UNITER$_{large}$	303M	29.1	28.6	8.4	45.4	-	-	-	**58.5**	40.7	24.2	-
VL-T5	224M	29.7	28.0	8.7	**56.9**	**48.8**	**72.8**	**78.6**	48.7	-	35.6	18.9
FewVLM$_{base}$	224M	29.1	28.4	8.5	16.0	4.2	18.7	31.7	50.3	-	47.8	37.5
FewVLM$_{large}$	740M	30.0	30.1	9.3	17.4	5.1	22.7	38.0	51.3	-	**52.3**	**38.4**
GRILL	310M	**41.1**	**40.4**	**16.7**	48.1	25.4	61.3	76.0	56.2	**48.4**	46.8	37.1

Table 2: **Few-shot results.** We report performance on downstream tasks with 32 labeled examples for fine-tuning.

and choose the one with the highest score. VCR provides bounding boxes around entities, with explicit groundings between those entities and references in questions.

Referring Expression Comprehension Referring expression comprehension is to localize an object given a referring expression. We adopt the RefCOCOg dataset (Mao et al., 2016) for this task. We present a referring phrase and candidate regions from the image to our model; and our model finds the most plausible region to the given phrase by ranking the regions with $P(\text{true})/(P(\text{true}) + P(\text{false}))$. Following VL-T5 (Cho et al., 2021), we use Mask R-CNN (Anderson et al., 2018) to find region detections as candidates for inference. We consider the selected region to be correct if its intersection over union (IoU) with the ground truth region is greater than 0.5. The upper bound performance on the test set by the Mask R-CNN is 86.09%. We get the performance of the random predictor by randomly choosing the bounding box from the object detector.

Phrase Grounding Given one or more phrases, phrase grounding is to provide a set of bounding boxes for each phrase. We use the Flickr30k-entities dataset (Plummer et al., 2015) for this task. Following BAN (Kim et al., 2018) and VisualBERT (Li et al., 2019), we adopt Faster R-CNN (Ren et al., 2015) pre-trained on Visual Genome to detect regions as candidates for inference. The predicted region is correct if its intersection over union (IoU) with the ground-truth region is greater than 0.5. The upper bound performance on the test set by the Faster R-CNN is 87.45%. Similar to Refcoco, we pass a referring phrase and candidate regions from the image to our model; and our model finds the most plausible region to the given phrase by ranking the regions with $P(\text{true})/(P(\text{true}) + P(\text{false}))$. We use the any-box-protocol from MDETR (Kamath et al., 2021).

NLVR2 The task of NLVR2 (Suhr et al., 2019) is to determine whether a text description is true given two images. The task requires understanding two images and comparing them. To apply our model to this task, we create one image by concatenating the two images, and then our model generates text labels “true” and “false” for inference.

Visual Entailment Visual entailment, SNLI-VE (Xie et al., 2019) is to determine whether the image semantically entails the text given an image-sentence pair. The task is a 3-way classification where labels are “entailment”, “neutral”, and “contradiction.” We define label words for the classification as “entailment”: “true”, “neutral”: “maybe”, “contradiction”: “false.” We choose the classification label by measuring the probability of each word and picking the highest one.

Visual Question Answering The visual question answering task (Goyal et al., 2017) requires models to answer a question to a given context image. We approach the visual question answering task as a generation task so that the model can produce the answers without introducing any task-specific heads as in (Jin et al., 2021; Cho et al., 2021). We adopt the following input prompt, “*question:* {ques-

Method	**Size**	**RefCOCOg**		**Flickr30k-entities**	
		0	32	0	32
Random	-	19.0	19.0	6.5	6.5
UNITER$_{large}$ (Chen et al., 2019)	303M	10.0	45.4	-	-
VL-T5 (Cho et al., 2021)	224M	0.0	**56.9**	0.0	**48.8**
FewVLM$_{large}$ (Jin et al., 2021)	740M	0.0	17.4	0.0	5.1
CPT (Yao et al., 2021) (Yao et al., 2021)	113M	36.5	-	-	-
MDETR-ENB3 (Kamath et al., 2021)	152M	**54.0**†	-	84.8‡	-
GLIP-L (Li et al., 2022; Zhang et al., 2022)	231M	-	-	**87.1**‡	-
GRILL	310M	47.5	48.1	18.9	25.4

Table 3: **Results on RefCOCOg and Flickr30k-entities with 0 and 32 examples.** We report recall@1 for Flickr30k-entities. †This model used the RefCOCOg dataset in the pre-training. ‡These models used the Flickr30k-entities dataset in the pre-training while ours did not.

Model	**size**	**0-shot**	**32-shot**
Random	-	0.0	0.0
UNITER$_{large}$ (Chen et al., 2019)	303M	0.0	24.2
VL-T5 (Cho et al., 2021)	224M	13.5	35.6
FewVLM$_{large}$ (Jin et al., 2021)	740M	47.7	52.3
Flamingo-3B (Alayrac et al., 2022)	3B	49.2	57.1
Flamingo-80B	80B	**56.3**	**67.6**
GRILL	310M	42.3	46.8

Table 4: **VQA results with 0 and 32 examples.** We report zero-/32-shot performance on the VQAv2 dataset. Flamingo has 3B or 80B parameters and uses in-context examples for inference while our model has 310M parameters and use the examples for fine-tuning.

tion} *answer:* `<text_1>`," where `<text_1>` is a sentinel token, from (Jin et al., 2021) for the generation.

Captioning The captioning task is to generate a caption given an image. In Flickr30k (Young et al., 2014), we use "*an image of*" as our input prompt from (Jin et al., 2021).

4.5 Results

Zero-shot performance. We evaluate the existing models in a zero-shot manner, where models do not have access to any training data. Table 1 shows the performance on diverse tasks. Note that VCR, RefCOCOg, NLVR2, Flickr30k-entities require phrase or region grounding. Firstly, GRILL shows the best performance on all the grounding tasks while baselines show worse performance than the random predictor on many of the grounding tasks. This suggests that competitors have difficulty generalizing to grounding tasks that need phrase or region grounding in a zero-shot way. On Table 3, we additionally include baselines, GLIP-L and MDETR-ENB3, that are targeted for grounding tasks. These models include the corresponding task-specific datasets in pre-training so they demonstrate great performance without additional fine-tuning. Note that we do not include task-specific datasets in the pre-training. Our method also exhibits decent performance on VQAv2 and Flickr30k captioning. By comparing Flamingo, a 3B or 80B-sized vision-language model, our model demonstrates good accuracy considering our model size. On Flickr30 captioning, our model underperforms FewVLM$_{base}$ which is a bit smaller model than ours.

Few-shot performance. We observe interesting results on few-shot performance (Table 2). We use 32 labeled examples in total for fine-tuning. While our model, GRILL, improves the performance on all the tasks, baseline methods outperforms our model on RefCOCOg and Flickr30k-entities unlike zero-shot results. We conjecture that baseline methods include the phrase grounding task in their pre-training, so they achieves good performance. However, the models still struggle on the VCR task, while our model surpasses the models on the task. Interestingly, our model achieves the comparable result to FewVLM on the Flickr30k captioning on the few-shot setup.

4.6 Ablations

Here, we study ablations for our method. Table 5 and Fig. 5 show the ablations on the pre-training objectives and hybrid sequences, and different input formats during inference on the zero-shot setup, respectively.

Pre-training objectives and hybrid sequences. Firstly, we study the ablation of pre-training objectives and hybrid sequences in pre-training. On Table 5. our model without hybrid sequences affects the performance a lot on many tasks. Specifically, results on RefCOCOg and Flickr30k-entities are significantly degraded suggesting that hybrid sequences in pre-training play a vital role in improving grounding. Among pre-training objectives in GRILL, we notice that the discriminative objective is important for the results. We conjecture that the tasks in the table are classification tasks so the discriminative objective is the most useful for the tasks.

Input formats for inference. We investigate the input formats during inference on Fig. 5. On VCR, we replace the referring words with bounding boxes for text input (hybrid sequences), or we do not replace them and use original text input (original sequences). On NLVR2, we replace the "left" word with the left image and the "right" word with the right image (hybrid sequences), or we do not replace them and use the original text input (original).

Model	VCR	Ref-COCOg	NLVR2	Flickr30k-entities
Zero-shot				
GRILL	16.2	47.5	56.1	18.9
No hybrid sequences	12.9	18.9	55.7	5.7
No discriminative	6.8	30.5	50.4	12.7
No PrefixLM	14.4	48.5	55.8	18.5
No MLM	15.6	47.8	56.0	19.3
32-shot				
GRILL	16.7	47.4	56.2	25.4
No hybrid sequences	14.3	16.3	55.9	18.7
No discriminative	7.2	42.0	50.5	15.3
No PrefixLM	14.7	48.7	55.9	21.9
No MLM	16.3	47.9	56.1	23.5

Table 5: **Ablations on the pre-training objectives and hybrid sequences in pre-training.** We report Q → AR for VCR, and R@1 for Flick30k-entities.

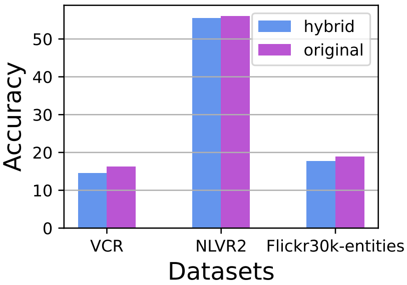


Figure 5: **Performance with different input formats for inference on the zero-shot setup.** We report Q → AR for VCR, and R@1 for Flick30k-entities.

On Flickr30k-entities, we do the same thing as on VCR (hybrid sequences), or we use the referring words and bounding boxes for model input (original). Counter-intuitively, we observe that our model with original input formats during inference. shows better performance on all the datasets. We conjecture that introducing the grounding information may disturb the model predictions since the model needs to judge whether the grounding information is correct or not. We leave the sophisticated design for future work.

5 Related Work

Vision-language few-shot learning. There have been attempts to address the challenge of data-hungry supervised learning in vision-language domains: FewVLM (Jin et al., 2021), Frozen (Tsimpoukelli et al., 2021), Flamingo (Alayrac et al., 2022), GLIP (Li et al., 2022; Zhang et al., 2022), FewVLM (Jin et al., 2021) improves the few-shot performance of VQA and captioning by prompting the model and its performance is on par with large few-shot learners. Frozen (Tsimpoukelli et al., 2021) adapts a few-shot language model (Radford et al., 2019) to vision-language tasks with soft prompting for images. Flamingo (Alayrac et al., 2022) achieves state-of-the-art results on few-shot VQA and captioning tasks by prompting the model with task-specific examples. While these models achieve improvement on few-shot tasks, they are not applicable to grounding tasks. Lastly, GLIP (Li et al., 2022; Zhang et al., 2022) unifies object detection and phrase grounding and it achieves great performance on zero-shot object detection and phrase grounding tasks. Unlike our method, GLIP used grounding datasets including Flickr30k-entities in pre-training so it achieved great performance on the phrase grounding without fine-tuning. Our method is not applicable to object detection since it requires bounding box regression. We leave this extension for future work.

Grounded Vision-language Learning. Grounded vision-language learning has been explored to learn grounding between objects in images and phrases in sentence (Li et al., 2020; Zhang et al., 2021; Kamath et al., 2021; Li et al., 2022; Zhang et al., 2022). MDETR is a modulated detector that detects objects in an image conditioned on a raw text query (Kamath et al., 2021). The model exhibits remarkable results on object detection, phrase grounding, and referring expression comprehension by pre-training the model on object detection data. GLIP followed a similar direction and it unifies object detection and phrase grounding (Li et al., 2022; Zhang et al., 2022). While the methods rely on object detection datasets to improve grounding, our method utilizes grounded sequences from image-caption datasets and an object Our model does not only work on grounding tasks but also on visual question answering and captioning tasks.

6 Conclusion

In this work, we proposed GRILL, a new VL model that can learn object grounding and localization during pre-training and generalize to a variety of VL tasks including grounding tasks. Our model is a sequence-to-sequence transformer model that uses a vision transformer for versatile image processing on zero-shot tasks. To pre-train our model, we introduced our dataset using object-word alignments and pre-train it with masked language modeling, prefix language modeling, and the discriminative objective. On the empirical analysis, we observed that our model demonstrated good zero-/few-shot generalization on diverse tasks. We also observed that the discriminative objective and hybrid sequences in pre-training were vital for better zero-/few-shot performance.

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, et al. 2022. Flamingo: a visual language model for few-shot learning. *arXiv preprint arXiv:2204.14198*.

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. 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*, pages 6077–6086.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Khyathi Raghavi Chandu, Yonatan Bisk, and Alan W Black. 2021. Grounding'grounding'in nlp. *arXiv preprint arXiv:2106.02192*.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. 2015. Microsoft coco captions: Data collection and evaluation server. *ArXiv preprint*, abs/1504.00325.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2019. Uniter: Learning universal image-text representations.

Jaemin Cho, Jie Lei, Hao Tan, and Mohit Bansal. 2021. Unifying vision-and-language tasks via text generation. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 1931–1942. PMLR.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 6325–6334. IEEE Computer Society.

Woojeong Jin, Yu Cheng, Yelong Shen, Weizhu Chen, and Xiang Ren. 2021. A good prompt is worth millions of parameters? low-resource prompt-based learning for vision-language models. *arXiv preprint arXiv:2110.08484*.

Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. 2021. Mdetr-modulated detection for end-to-end multi-modal understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1780–1790.

Andrej Karpathy and Fei-Fei Li. 2015. Deep visual-semantic alignments for generating image descriptions. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 3128–3137. IEEE Computer Society.

Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. 2018. Bilinear attention networks. *Advances in neural information processing systems*, 31.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. Visualbert: A simple and performant baseline for vision and language. *ArXiv preprint*, abs/1908.03557.

Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. 2022. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10965–10975.

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. 2020. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*, pages 121–137. Springer.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022.

Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. 2016. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 11–20.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26.

George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67.

Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. 2015. Faster R-CNN: towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 91–99.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, et al. 2021. Multitask prompted training enables zero-shot task generalization. *arXiv preprint arXiv:2110.08207*.

Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2020. VL-BERT: pre-training of generic visual-linguistic representations. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. 2019. A corpus for reasoning about natural language grounded in photographs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6418–6428, Florence, Italy. Association for Computational Linguistics.

Maria Tsimpoukelli, Jacob Menick, Serkan Cabi, SM Eslami, Oriol Vinyals, and Felix Hill. 2021. Multimodal few-shot learning with frozen language models. *ArXiv preprint*, abs/2106.13884.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 4566–4575. IEEE Computer Society.

Ning Xie, Farley Lai, Derek Doran, and Asim Kadav. 2019. Visual entailment: A novel task for fine-grained image understanding. *arXiv preprint arXiv:1901.06706*.

Yuan Yao, Ao Zhang, Zhengyan Zhang, Zhiyuan Liu, Tat-Seng Chua, and Maosong Sun. 2021. Cpt: Colorful prompt tuning for pre-trained vision-language models. *arXiv preprint arXiv:2109.11797*.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78.

Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. From recognition to cognition: Visual commonsense reasoning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 6720–6731. Computer Vision Foundation / IEEE.

Haotian Zhang, Pengchuan Zhang, Xiaowei Hu, Yen-Chun Chen, Liunian Harold Li, Xiyang Dai, Lijuan Wang, Lu Yuan, Jenq-Neng Hwang, and Jianfeng Gao. 2022. Glipv2: Unifying localization and vision-language understanding. *arXiv preprint arXiv:2206.05836*.

Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. 2021. Vinvl: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5579–5588.