

# OFA: UNIFYING ARCHITECTURES, TASKS, AND MODALITIES THROUGH A SIMPLE SEQUENCE-TO-SEQUENCE LEARNING FRAMEWORK

Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai

Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, Hongxia Yang

DAMO Academy, Alibaba Group \*

{zheluo.wp, ya235025, menrui.mr, junyang.ljy, baishuai.bs, zhikang.lzk, jason.mjx, ericzhou.zc, jingren.zhou, yang.yhx}@alibaba-inc.com

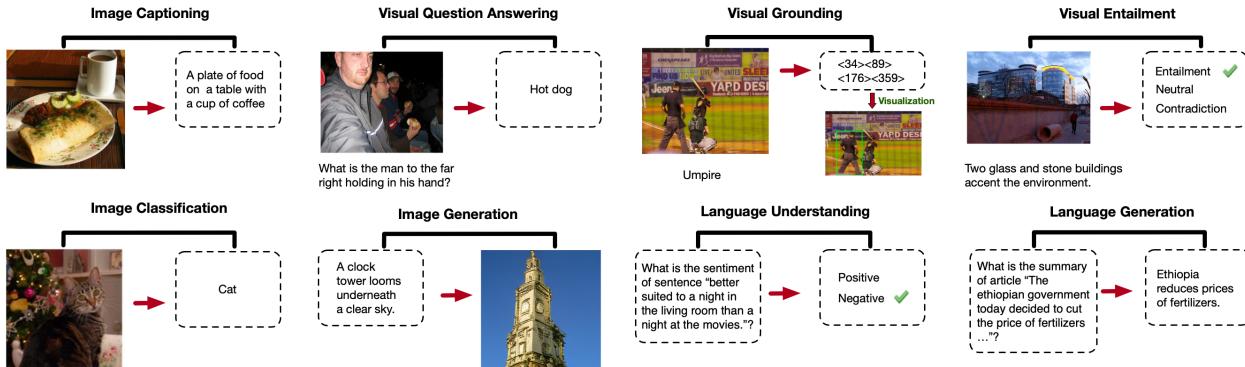


Figure 1: Examples of various tasks supported by OFA.

## ABSTRACT

In this work, we pursue a unified paradigm for multimodal pretraining to break the scaffolds of complex task/modality-specific customization. We propose OFA, a Task-Agnostic and Modality-Agnostic framework that supports Task Comprehensiveness. OFA unifies a diverse set of cross-modal and unimodal tasks, including image generation, visual grounding, image captioning, image classification, language modeling, etc., in a simple sequence-to-sequence learning framework. OFA follows the instruction-based learning in both pretraining and finetuning stages, requiring no extra task-specific layers for downstream tasks. In comparison with the recent state-of-the-art vision & language models that rely on extremely large cross-modal datasets, OFA is pretrained on only 20M publicly available image-text pairs. Despite its simplicity and relatively small-scale training data, OFA achieves new SOTAs in a series of cross-modal tasks while attaining highly competitive performances on uni-modal tasks. Our further analysis indicates that OFA can also effectively transfer to unseen tasks and unseen domains. Our code and models are publicly available at <https://github.com/OFA-Sys/OFA>.

**Keywords** Unified frameworks · Multimodal pretraining · Multitask learning · Zero-shot learning

\*Correspondence to: Chang Zhou<ericzhou.zc@alibaba-inc.com>.

## 1 Introduction

Building an omnipotent model that handles as many tasks and modalities as human beings is an attractive goal in the AI community. The possibilities of achieving this goal may largely depend on whether massive varieties of modalities, tasks and training regimes can be represented with only a few forms that can be unified and managed by a single model or system.

Recent developments of the Transformer [1] architecture have shown its potential for being a universal computation engine [2, 3, 4, 5, 6, 7, 8]. In the settings of supervised learning, the “pretrain-finetune” paradigm achieves excellent success in many domains. In the regimes of few-/zero-shot learning, language models with prompt / instruction tuning prove powerful zero-/few-shot learners [3, 9, 10]. These advances have provided more significant than ever opportunities for the emergence of an omni-model.

To support better generalization for open-ended problems while maintaining multitask performance and ease of use, we advocate that an omnipotent model should have the following three properties: 1. Task-Agnostic (TA): unified task representation to support different types of tasks, including classification, generation, self-supervised pretext tasks, etc., and to be agnostic to either pretraining or finetuning. 2. Modality-Agnostic (MA): unified input and output representation shared among all tasks to handle different modalities. 3. Task Comprehensiveness (TC): enough task variety to accumulate generalization ability robustly.

However, it is challenging to satisfy these properties while maintaining superior performance in downstream tasks. Current language and multimodal pretrained models readily fail at parts of these properties, due to their following design choices. 1. Extra learnable components for finetuning, e.g., task-specific heads [2], adapters [11], soft prompts [12]. This makes the model structure task-specific and poses discrepancy between pretraining and finetuning. Such designs are also not friendly to supporting unseen tasks in a zero-shot manner. 2. Task-specific formulation. For most current methods, pretraining, finetuning and zero-shot tasks usually differ in task formulation and training objectives. This violates TA and it is burdensome to scale up the task population to achieve TC. 3. Entangling modality representation with downstream tasks. It is a common practice for Vision-Language models to take the detected objects as part of the image input features [8, 13, 14, 15, 16, 17]. Though it demonstrates better downstream task performance on some closed-domain datasets, it depends on an extra object detector which usually fails at open-domain data.

Therefore, we explore an omni-model for multimodal pretraining and propose **OFA**, hopefully “One For All”, which achieves the objectives of unifying architectures, tasks, and modalities, and supports the three properties above.<sup>2</sup> We formulate both pretraining and finetuning tasks in a unified sequence-to-sequence abstraction via handcrafted instructions [9, 10] to achieve Task-Agnostic. A Transformer is adopted as the Modality-Agnostic compute engine, with a constraint that no learnable task- or modality-specific components will be added to downstream tasks. It is available to represent information from different modalities within a globally shared multimodal vocabulary across all tasks. We then support Task Comprehensiveness by pretraining on varieties of uni-modal and cross-modal tasks.

To summarize:

- We propose OFA, a Task-Agnostic and Modality-Agnostic framework that supports Task Comprehensiveness. OFA is the first attempt to unify the following vision & language, vision-only and language-only tasks, including understanding and generation, e.g., text-to-image generation, visual grounding, visual question answering (VQA), image captioning, image classification, language modeling, etc., via a simple sequence-to-sequence learning framework with a unified instruction-based task representation.
- OFA is pretrained on the publicly available datasets of 20M image-text pairs, in comparison with recent models that rely on paired data of a much larger scale [22, 23]. OFA achieves state-of-the-art performances in a series of vision & language downstream tasks, including image captioning, visual question answering, visual entailment, referring expression comprehension, etc.
- OFA, as a multimodal pretrained model, achieves comparable performances on unimodal tasks with SOTA pretrained models in language or vision, e.g., RoBERTa, ELECTRA and DeBERTa for natural language understanding, UniLM, Pegasus and ProphetNet for natural language generation, and MoCo-v3, BEiT and MAE for image classification.
- We verify that OFA achieves competitive performance in zero-shot learning. Also, it can transfer to unseen tasks with new task instructions and adapt to out-of-domain information without finetuning.

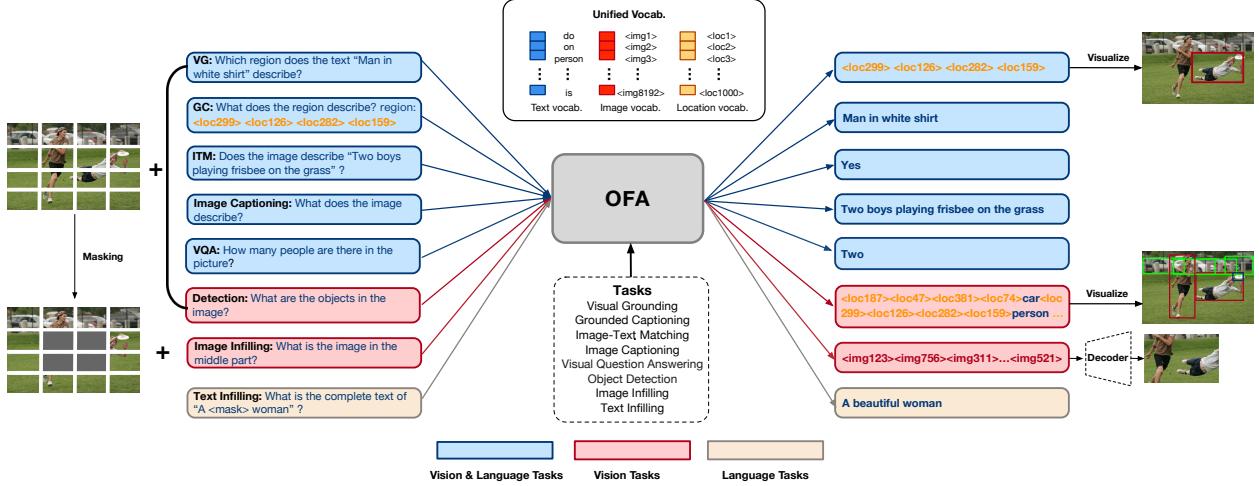


Figure 2: A demonstration of the pretraining tasks, including visual grounding, grounded captioning, image-text matching, image captioning, VQA, object detection, image infilling as well as text infilling.

## 2 Related Work

**Language Pretraining & Vision Pretraining** Natural language pretraining has revolutionized the whole NLP research community. A representation of this track is the birth of BERT [2] and GPT [24]. A number of studies have been progressively advancing pretraining by improving pretraining tasks and designing more sophisticated model architectures [25, 26, 27, 28, 29, 30, 31]. Having witnessed the success of natural language pretraining, researchers have promoted self-supervised learning (SSL) in computer vision [32, 33, 34, 35]. Recently, mirroring masked language modeling (MLM) in language pretraining, generative pretraining [36, 37] with ViT architecture [6] further boosts downstream performance.

**Multimodal Pretraining** Multimodal pretraining has been developing rapidly [38, 13, 39, 40, 14, 41, 42, 43, 44, 15, 16, 17, 45, 46, 47]. Researchers have applied the masking strategies and the encoder-decoder architecture to adapt models to generation tasks [15, 17, 18, 22]. Besides, to simplify preprocessing, patch projection has received attention and helped Transformer achieve SOTA performance in downstream tasks [22, 48]. To make full use of large-scale weakly supervised data, [49] trains a bi-encoder on 400 million pairs and demonstrates excellent performance in retrieval tasks. Another line of work is text-to-image synthesis. A bunch of works [50, 51, 18, 52] incorporate Transformer with VQVAE [53] or VQGAN [54] to generate high-quality images with high resolution. However, the previously mentioned methods are limited in processing a single type of data, such as cross-modal data only or limited in their capabilities. Also, the discrepancy between pretraining and finetuning behaviors limits the transferability to open-ended data.

**Unified Frameworks** To pursue the unified models, [55] demonstrate a uniform format to represent tasks. In NLP, recent studies unify diverse tasks covering natural language understanding and generation to text-to-text transfer [30] or language modeling [3]. Following this idea, [56] and [57] demonstrate text-generation-based multimodal pretrained models. [7] and [58] propose a simple framework that can process information from multiple modalities with a uniform byte-sequence representation. [59] and [60] unify tasks of different modalities by designing various task-specific layers. [61] explores to employ a retrieval-based unified paradigm. However, these multimodal pretrained models suffer from performance degradation in downstream tasks, e.g., VQA, image captioning, etc., and they have no image generation capability.

## 3 OFA

In this work, we propose OFA, a unified Seq2Seq framework for the unification of I/O & architectures, tasks, and modalities. The overall framework is illustrated in Figure 2.

<sup>2</sup>This work is the latest one of our M6 series [18, 19, 20, 21].

### 3.1 I/O & Architecture

**I/O** The most common practice of multimodal pretraining is the pretraining of Transformer models on image-text pair corpus at scale. This requires data preprocessing or modality-specific adaptors to enable the joint training of both visual and linguistic information with the Transformer architecture. Compared with the complex, resource&time-consuming object feature extraction, we aim for simplicity and directly use ResNet modules to convolve  $x_v \in \mathbb{R}^{H \times W \times C}$  to  $P$  patch features of the hidden size, following [62] and [22]. As to processing the linguistic information, we follow the practice of GPT [24] and BART [31] that we apply byte-pair encoding (BPE) [63] to the given text sequence to transform it into a subword sequence and then embed them to features.

To process different modalities without task-specific output schema, it is essential to represent data of various modalities in a unified space. A possible solution is to discretize text, image, and object and represent them with tokens in a unified vocabulary. Recent advances in image quantization [53, 54] have demonstrated effectiveness in text-to-image synthesis [50, 18, 51, 19], and thus we utilize this strategy for the target-side image representations. Sparse coding is effective in reducing the sequence length of image representation. For example, an image of the resolution of  $256 \times 256$  is represented as a code sequence of the length of  $16 \times 16$ . Each discrete code strongly correlates with the corresponding patch [36].

Apart from representing images, it is also essential to represent objects within images as there are a series of region-related tasks. Following [64], we represent objects as a sequence of discrete tokens. To be more specific, for each object, we extract its label and its bounding box. The continuous corner coordinates (the top left and the bottom right) of the bounding box are uniformly discretized to integers as location tokens  $\langle x_1, y_1, x_2, y_2 \rangle$ . As to the object labels, they are intrinsically words and thus can be represented with BPE tokens.

Finally, we use a unified vocabulary for all the linguistic and visual tokens, including subwords, image codes, and location tokens.

**Architecture** Following the previous successful practices in multimodal pretraining [14, 17, 22], we choose Transformer as the backbone architecture, and we adopt the encoder-decoder framework as the unified architecture for all the pretraining, finetuning, and zero-shot tasks. Specifically, both the encoder and the decoder are stacks of Transformer layers. A Transformer encoder layer consists of a self attention and a feed-forward network (FFN), while a Transformer decoder layer consists of a self attention, an FFN and a cross attention for building the connection between the decoder and the encoder output representations. To stabilize training and accelerate convergence, we add head scaling to self attention, a post-attention layer normalization (LN) [65], and an LN following the first layer of FFN [66]. For positional information, we use two absolute position embeddings for text and images, respectively. Instead of simply adding the position embeddings, we decoupling the position correlation from token embeddings and patch embeddings [67]. In addition, we also use 1D relative position bias for text [30] and 2D relative position bias for image [22, 62].

### 3.2 Tasks & Modalities

A unified framework is designed to provide architecture compatibility across different modalities and downstream tasks so that opportunities can arise to generalize to unseen tasks within the same model. Then we have to represent the possible downstream tasks concerning different modalities in a unified paradigm. Therefore, an essential point for the design of pretraining tasks is the consideration of multitask and multimodality.

To unify tasks and modalities, we design a unified sequence-to-sequence learning paradigm for pretraining, finetuning, and inference on all tasks concerning different modalities. Both pretraining tasks and downstream tasks of cross-modal and uni-modal understanding and generation are all formed as Seq2Seq generation. It is available to perform multitask pretraining on multimodal and uni-modal data to endow the model with comprehensive capabilities. Specifically, we share the identical schema across all tasks, while we specify handcrafted instructions for discrimination [9].

For cross-modal representation learning, we design 5 tasks, including visual grounding (VG), grounded captioning (GC), image-text matching (ITM), image captioning (IC), and visual question answering (VQA). For VG, the model learns to generate location tokens specifying the region position  $\langle x_1, y_1, x_2, y_2 \rangle$  based on the input of the image  $x^i$  and the instruction “Which region does the text  $x^t$  describe?” where  $x^t$  refers to the region caption. GC is an inverse task of VG. The model learns to generate a description based on the input image  $x^i$  and the instruction “What does the region describe? region:  $\langle x_1, y_1, x_2, y_2 \rangle$ ”. For ITM, we use each original image-text pair as the positive sample and construct a new one as the negative by pairing the image with a randomly substituted caption. The model learns to discriminate whether the given image and text are paired by learning to generate “Yes” or “No” based on the input image  $x^i$  and the instruction “Does the image describe  $x^t$ ?”. As to image captioning, this task can naturally adapt to the sequence-to-sequence format. The model learns to generate the caption based on the given image and the instruction

Table 1: Detailed hyperparameters of OFA model configuration. We list the configuration for OFA of 5 different sizes.

Model	#Param.	Backbone	Hidden size	Intermediate Size	#Head	#Enc. Layers	#Dec. Layers
OFA <sub>Tiny</sub>	33M	ResNet50	256	1024	4	4	4
OFA <sub>Medium</sub>	93M	ResNet101	512	2048	8	4	4
OFA <sub>Base</sub>	182M	ResNet101	768	3072	12	6	6
OFA <sub>Large</sub>	472M	ResNet152	1024	4096	16	12	12
OFA <sub>Huge</sub>	930M	ResNet152	1280	5120	16	24	12

“What does the image describe?”. For VQA, we send the image and the question as the input and require the model to learn to generate correct answers.

For uni-modal representation learning, we design 2 tasks for vision and 1 task for language, respectively. The model is pretrained with image infilling and object detection for vision representation learning. Recent advances in generative self-supervised learning for computer vision show that masked image modeling is an effective pretraining task [36, 37]. In practice, we mask the middle part of the images as the input. The model learns to generate the sparse codes for the central part of the image based on the corrupted input and the specified instruction “What is the image in the middle part?”. We additionally add object detection to pretraining following [44]. The model learns to generate human-annotated object representations, i.e., the sequence of object position and label, based on the input image and the text “What are the objects in the image?” as the instruction. Both tasks strengthen the representation learning on both pixel and object levels. For language representation learning, following the practice of [31], we pretrain the unified model on plain text data with text infilling.

In this way, we unify multiple modalities and multiple tasks to a single model and pretraining paradigm. OFA is pretrained jointly with those tasks and data. Thus, it can perform different tasks concerning natural language, vision, and cross-modality.

### 3.3 Pretraining Datasets

We construct pretraining datasets by incorporating Vision & Language data (i.e., image-text pairs), Vision data (i.e., raw image data, object-labeled data), and Language data (i.e., plain texts). For replication, we only use datasets that are publicly available. We carefully filter our pretraining data and exclude images that appear in the validation and test sets of downstream tasks to avoid data leakage. We provide more details about pretraining datasets in Appendix A.1.

### 3.4 Training & Inference

We optimize the model with the cross-entropy loss. Given an input  $x$ , an instruction  $s$  and an output  $y$ , we train OFA by minimizing  $\mathcal{L} = -\sum_{i=1}^{|y|} \log P_\theta(y_i | y_{<i}, x, s)$ , where  $\theta$  refers to the model parameters. For inference, we apply the decoding strategies, e.g., beam search, to enhance the quality of generation. However, this paradigm has several problems in classification tasks: 1. optimizing on the entire vocabulary is unnecessary and inefficient; 2. the model may generate invalid labels out of the closed label set during inference. To overcome these issues, we introduce a search strategy based on prefix tree (Trie, [68]). Experimental results show that the Trie-based search can enhance the performance of OFA on classification tasks. See Appendix B for more details.

### 3.5 Scaling Models

In order to investigate how OFA of different model sizes perform in downstream tasks, we have developed 5 versions of OFA models, scaling from 33M to 940M parameters, and we list their detailed hyperparameters in Table 1.

To be more specific, we have built basic models of Base and Large sizes, OFA<sub>Base</sub> and OFA<sub>Large</sub>. As our network configuration is similar to BART [31], their sizes are similar to those of BART<sub>Base</sub> and BART<sub>Large</sub>. Additionally, we have developed OFA of a larger size, which we name it OFA<sub>Huge</sub>, or OFA without specific mentioning in the tables. Its size is comparable to that of SimVLM<sub>Huge</sub> or ViT<sub>Huge</sub>. To investigate whether smaller OFA can still reach satisfactory performance, we have developed OFA<sub>Medium</sub> and OFA<sub>Tiny</sub>, which are solely around half and less than 20% as large as OFA<sub>Base</sub>.

Table 2: Experimental results on cross-modal understanding tasks including VQA and visual entailment. Note that we report the best results from the previous SOTAs, and specifically SimVLM is a huge-size model comparable to ViT-Huge pretrained on 1.8B image-text pairs, and Florence is built with CoSwin-H and RoBERTa and it is pretrained on 900M image-text pairs.

Model	VQA		SNLI-VE	
	test-dev	test-std	dev	test
UNITER [14]	73.8	74.0	79.4	79.4
OSCAR [15]	73.6	73.8	-	-
VILLA [16]	74.7	74.9	80.2	80.0
VL-T5 [56]	-	70.3	-	-
VinVL [17]	76.5	76.6	-	-
UNIMO [46]	75.0	75.3	81.1	80.6
ALBEF [69]	75.8	76.0	80.8	80.9
METER [70]	77.7	77.6	80.9	81.2
VLMo [48]	79.9	80.0	-	-
SimVLM [22]	80.0	80.3	86.2	86.3
Florence [23]	80.2	80.4	-	-
OFA <sub>Tiny</sub>	70.3	70.4	85.3	85.2
OFA <sub>Medium</sub>	75.4	75.5	86.6	87.0
OFA <sub>Base</sub>	78.0	78.1	89.3	89.2
OFA <sub>Large</sub>	80.3	80.5	90.3	90.2
OFA	<b>82.0</b>	<b>82.0</b>	<b>91.0</b>	<b>91.2</b>

Table 3: Experimental results on MSCOCO Image Captioning. We report the results on the Karpathy test split. Note that SimVLM and LEMON are huge-size models.

Model	Cross-Entropy Optimization				CIDEr Optimization			
	BLEU@4	METEOR	CIDEr	SPICE	BLEU@4	METEOR	CIDEr	SPICE
VL-T5 [56]	34.5	28.7	116.5	21.9	-	-	-	-
OSCAR [15]	37.4	30.7	127.8	23.5	41.7	30.6	140.0	24.5
UNICORN [57]	35.8	28.4	119.1	21.5	-	-	-	-
VinVL [17]	38.5	30.4	130.8	23.4	41.0	31.1	140.9	25.2
UNIMO [46]	39.6	-	127.7	-	-	-	-	-
LEMON [71]	41.5	30.8	139.1	24.1	42.6	31.4	145.5	25.5
SimVLM [22]	40.6	<b>33.7</b>	143.3	<b>25.4</b>	-	-	-	-
OFA <sub>Tiny</sub>	35.9	28.1	119.0	21.6	38.1	29.2	128.7	23.1
OFA <sub>Medium</sub>	39.1	30.0	130.4	23.2	41.4	30.8	140.7	24.8
OFA <sub>Base</sub>	41.0	30.9	138.2	24.2	42.8	31.7	146.7	25.8
OFA <sub>Large</sub>	42.4	31.5	142.2	24.5	43.6	32.2	150.7	26.2
OFA	<b>43.9</b>	31.8	<b>145.3</b>	24.8	<b>44.9</b>	<b>32.5</b>	<b>154.9</b>	<b>26.6</b>

## 4 Experiments

This section provides experimental details and analyses to demonstrate our model’s effectiveness. See Appendix A for implementation details.

### 4.1 Results on Cross-modal Tasks

We evaluate our models on different cross-modal downstream tasks, covering cross-modal understanding and generation. Specifically, we implement experiments on multimodal understanding datasets including VQAv2 for visual question answering and SNLI-VE [73] for visual entailment, and multimodal generation including MSCOCO Image Caption [74] for image captioning, RefCOCO / RefCOCO+ / RefCOCOg [75, 76] for referring expression comprehension as this

Table 4: Experimental results on the 3 datasets of referring expression comprehension, namely RefCOCO, RefCOCO+, and RefCOCOg. We report the Acc@0.5 on different test splits of the datasets.

Model	RefCOCO			RefCOCO+			RefCOCOg	
	val	testA	testB	val	testA	testB	val-u	test-u
VL-T5 [56]	-	-	-	-	-	-	-	71.3
UNITER [14]	81.41	87.04	74.17	75.90	81.45	66.70	74.86	75.77
VILLA [16]	82.39	87.48	74.84	76.17	81.54	66.84	76.18	76.71
MDETR [72]	86.75	89.58	81.41	79.52	84.09	70.62	81.64	80.89
UNICORN [57]	88.29	90.42	83.06	80.30	85.05	71.88	83.44	83.93
OFA <sub>Tiny</sub>	80.20	84.07	75.00	68.22	75.13	57.66	72.02	69.74
OFA <sub>Medium</sub>	85.34	87.68	77.92	76.09	83.04	66.25	78.76	78.58
OFA <sub>Base</sub>	88.48	90.67	83.30	81.39	87.15	74.29	82.29	82.31
OFA <sub>Large</sub>	90.05	92.93	85.26	85.80	89.87	79.22	85.89	86.55
OFA	<b>92.04</b>	<b>94.03</b>	<b>88.44</b>	<b>87.86</b>	<b>91.70</b>	<b>80.71</b>	<b>88.07</b>	<b>88.78</b>

task can be viewed as bounding box generation, and MSCOCO Image Caption for text-to-image generation. More details are provided in Appendix A.3.

Table 2 presents the performance of OFA and baseline models on VQA and SNLI-VE. In general, OFA achieves the best performance in both tasks with 82.0 on the VQA test-std set and 91.2 on the SNLI-VE test set. For smaller-size models, OFA<sub>Large</sub> can outperform the recent SOTAs, e.g., VLMo and SimVLM, and OFA<sub>Base</sub> can beat the SOTAs before the aforementioned two models in both tasks. This demonstrates that OFA can achieve superior performance on cross-modal understanding tasks and scaling up OFA can bring significant improvements, reflecting the strong potential of large-scale pretrained models.

Table 3 presents the performance of OFA and baseline models on the MSCOCO image captioning dataset. We report the results on the Karpathy test split, and we demonstrate the performance of models trained with Cross-Entropy optimization and additionally with CIDEr optimization based on reinforcement learning. In comparison with the previous SOTA SimVLM<sub>Huge</sub> for Cross-Entropy optimization, OFA outperforms it by around 2 points in CIDEr evaluation. For CIDEr optimization, OFA of the 3 sizes all outperform the huge-size LEMON, and OFA demonstrates a new SOTA of 154.9 CIDEr score. By May 31 2022, the single-model OFA had topped the MSCOCO Image Caption Leaderboard.<sup>3</sup>

To evaluate the capability of visual grounding, we conduct experiments on RefCOCO, RefCOCO+, and RefCOCOg. While we unify locations to the vocabulary, visual grounding can be viewed as a sequence generation task. As there is only one target for each query, we limit the generation length to 4 in order to generate a bounding box by  $\langle x_1, y_1, x_2, y_2 \rangle$ . Experimental results in Table 4 show that OFA reaches the SOTA performance on the 3 datasets. Compared with the previous SOTA UNICORN [57], OFA achieves significant improvement with a gain of 3.61, 6.65 and 4.85 points on the testA sets of RefCOCO and RefCOCO+ as well as the test-u set of RefCOCOg.

Text-to-image generation is a challenging task even for pretrained model. As we pretrain OFA with the task “image-infilling”, i.e., recovering masked patches by generating the corresponding codes [36], and thus OFA is able to generate code. We thus directly finetune OFA on the MSCOCO Image Caption dataset for text-to-code generation. At the inference stage, we additionally transform the generated codes to an image with the code decoder. Specifically, we use the codes from VQGAN [54] following [52]. Experimental results show that OFA outperforms the baselines in all the metrics. Note that increasing the sampling size during inference is expected to bring clear improvements on FID and IS. Compared with DALLE [50], CogView [51] and NUWA [52], whose sampling sizes are 512, 60 and 60, respectively, OFA outperforms these SOTA methods on FID and IS with a much smaller sampling size 24. This illustrates that OFA has learned better correspondence among the query text, the image and the image codes.

We compare OFA with CogView and GLIDE on generation quality with normal and counterfactual queries.<sup>4</sup> Normal queries describe existing things in the real world, while counterfactual queries refer to those describing things that could only exist in our imagination. For normal queries, both CogView and OFA generate images semantically consistent with the given texts, in comparison with GLIDE. The generated examples from our model can provide more sophisticated

<sup>3</sup><https://competitions.codalab.org/competitions/3221#results>

<sup>4</sup>For more implementation details, please refer to Appendix A.3

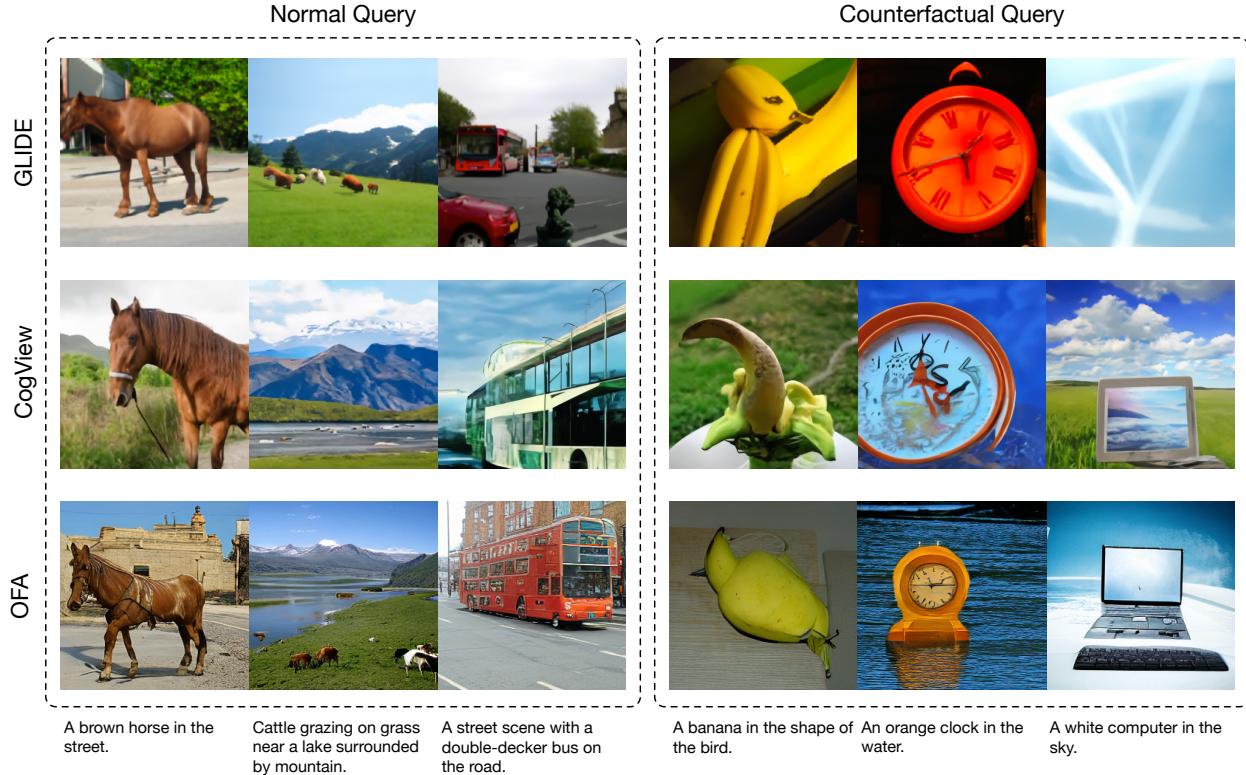


Figure 3: Qualitative comparison with state-of-the-art models for text-to-image generation task. We present more qualitative examples of text-to-image generation for better demonstration in Appendix C.

Table 5: Experimental results on text-to-image generation. Models are evaluated on FID, CLIPSIM, and IS scores. OFA outperforms the baselines, including the concurrent SOTA NÜWA. We report the results of OFA<sub>Large</sub>. Note that GLIDE additionally has 1.5B parameters for upsampling except for the 3.5B parameters.

Model	FID↓	CLIPSIM↑	IS↑
DALLE [50]	27.5	-	17.9
CogView [51]	27.1	33.3	18.2
GLIDE [77]	12.2	-	-
Unifying [78]	29.9	30.9	-
NÜWA [52]	12.9	34.3	27.2
<b>OFA</b>	<b>10.5</b>	<b>34.4</b>	<b>31.1</b>

details of objects, say the horse and the double-decker bus. For counterfactual queries, we find that OFA is the only one that can generate the three imaginary scenes, which indicates its imaginative power based on its strong capability to align text to the image. See Appendix C for more qualitative examples.

## 4.2 Results on Uni-modal Tasks

As the design of OFA unifies different modalities, we evaluate its performance on unimodal tasks, namely tasks of natural language and computer vision. For natural language tasks, we evaluate OFA on 6 tasks of the GLUE benchmark [79] for natural language understanding and Gigaword abstractive summarization [80] for natural language generation. For computer vision, we evaluate OFA on the classic ImageNet-1K [81] dataset for image classification. More details are provided in Appendix A.3.

As OFA has been pretrained on plain text data, it can be directly transferred to natural language downstream tasks. For natural language generation, it is essentially a sequence-to-sequence generation task, and for natural language

Table 6: Experimental results on the GLUE benchmark datasets [79]. For comparison, we list the performance of multimodal pretrained models as well the recent SOTA models that were pretrained on natural language data only. Following [28], we finetune RTE and MRPC starting from the checkpoint finetuned on MNLI.

Model	SST-2	RTE	MRPC	QQP	MNLI	QNLI
<i>Multimodal Pretrained Baseline Models</i>						
VisualBERT [38]	89.4	56.6	71.9	89.4	81.6	87.0
UNITER [14]	89.7	55.6	69.3	89.2	80.9	86.0
VL-BERT [8]	89.8	55.7	70.6	89.0	81.2	86.3
VilBERT [13]	90.4	53.7	69.0	88.6	79.9	83.8
LXMERT [40]	90.2	57.2	69.8	75.3	80.4	84.2
Uni-Perceiver [61]	90.2	64.3	86.6	87.1	81.7	89.9
SimVLM [22]	90.9	63.9	75.2	90.4	83.4	88.6
FLAVA [60]	90.9	57.8	81.4	90.4	80.3	87.3
UNIMO [46]	96.8	-	-	-	89.8	-
<i>Natural-Language-Pretrained SOTA Models</i>						
BERT [2]	93.2	70.4	88.0	91.3	86.6	92.3
RoBERTa [28]	96.4	86.6	90.9	92.2	90.2	93.9
XLNet [25]	<b>97.0</b>	85.9	90.8	92.3	90.8	94.9
ELECTRA [82]	96.9	88.0	90.8	92.4	90.9	95.0
DeBERTa [83]	96.8	88.3	<b>91.9</b>	92.3	<b>91.1</b>	<b>95.3</b>
<i>Ours</i>						
OFA	96.6	<b>91.0</b>	91.7	<b>92.5</b>	90.2	94.8

Table 7: Experimental results on Gigaword abstractive summarization. We report performance on the ROUGE evaluation [84]

Model	Gigaword		
	ROUGE-1	ROUGE-2	ROUGE-L
BERTSHARE [85]	38.13	19.81	35.62
MASS [86]	38.73	19.71	35.96
UniLM [29]	38.45	19.45	35.75
PEGASUS [87]	39.12	19.86	36.24
ProphetNet [88]	39.55	20.27	36.57
UNIMO [46]	39.71	20.37	36.88
OFA	<b>39.81</b>	<b>20.66</b>	<b>37.11</b>

understanding, typically text classification, we regard them as generation tasks where labels are essentially word sequences. Additionally, for each task, we design a manual instruction to indicate the model what types of questions it should answer. We list our instruction design in Appendix A.3.

We demonstrate that even a unified multimodal pretrained model can achieve highly competitive performance in natural language tasks. Specifically, in the evaluation of natural language understanding, OFA surpasses multimodal pretrained models by large margins in all tasks. In comparison with the state-of-the-art natural language pretrained models, including RoBERTa [28], XLNET [25], ELECTRA [82], and DeBERTa [83], OFA reaches a comparable performance. In the evaluation of natural language generation, OFA even reaches a new state-of-the-art performance on the Gigaword dataset.

Also, OFA can reach a competitive performance in image classification. Table 8 shows the performance of OFA on image classification. OFA<sub>Large</sub> achieves higher accuracy than previous backbone models such as EfficientNet-B7 [89] and ViT-L [6]. We also compare OFA with self-supervised pretraining models based on contrastive learning and masked image modeling. OFA outperforms contrastive-based models such as SimCLR [32] and MoCo-v3 [33, 35] with similar parameters. Compared with pretrained models based on masked image modeling, e.g., BEiT-L [36] and MAE-L [37], OFA can achieve similar performance.

Table 8: ImageNet-1K finetuning results. All the listed models do not use extra labeled image classification samples during training for fair comparison. We report the results of OFA<sub>Large</sub>.

Model	Top-1 Acc.
EfficientNet-B7 [89]	84.3
ViT-L/16 [6]	82.5
DINO [90]	82.8
SimCLR v2 [32]	82.9
MoCo v3 [35]	84.1
BEiT <sub>384</sub> -L/16 [36]	<b>86.3</b>
MAE-L/16 [37]	85.9
OFA	85.6

Table 9: Zero-shot performance on 6 GLUE tasks and SNLI-VE.

Model	SST-2 Acc.	RTE Acc.	MRPC F1	QQP F1	QNLI Acc.	MNLI Acc.	SNLI-VE Acc. (dev/test)
Uni-Perceiver	70.6	55.6	76.1	53.6	51.0	<b>49.6</b>	-
OFA <sub>Base</sub>	<b>71.6</b>	<b>56.7</b>	<b>79.5</b>	<b>54.0</b>	<b>51.4</b>	37.3	<b>49.71 / 49.18</b>

These aforementioned results in both natural language and vision tasks indicate that a unified multimodal pretrained model is not only effective in multimodal tasks but also capable of tackling unimodal tasks, and in the future, it might be sufficient for such a model to solve complex tasks concerning different modality combinations.

### 4.3 Zero-shot Learning & Task Transfer

The instruction-guided pretraining enables OFA to perform zero-shot inference. Following Uni-Perceiver [61], we evaluate our model on the 6 tasks of the GLUE benchmark, including single-sentence classification and sentence pair classification. Table 9 demonstrates that OFA generally outperforms Uni-Perceiver. However, both models do not achieve satisfactory performance in sentence-pair classification (with Acc. < 60%). We hypothesize that the missing sentence-pair data in the pretraining dataset attributes to the performance.

Also, we find that the model performance is highly sensitive to the design of instructions. To obtain the best result, one should search a proper instruction template possibly from a large pool of candidates. A slight change to manual prompts or model parameters may drastically influence the model performance, which is not robust. We leave this issue to the future work.

We observe that the model can transfer to unseen tasks well with new task instructions. We design a new task called grounded question answering and present examples in Figure 4. In this scenario, given a question about a certain region on the image, the model should provide a correct answer. We find that the model can achieve a satisfactory performance in this new task, which reflects its strong transferability. Besides, OFA can solve tasks with the out-of-domain input data. For example, OFA without finetuning achieves satisfactory performance in VQA for the out-of-domain images. Examples are demonstrated in Figure 5. OFA can also perform accurate visual grounding on the out-of-domain images, e.g., anime pictures, synthetic images, etc., and we demonstrate more examples on Figure 11 in Appendix C.

### 4.4 Ablation on Multitask Pretraining

Thanks to the unified framework, OFA has been pretrained on multiple tasks and thus endowed with comprehensive capabilities. However, the effects of each task are still undiscovered. We verify their effects on multiple downstream tasks, including image captioning, VQA, image classification, and text-to-image generation.

We first evaluate how uni-modal pretraining tasks influence the performance in both cross-modal and uni-modal tasks. Table 10 demonstrates our experimental results. We observe some interesting phenomena about the effects of uni-modal pretraining tasks. Text infilling brings improvement on image caption (+0.8 CIDEr) and VQA (+0.46 Acc.). Natural language pretraining betters the contextualized representation of language and thus enhances performance in cross-modal tasks. However, it is noticed that the language pretraining task may degrade the performance in image



**Q:** what color is the car in the region? region: <loc301> <loc495> <loc501> <loc596>

**A:** tan



**Q:** what color is the car in the region? region: <loc512> <loc483> <loc675> <loc576>

**A:** gray

Figure 4: Qualitative results on an unseen task grounded QA. We design a new task called grounded question answering, where the model should answer a question about a certain region in the image. More samples are provided in Figure 10 in Appendix C.



**Q:** what is grown on the plant?

**A:** money



**Q:** what does the red-roofed building right to the big airship look like?

**A:** a mushroom

Figure 5: Qualitative results on unseen domain VQA. During pretraining, only real-world photographs are used for VQA. We present cases of VQA on out-of-domain images, i.e., the iconic and sci-fi images, and demonstrate their capability of transferring to unseen domains. More samples are provided in Figure 9 in Appendix C.

classification, leading to the decrease in ImageNet-1K ( $-1.0$  Acc.). Also, it is interesting to find that it does not encourage improvement in text-to-image generation ( $-0.1$  CLIPSIM). It may attribute to the simplicity of text in this task, which indicates that improved representation of language does not affect the performance. As to image infilling, it significantly improves the performance in image classification ( $+1.0$  Acc.) and text-to-image generation ( $+0.6$  CLIPSIM). Learning to recover images is an effective self-supervised task for image representation, and it also encourages the decoder’s ability to generate image codes. However, it hurts the performance in image captioning and VQA. Both tasks require a strong capability in generating texts, and the decoder’s learning of image generation naturally brings performance degradation in captioning ( $-0.7$  CIDEr) and VQA ( $-0.3$  Acc.).

Furthermore, we evaluate how multimodal tasks impact the performance. Previous studies have provided evidence of the contribution of conventional pretraining tasks, e.g., MLM, MOC, ITM, VQA, image captioning, etc. [14, 17]. However, they miss other tasks, e.g., detection and visual grounding & grounded captioning. We conduct experiments on these tasks and find that tasks predicting regions are crucial to multimodal tasks, with a performance increase in image captioning ( $+2.3$  CIDEr &  $+1.4$  CIDEr) and VQA ( $+0.6$  Acc. &  $+0.5$  Acc.). It suggests that detection and visual grounding & grounded captioning help the model grasp fine-grained alignments between vision and language.

Table 10: Ablation results of OFA. All models are pretrained for 250k steps. *w/o ground.* represents the removal of both visual grounding and grounded captioning tasks. Note that all models are only finetuned with the cross-entropy loss in image captioning.

Model	Caption CIDEr	VQA Test-dev	ImageNet Top-1 Acc.	Image Generation FID / CLIPSIM / IS
OFA <sub>Base</sub>	135.6	76.0	82.2	20.8 / 31.6 / 21.5
<i>w/o text infill.</i>	134.8	75.6	83.2	20.3 / 31.7 / 21.8
<i>w/o image infill.</i>	136.3	76.3	81.8	23.2 / 31.0 / 20.0
<i>w/o det.</i>	133.3	75.4	81.4	20.9 / 31.5 / 21.6
<i>w/o ground.</i>	134.2	75.5	82.0	21.2 / 31.5 / 21.5

Region information contributes little to text-to-image generation (+0.1 CLIPSIM & +0.1 CLIPSIM), as this task requires far less text-region alignment information. We surprisingly find that detection can encourage the performance in visual understanding (+0.8 Acc.). It indicates that incorporating region information might be essential to visual understanding, especially on images with complex objects.

## 5 Conclusion

In this work, we propose **OFA**, a Task-Agnostic and Modality-Agnostic framework supporting Task Comprehensiveness. OFA achieves the unification in architecture, tasks and modalities, and thus is capable of multimodal & uni-modal understanding and generation, without specification in additional layers or tasks. Our experiments show that OFA creates new SOTAs in a series of tasks, including image captioning, VQA, visual entailment, and referring expression comprehension. OFA also demonstrates a comparable performance with language / vision pretrained SOTA models in uni-modal understanding and generation tasks, e.g., GLUE, abstractive summarization, and image classification. We provide a further analysis to demonstrate its capability in zero-shot learning and domain & task transfer, and we also verify the effectiveness of pretraining tasks.

In the future, we will continue exploring the issues discovered in this work. Also, we endeavor to figure out a reasonable solution to building an omni-model essentially generalizable to the complex real world.

## Acknowledgments

We would like to thank Jie Zhang, Yong Li, Jiamang Wang, Shao Yuan, and Zheng Cao for their support to this project, and we would like to thank Guangxiang Zhao and Fei Sun for their insightful comments to our paper.

## References

- [1] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS 2017*, pages 5998–6008, 2017.
- [2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *NAACL-HLT 2019*, pages 4171–4186. Association for Computational Linguistics, 2019.
- [3] Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.
- [4] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.
- [5] Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. wav2vec: Unsupervised pre-training for speech recognition. *arXiv preprint arXiv:1904.05862*, 2019.
- [6] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

- [7] Andrew Jaegle, Felix Gimeno, Andrew Brock, Andrew Zisserman, Oriol Vinyals, and Joao Carreira. Perceiver: General perception with iterative attention. [arXiv preprint arXiv:2103.03206](#), 2021.
- [8] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. Vi-bert: Pre-training of generic visual-linguistic representations. In [International Conference on Learning Representations](#), 2019.
- [9] Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. Finetuned language models are zero-shot learners. [arXiv preprint arXiv:2109.01652](#), 2021.
- [10] Victor Sanh, Albert Webson, Colin Raffel, Stephen H Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, et al. Multitask prompted training enables zero-shot task generalization. [arXiv preprint arXiv:2110.08207](#), 2021.
- [11] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In [International Conference on Machine Learning](#), pages 2790–2799. PMLR, 2019.
- [12] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. [arXiv preprint arXiv:2104.08691](#), 2021.
- [13] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In [NeurIPS](#), 2019.
- [14] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In [ECCV](#), 2020.
- [15] Xiuju Li, Xi Yin, Chunyuan Li, Xiaowei Hu, Pengchuan Zhang, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. Oscar: Object-semantics aligned pre-training for vision-language tasks. In [ECCV](#), 2020.
- [16] Zhe Gan, Yen-Chun Chen, Linjie Li, Chen Zhu, Yu Cheng, and Jingjing Liu. Large-scale adversarial training for vision-and-language representation learning. [ArXiv](#), abs/2006.06195, 2020.
- [17] Pengchuan Zhang, Xiuju Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinvl: Revisiting visual representations in vision-language models. [2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition \(CVPR\)](#), pages 5575–5584, 2021.
- [18] Junyang Lin, Rui Men, An Yang, Chang Zhou, Ming Ding, Yichang Zhang, Peng Wang, Ang Wang, Le Jiang, Xianyan Jia, et al. M6: A chinese multimodal pretrainer. [arXiv preprint arXiv:2103.00823](#), 2021.
- [19] Zhu Zhang, Jianxin Ma, Chang Zhou, Rui Men, Zhikang Li, Ming Ding, Jie Tang, Jingren Zhou, and Hongxia Yang. M6-ufc: Unifying multi-modal controls for conditional image synthesis. [arXiv preprint arXiv:2105.14211](#), 2021.
- [20] An Yang, Junyang Lin, Rui Men, Chang Zhou, Le Jiang, Xianyan Jia, Ang Wang, Jie Zhang, Jiamang Wang, Yong Li, et al. Exploring sparse expert models and beyond. [arXiv preprint arXiv:2105.15082](#), 2021.
- [21] Junyang Lin, An Yang, Jinze Bai, Chang Zhou, Le Jiang, Xianyan Jia, Ang Wang, Jie Zhang, Yong Li, Wei Lin, et al. M6-10t: A sharing-delinking paradigm for efficient multi-trillion parameter pretraining. [arXiv preprint arXiv:2110.03888](#), 2021.
- [22] Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. Simvlm: Simple visual language model pretraining with weak supervision. [ArXiv](#), abs/2108.10904, 2021.
- [23] Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel C. F. Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, Ce Liu, Mengchen Liu, Zicheng Liu, Yumao Lu, Yu Shi, Lijuan Wang, Jianfeng Wang, Bin Xiao, Zhen Xiao, Jianwei Yang, Michael Zeng, Luowei Zhou, and Pengchuan Zhang. Florence: A new foundation model for computer vision. [ArXiv](#), abs/2111.11432, 2021.
- [24] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. [URL https://s3-us-west-2.amazonaws.com/openai-assets/researchcovers/languageunsupervised/language-understanding-paper.pdf](https://s3-us-west-2.amazonaws.com/openai-assets/researchcovers/languageunsupervised/language-understanding-paper.pdf), 2018.
- [25] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. Xlnet: Generalized autoregressive pretraining for language understanding. In [NeurIPS 2019](#), pages 5754–5764, 2019.
- [26] Yu Sun, Shuohuan Wang, Yu-Kun Li, Shikun Feng, Xuyi Chen, Han Zhang, Xin Tian, Danxiang Zhu, Hao Tian, and Hua Wu. ERNIE: enhanced representation through knowledge integration. [CoRR](#), abs/1904.09223, 2019.
- [27] Yu Sun, Shuohuan Wang, Yu-Kun Li, Shikun Feng, Hao Tian, Hua Wu, and Haifeng Wang. ERNIE 2.0: A continual pre-training framework for language understanding. [CoRR](#), abs/1907.12412, 2019.

- [28] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019.
- [29] Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. Unified language model pre-training for natural language understanding and generation. In *NeurIPS 2019*, pages 13042–13054, 2019.
- [30] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020.
- [31] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *ACL 2020*, July 2020.
- [32] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.
- [33] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020.
- [34] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent: A new approach to self-supervised learning. *arXiv preprint arXiv:2006.07733*, 2020.
- [35] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15750–15758, 2021.
- [36] Hangbo Bao, Li Dong, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021.
- [37] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. *arXiv preprint arXiv:2111.06377*, 2021.
- [38] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *ArXiv*, abs/1908.03557, 2019.
- [39] Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason J. Corso, and Jianfeng Gao. Unified vision-language pre-training for image captioning and VQA. In *AAAI 2020*, pages 13041–13049, 2020.
- [40] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5100–5111, 2019.
- [41] Gen Li, Nan Duan, Yuejian Fang, Dixin Jiang, and Ming Zhou. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. *CoRR*, abs/1908.06066, 2019.
- [42] Junyang Lin, An Yang, Yichang Zhang, Jie Liu, Jingren Zhou, and Hongxia Yang. Interbert: Vision-and-language interaction for multi-modal pretraining. *arXiv preprint arXiv:2003.13198*, 2020.
- [43] Jiasen Lu, Vedanuj Goswami, Marcus Rohrbach, Devi Parikh, and Stefan Lee. 12-in-1: Multi-task vision and language representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10437–10446, 2020.
- [44] Haiyang Xu, Ming Yan, Chenliang Li, Bin Bi, Songfang Huang, Wenming Xiao, and Fei Huang. E2e-vlp: End-to-end vision-language pre-training enhanced by visual learning. *arXiv preprint arXiv:2106.01804*, 2021.
- [45] Fei Yu, Jiji Tang, Weichong Yin, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. Ernie-vil: Knowledge enhanced vision-language representations through scene graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 3208–3216, 2021.
- [46] Wei Li, Can Gao, Guocheng Niu, Xinyan Xiao, Hao Liu, Jiachen Liu, Hua Wu, and Haifeng Wang. UNIMO: towards unified-modal understanding and generation via cross-modal contrastive learning. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *ACL/IJCNLP 2021*, pages 2592–2607. Association for Computational Linguistics, 2021.
- [47] Zhicheng Huang, Zhaoyang Zeng, Bei Liu, Dongmei Fu, and Jianlong Fu. Pixel-bert: Aligning image pixels with text by deep multi-modal transformers. *ArXiv*, abs/2004.00849, 2020.
- [48] Wenhui Wang, Hangbo Bao, Li Dong, and Furu Wei. Vlmo: Unified vision-language pre-training with mixture-of-modality-experts. *ArXiv*, abs/2111.02358, 2021.

- [49] 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. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang, editors, *ICML 2021*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 2021.
- [50] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. *arXiv preprint arXiv:2102.12092*, 2021.
- [51] Ming Ding, Zhuoyi Yang, Wenqi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, et al. Cogview: Mastering text-to-image generation via transformers. *arXiv preprint arXiv:2105.13290*, 2021.
- [52] Chenfei Wu, Jian Liang, Lei Ji, Fan Yang, Yuejian Fang, Dixin Jiang, and Nan Duan. N\ " uwa: Visual synthesis pre-training for neural visual world creation. *arXiv preprint arXiv:2111.12417*, 2021.
- [53] Aäron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. In *NIPS*, 2017.
- [54] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12873–12883, 2021.
- [55] Lukasz Kaiser, Aidan N Gomez, Noam Shazeer, Ashish Vaswani, Niki Parmar, Llion Jones, and Jakob Uszkoreit. One model to learn them all. *arXiv preprint arXiv:1706.05137*, 2017.
- [56] Jaemin Cho, Jie Lei, Haochen Tan, and Mohit Bansal. Unifying vision-and-language tasks via text generation. In *ICML*, 2021.
- [57] Zhengyuan Yang, Zhe Gan, Jianfeng Wang, Xiaowei Hu, Faisal Ahmed, Zicheng Liu, Yumao Lu, and Lijuan Wang. Crossing the format boundary of text and boxes: Towards unified vision-language modeling. *ArXiv, abs/2111.12085*, 2021.
- [58] Andrew Jaegle, Sebastian Borgeaud, Jean-Baptiste Alayrac, Carl Doersch, Catalin Ionescu, David Ding, Skanda Koppula, Daniel Zoran, Andrew Brock, Evan Shelhamer, et al. Perceiver io: A general architecture for structured inputs & outputs. *arXiv preprint arXiv:2107.14795*, 2021.
- [59] Ronghang Hu and Amanpreet Singh. Unit: Multimodal multitask learning with a unified transformer. *arXiv preprint arXiv:2102.10772*, 2021.
- [60] Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. Flava: A foundational language and vision alignment model. *arXiv preprint arXiv:2112.04482*, 2021.
- [61] Xizhou Zhu, Jingyu Zhu, Hao Li, Xiaoshi Wu, Xiaogang Wang, Hongsheng Li, Xiaohua Wang, and Jifeng Dai. Uni-perceiver: Pre-training unified architecture for generic perception for zero-shot and few-shot tasks. *arXiv preprint arXiv:2112.01522*, 2021.
- [62] Zihang Dai, Hanxiao Liu, Quoc V Le, and Mingxing Tan. Coatnet: Marrying convolution and attention for all data sizes. *arXiv preprint arXiv:2106.04803*, 2021.
- [63] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, 2016.
- [64] Ting Chen, Saurabh Saxena, Lala Li, David J Fleet, and Geoffrey Hinton. Pix2seq: A language modeling framework for object detection. *arXiv preprint arXiv:2109.10852*, 2021.
- [65] Lei Jimmy Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization. *CoRR, abs/1607.06450*, 2016.
- [66] Sam Shleifer, Jason Weston, and Myle Ott. Normformer: Improved transformer pretraining with extra normalization. *arXiv preprint arXiv:2110.09456*, 2021.
- [67] Guolin Ke, Di He, and Tie-Yan Liu. Rethinking positional encoding in language pre-training. In *International Conference on Learning Representations*, 2020.
- [68] Thomas H Cormen, Charles E Leiserson, Ronald L Rivest, and Clifford Stein. *Introduction to algorithms*. MIT press, 2009.
- [69] Junnan Li, Ramprasaath R Selvaraju, Akhilesh Deepak Gotmare, Shafiq Joty, Caiming Xiong, and Steven Hoi. Align before fuse: Vision and language representation learning with momentum distillation. In *Thirty-Fifth Conference on Neural Information Processing Systems*, 2021.

- [70] Zi-Yi Dou, Yichong Xu, Zhe Gan, Jianfeng Wang, Shuohang Wang, Lijuan Wang, Chenguang Zhu, Nanyun Peng, Zicheng Liu, and Michael Zeng. An empirical study of training end-to-end vision-and-language transformers. *ArXiv*, abs/2111.02387, 2021.
- [71] Xiaowei Hu, Zhe Gan, Jianfeng Wang, Zhengyuan Yang, Zicheng Liu, Yumao Lu, and Lijuan Wang. Scaling up vision-language pre-training for image captioning. *CoRR*, abs/2111.12233, 2021.
- [72] Aishwarya Kamath, Mannat Singh, Yann LeCun, Ishan Misra, Gabriel Synnaeve, and Nicolas Carion. Mdetr - modulated detection for end-to-end multi-modal understanding. *ArXiv*, abs/2104.12763, 2021.
- [73] Ning Xie, Farley Lai, Derek Doran, and Asim Kadav. Visual entailment: A novel task for fine-grained image understanding. *arXiv preprint arXiv:1901.06706*, 2019.
- [74] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015.
- [75] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *European Conference on Computer Vision*, pages 69–85. Springer, 2016.
- [76] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 11–20, 2016.
- [77] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021.
- [78] Yupan Huang, Hongwei Xue, Bei Liu, and Yutong Lu. Unifying multimodal transformer for bi-directional image and text generation. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 1138–1147, 2021.
- [79] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*, 2018.
- [80] Alexander M Rush, Sumit Chopra, and Jason Weston. A neural attention model for abstractive sentence summarization. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 379–389, 2015.
- [81] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.
- [82] Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. ELECTRA: pre-training text encoders as discriminators rather than generators. In *8th International Conference on Learning Representations, ICLR 2020*. OpenReview.net, 2020.
- [83] Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. Deberta: decoding-enhanced bert with disentangled attention. In *9th International Conference on Learning Representations, ICLR 2021*. OpenReview.net, 2021.
- [84] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, Barcelona, Spain, July 2004. Association for Computational Linguistics.
- [85] Sascha Rothe, Shashi Narayan, and Aliaksei Severyn. Leveraging pre-trained checkpoints for sequence generation tasks. *Transactions of the Association for Computational Linguistics*, 8:264–280, 2020.
- [86] Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. MASS: masked sequence to sequence pre-training for language generation. In *ICML 2019*, pages 5926–5936, 2019.
- [87] Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International Conference on Machine Learning*, pages 11328–11339. PMLR, 2020.
- [88] Weizhen Qi, Yu Yan, Yeyun Gong, Dayiheng Liu, Nan Duan, Jiusheng Chen, Ruofei Zhang, and Ming Zhou. Prophetnet: Predicting future n-gram for sequence-to-sequence pre-training. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 2401–2410, 2020.
- [89] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*, pages 6105–6114. PMLR, 2019.
- [90] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. *arXiv preprint arXiv:2104.14294*, 2021.

- [91] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3558–3568, 2021.
- [92] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *ACL 2018*, pages 2556–2565, 2018.
- [93] Vicente Ordonez, Girish Kulkarni, and Tamara L. Berg. Im2text: Describing images using 1 million captioned photographs. In *NeurIPS 2011*, pages 1143–1151, 2011.
- [94] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73, 2017.
- [95] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6904–6913, 2017.
- [96] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *CVPR 2019*, pages 6700–6709, 2019.
- [97] Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73, 2016.
- [98] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Mallocci, Alexander Kolesnikov, et al. The open images dataset v4. *International Journal of Computer Vision*, 128(7):1956–1981, 2020.
- [99] Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. Objects365: A large-scale, high-quality dataset for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8430–8439, 2019.
- [100] Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*, 2020.
- [101] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR 2016*, pages 770–778, 2016.
- [102] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR 2019*, 2019.
- [103] Gao Huang, Yu Sun, Zhuang Liu, Daniel Sedra, and Kilian Q. Weinberger. Deep networks with stochastic depth. In *ECCV*, 2016.
- [104] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.
- [105] Satanjeev Banerjee and Alon Lavie. 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*, pages 65–72, 2005.
- [106] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575, 2015.
- [107] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. In *European conference on computer vision*, pages 382–398. Springer, 2016.
- [108] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137, 2015.
- [109] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.
- [110] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *Advances in neural information processing systems*, 29:2234–2242, 2016.

- [111] Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. Self-critical sequence training for image captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7008–7024, 2017.
- [112] Guangxiang Zhao, Wenkai Yang, Xuancheng Ren, Lei Li, and Xu Sun. Well-classified examples are underestimated in classification with deep neural networks. *CoRR*, abs/2110.06537, 2021.
- [113] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 702–703, 2020.
- [114] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 13001–13008, 2020.
- [115] Hongyi Zhang, Moustapha Cissé, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018.
- [116] Sangdoo Yun, Dongyoon Han, Sanghyuk Chun, Seong Joon Oh, Youngjoon Yoo, and Junsuk Choe. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 6022–6031. IEEE, 2019.
- [117] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021.

## A Implementation Details

### A.1 Pretraining Datasets

We construct pretraining datasets by incorporating Vision & Language data (i.e., image-text pairs), Vision data (i.e., raw image data, object-labeled data), and Language data (i.e., plain texts). For replication, the pretraining datasets are publicly available. We carefully filter our pretraining data and exclude images that appear in the validation and test sets of downstream tasks to avoid data leakage. The statistics on the pretraining datasets are listed in Table 11.

**Cross-modal Data** For vision & language pretraining, we mainly apply image-text pairs, including image-caption pairs, image-QA pairs, and image-region pairs, as the pretraining data. For the pretraining tasks of image captioning and image-text matching, we collect Conceptual Caption 12M (CC12M) [91], Conceptual Captions (CC3M) [92], SBU [93], MSCOCO image captions (COCO) [74], and Visual Genome Captions (VG Captions) [94]. Specifically, the part of data from VG requires some additional processing. As texts in VG captions describe local regions on the images, we retrieve regions with area larger than 16,384 pixels and construct region-caption pairs. For visual question answering, we collect VQAv2 [95], VG-QA [94], as well as GQA [96]. VQAv2 is a visual question answering dataset with real-world photographs from COCO. VG-QA is also a visual question answering dataset with real-world photographs from VG. The questions of VG-QA are related to specific regions on the images. GQA is a large VQA dataset featuring compositional questions. The images of GQA are also collected from VG. For visual grounding and grounded captioning, we collect data from RefCOCO [75], RefCOCO+ [75], RefCOCOg [76] and VG captions. Additional processing is applied to VG Captions for this task. Specifically, we use the data of VG that contains regions with area smaller than 16,384 pixels for Visual Grounding, in order to encourage model to grasp fine-grained alignments between vision and language.

**Uni-modal Data** Uni-modal data includes vision and language data. Vision data consists of raw images for image infilling and object-labeled images for object detection. For image infilling, we collect raw images from OpenImages, YFCC100M [97] and ImageNet-21K [81], and exclude annotations. Thus the model is unable to access labels in the pretraining stage. For object detection, we collect OpenImages [98], Object365 [99], VG and COCO for object detection. Language data consists of plain texts, i.e., passages consisting of sentences. We use around 140GB of data from Pile [100] to leverage its diversity. Specifically, we extract natural language data and implement preprocessing methods, including truncation to the length of 512.

Table 11: Statistics on the datasets of pretraining tasks. “#Image” denotes the number of distinct images, and “#Sample” denotes the number of samples. \*For language data, we report its storage following the previous studies [2, 28].

Type	Pretraining Task	Source	#Image	#Sample
Vision & Language	Image Captioning	CC12M, CC3M, SBU, COCO, VG-Cap	14.78M	15.25M
	Image-Text Matching			
	Visual Question Answering	VQAv2, VG-QA, GQA	178K	2.92M
	Visual Grounding Grounded Captioning	RefCOCO, RefCOCO+, RefCOCOg, VG-Cap	131K	3.20M
Vision	Detection	OpenImages, Object365, VG, COCO	2.98M	3.00M
	Image Infilling	OpenImages, YFCC100M, ImageNet-21K	36.27M	-
Language	Masked Language Modeling	Pile (Filtered)	-	140GB*

### A.2 Pretraining Details

For the image processing, we first resize and crop the images into different resolutions,  $256 \times 256$  for OFATiny and OFAMedium,  $384 \times 384$  for OFABase,  $480 \times 480$  for OFALarge and OFAHuge, with a fixed patch size of  $16 \times 16$ . Note that training OFALarge and OFAHuge are time and computation consuming, we first train them with images of the resolution of  $384 \times 384$  and  $256 \times 256$ , and continue pretraining with images of the resolution of  $480 \times 480$ .

For each patch, we obtain its feature vector with the first three blocks of ResNet [101]. The ResNet module is jointly trained along with the transformer module. Note that through extensive experiments we find that random sampling patches [47] does not bring additional benefits in our scenario. For the text processing, we tokenize the texts with the

same BPE Tokenizer [63] as BART [31]. The maximum text sequence length of both encoder and decoder is set to 256. We share parameters between the embedding and the decoder softmax output layer.

From our preliminary experiments, we find that the initialization for Transformer plays an important role. For OFA<sub>Base</sub> and OFA<sub>Large</sub>, we initialize the transformer with most of the weights of BART<sub>Base</sub> and BART<sub>Large</sub> considering the slight difference between OFA Transformer and BART as described in Sec 3.1. For OFA of the other sizes, we pretrain language models with the same pretraining strategy with BART and use the pretrained weights to initialize the Transformer in OFA.

We use the AdamW [102] optimizer with  $(\beta_1, \beta_2) = (0.9, 0.999)$  and  $\epsilon = 1e-8$  to pretrain our models. We set the peak learning rate to  $2e-4$ , and apply a scheduler with linear decay with a warmup ratio of 0.01 to control the learning rate. For regulation, we set dropout to 0.1 and use weight decay with 0.01. We employ stochastic depth [103] with a 0.1 rate (applied to encoder and decoder except for convolution blocks). We mix all the pretraining data within each batch, which contains 2,048 vision&language samples, 256 object detection samples, 256 image-only samples and 512 text-only samples. All models are pretrained for at least  $300K$  steps except the models used for ablation study.

### A.3 Details of Downstream Tasks

We verify the capability of OFA on various downstream tasks in both finetuning and zero-shot settings. We design various task-specific instructions to transfer the knowledge learned from pretraining to downstream tasks effectively. The instructions of different tasks are listed in Table 12. For finetuning, if not specified, the input image resolution is set to  $480 \times 480$ , and the other hyper-parameters remain the same as for pretraining. The experimental details of different downstream tasks, including both multimodal and uni-modal tasks, are listed below:

**Image Captioning** Image captioning is a standard vision&language task that requires models to generate an appropriate and fluent caption for an image. We adopt the most widely used MSCOCO Image Caption dataset [74] to evaluate the multi-modal generation capability of OFA. We report BLEU-4 [104], METEOR [105], CIDEr [106], and SPICE [107] scores on the Karpathy test split [108]. Following the previous standard practice, we first finetune OFA with cross-entropy loss for 2 epochs with a batch size of 128 and a learning rate of  $1e - 5$ , and label smoothing is set to 0.1. We then finetune the model with CIDEr optimization for 3 epochs with a batch size of 64, and disable dropout and stochastic depth. We report both scores at the two stages.

**Visual Question Answering** Visual question answering (VQA) is a cross-modal task that requires the models to answer the question given an image. Previous works such as VLMo [48] or SimVLM [22] define VQA as a classification task. They use a linear output layer to predict the probability of each candidate answer on a given set. In contrast with these studies, to adapt the generative OFA model to VQA benchmark, we use the Trie-based search strategy mentioned in Sec. 3.4 to ensure that the answer generated by OFA is constrained in the candidate set. We evaluate our model with other baselines on the commonly used VQAv2 dataset [95]. Accuracy scores on both test-dev and test-std sets are reported. The OFA models of all the reported sizes are finetuned for 40,000 steps with a batch size of 512. The learning rate is  $5e - 5$  with the label smoothing of 0.1. When finetuning OFA<sub>Large</sub> and OFA<sub>Huge</sub>, we increase the image resolution from 480 to 640. Linear interpolation of the image absolute positional embedding proposed in [6] is employed when transferring the pretrained OFA to VQA finetuning. During Trie-based searching, we constrain the generated answers over the most frequent 3,129 answer candidates. Exponential moving average (EMA) with decay rate 0.9999 is employed in finetuning.

**Visual Entailment** Visual entailment requires the model to evaluate how the given image and text are semantically correlated, i.e., entailment, neutral, or contradiction. We perform experiments on the SNLI-VE dataset [73]. The image premise, text premise and text hypothesis are fed to the encoder, and the decoder generates appropriate labels. To transfer the knowledge learned by pretraining to this task, we convert the labels entailment/neutral/contradiction to yes/maybe/no. We also use the Trie-based search strategy to constrain the generated labels over the candidate set. We report accuracy on both dev and test sets. The OFA model is finetuned for 6 epochs with a learning rate of  $2e - 5$  and a batch size of 256.

**Referring Expression Comprehension** Referring expression comprehension requires models to locate an image region described by a language query. Different from the approach taken by most previous methods [13, 14] which ranks a set of candidate bounding boxes detected by a pretrained object detector, our method directly predicts the best matching bounding box without any proposals. We perform experiments on RefCOCO [75], RefCOCO+ [75], and RefCOCOg [76]. Consistent with other downstream tasks, we formulate referring expression comprehension as a conditional sequence generation task. In detail, given an image and a language query, OFA generates the box sequence (e.g.,  $\langle x_1, y_1, x_2, y_2 \rangle$ ) in an autoregressive manner. We report the standard metric Acc@0.5 on the validation and test

Table 12: Instructions for downstream tasks.

Task	Dataset	Instruction	Target
Image Captioning	COCO	[Image] What does the image describe?	{Caption}
Visual Question Answering	VQA	[Image] {Question}	{Answer}
Visual Entailment	SNLI-VE	[Image] Can image and text1 "{Text1}" imply text2 "{Text2}"?	Yes/No/Maybe
Referring Expression Comprehension	RefCOCO, RefCOCO+, RefCOCOg	[Image] Which region does the text "{Text}" describe?	{Location}
Image Generation	COCO	What is the complete image? caption: {Caption}	{Image}
Image Classification	ImageNet-1K	[Image] What does the image describe?	{Label}
Single-Sentence Classification	SST-2	Is the sentiment of text "{Text}" positive or negative?	Positive/Negative
Sentence-Pair Classification	RTE	Can text1 "{Text1}" imply text2 "{Text2}"?	Yes/No
	MRPC	Does text1 "{Text1}" and text2 "{Text2}" have the same semantics?	Yes/No
	QQP	Is question "{Question1}" and question "{Question2}" equivalent?	Yes/No
	MNLI	Can text1 "{Text1}" imply text2 "{Text2}"?	Yes/No/Maybe
	QNLI	Does "{Text}" contain the answer to question "{Question}"?	Yes/No
Text Summarization	Gigaword	What is the summary of article "{Article}"?	{Summary}

sets. For finetuning, the input image resolution is set to  $512 \times 512$ . We finetune the OFA model on each dataset for about 10 epochs with a batch size of 128. The learning rate is  $3e - 5$  with the label smoothing of 0.1. Each query only corresponds to an image region, so we limit the maximum generated length to 4 during inference.

**Image Generation** Following the same setting with [52], we train our model on the MS COCO train split and evaluate our model on the validation split by randomly sampling 30,000 images. We use Fréchet Inception Distance (FID) [109] and Inception Score (IS) [110] to evaluate the quality of the images. Following the previous studies [78, 52], we also compute CLIP Similarity Score (CLIPSIM) to evaluate the semantic similarity between the query text and the generated images. During finetuning, OFA learns to generate the image code sequence according to the given text query only. The model is first finetuned with cross-entropy and then with CLIPSIM optimization following [111, 78]. In the first stage, we finetune the OFA model for about 50 epochs with a batch size of 512 and a learning rate of  $1e - 3$ . In the second stage, the model is finetuned for extra 5000 steps with a batch size of 32 and a learning rate of  $1e - 6$ . During the evaluation, we sample 24 images with the resolution of  $256 \times 256$  for each query and choose the best one using the pretrained CLIP model [49].

For case study, we compare OFA with CogView and GLIDE. CogView provides an API website<sup>5</sup>. Note that this API samples 8 images of resolution of  $512 \times 512$  for each query. We select the first one of generated images and resize it to the resolution of  $256 \times 256$ . GLIDE provides a Colab notebook.<sup>6</sup>. Note that the only publicly available GLIDE model is of *base* size ( $\sim 385M$ ).

**Image Classification** We provide finetuning results on ImageNet-1K [81] following recent studies in self-supervised learning for computer vision. During finetuning and inference, a Trie-based search strategy is employed to constrain the generated text into the set of 1,000 candidate labels. We finetune OFA for 32 epochs and a batch size of 256. The learning rate is  $5e - 5$ . The ratio for label smoothing is 0.1. The encouraging loss proposed in [112] is employed with the hyperparameter LE set to 0.75. Following [36], we use the same random resize cropping, random flipping, RandAug [113] and random erasing [114] transformations as data augmentation strategies. Mixup [115] and CutMix [116] are used with overall 0.5 probability to be performed on each batch and alpha is 0.8 and 1.0, respectively. To adapt the mixed soft target of Mixup and CutMix into generation paradigm during finetuning, we run the decoder twice each with one of the target sequences to be mixed and sum the loss weighted by the mixing ratio.

**Natural Language Understanding** To verify the natural language understanding ability of OFA, we select 6 language understanding tasks from GLUE benchmark [79], including both single-sentence classification tasks and sentence-pair

<sup>5</sup><https://wudao.aminer.cn/CogView/index.html>

<sup>6</sup><https://colab.research.google.com/drive/1q6tJ58UKod1eC0kbaUNGzF3K5BbX1B5m>

classification tasks. To adapt to sentence-pair classification, previous models [2, 28] usually use segment embeddings to distinguish different sentences. Unlike those models, OFA can apply the model to sentence-pair classification tasks by constructing appropriate instructions without introducing additional segment embeddings. For the hyper-parameters of finetuning, we tune the training epochs among {5, 7, 10}, learning rate among { $3e - 5$ ,  $5e - 5$ ,  $6e - 5$ ,  $7e - 5$ ,  $1e - 4$ }, batch size among {32, 64, 128}, weight decay among {0.01, 0.05}, and dropout rate among {0.0, 0.1}. We report the best performance on the development set for each task.

**Natural Language Generation** We verify the natural language generation ability of OFA in the Gigaword dataset [80]. We report ROUGE-1/ROUGE-2/ROUGE-L to evaluate the generation results following [80]. We finetune the OFA models for 6 epochs with a batch size of 512. The learning rate is  $1e - 4$  with the label smoothing of 0.1, and the maximum input text sequence length is set to 512. During inference, we set the length penalty to 0.7 and beam size to 6, and limit the maximum generated length to 32.

## B Trie-based Search

This section describes how to use Trie-based search to improve model performance on downstream classification tasks. When dealing with classification tasks, we first construct a Trie where nodes are annotated with tokens from the candidate label-set. During finetuning, the model computes the log-probabilities of the target tokens based on their positions on the Trie. As shown in Figure 6, when computing the log-probabilities of the target token “sky”, we only consider tokens in {"sky", "ocean"} and forcefully set the logits for all invalid tokens to  $-\infty$ . During inference, we constrain the generated labels over the candidate set. As shown in Table 13, Trie-based search strategy can boost the performance of OFA in various downstream classification tasks.

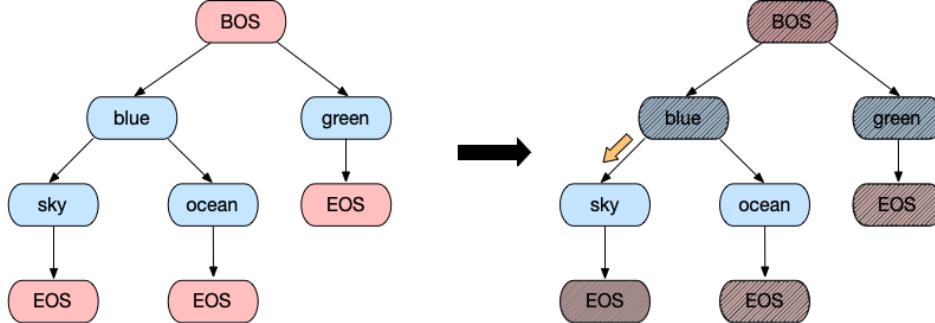


Figure 6: Example of Trie-based search where the constraint labels are “blue sky”, “blue ocean” and “green”. When computing the log-prob of token “sky”, we only consider tokens in {"sky", "ocean"} and forcefully set the logits for all invalid tokens to  $-\infty$ .

Table 13: Ablation results of Trie. The removal of Trie-based search degenerates the performance on downstream tasks. Note that the baseline OFA<sub>Base</sub> is only pre-trained for 250k steps, which is also used in Table 10.

Model	VQA Test-dev Acc.	SNLI-VE Dev Acc.	ImageNet Top-1 Acc.	MRPC F1	QQP F1
OFA <sub>Base</sub>	76.03	89.2	82.2	90.6	88.4
w/o Trie	75.86(-0.17)	89.0(-0.2)	81.9(-0.3)	90.1(-0.5)	88.2(-0.2)

## C Qualitative Examples

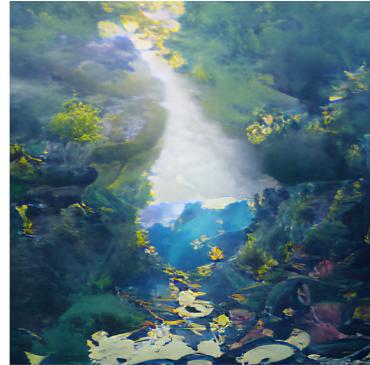
This section provides more qualitative examples of multiple tasks, including text-to-image generation, open-domain VQA, grounded question answering, and open-domain visual grounding, from the generation of OFA. By reading this section, we hope that readers can better perceive OFA.



An eagle view of a magic city.



A pathway to a temple with sakura trees in full bloom, HD.



A beautiful painting of native forest landscape photography, HD.



An art painting of a soldier, in the style of cyberspace.



The golden palace of the land of clouds.



Rustic interior of an alchemy shop.



An art painting of a dog, in the style of steampunk, white background.

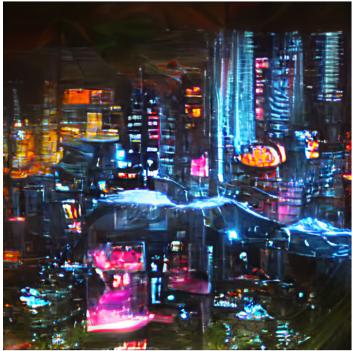


A strawberry splashing in the coffee in a mug under the starry sky.



Elf elk in the forest illustration, HD, fantasy art.

Figure 7: Examples of text-to-image generation. For better demonstration, we continue finetuning OFA on a subset of LAION-400M [117].



An art painting of a city, in the style of cyberpunk.



A painting of the sunset cliffs in the style of fantasy art.



A painting of the superman.



An art painting of a city, in the style of steampunk.



A painting of the sunset cliffs in the style of dark fantasy art.



A painting of the superman, in the dark style.



An oil painting of a pizza on the beach.



An oil painting of a computer in the sky.



A painting of the ship in the style of etching, HD.

Figure 8: Examples of text-to-image generation.



Figure 9: More samples of VQA task on unseen domains. The answers are generated by pretrained OFA without finetuning. The datasets used in VQA pretraining task only contain real-world photographs. We present more cases of VQA task on out-of-domain (non-photographic) images and demonstrate the capability of transferring OFA to these unseen domains.



Figure 10: Samples of the unseen grounded question answering task. In this task, the model should answer a question about a particular region in the image. This task is unseen in pretraining. We demonstrate that directly transferring pretrained OFA to this new task without finetuning works well.

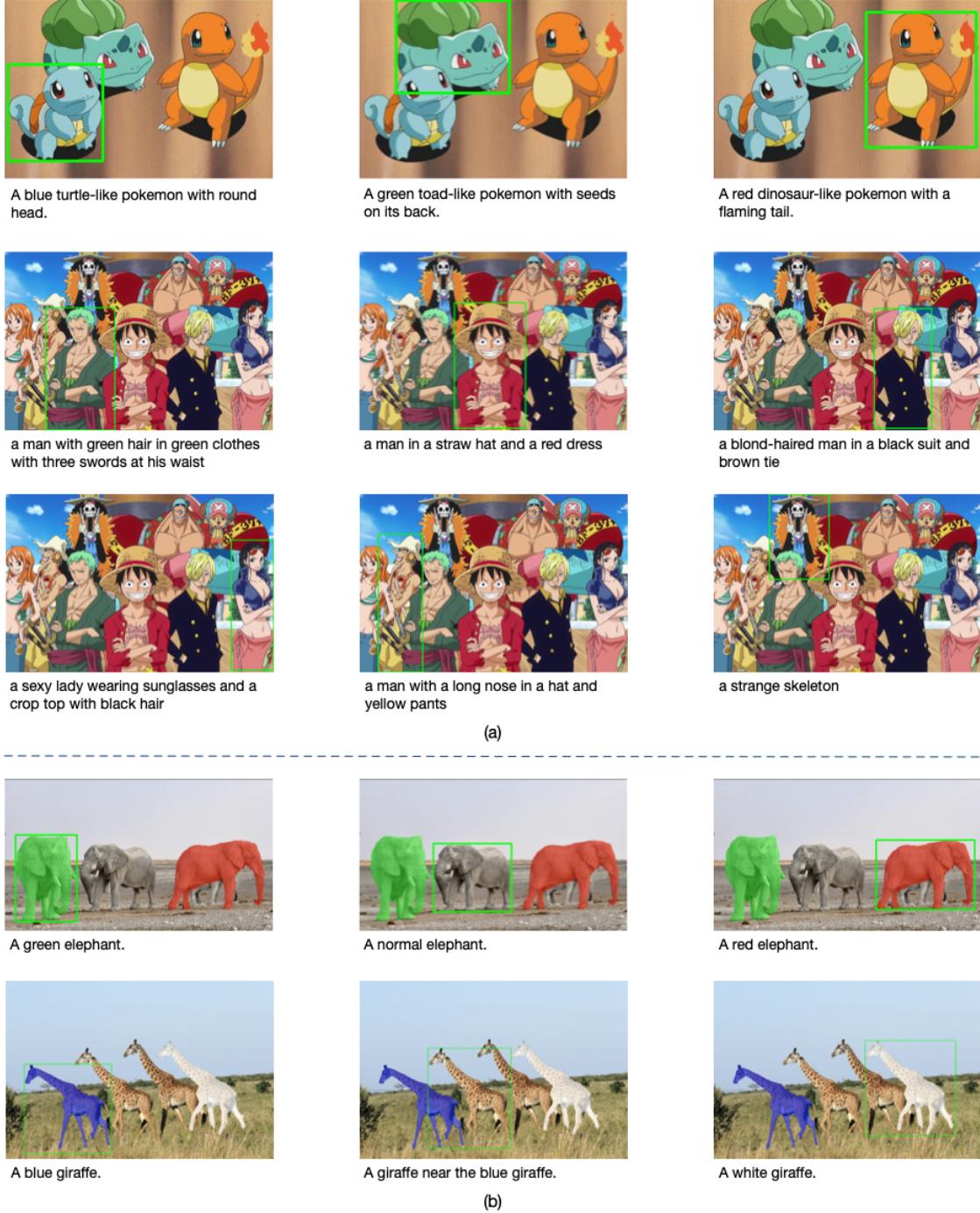


Figure 11: Samples of visual grounding task generated by OFA for various unseen domains: (a) anime (the corresponding animations are *Pokemon* and *One Piece*); (b) synthetic images with attribute combinations.