

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

OFA (One-For-All)

Task Agnostic

Unified task
representation
to support
different types
of tasks

Modality Agnostic

Unified input
and output
representation
shared among
all tasks to
handle
different
modalities

Task Comprehensiv eness

Enough task
variety to
accumulate
generalization
ability robustly

3 Unification

I/O

Shard I/O
across
different
modalities and
tasks

Architecture

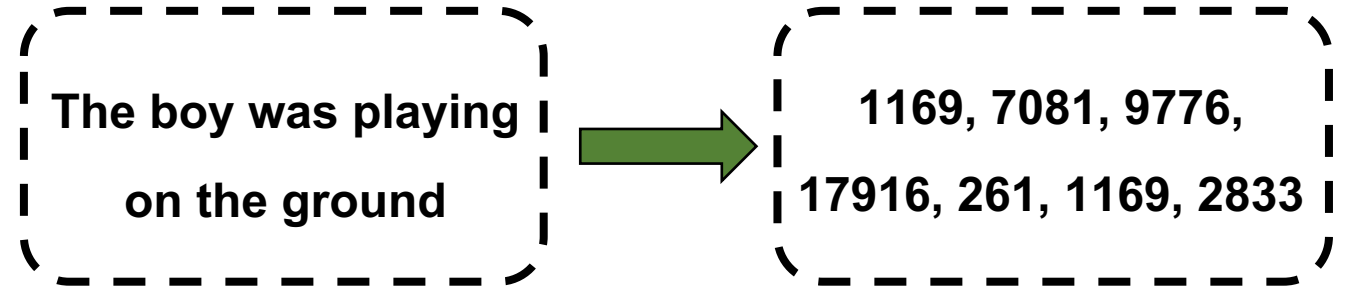
A encoder-
decoder
framework
without task-
specific layers

Task

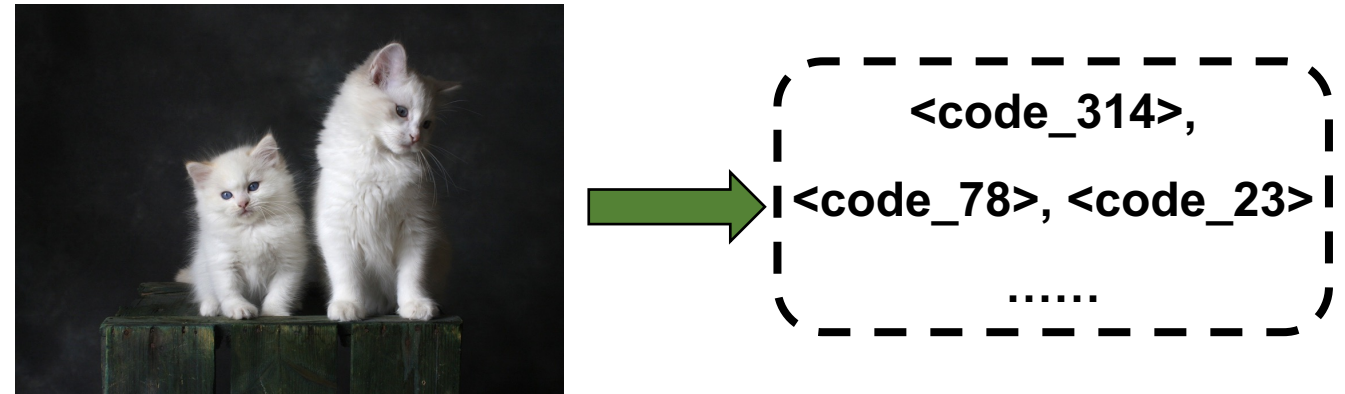
Varieties of
tasks are
unified to the
sequence-to-
sequence
format

I/O

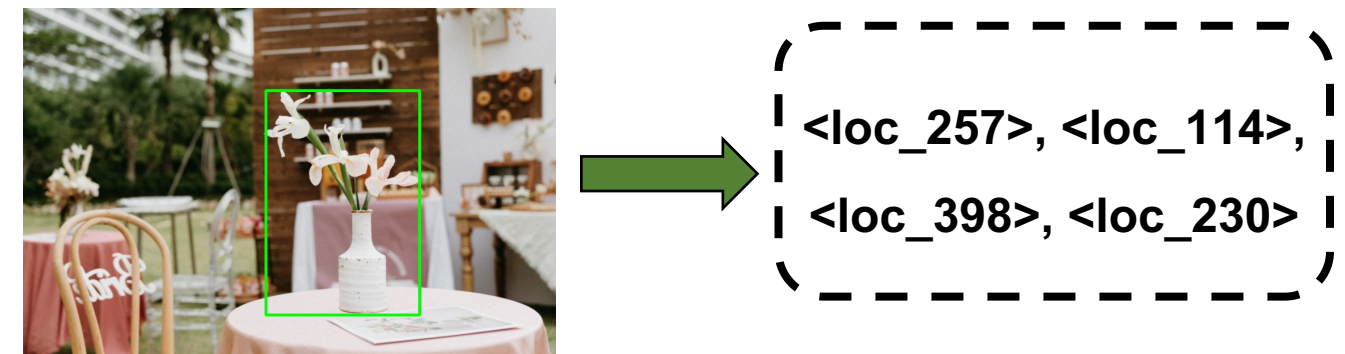
Byte-Pair Encoding
for texts



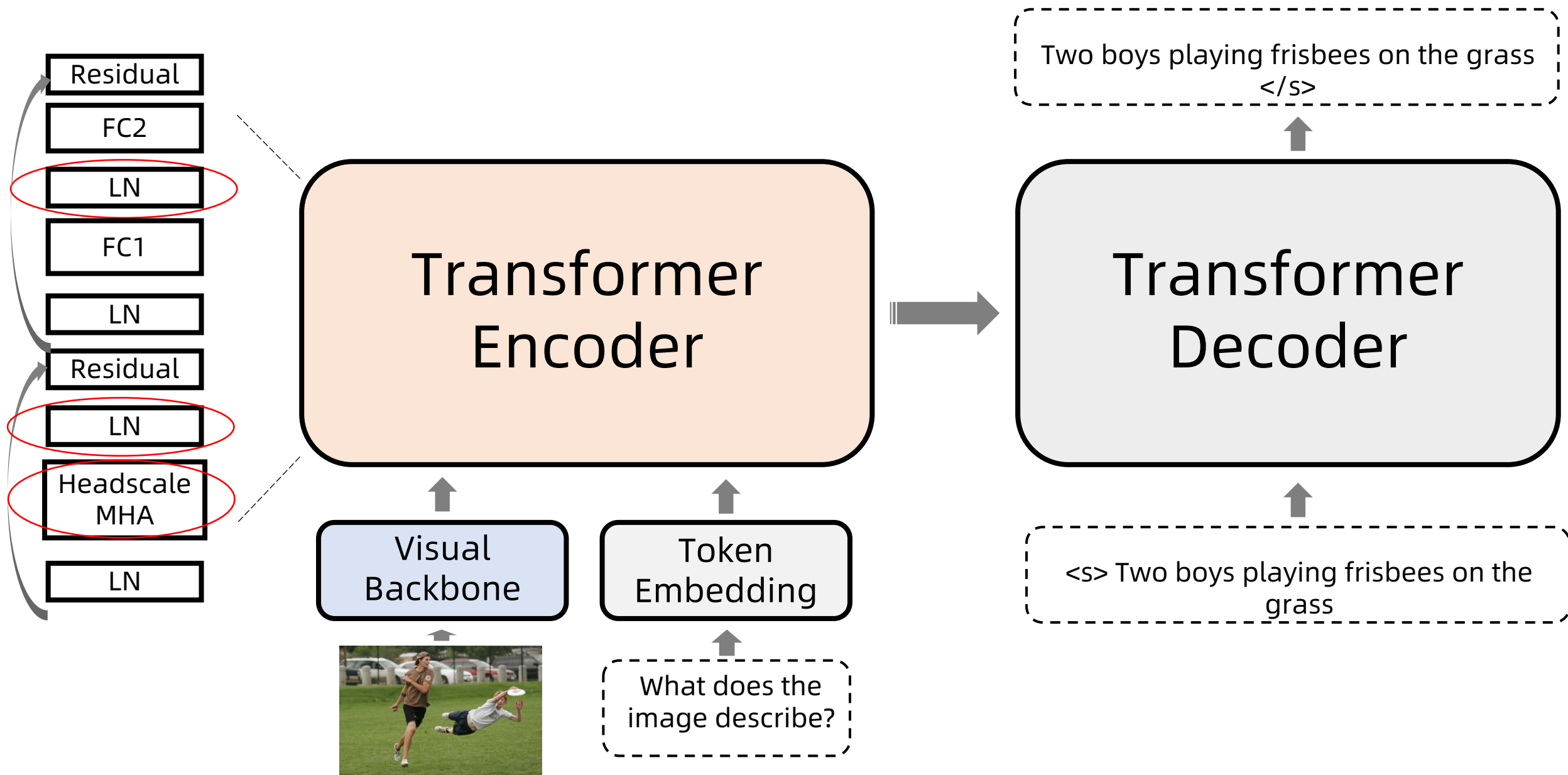
Vector Quantization
for images



Discretization
for bounding boxes



Architecture



Unified Tasks

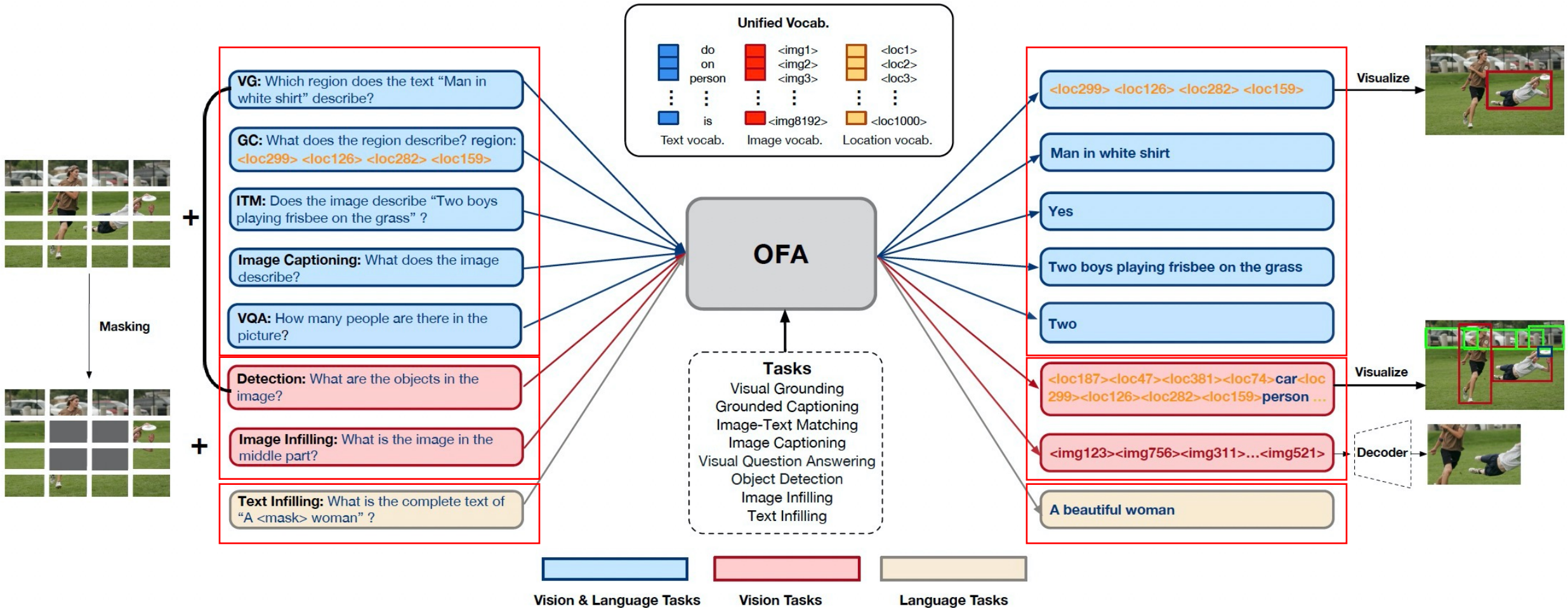


Figure 2: A demonstration of the pretraining tasks, including visual grounding, grounded captioning, image-text matching, image captioning, VQA, object detection, image infilling as well as text infilling.

Cross-modal Understanding

SOTA performance in VQA and SNLI-VE

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 _{Tiny}	70.3	70.4	85.3	85.2
OFA _{Medium}	75.4	75.5	86.6	87.0
OFA _{Base}	78.0	78.1	89.3	89.2
OFA _{Large}	80.3	80.5	90.3	90.2
OFA	82.0	82.0	91.0	91.2

Image-to-Text Generation

SOTA performance on the MSCOCO Image Caption

Model	Cross-Entropy Optimization				CIDEr Optimization			
	BLEU@4	METEOR	CIDEr	SPICE	BLEU@4	METEOR	CIDEr	SPICE
VL-T5 [57]	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 [58]	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 [47]	39.6	-	127.7	-	-	-	-	-
LEMON [24]	41.5	30.8	139.1	24.1	42.6	31.4	145.5	25.5
SimVLM [22]	40.6	33.7	143.3	25.4	-	-	-	-
OFA _{Tiny}	35.9	28.1	119.0	21.6	38.1	29.2	128.7	23.1
OFA _{Medium}	39.1	30.0	130.4	23.2	41.4	30.8	140.7	24.8
OFA _{Base}	41.0	30.9	138.2	24.2	42.8	31.7	146.7	25.8
OFA _{Large}	42.4	31.5	142.2	24.5	43.6	32.2	150.7	26.2
OFA	43.9	31.8	145.3	24.8	44.9	32.5	154.9	26.6

Image-to-Text Generation

On the top of the MSCOCO official leaderboard

Results																	
#	User	Entries	Date of Last Entry	BLEU-1		BLEU-2		BLEU-3		BLEU-4		METEOR		ROUGE-L		CIDEr-D	
				c5 ▲	c40 ▲	c5 ▲	c40 ▲	c5 ▲	c40 ▲	c5 ▲	c40 ▲	c5 ▲	c40 ▲	c5 ▲	c40 ▲	c5 ▲	c40 ▲
1	OFA-Sys_OFA	4	05/31/22	0.845 (1)	0.981 (1)	0.701 (1)	0.944 (2)	0.559 (2)	0.878 (2)	0.436 (1)	0.787 (1)	0.321 (1)	0.427 (1)	0.625 (1)	0.790 (1)	1.472 (1)	1.496 (1)
2	MS_Cog_Svcs-GIT-Single_Model	2	05/31/22	0.842 (2)	0.980 (2)	0.700 (2)	0.945 (1)	0.559 (1)	0.878 (1)	0.435 (2)	0.786 (2)	0.320 (2)	0.422 (2)	0.621 (2)	0.785 (2)	1.465 (2)	1.495 (2)
3	CMG	3	04/04/22	0.840 (4)	0.975 (5)	0.692 (7)	0.932 (6)	0.545 (7)	0.857 (7)	0.421 (8)	0.761 (7)	0.305 (8)	0.402 (8)	0.604 (8)	0.756 (14)	1.414 (3)	1.439 (3)
4	tohoku_cvlab	2	03/06/22	0.841 (3)	0.976 (4)	0.694 (3)	0.935 (3)	0.549 (3)	0.863 (3)	0.425 (4)	0.768 (3)	0.309 (3)	0.410 (3)	0.612 (3)	0.771 (3)	1.413 (4)	1.438 (4)
5	hwy	2	08/30/21	0.840 (5)	0.977 (3)	0.693 (4)	0.935 (4)	0.548 (5)	0.861 (4)	0.424 (5)	0.765 (5)	0.308 (4)	0.405 (7)	0.610 (5)	0.764 (7)	1.413 (5)	1.438 (5)

Visual Grounding

SOTA performance in Referring Expression Comprehension

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 _{Tiny}	80.20	84.07	75.00	68.22	75.13	57.66	72.02	69.74
OFA _{Medium}	85.34	87.68	77.92	76.09	83.04	66.25	78.76	78.58
OFA _{Base}	88.48	90.67	83.30	81.39	87.15	74.29	82.29	82.31
OFA _{Large}	90.05	92.93	85.26	85.80	89.87	79.22	85.89	86.55
OFA	92.04	94.03	88.44	87.86	91.70	80.71	88.07	88.78

Text-to-Image Generation

Competitive performance on the MSCOCO Dataset

Model	#Param.	FID↓	CLIPSIM↑	IS↑
DALLE [51]	12B	27.5	-	17.9
CogView [52]	4B	27.1	33.3	18.2
GLIDE [77]	3.5B	12.2	-	-
Unifying [78]	228M	29.9	30.9	-
NÜWA [53]	870M	12.9	34.3	27.2
OFA	472M	10.5	34.4	31.1

Cross-modal Generation

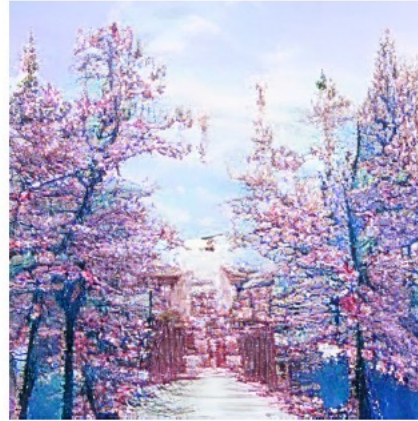


Figure 3. Qualitative comparison with state-of-the-art models for text-to-image generation. We present more qualitative examples of text-to-image generation for better demonstration in Appendix C.

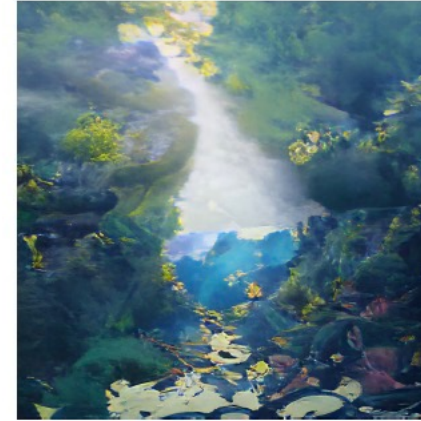
Cases of Text-to-Image Generation



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 cyperpunk.



The golden palace of the land of clouds.



Rustic interior of an alchemy shop.

Cases of Text-to-Image Generation



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.

Uni-modal Tasks

Competitive performance on uni-modal tasks

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]	97.0	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	91.9	92.3	91.1	95.3
<i>Ours</i>						
OFA	96.6	91.0	91.7	92.5	90.2	94.8

Model	ROUGE-1	Gigaword 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	39.81	20.66	37.11

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 ₃₈₄ -L/16 [36]	86.3
MAE-L/16 [37]	85.9
OFA	85.6

Zero-shot Learning & Task Transfer

Table 6: Zero-shot performance on 6 GLUE subtasks 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	49.6	-
OFA _{Base}	71.6	56.7	79.5	54.0	51.4	37.3	49.71 / 49.18



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

A: tan

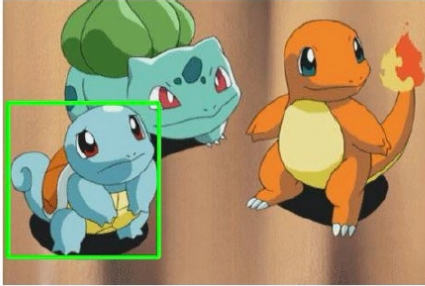


Q: what color is the car in the region? region:
<loc512> <loc483> <loc675> <loc576>

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.

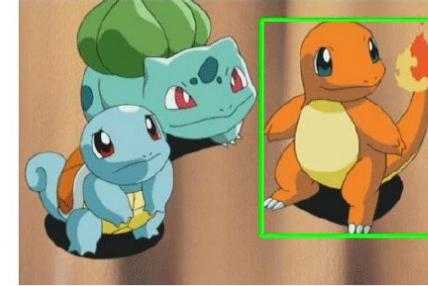
Domain Transfer



A blue turtle-like pokemon with round head.



A green toad-like pokemon with seeds on its back.



A red dinosaur-like pokemon with a flaming tail.



a man with green hair in green clothes with three swords at his waist



a man in a straw hat and a red dress



a blond-haired man in a black suit and brown tie



a sexy lady wearing sunglasses and a crop top with black hair



a man with a long nose in a hat and yellow pants



a strange skeleton

Thanks!

Code: <https://github.com/OFA-Sys/OFA>

Demo: <https://huggingface.co/ofa-sys>

Contact us: zheluo.wp@alibaba-inc.com