UniTAB: Unifying Text and Box Outputs for Grounded Vision-Language Modeling (Supplementary Material)

Zhengyuan Yang, Zhe Gan, Jianfeng Wang, Xiaowei Hu,
Faisal Ahmed, Zicheng Liu, Yumao Lu, Lijuan Wang

Microsoft Cloud and AI
{zhengyang,zhe.gan,jianfw,xiaowei.hu,
fiahmed,zliu,yumaolu,lijuanw}@microsoft.com

A Experiment Details

Hyper-parameter. We summarize the detailed experiment settings of UniTAB in Table A. In the UniTAB decoder, we encode previous target token inputs $s_{<t}$ with token and position embedding, and do not use type embedding to differentiate text and box tokens.

Training corpus. We introduce the "200K" pre-training corpus in the main paper, which contains both image-text pairs and grounded box annotations. In the main paper's Tables 4-6, we refer to the training corpus used in previous studies by their contained image numbers. Specifically, the "180K" corpus [5,32] aggregate images and annotations from COCO [20] and Visual Genome [16].

The "3M" corpus [38,23] contains image-text pairs from the Conceptual Captions dataset [30]. The "4M" corpus [4,19,10] consists of the COCO [20], Visual Genome [16], Conceptual Captions [30], and SBU Captions [25] image-text pairs.

Downstream task post-processing and evaluation. We detail the post-processing and downstream task evaluation in UniTAB inference. The first step shared among different tasks is to extract text, box, and word-box alignment predictions from the unified output sequence, as visualized in the main paper's Figure 3. We then use the extracted outputs for downstream task evaluations. We next detail the evaluation process of specific downstream tasks. **1). Grounded captioning.** We use the extracted text, box, and alignment predictions to compute caption and grounding evaluations following the standard benchmark [37]. **2). Phrase grounding.** We require the model to repeat the input text query and ground boxes as box tokens inline in the output sequence. For phrase grounding, the model needs to predict object boxes and align the box with words in the input text query. Instead of separately predicting the alignments between predicted boxes and input words [15,18], UniTAB repeats the input text and extracts alignments with the $<obj>$ token from the unified output sequence. If the repeated text output is wrong, the alignment will be disarrayed, thus leading to wrong phrase grounding predictions. **3). Referring expression comprehension.** Since the referring expression comprehension task [36,24] does not require the alignment prediction, we take the first predicted box in the output

Table A. The detailed experiment settings of UniTAB.

Hyper-parameter	Value
(a) Optimizer hyper-parameters	
optimizer	AdamW [22]
base learning rate	1e-4
backbone learning rate	2e-5
learning rate schedule	Step *0.1 for final 5 epochs
weight decay	1e-4
batch size	64
pre-training epochs	40
multi-task finetuning epochs	20
task-specific finetuning epochs	20
exp. moving average	0.9998
(b) Model hyper-parameters	
encoder layer number	6
encoder hidden size	256
encoder intermediate size	2048
encoder head number	8
decoder layer number	6
decoder hidden size	256
decoder intermediate size	2048
decoder head number	8
max input words	256
input visual tokens	$\frac{H_0}{32} \times \frac{W_0}{32}$ [11]
max decoding steps	256
number of bins	200
augmentation	RandomResizedCrop [2]
image size	800-1333
encoder vocab size	50265 (RoBERTa [21])
decoder vocab size	50265+200+2=50467

sequence as the final grounding prediction. **4). COCO captioning and VQA.** We use the extracted text outputs for final evaluations (*i.e.*, captioning metrics for COCO and exact match for VQA accuracy).

B Ablation Studies on Decoding Design

In this section, we present ablation studies on UniTAB decoder and output sequence design, starting with "<*obj*> token," "decoder type embedding," and "number of object tokens." For these three ablation studies on the decoder, we initialize single-modality and transformer encoders with pre-trained UniTAB weights, and finetune model variants that have different decoder designs on the experimented task. We then discuss different inference-time "decoding sampling method," and the experiment on "decoding syntactic restrictions."

<obj> token. UniTAB's special <*obj*> token naturally represents the word-box alignments in the output sequence. In addition to indicating the alignments, we examine if <*obj*> simplifies the sequence prediction and thus improves the model performance. The referring expression comprehension task requires a single box output and does not need the word-box alignment. Thus, we could remove the <*obj*> token while still being able to perform the task. Table B shows the

Table B. Ablation study of the <**obj**> token on the Refcocog [24] dataset.

$\text{UniTAB}_{\text{Separate}}$	Refcocog
Baseline	80.23
w/o <\obj>	79.41
w/o <obj>, <\obj>	79.31

Table C. Ablation study of the decoder type embedding. We experiment on the Refcocog [24] and Flickr30k Entities [27] grounding tasks.

$\text{UniTAB}_{\text{Separate}}$	Refcocog	Flickr
Baseline (w/o type emd.)	80.23	78.92
<obj> as text tokens	80.47	78.65
<obj> as box tokens	80.39	79.40
<obj> as a third type	80.54	78.83

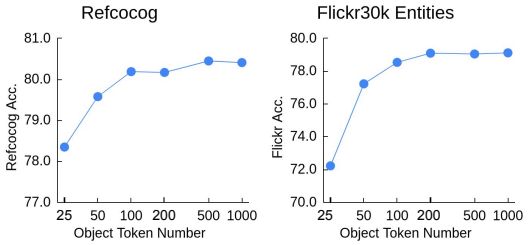

Fig. A. Ablation studies on the box token number. We experiment on the Refcocog [24] and Flickr [27] grounding tasks.

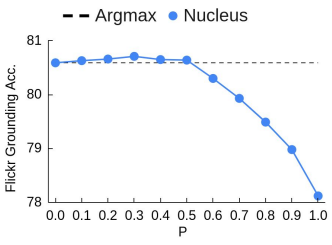

Fig. B. Ablation studies on decoding sampling method on the phrase grounding task.

experiments on the Refcocog dataset [24]. The $\text{UniTAB}_{\text{Separate}}$ baseline inserts a pair of <*obj*> and <\obj> tokens before and after a word-box token segment. We experiment with removing the <\obj> token, or both special tokens. We observe an around 1% accuracy improvement by adding <*obj*> tokens.

Decoder type embedding. Table C shows the ablation study on decoder type embedding, *i.e.*, whether to use type embedding to differentiate text and box tokens. We experiment with the following variants of decoder embedding [14,34]. The $\text{UniTAB}_{\text{Separate}}$ baseline does not use type embedding. "<obj> as text/box tokens" uses two type embedding to differentiate text and box tokens, where the <obj> token is tagged as text or box token. "<obj> as a third type" introduces an extra type embedding specialized for <obj> and <\obj>. We experiment on the Refcocog [24] and Flickr30k Entities [27] grounding tasks. We empirically observe that the decoder type embedding has no major influence on model performance, and thus do not use type embedding in UniTAB.

Number of object tokens. Figure A shows the influence of object token number on the grounding performance. We observe a steady performance when the object token number is large enough for a dataset to avoid quantization error. The token number is around 200 for the experimented VL datasets.

Decoding sampling method. Figure B shows the ablation study of the decoding sampling method on Flickr phrase grounding [27]. Compared with the simple argmax decoding sampling, we observe a marginal improvement from nucleus sampling [13,3]. The improvement from nucleus sampling is smaller on

Table D. UniTAB pre-training with additional bounding box annotations. "Separate^{box}" adopts the extra bounding box annotations from the COCO [20], VG [16], Objects365 [29], OpenImages [17] object detection datasets.

UniTAB	Visual grounding				Grounded caption		COCO	VQAv2
	Refcoco	Refcoco+	Refcocog	Flickr	Cider	$F1_{all}$	test-Cider	KP-test
Separate	86.32	78.70	79.96	79.39	65.6	11.46	119.3	66.6
Separate^{box}	88.27	80.98	83.78	81.90	70.0	13.46	120.7	68.4

UniTAB's experimented VL tasks, compared with previous explorations on object detection [3]. We suspect that the smaller gain is due to the difference in target sequences. Specifically, object detection [3] has multiple correct decoding sequences, as object order doesn't matter in object detection outputs. In contrast, UniTAB only has one fixed decoding target of the constructed text+box sequence. Thus, the diversity brought by nucleus sampling helps less in UniTAB.

Decoding syntactic restriction. We apply no decoding syntactic restriction in UniTAB training. We scan UniTAB predictions for two types of failure cases that break the syntactic restrictions in output decoding sequences: **(1)** before $<\backslash obj>$ there are not exactly four consecutive box tokens; **(2)** $<obj>$, $<\backslash obj>$ tokens are followed by box tokens, or are not paired. We scan the Refcocog grounding and Flickr grounded captioning predictions generated by $\text{UniTAB}_{\text{Pre-finetuning}}$, and COCO captions generated by $\text{UniTAB}_{\text{Shared}}$ (for generalized grounded captioning in Figure 3 (d)). We observe **zero** syntactic failure cases in all scanned outputs, implying that the decoding token type pattern is easy to learn.

We then incorporate these two syntactic restrictions into the model training, and examine if the restrictions ease the training and improve model performance. Specifically, we compute the softmax language modeling loss over a subset of all tokens in applicable decoding positions, such as masking out box logits after $<obj>$. We experiment on Refcoco and Flickr grounded captioning, based on $\text{UniTAB}_{\text{Separate}}$. We empirically observe that the syntactic restriction has no major influence on the model performance, with +0.9 accuracy gain on Refcoco and a slight drop on Flickr grounded captioning (-0.2 CIDEr, -0.15 $F1_{all}$).

C Discussions

Pre-training with additional box annotations. In addition to the "200K" pre-training corpus and the additional image-text data [30,25] introduced in the main paper, we further explore using additional box annotations with no caption texts for UniTAB pre-training. We aggregate object detection annotations from COCO [20], VG [16], Objects365 [29], and OpenImages [17]. Each sample is an image with object box and class annotations. For pre-training, we randomly select up to 32 objects and shuffle the object order. We concatenate the object class name as the input text, and train the model to generate the text+box sequence to ground the selected objects. We refer to $\text{UniTAB}_{\text{Separate}}$ with those extra box annotations as "Separate^{box}."

Table E. Experiment results of $\text{UniTAB}_{\text{Prefix}}$ that adds task-specific prefix in multi-task finetuning.

Method	#Pre-train	Visual grounding				Grounded caption		COCO	VQAv2
		Refcoco	Refcoco+	Refcocog	Flickr	Cider	$F1_{all}$	test-Cider	KP-test
MDETR [15]	200K	86.75	79.52	81.64	83.8	-	-	-	-
UNITER [4]	4M	81.24	75.31	74.31	-	-	-	-	70.5
GVD [37]	-	-	-	-	-	62.3	7.55	-	-
VL-T5 [5]	180K	-	-	71.2	-	-	-	116.5	67.9
OSCAR [19]	4M	-	-	-	-	-	-	123.7	-
$\text{UniTAB}_{\text{Shared-Scratch}}$	None	82.06	70.72	73.39	65.67	61.1	7.85	111.8	63.1
$\text{UniTAB}_{\text{Prefix-Scratch}}$	None	82.38	70.96	75.43	69.58	62.1	8.51	112.8	64.3
$\text{UniTAB}_{\text{Shared}}$	200K	88.50	80.98	84.46	79.23	63.4	9.18	115.8	66.6
$\text{UniTAB}_{\text{Prefix}}$	200K	87.60	79.72	83.41	80.13	62.4	10.54	115.6	66.0

Table D shows the experiment results of adding additional box annotations. On VL tasks that require box prediction, such as the visual grounding task and the grounding evaluation in grounded captioning, "$\text{Separate}^{\text{box}}$" consistently outperforms $\text{UniTAB}_{\text{Separate}}$ on the grounding accuracy and grounded captioning F1 score. More interestingly, we empirically observe that extra box annotations could also help the text output quality. For example, "$\text{Separate}^{\text{box}}$" improves grounded captioning CIDEr score from 65.6 to 70.0, COCO captioning CIDEr score from 119.3 to 120.7, and VQA accuracy from 66.6% to 68.4%.

Multi-task finetuning with prefix. In the main paper, we discuss the effectiveness of multi-task finetuning, which gathers training data from all downstream tasks and learns a single set of parameters for different VL tasks. By unifying all considered downstream tasks as a sequence generation problem, a single $\text{UniTAB}_{\text{Shared}}$ model could perform well on different tasks, meanwhile being parameter efficient and showing promise in zero-shot generalization.

One variant of $\text{UniTAB}_{\text{Shared}}$ is to add a task-specific input text string to identify the task for each sample [5], such as "visual grounding:". The extra input text string is known as the prefix. We experiment with a variant of UniTAB multi-task finetuning with prefix, namely $\text{UniTAB}_{\text{Prefix}}$. $\text{UniTAB}_{\text{Prefix}}$ adds a task-specific prefix at the beginning of each input text string, *e.g.*, "Visual grounding: the coffee mug next to the plate." We use the task name as the prefix, *i.e.*, "visual grounding:", "phrase grounding:", "grounded captioning:", "image captioning:", "question answering:", *etc*. We then train the model with multi-task finetuning, the same as $\text{UniTAB}_{\text{Shared}}$. Table E compares $\text{UniTAB}_{\text{Prefix}}$ with $\text{UniTAB}_{\text{Shared}}$. We observe a comparable performance with and without prefix on the experimented tasks and datasets.

Robustness and bias analyses. We conduct robustness and bias analyses to better understand the limitation of UniTAB. Tables F,G show initial robustness and bias analyses. We retrain UniTAB on the splits (VQA-train, COCO-14) used in the established analysis setups [12,6]. In Table F, we follow VQA-CE [6] and compare the gain over the UpDown baseline on two subsets. UniTAB achieves a larger gain on "counterexamples" (+9.76%) compared with "easy" (+6.17%), indicating better robustness against shortcuts compared with the UpDown base-

Table F. VQA-CE [6] robustness analyses.

Accuracy(%)	UpDown	UniTAB
Overall	63.52 (+0.00)	70.78 (+7.26)
Counterexamples	33.91 (+0.00)	43.67 (+9.76)
Easy	76.69 (+0.00)	82.86 (+6.17)

Table G. Gender error analyses [12].

Error rate(%)	COCO-Bias	COCO-Balanced
BaselineFT	12.83	19.30
Balanced	12.85	18.30
UpWeight	13.56	16.30
Equalizer	7.02	8.10
UniTAB	9.87	9.21

line, as discussed in VQA-CE [6]. Table G evaluates gender bias with error rate [12]. UniTAB achieves a lower error rate than general captioning models (*cf*., Baseline-FT of 19.30% *vs*. UniTAB of 9.21%), and is only slightly worse than the specialized method Equalizer [12]. We hypothesize that the reasonable robustness and bias performance is related to UniTAB's grounded training, which better binds visual concepts with text words. Despite the reasonable performance on the standard analyses, building robust and unbiased models remains a challenging problem and could be further improved.

D Qualitative Results

In this section, we present additional qualitative results made by $\text{UniTAB}_{\text{Shared}}$. We start with the captioning tasks in Figure C. Figure C**(a)** presents the grounded captioning results on Flickr30k Entities, where the predictions are evaluated by both the caption and grounding metrics. UniTAB performs well in both generating captions and grounding noun phrases to image regions. For captioning, the model generates a smooth and accurate image description, and properly includes attribute words to produce an informative caption, *e.g*., "young boy" and "blue shirt" in the top left example. UniTAB is also capable of providing a comprehensive description of the scene. For example, in the bottom right sub-figure of (a), the caption consists of the foreground object and its detailed attributes "man in red shirt and blue jeans," scene descriptions "a red door" and "on the street," and the nearby object "a dog." The model also performs well in grounding. Noticeably, UniTAB performs well on grounding and describing tiny objects, *e.g*., the "a bat" and "a baseball" in the top left example and the "a red ball" in the bottom left example.

Figure C**(b)** shows UniTAB's prediction on the MSCOCO image captioning task. With the same inputs as Flickr30k grounded captioning, UniTAB learns to transfer the grounded captioning ability learned on Flirckr30k Entities to MSCOCO, although COCO captioning does not have grounding annotations. For evaluation, we extract the text tokens and compute the standard COCO captioning metrics [26,9,1,31]. We note that UniTAB achieves comparable caption performance to the state of the art, and meanwhile being capable of grounding all noun phrases in the caption. As shown in Figure C(b), UniTAB generates an informative caption and accurately grounds all noun phrases in the caption to visual regions. Such grounded captioning ability is important for reducing object hallucination [28], boosting the model's interpretability and fairness [28,12],

Fig. C. Additional qualitative results from $\text{UniTAB}_{\text{Shared}}$ on captioning tasks.

and facilitating applications in robotics and human-computer interaction. We also visualize additional captioning examples on ImageNet [7]. We observe that UniTAB generalizes well onto the ImageNet images. The ImageNet caption and grounding predictions in Figure C(**c**) are of similar qualities as on Flickr30k Entities and MSCOCO.

Figure D shows $\text{UniTAB}_{\text{Shared}}$'s predictions on grounding tasks. Figures D(a,b) are from the Refcoco [36] and Refcocog [24] datasets, for the referring expression comprehension task. We observe that the model learns to identify different objects in the same image conditioned on different input queries. For example, in Figure D(**a**), the targets of "yellow sleeve guy" on the left and "blue" skier in the background. Similarly, in Figure D(**b**), UniTAB correctly differentiates the four people in the image. UniTAB also correctly localizes the head noun in a long referring expression and predicts the box on the corresponding phrase. For example, in Figure D(b), grounding boxes are predicted on the words "girl" and "person," instead of the entire query as in previous studies [35,33,8]. Another

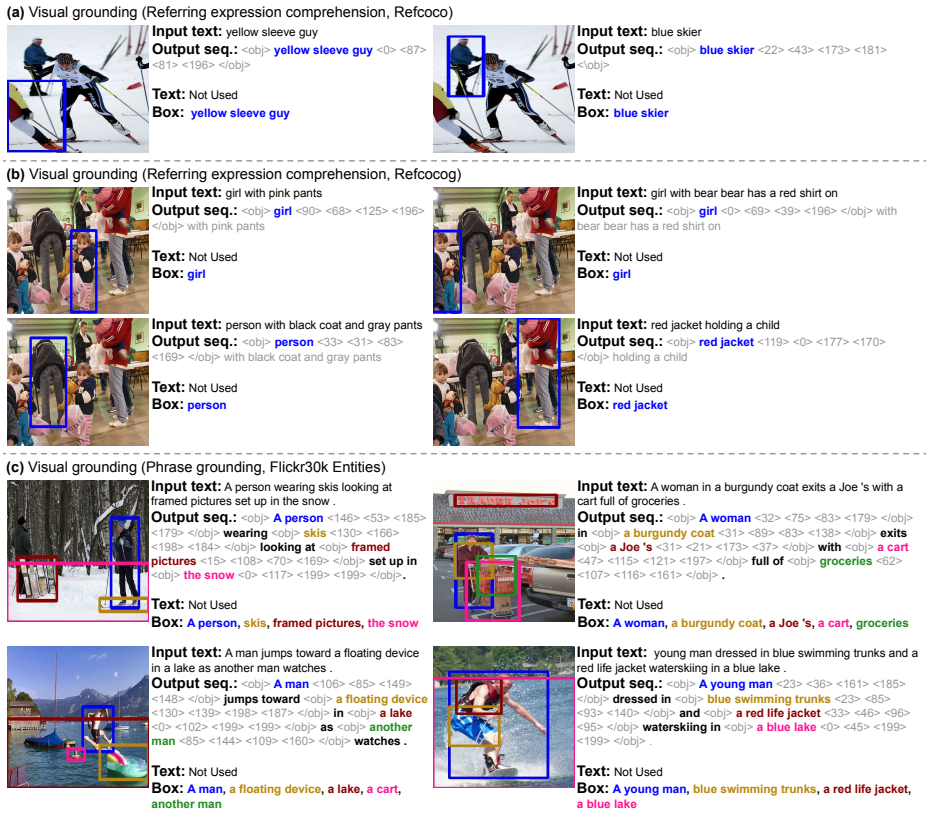

Fig. D. Additional qualitative results from UniTAB$_{\text{Shared}}$ on grounding tasks.

observation is that UniTAB usually predicts a single box in the output sequence for referring expression comprehension samples. For example, in the top left sub-figure of Figure D(b), the model only grounds the head noun "girl" and does not generate a box for the remaining phrase like "pink pants." We conjecture that UniTAB learns to identify the referring expression comprehension task based on the input text (*e.g.*, a short referring query *vs*. a complete sentence), and generates a single box when performing the task.

Figure D**(c)** shows the phrase grounding examples on the Flickr30k Entities dataset [27]. Phrase grounding requires the model to identify all noun phrases in a sentence and ground them to corresponding image regions. UniTAB correctly grounds all types of phrases referred to in the sentence, including foreground objects "person" and "woman," smaller background objects "skies" in the top left example and "another man" in the bottom left example, and scene regions "the snow," "a lake," and "a blue lake." The model even correctly predicts challenging regions such as the "trader joe's" logo in the top right sub-figure.

References

1. Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. In *ECCV*, 2016. 6
2. Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, 2020. 2
3. Ting Chen, Saurabh Saxena, Lala Li, David J Fleet, and Geoffrey Hinton. Pix2seq: A language modeling framework for object detection. In *ICLR*, 2022. 3, 4
4. Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Learning universal image-text representations. In *ECCV*, 2020. 1, 5
5. Jaemin Cho, Jie Lei, Hao Tan, and Mohit Bansal. Unifying vision-and-language tasks via text generation. In *ICML*, 2021. 1, 5
6. Corentin Dancette, Remi Cadene, Damien Teney, and Matthieu Cord. Beyond question-based biases: Assessing multimodal shortcut learning in visual question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1574–1583, 2021. 5, 6
7. Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 7
8. Jiajun Deng, Zhengyuan Yang, Tianlang Chen, Wengang Zhou, and Houqiang Li. Transvg: End-to-end visual grounding with transformers. In *ICCV*, 2021. 7
9. Michael Denkowski and Alon Lavie. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the ninth workshop on statistical machine translation*, 2014. 6
10. Zhe Gan, Yen-Chun Chen, Linjie Li, Chen Zhu, Yu Cheng, and Jingjing Liu. Large-scale adversarial training for vision-and-language representation learning. In *NeurIPS*, 2020. 1
11. Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 2
12. Lisa Anne Hendricks, Kaylee Burns, Kate Saenko, Trevor Darrell, and Anna Rohrbach. Women also snowboard: Overcoming bias in captioning models. In *ECCV*, 2018. 5, 6
13. Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. In *ICLR*, 2020. 3
14. Ronghang Hu, Amanpreet Singh, Trevor Darrell, and Marcus Rohrbach. Iterative answer prediction with pointer-augmented multimodal transformers for textvqa. In *CVPR*, 2020. 3
15. Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. Mdetr-modulated detection for end-to-end multi-modal understanding. In *ICCV*, 2021. 1, 5
16. Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV*, 2017. 1, 4
17. Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. The open images dataset v4. *International Journal of Computer Vision*, 128(7):1956–1981, 2020. 4
18. Liunian Harold Li*, Pengchuan Zhang*, Haotian Zhang*, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, Kai-Wei Chang, and Jianfeng Gao. Grounded language-image pre-training. In *arXiv preprint arXiv:2112.03857*, 2021. 1

19. Xiujun Li, Xi Yin, Chunyuan Li, Xiaowei Hu, Pengchuan Zhang, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *ECCV*, 2020. 1, 5
20. Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 1, 4
21. 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. *arXiv preprint arXiv:1907.11692*, 2019. 2
22. Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 2
23. Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *NeurIPS*, 2019. 1
24. Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *CVPR*, 2016. 1, 3, 7
25. Vicente Ordonez, Girish Kulkarni, and Tamara L Berg. Im2text: Describing images using 1 million captioned photographs. In *NeurIPS*, 2011. 1, 4
26. Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *ACL*, 2002. 6
27. Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *ICCV*, 2015. 3, 8
28. Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. In *EMNLP*, 2018. 6
29. 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. 4
30. 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. 1, 4
31. Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *CVPR*, 2015. 6
32. 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. In *ACL*, 2021. 1
33. Zhengyuan Yang, Boqing Gong, Liwei Wang, Wenbing Huang, Dong Yu, and Jiebo Luo. A fast and accurate one-stage approach to visual grounding. In *ICCV*, 2019. 7
34. Zhengyuan Yang, Yijuan Lu, Jianfeng Wang, Xi Yin, Dinei Florencio, Lijuan Wang, Cha Zhang, Lei Zhang, and Jiebo Luo. Tap: Text-aware pre-training for text-vqa and text-caption. In *CVPR*, pages 8751–8761, 2021. 3
35. Licheng Yu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, Mohit Bansal, and Tamara L Berg. Mattnet: Modular attention network for referring expression comprehension. In *CVPR*, 2018. 7
36. Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *ECCV*, 2016. 1, 7
37. Luowei Zhou, Yannis Kalantidis, Xinlei Chen, Jason J Corso, and Marcus Rohrbach. Grounded video description. In *CVPR*, 2019. 1, 5

38. 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. 1