Appendix

A. Pre-training Details

The statistics of pre-training datasets are presented in Table 1. The COCO Captions dataset comprises manually generated captions where multiple captions are assigned to each image. For the Visual Genome dataset, the region description serves as the image caption, yielding several captions for each image. The SBU Captions and Conceptual Caption datasets contain a single caption per image. It should be noted that a considerable number of the image links in these two datasets have become invalid because they are collected from the Internet.

	COCO	**VG**	**SBU**	**CC3M**
#Images	113K	108K	855K	2.98M
#Captions	567K	5.4M	855K	2.98M

Table 1. Statistics of the pre-training datasets.

The default architecture of ViLTA contains a dual-encoder architecture (a pre-trained vision encoder and a pre-trained language encoder) and a multimodal encoder. Table 2 reports the hyperparameters used in our pre-training model. For ViLTA$_{\text{BASE}}$, we leverage a 12-layer transformer-based structure as language/vision encoder and 6-layer for the multimodal encoder respectively. The number of transformer layers for the language and vision encoders is set to 24 for ViLTA$_{\text{LARGE}}$. The number of the multimodal encoder also maintains the default setup of 6-layer transformer-based structure. Here, we initialize the language encoder with weights from the pre-trained RoBERTa [9] and the vision encoder with weights from the pre-trained CLIP-ViT-224/16 [11].

Hyperparameters	**ViLTA$_{\text{BASE}}$**	**ViLTA$_{\text{LARGE}}$**
Total steps	36k	24k
Warmup steps	21.6k	14.4k
Batch size	1024	1024
Learning rate	$1e^{-5}$	$4e^{-6}$
Learning rate decay	Linear	
Weight decay	0.01	
Dropout ratio	0.1	
AdamW ϵ	$1e^{-8}$	
AdamW β	(0.9, 0.98)	
Textual encoder	RoBERTa$_{\text{BASE}}$	RoBERTa$_{\text{LARGE}}$
Visual encoder	CLIP-ViT-B-224	CLIP-ViT-L-336
Patch size	16	14
Input resolution	288	224
Number of layers	6	6
Hidden size	768	1024
FFN inner hidden size	3072	4096
Number of attention heads	12	16

Table 2. Hyperparameters for pre-training model. The last block is the hyperparameters for the multimodal encoder.

B. Fine-tuning Details

We fine-tune ViLTA on 5 downstream tasks using the hyperparameters reported in Table 3 for VL classification tasks, Table 4 for VL retrieval tasks, Table 5 for image captioning. In the following sections, we provide a comprehensive description of the fine-tuning configurations employed for each task.

- *Visual Question Answering (VQA)* [2] aims to predict a natural language answer based on the given image and question. Following the previous works [6, 5, 3, 13], we treat VQA as a multi-label classification task with 3,129 possible answers. We concatenate the image representation v_{cls} and text representation w_{cls} obtained from the multimodal model, and then pass it through a 2-layer MLP layer to perform a classification task. We use GELU activation function and a binary cross-entropy loss function on the soft target scores to optimize the model.

- *Visual Reasoning* focuses on predicting whether the caption is true or false for a pair of images. Here, we employ a pairwise strategy to effectively process the input in NLVR^2 [12] dataset, where each data sample is divided into *(image1, statement)* and *(image2, statement).* We then feed them separately into the model to obtain two representations and concatenate them together to pass through a binary classification head.

- *Visual Entailment* aims to predict whether a natural language hypothesis is entailed, neutral or contradicted by the image premise. We train and evaluate our model on SNLI-VE [14] dataset and treat it as a three-class classification problem.

- *Image-Text Retrieval* contains two sub tasks: image-to-text retrieval (TR) and text-to-image retrieval (IR). COCO [8] and Flickr30K [10] serve as evaluation datasets. Following the standard setting in ViLT [6], We use the pre-trained ITM head, specifically the component that calculates the true-pair logits, to initialize the similarity score head. We then sample 15 random texts as negative examples and use a cross-entropy loss that maximizes the scores for positive pairs.

- *Image Captioning* is a generative task and we investigate whether our encoder-only model is suitable for such generative tasks. To adapt our model for image captioning, we modify the encoder on the text side of the model by transforming it into a causal decoder through the adjustment of the attention mask. Subsequently, we fine-tune the model on the COCO Captions [8] dataset using cross-entropy loss and evaluate it on the NoCaps [1] dataset without additional training.

Hyperparameters	**VQAv2**	**NLVR^2**	**SNLI-VE**
Learning rate	$1e^{-5}$	$1e^{-5}$	$2e^{-6}$
Epochs	10	10	5
Batch size	512	256	64
AdamW ϵ		$1e^{-8}$	
AdamW β		(0.9, 0.98)	
Weight decay	0.05	0.01	0.01
Dropout ratio		0.1	
Input resolution	576^2	384^2	288^2

Table 3. Hyperparameters for fine-tuning ViLTA on VL classification tasks.

Hyperparameters	**COCO Flickr**
Learning rate	$5e^{-6}$
Epochs	10
Batch size	64
AdamW ϵ	$1e^{-8}$
AdamW β	$(0.9, 0.98)$
Weight decay	0.01
Dropout ratio	0.1
Input resolution	576^2

Table 4. Hyperparameters for fine-tuning ViLTA on VL retrieval tasks.

C. Scaling Ability

To show the effectiveness of ViLTA on extensive datasets, we expand the training of ViLTA-base and ViLTA-large on a subset of the LAION-2B and CC12M datasets employing 64 A100 GPUs in Table 6. The total volume of data was roughly 150M, comparable to the 129M dataset used in BLIP. All performance metrics for retrieval tasks show substantial enhancements, ranging from 73.3 to 80.5 on the COCO dataset for text retrieval in terms of recall@1. However, the gain in VL understanding (VLU) tasks is not as prominent as the increase in retrieval tasks, which is consistent with the findings in previous studies [7, 4]. Such discrepancy arises due to the challenges associated with the considerable noise present in large-scale web data, which are integral to VLU tasks. As shown in Table 7, in the context of a large-scale dataset, ViLTA achieves a better gain, while, in contrast, BLIP brings about performance degradation.

Hyperparameters	**COCO Captioning**
Learning rate	$1e^{-5}$
Epochs	10
Batch size	512
AdamW ϵ	$1e^{-8}$
AdamW β	(0.9, 0.98)
Weight decay	0.01
Dropout ratio	0.1
Input resolution	576^2
Label smoothing ε	0.1
Beam size	5

Table 5. Hyperparameters for fine-tuning ViLTA on image captioning.

Dataset	**Flickr**			**COCO**		
	TR@1	**TR@5**	**TR@10**	**TR@1**	**TR@5**	**TR@10**
4M	94.5	99.8	99.8	73.3	91.8	95.9
150M	**95.7**	**99.9**	**99.9**	**80.5**	**94.6**	**97.3**

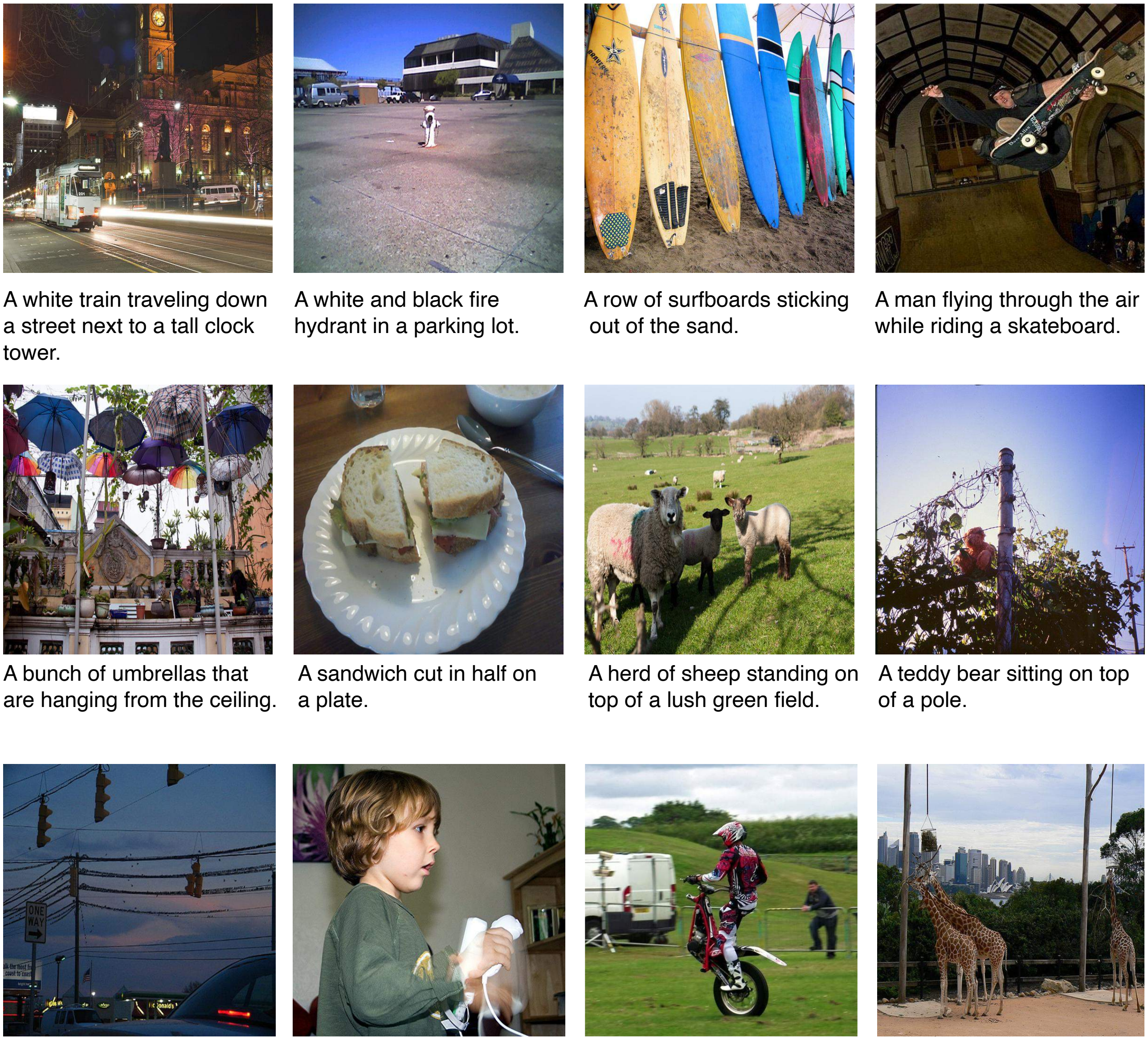
Table 6. Experimental results on retrieval task.

Dataset	**BLIP**		**ViLTA**	
	14M	**129M**	**4M**	**129M**
NLVR2-dev	82.67	82.15	85.16	86.33
NLVR2-test	82.30	82.24	86.13	87.25

Table 7. Results on NLVR2 dataset. Large scale data may not have significant benefits for VLU tasks.

D. Additional Results

In this section, we present additional results generated by ViLTA. Specifically, we show the efficacy of ViLTA in image captioning. The case study in Figure 1 shows the generated image captions on a series of samples. Notably, ViLTA generates diverse and descriptive captions, which can effectively encapsulate the content of the corresponding images. These results verify the effectiveness of ViLTA in different VL tasks.



A white train traveling down a street next to a tall clock tower.

A white and black fire hydrant in a parking lot.

A row of surfboards sticking out of the sand.

A man flying through the air while riding a skateboard.

A bunch of umbrellas that are hanging from the ceiling.

A sandwich cut in half on a plate.

A herd of sheep standing on top of a lush green field.

A teddy bear sitting on top of a pole.

A street scene with cars and traffic lights.

A young boy holding a Nintendo Wii game controller.

A man riding a dirt bike on top of a lush green field.

Three giraffes are standing in a grassy field.

Figure 1. Case study of ViLTA on image captioning task.

References

[1] Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. Nocaps: Novel object captioning at scale. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8948–8957, 2019.

[2] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015.

[3] Hangbo Bao, Wenhui Wang, Li Dong, Qiang Liu, Owais Khan Mohammed, Kriti Aggarwal, Subhojit Som, and Furu Wei. Vlmo: Unified vision-language pre-training with mixture-of-modality-experts. *arXiv preprint arXiv:2111.02358*, 2021.

[4] Emanuele Bugliarello, Laurent Sartran, Aishwarya Agrawal, Lisa Anne Hendricks, and Aida Nematzadeh. Measuring progress in fine-grained vision-and-language understanding. *arXiv preprint arXiv:2305.07558*, 2023.

[5] Zi-Yi Dou, Yichong Xu, Zhe Gan, Jianfeng Wang, Shuohang Wang, Lijuan Wang, Chenguang Zhu, Pengchuan Zhang, Lu Yuan, Nanyun Peng, et al. An empirical study of training

end-to-end vision-and-language transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18166–18176, 2022.

[6] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*, pages 5583–5594. PMLR, 2021.

[7] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR, 2022.

[8] 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 *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.

[9] 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.

[10] 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 *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649, 2015.

[11] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[12] Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. A corpus for reasoning about natural language grounded in photographs. *arXiv preprint arXiv:1811.00491*, 2018.

[13] Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, et al. Image as a foreign language: Beit pretraining for all vision and vision-language tasks. *arXiv preprint arXiv:2208.10442*, 2022.

[14] 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.